МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА
ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ

На правах рукописи

Васильев Юлий Алексеевич

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ АНАЛИЗА ВЫЖИВАЕМОСТИ

Специальность 2.3.5 —
«Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

ДИССЕРТАЦИЯ
на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
кандидат физико-математических наук,
доцент Петровский Михаил Игоревич

Москва, 2024

Оглавление

Стр.

ВВЕДЕНИЕ

Актуальность темы исследования

Интеллектуальные системы анализа событий широко используются в медицине, биостатистике, социологии и анализе технологических процессов. Например, в медицине прогнозируется ожидаемое время и вероятность летального исхода, а в анализе надежности время технического сбоя или поломки оборудования. Интеллектуальные модели позволяют описывать контекст события, интерпретировать зависимости и прогнозировать время наступления события на основе характеристик объектов исследования (наблюдений).

Модели анализа выживаемости позволяют оценивать вероятность и время до наступления определенного события. Для сбора данных определяется целевое событие и фиксируется интервал исследования, в рамках которого могут появляться новые наблюдения. Каждому наблюдению сопоставляется вектор признаков X, полученный на момент начала исследования, а также время наступления события T.

Наблюдения, для которых наступает целевое событие, называются терминальными. Однако, полные данные могут быть недоступны и истинное время наступления события неизвестно в случае потери наблюдения или раннего прекращения исследования. Наблюдения с неизвестным временем события называются цензурированными. Например, в исследованиях летального исхода причиной цензурирования может быть перевод пациентов в другое учреждение, выписка или отказ пациента от исследования. Важно отметить, что наиболее распространено правое цензурирование, при котором известен момент выхода из исследования до наступления целевого события. Таким образом, уникальность анализа выживаемости заключается в использовании двух целевых переменных: времени события T и флага цензурирования δ.

Особенностью моделей анализа выживаемости является возможность прогнозирования функций вероятности наступления события для каждого момента времени. Функция выживания (survival function) определяет вероятность ненаступления события по истечении определенного времени $S(t) = P(T \geq t)$, где t – время наблюдения, T – случайная величина времени события. Функция плотности (density function) определяет риск наступления события $f(t) = (1 - S(t))'$ в момент времени t. Функция риска (hazard function) определяет относительный риск события $h(t) = f(t)/S(t)$ в момент времени t при условии, что событие не наступило ранее. Системы интеллектуального анализа событий должны обеспечивать прогноз данных функций в зависимости от характеристик наблюдения для каждого момента времени.

Построение прикладных интеллектуальных систем анализа событий напрямую связано со следующими особенностями реальных данных:

- **Гетерогенность признакового пространства**. Для описания состояния наблюдения используются непрерывные и категориальные показатели, которые могут содержать пропущенные значения из-за ограниченности информации или наличия ошибок;
- **Распределение вероятностей времени событий**. Постановка задачи и формат исследования влияют на распределение вероятностей времени и соотношение терминальных и цензурированных событий;

- **Информативность цензурирования**. Если причина цензурирования не связана с условиями проведения исследования, то говорят о неинформативном цензурировании, в противном случае существуют неучтенные факторы и цензурирование считается информативным.

Классические модели анализа выживаемости не позволяют работать с представленными особенностями данных и используют строгие предположения. Таким образом, актуальным является разработка интеллектуальных систем анализа выживаемости, не использующих строгие статистические предположения и применимых к особенностям реальных данных.

Степень разработанности темы

Построение интеллектуальных систем анализа выживаемости является перспективным направлением исследований и применяется в здравоохранении, анализе надежности и биостатистике. Большинство исследований посвящены применению классических статистических подходов и методов машинного обучения для анализа событий. Существующие решения основаны на следующих концепциях:

- Непараметрические методы не учитывают связь между признаками наблюдения и целевыми переменными и предполагают неинформативность цензурирования. Полупараметрические методы основаны на идее масштабирования непараметрической функции риска по индивидуальному для каждого наблюдения коэффициенту масштабирования. Параметрические методы предполагают теоретическое распределение времени, описывая индивидуальный прогноз как сдвиг функции во времени.
- Дискретные модели машинного обучения прогнозируют вектор вероятностей наступления события в фиксированные моменты времени. Регрессионные методы прогнозируют одну целевую переменную, но учитывают полную информацию при расчете функции потерь. Ансамбли регрессионных моделей строят отдельную модель для каждого момента времени. Нейросетевые модели устанавливают размер выходного слоя количеству точек фиксированной временной шкалы, минимизируя отклонения между прогнозом и теоретической дискретной функцией.
- Непрерывные модели машинного обучения основаны на расширении статистических моделей. Регрессионные модели масштабируют базовую функцию на основе точечного прогноза относительного риска события. Модели деревьев выживаемости рекурсивно разбивают признаковое пространство по статистическому критерию на группы с максимально различной выживаемостью. Модели ансамблирования деревьев агрегируют прогнозы множества моделей, повышая качество прогнозирования, но теряя интерпретируемость. В таком случае строгость предположений зависит от критерия разбиения и непараметрических оценок в листах дерева.
- Для оценки качества прогнозирования используются точечные и интегральные метрики. Точечные метрики основаны на сравнении ожидаемой вероятности и времени события, а также единичных значений функций. Интегральные метрики оценивают значения функций для всех моментов времени путем сравнения с эталонной функцией или ранжирования наблюдений по риску наступления события. Наибольшую популярность получили метрики: правдоподобие, индекс согласованности и интегральный показатель Браера.

Существующие модели анализа выживаемости обладают несколькими недостатками. Статистические модели основаны на строгих предположениях, которые могут не выполняться на реальных данных. Дискретные модели прогнозируют значения функций в рамках ограниченной временной шкалы. Непрерывные модели используют статистические предположения для прогнозирования функций. Важно отметить, что существующие модели анализа выживаемости не позволяют непосредственно обрабатывать категориальные и пропущенные значения и требуют предварительной обработки данных.

Возможным путем преодоления существующих недостатков является разработка подхода построения моделей анализа выживаемости на основе деревьев решений, поскольку в задачах машинного обучения они позволяют определять зависимости без необходимости предварительного определения предположений модели и обработки данных. Модификация этапа поиска разбиения и построения листовых оценок позволит обрабатывать категориальные и пропущенные значения для обучения и применения модели, а также преодолеть строгие предположения критериев разбиения и непараметрических оценок. Также, необходимо разработать программную библиотеку анализа выживаемости.

Целью диссертационной работы является разработка математического и программного обеспечения интеллектуальной системы для решения задач анализа выживаемости с использованием методов машинного обучения на основе деревьев решений.

Объектом исследования диссертационной работы являются модели анализа выживаемости, позволяющие прогнозировать время и вероятность наступления события, а также функции выживания и риска. Предметом исследования диссертационной работы является разработка алгоритмов построения моделей анализа выживаемости, применимых к неполным непрерывным и категориальным данным, а также к случаям информативного цензурирования без использования строгих статистических предположений.

Для достижения цели необходимо решение следующих **задач**:

1. Разработать методы построения интерпретируемых моделей анализа выживаемости на основе деревьев решений, учитывающих особенности реальных данных.
2. Исследовать и разработать методы оценки качества прогнозирования моделей анализа выживаемости;
3. Разработать алгоритмы ансамблирования предложенных деревьев выживаемости, позволяющих повысить качество прогнозирования;
4. Реализовать интеллектуальную программную систему на основе разработанного комплекса алгоритмов анализа выживаемости и провести её апробацию на прикладной задаче анализа медицинских данных.

Диссертация соответствует специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей» в части направления разработки интеллектуальных систем машинного обучения и инструментальных средств разработки цифровых продуктов, поскольку целью работы является исследование, разработка и программная реализация комплекса алгоритмов для построения интеллектуальной системы анализа событий, применимой для решения задач анализа выживаемости.

Необходимость работы с реальными данными низкого качества существенно ограничивает возможность применения в таких системах классических моделей, использующих статистические подходы. Для решения этой проблемы необходимо разработать новые оригиналь-

ные методы или предложить модификации существующих методов машинного обучения, а также реализовать их в виде программной библиотеки с открытым кодом, которая может быть использована для построение интеллектуальных систем анализа событий для широкого спектра прикладных областей. В рамках настоящей работы библиотека будет использоваться для решения ряда прикладных задач анализа выживаемости из области медицины.

Научная новизна

Разработан алгоритм поиска разбиений в данных с цензурированием, основанный на гистограммном вычислении взвешенных критериев log-rank и учитывающий категориальные и пропущенные значения. Предложенный подход регуляризации критерия позволяет обрабатывать случаи информативного цензурирования, учитывая информацию об априорном распределении событий, в том числе в случае малых выборок, когда в процессе построения дерева возникает разреженная область в пространстве признаков. На основе алгоритма поиска разбиений предложен метод построения интерпретируемых деревьев выживаемости с модифицированными непараметрическими оценками функций выживания и риска. Для оценки качества прогнозирования исследованы существующие и предложены модифицированные метрики с равным вкладом событий и временных интервалов. Также предложены методы ансамблирования деревьев выживаемости, основанные на построении независимых базовых бутстеп моделей, а также на подходе усиления слабых моделей с использованием адаптивного бустинга с перевыборкой.

Теоретическая и практическая значимость

Разработанная программная библиотека анализа выживаемости призвана упростить процесс построения и применения моделей анализа выживаемости, оценки качества прогнозирования и проведения экспериментального исследования. Разработанные методы построения моделей могут использоваться для решения различных прикладных задач, основанных на анализе выживаемости. Апробация библиотеки проводилась на прикладных задачах анализа медицинских данных.

Комплекс предложенных алгоритмов позволяет строить модели анализа выживаемости, применимые к реальным данным. Метод построения деревьев выживаемости позволяет строить интерпретируемые прогнозы, а ансамбли деревьев имеют высокое качество прогнозирования. По результатам экспериментального исследования, предложенные методы превзошли по качеству существующие методы анализа выживаемости. Полученные результаты диссертационной работы могут послужить основой для построения перспективных современных систем анализа событий, которые будут включать в себя средства анализа выживаемости наблюдений. При этом, могут использоваться как все разработанные модули, так и отдельные из них.

Методология и методы исследования

При получении основных результатов диссертационной работы использовались методы машинного обучения и математической статистики. При разработке модулей программной библиотеки анализа выживаемости использовались методы объектно-ориентированного проектирования, а также методы векторизации и параллелизации вычислений.

Основные положения, выносимые на защиту:

1. Предложенный метод построения деревьев выживаемости, учитывающий особенности реальных данных: наличие категориальных признаков и пропущенных значений, рас-

пределения вероятностей времени наступления событий и информативность цензурирования. Алгоритм поиска разбиений в данных с цензурированием основан на взвешенных регуляризованных критериях log-rank и реализован в виде гистограммного метода. Метод позволяет строить интерпретируемые прогнозы времени и вероятности события, функций выживания и риска;

2. Предложенные методы построения бутстреп и бустинг ансамблей деревьев выживания позволяют достичь высокого качества прогнозирования за счет использования независимой и адаптивной схем агрегации прогнозов базовых моделей, формирующих ансамбль. В качестве функций потерь используются модифицированные метрики, которые обеспечивают равенство вкладов относительно целевого времени события, флага цензурирования и временной шкалы;
3. Разработанная программная библиотека `survivors` включает предложенный комплекс алгоритмов для построения интеллектуальных систем анализа выживаемости. По результатам экспериментального применения библиотеки на медицинских данных, предложенные методы превзошли по качеству прогнозирования существующие методы.

Достоверность полученных результатов обеспечивается проведенными экспериментами, открытым кодом реализованных методов и подходов, обоснованием принимаемых решений, публикациями в рецензируемых журналах и апробацией на российских и международных конференциях.

Апробация работы

Основные результаты работы докладывались на:

- Научная конференция «Тихоновские чтения» (Россия, Москва, 2021).
- 11th International Conference on Pattern Recognition Applications and Methods (Австрия, 2022).
- Научная конференция «Ломоносовские чтения» (Россия, Москва, 2022).
- XXIX Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов 2022» (Россия, Москва, 2022).
- Научная конференция «Ломоносовские чтения» (Россия, Москва, 2023).
- XXX Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов 2023» (Россия, Москва, 2023).
- IV кафедральная студенческая конференция «Artificial Intelligence and Creativity» (Россия, Москва, 2023).
- Научная конференция «Ломоносовские чтения» (Россия, Москва, 2024).
- Научный семинар МС ВМК МГУ под руководством В.Ю. Королева.

Результаты выступлений были изданы в 7 работах в сборниках тезисов и трудов конференций [1–7]. Получено свидетельство о государственной регистрации программы для ЭВМ [8].

Результаты диссертационной работы использовались в следующих НИР:

- «Исследование, разработка и применение инновационных технологий построения интеллектуальных программных систем» (Номер договора: 6.2.18), 2018–2027 гг.
- «Выполнение работ в области разработки и внедрения методов искусственного интеллекта и анализа больших данных в сфере здравоохранения» (Номер договора: ЦАРСС-12/20-03/У), 2020–2021 гг.

– «Выполнение части работ по развитию прикладного программного обеспечения государственной информационной системы обязательного медицинского страхования» (Номер договора: № С/01-ПО7/02731000011210000030001), 2021–2022 гг.

Личный вклад автора заключается в выполнении основного объема теоретических и экспериментальных исследований, а также в разработке архитектуры и реализации открытой библиотеки анализа выживаемости. Подготовка части материалов к публикации проводилась совместно с соавторами, причем вклад диссертанта был определяющим. В работах [9, 10] М.И. Петровскому принадлежит постановка задачи применения моделей к категориальным и пропущенным значениям, а И.В. Машечкину принадлежат рекомендации к методологии исследований. В работе [11] М.И. Петровский и И.В. Машечкин участвовали в постановке задачи и анализе результатов. Создание программных реализаций алгоритмов и проведение всех численных экспериментов было выполнено автором полностью самостоятельно. Диссертационное исследование является самостоятельным и законченным трудом автора.

Публикации

Основные результаты по теме диссертации изложены в 4 публикациях [9–12], изданных в рецензируемых научных изданиях, определенных в п. 2.3 Положения о присуждении ученых степеней в Московском государственном университете имени М.В. Ломоносова.

Объем и структура работы

Диссертационная работа состоит из введения, пяти глав, заключения и списка литературы. Полный объём диссертации составляет 142 страницы, включая 57 рисунков и 29 таблиц. Список литературы содержит 123 наименования.

Первая глава посвящена исследованию особенностей событийных данных и существующих подходов анализа выживаемости. Рассматриваются существующие методы построения статистических моделей и моделей машинного обучения, их достоинства и недостатки. Также, рассматриваются точечные и интегральные метрики оценки качества прогнозирования величин анализа выживаемости. На основе проведенного аналитического обзора формулируются направления дальнейших исследований в части оценки влияния особенностей данных на построение прогнозных моделей и вычисление метрик качества.

Вторая глава посвящена исследованию и разработке методов построения деревьев выживаемости, применимых к категориальным и непрерывным данным, пропущенным значениям, различным распределениям вероятностей времени событий и случаям информативного цензурирования. Предложен гистограммный метод поиска лучшего разбиения в неполных данных с цензурированием по категориальным и непрерывным признакам со сравнением выборок по взвешенному критерию log-rank. Предложен метод построения интерпретированной модели дерева выживаемости. Для применения моделей к случаям информативного цензурирования предложен подход регуляризации критерия на этапе поиска разбиения и модификации непараметрических листовых моделей.

Третья глава посвящена исследованию и разработке методов оценки качества прогнозирования моделей анализа выживаемости. Выделены четыре случая избыточной чувствительности метрик качества к вкладу отдельных событий, временных компонент, временных интервалов и дисбалансу цензурирования. Разработаны модификации метрик качества, преодолевающих избыточную чувствительность существующих метрик в рассмотренных случаях, для обеспечения равного вклада событий при валидации моделей. На основе модифици-

рованных метрик качества проводится экспериментальное исследование методов построения деревьев выживаемости.

Четвертая глава посвящена исследованию и разработке методов ансамблирования деревьев выживаемости. Предложен метод построения бутстреп ансамбля независимых деревьев выживаемости с определением размера ансамбля на out-of-back выборке. Предложен метод построения адаптивного бустинга деревьев выживаемости с перевыборкой, в котором каждая последующая базовая модель строится по выборке с наблюдениями, имеющими низкое качество прогноза на предыдущих итерациях ансамбля. Проводится экспериментальное исследование влияние функции потерь на качество предложенных ансамблей и сравнение с существующими статистическими подходами и методами машинного обучения.

В пятой главе проводятся разработка и реализация открытой программной библиотеки анализа выживаемости, использующей предложенный комплекс алгоритмов. Приводится детальное описание архитектуры и программной реализации разработанной библиотеки. Также, проводится экспериментальная оценка производительности предложенных моделей.

1 ОБЗОР СОВРЕМЕННЫХ МЕТОДОВ АНАЛИЗА ВЫЖИВАЕМОСТИ

Анализ событийных данных применяется для решения множества прикладных задач медицины, биостатистики, социлогии, анализа технологических процессов и многих других областей. В отличии от классических методов машинного обучения, методы анализа выживаемости позволяют оценивать изменение вероятности наступления события во времени.

Для проведения полного и всестороннего обзора существующих подходов анализа выживаемости, были рассмотрены следующие этапы решения задачи: сбор и обработка событийных данных, построение моделей прогнозирования и оценка качества. В разделе 1.1 рассматриваются особенности событийных данных, методы обработки цензурированных наблюдений и возможные прогнозируемые величины. Также, выделяются особенности данных, влияющие как на качество построения описательных и прогнозных моделей. В разделе 1.2 рассматриваются существующие метрики оценки качества прогнозируемых величин, определяются их недостатки и преимущества.

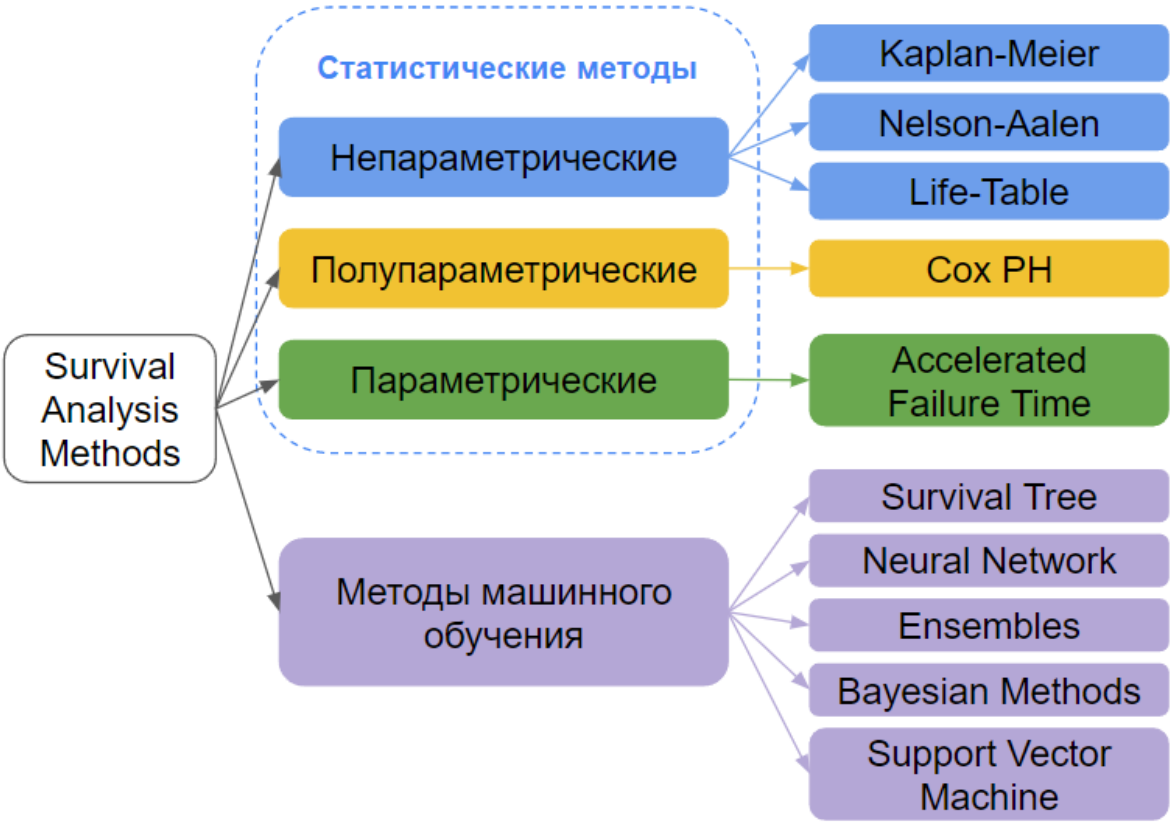


Рис. 1: Схема методов анализа выживаемости

Также, проводится обширный обзор существующих методов построения прогнозных моделей. В работе [13] представлена классификация методов анализа выживаемости на статистические методы и методы машинного обучения (Рисунок 1). Статистические методы включают дополнительные предположения относительно распределения времени события, а также основываются на статистических свойствах оценки параметров. Методы машинно-

го обучения сосредоточены на прогнозировании возникновения события на основе методов обучения с учителем. В разделе 1.3 рассматриваются предположения и параметры традиционных статистических методов анализа выживаемости. В частности, выделяются классы непараметрических, полупараметрических и параметрических моделей. В рамках алгоритмов машинного обучения рассмотриваются деревья выживания (раздел 1.4), нейронные сети (раздел 1.5.1), метод опорных векторов (раздел 1.5.2) и ансамблевое обучение (раздел 1.6).

1.1 Анализ событийных данных

1.1.1 Источники данных

Методы выживаемости широко применяются для решения задач анализа данных в здравоохранении, анализе надежности, биоинформатике, маркетинге и других областей [13]. Рассмотрим особенности постановки задач в четырех прикладных областях.

В области здравоохранения анализируется состояние пациентов. Вход наблюдения в исследование обычно сопровождается медицинским вмешательством, таким как госпитализация, начало приема определенного лекарства или постановка диагноза определенного заболевания. В качестве события рассматриваются летальный исход [14–16], рецидив болезни [17, 18], факт выписки из стационара [19]. Признаковое пространство включает анамнез пациента, клинические и лабораторные показатели, стратегии лечения [16, 19]. Использование моделей анализа выживаемости позволяет решать следующие задачи: оценка вероятности события и распределения вероятности во времени, сравнение стратегий лечения, оценка эффективности схемы лечения.

В области надежности анализируется состояние используемого оборудования. Этап сбора данных имеет фиксированные временные рамки и часто основан на периодической проверке показателей наблюдений. В качестве события рассматривается выход из строя. Признаковое пространство определяется характеристиками конкретного устройства. Например, для анализа жестких дисков используются показатели S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology) [20]. Анализ надежности сосредоточен на разработке методов прогнозирования отказа электронных систем или их отдельных компонент, а также на оценке надежности новых продуктов [21]. Использование моделей анализа выживаемости позволяет предупреждать о потенциальных сбоях и оптимизировать нагрузку на оборудование.

В биоинформатике модели выживаемости применяются для задачи экспрессии генов. Экспрессия генов – это процесс синтеза функционального генного продукта из генной информации РНК [22]. В качестве события рассматривается наступление определенного заболевания и ставится задача оценки вероятности её развития на основе измерений экспрессии генов [23]. Признаковое пространство включает десятки тысяч измерений молекул информационной РНК. Модели выживаемости позволяют оценивать влияние отдельного гена на прогноз выживаемости [22], а также выявлять наиболее значимые гены в качестве биомаркеров для пациентов [24].

В области маркетинга анализируется отток клиентов магазина. Основная цель моделей оттока – предупреждение об уходе клиента и выявление потребностей для его удержания. Компании важно удовлетворить текущих клиентов, чтобы сохранить или увеличить свою прибыль в долгосрочной перспективе. В качестве события рассматривается факт соверше-

ния покупки или отказ клиента от услуг магазина [25]. Признаковое пространство включает информацию о пользователе и истории покупок. Модели выживаемости позволяют прогнозировать время следующей покупки, вероятность ухода клиента и строить индивидуальные модели покупок ценных клиентов для рекомендации услуг пользователям с близкими интересами [26].

1.1.2 Сбор данных с цензурированием

Этап сбора данных имеет значимую роль в анализе выживаемости, поскольку качество и достоверность данных напрямую влияют на результаты и выводы исследования. При сборе данных необходимо учитывать не только объем и разнообразие информации, но и их согласованность с целью исследования. В зависимости от поставленной задачи, необходимо конкретизировать условия исследования и предварительно определить рассматриваемый вид цензурирования, тип и временные рамки исследования за наблюдениями.

Для сбора данных анализа выживаемости проводится исследование за множеством наблюдений. Наблюдения описываются вектором признаков и моментом входа. Целью исследования является определение времени T наступления события. На практике, полные данные могут быть недоступны из-за ограниченности контроля за наблюдениями или наличия временных рамок. В неполных данных время до наступления события может быть неизвестно по нескольким причинам (например, выход из исследования или потеря наблюдения). Наблюдения с известным истинным временем называются терминальными, а с неопределенным временем – цензурированными.

В зависимости от постановки задачи, на этапе сбора данных определяются 3 вида цензурирования наблюдений (Рисунок 2): правое, левое и интервальное. При правом цензурировании известна информация о наблюдении до наступления целевого события (в момент C_R), а истинное время $T \geq C_R$. При левом цензурировании известна информация о наблюдении после наступления целевого события (в момент C_L), а истинное время $T \leq C_L$. При интервальном цензурировании информация о наблюдении собрана до и после наступления целевого события (в моменты C_R и C_L), а истинное время $C_R \leq T \leq C_L$.

Рассмотрим пример из области эпидемиологии, в которой целевым событием является заболевание пациента. Для диагностики заболевания, пациент проходит тестирование в разные моменты времени. Если для пациента наблюдается только один отрицательный тест (отсутствие заболевания), то можно говорить о правом цензурировании. В случае одного положительного теста (наличие заболевания) наблюдается левое цензурирование. Если для пациента получено множество результатов тестирования, то интервал заболевания определяется между последним отрицательным тестом и первым положительным тестом, а задача решается в рамках интервального цензурирования.

Следует отметить, что истинное время наступления события неизвестно во всех трех случаях. Для многих практических задач, наиболее распространенным сценарием является правая цензура [14–16, 21, 22]. В данной работе основное внимание уделяется задачам анализа выживаемости с правым цензурированием.

Для сбора данных с правым цензурированием применяются 2 типа исследований. Исследования первого типа характеризуются строгим временным интервалом. В течение заданного интервала в исследование поступают новые наблюдения. При входе наблюдения, для

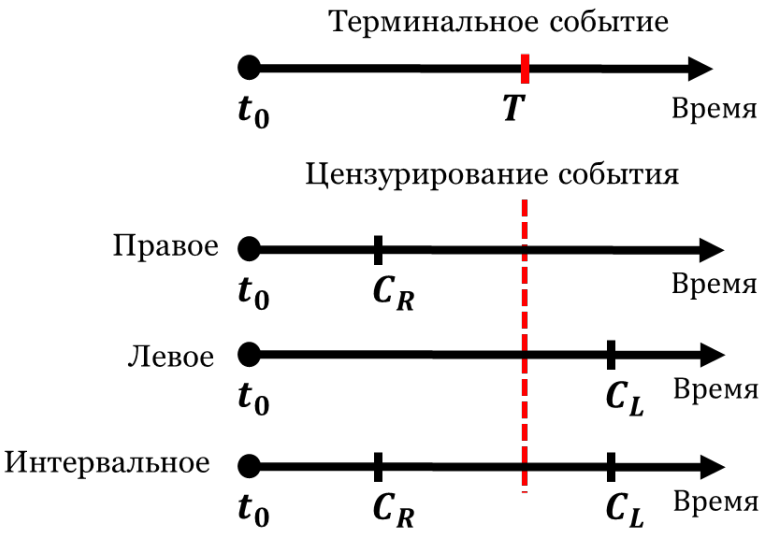


Рис. 2: Демонстрация типов событий в анализе выживаемости. Для терминальных наблюдений известно истинное время события T. При цензурировании, истинное время события неизвестно (пунктирная линия). При правом цензурировании известна информация в момент C_R (до наступления события), при левом – в момент C_L (после наступления), при интервальном – до и после наступления события.

него фиксируются значения признаков X и назначается время входа T_s. Для наблюдения возможны 3 варианта исхода. При наступлении события в момент T_e, истинное время события равно $T_t = T_e - T_s$. При выходе наблюдения из исследования в момент C_e, время цензурирования равно $C = C_e - T_s$. При завершении исследования в момент T_{tr}, время цензурирования с обрезкой (truncated) равно $C = T_{tr} - T_s$.

На Рисунке 3 представлен пример исследования первого типа на 5 наблюдениях (A, B, C, D, E). Для наблюдений B, C событие наступило в рамках исследования. Наблюдения A, D были цензурированы по внешним причинам. Наблюдение E было цензурировано с обрезкой из-за ограниченности интервала исследования. На правом графике представлено преобразование наблюдений на относительную временную шкалу (с общим временем входа в исследование).

Исследования второго типа (часто используемые в инженерном деле) рассматривают n наблюдений до тех пор, пока не наступит событие для r-ого наблюдения. При $r = n$ исследование проводится до тех пор, пока для всех наблюдений не наступит целевое событие. На практике, исследования второго типа редко проводятся в области биомедицины или общественного здравоохранения по причине жесткого контроля процесса цензурирования и гибкого временного интервала.

1.1.3 Формализация задачи

Результатом этапа сбора данных является множество троек значений для каждого наблюдения $i : (X_i, T_i, \delta_i)$. Вектор признаков X_i вычисляется при входе наблюдения в исследо-

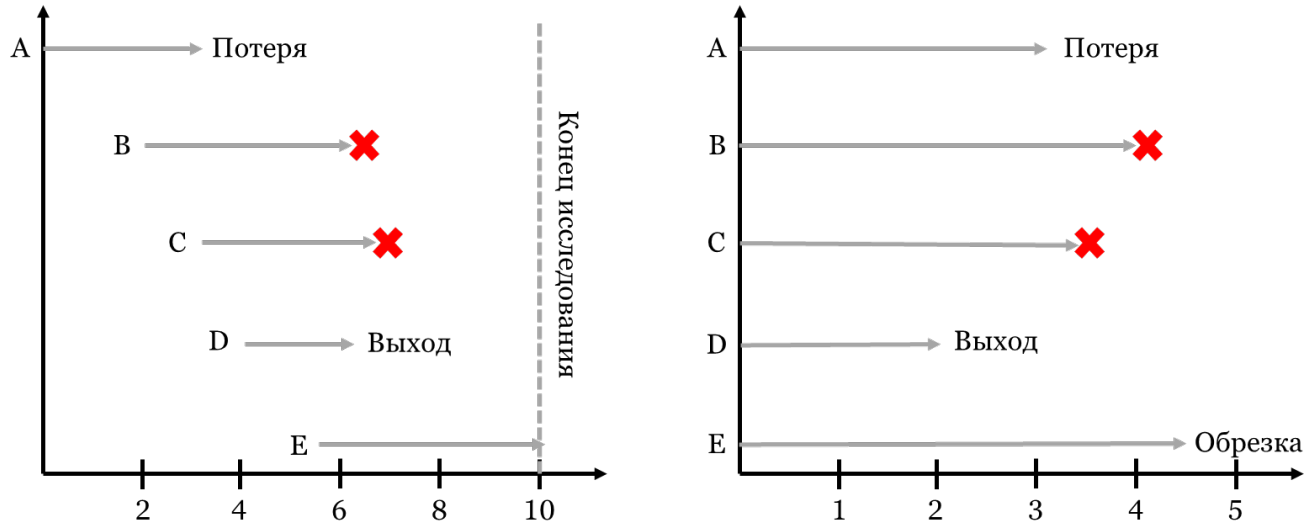


Рис. 3: Демонстрация исследования первого типа на этапе сбора (левый график) и подготовки данных (правый график).

вание. Целевыми переменными являются флаг цензурирования δ_i и время события T_i:

$$T_i = \begin{cases} T_{t,i}, & \text{if } \delta_i = 1, \\ C_i, & \text{if } \delta_i = 0. \end{cases} \tag{1}$$

В практических задачах, реальные данные содержат информацию о признаках (возраст, пол и т.д.) и ставится вопрос определения влияния факторов на функцию выживания, плотность смертности и функцию риска. Целью анализа выживаемости является оценка времени до интересующего события $T_{t,j}$ для нового экземпляра j с вектором признаков X_j.

Функции выживания и риска

Задача анализа выживаемости сводится к оценке трех функций [13]: функция выживания, функция плотности смертности и функции риска. Функция выживания (survival function) определяет вероятность ненаступления события по истечении определенного времени:

$$S(t) = P(T \geq t).$$

Функция плотности смертности (death density function) определяет риск наступления события в конкретный момент времени t:

$$f(t) = (1 - S(t))'.$$

Наиболее распространено использование функции риска (hazard function), которая определяет относительный риск события в конкретный момент времени t при условии, что событие не наступило ранее:

$$h(t) = \frac{f(t)}{S(t)}.$$

Дискретные и непрерывные задачи

В зависимости от поставленной задачи, функции анализа выживаемости формулируются в непрерывном и дискретном виде [13, 27]. Для задач непрерывного времени, рассматривается вся временная шкала, на которой исходным наблюдениям X_i соответствует время до события T_i. Тогда, функции анализа выживаемости имеют следующий вид:

$$\begin{aligned} S(t) &= P(T \geq t), \\ f(t) &= -\frac{d}{dt}S(t), \\ h(t) &= -\frac{d}{dt}[lnS(t)]. \end{aligned} \tag{2}$$

При построении моделей непрерывного времени, часто задаются строгие допущения о распределении времени событий и дифференцируемости функции выживания на всей временной шкале.

В случае задачи дискретного времени, временная шкала дискретизируется по заданным временным интервалам (bins). Предполагается, что заранее заданы моменты времени $0 = \tau_0 < \tau_1 < ...\tau_n$ и выборочные моменты времени T сводятся к множеству τ. Тогда, функции анализа выживаемости имеют следующий вид:

$$\begin{aligned} f(t_j) &= \frac{S(t_j) - S(t_j + \Delta t)}{\Delta t}, \\ S(t_j) &= \prod_{k=1}^{j}[1 - h(t_k)] = \sum_{k>j} f(t_k), \\ h(t_j) &= \frac{f(t_j)}{S(t_{j-1})}. \end{aligned} \tag{3}$$

Хотя формулировка с дискретным временем удобна и потенциально требует меньше допущений, чем подход с непрерывным временем, она также приводит к потере информации из-за дискретизации времени. Кроме того, количество временных интервалов для дискретизации является гиперпараметром и сильно влияет на точность и вычислительную сложность построения модели. Наконец, задача дискретого времени накладывает серьезное функциональное ограничение на допустимые моменты времени для прогнозирования.

1.1.4 Особенности данных

Информативность цензурирования

Процесс цензурирования наблюдений напрямую зависит от постановки исследования [28]. Например, в области здравоохранения, цензурирование пациентов связано с потерей наблюдения (например, при переводе в другую больницу) или самостоятельным уходом из исследования. Причиной ухода пациента могут служить побочные реакции, токсичность лечения, отсутствие улучшения, раннее выздоровление или другие факторы [29].

Отметим, что для построения моделей анализа выживаемости используется только наблюдаемые признаки при входе в исследование. Под внешними факторами будем подразумевать признаки наблюдения, недоступные для сбора из-за ограничений ресурсов исследования.

Цензурирование называется неинформативным, если причины цензурирования не связаны с проведением исследования, и информативным, если причины связаны с неучтенными факторами исследования. В частности, в литературе рассматривается три вида связи между временем цензурирования и временем наступления события: полная независимость, условная независимость и зависимость. Предположение о полной независимости означает, что вероятность цензурирования постоянна для всех моментов времени и не зависит от внешних факторов. При условной независимости вероятность цензурирования зависит только от времени и признаков наблюдения [30, 31]. Предположение зависимости основано на влиянии внешних факторов на вероятность цензурирования [32].

Предположение неинформативности цензурирования широко распространено при построении статистических моделей анализа выживаемости и может приводить к смещению прогнозов моделей [29, 31, 33].

Гетерогенность признакового пространства

На этапе извлечения признаков наблюдений необходимо учесть множество разнородных факторов, позволяющих достаточно полно описать состояние субъекта на момент входа в исследование. Непрерывные признаки однозначно отображаются на числовую шкалу и описывают дискретные и вещественные показатели. В области здравоохранения, к непрерывным признакам относятся: возраст, температура, артериальное давление, клинические и биохимические показатели [15].

Категориальные признаки не имеют численного представления и определяют значение показателя на основе конечного набора категорий. Для номинальных категорий не определена операция сравнения, например, для диагноза пациента или группы риска. Для порядковых категорий определены операции сравнения и ранжирования, но не определены арифметические операции, например, для степени поражения легких [34]. Отдельно стоит упомянуть бинарные признаки, принимающие значения 0 и 1 (отсутствие и наличие фактора), которые используются для описания анамнеза пациента (например, наличие вредных привычек и перенесенных болезней) и некоторых схем лечения (например, проведение химиотерапии).

Некоторые модели машинного обучения работают только с непрерывными и бинарными признаками и неспособны обрабатывать категориальные значения напрямую [35]. В таких случаях применяется процесс кодирования категорий в числовые значения. Существует несколько способов кодирования категориальных переменных, таких как прямое кодирование (one-hot encoding), кодирование метками (label encoding) и другие.

Другой проблемой событийных данных является наличие пропусков. Отстутствие данных может быть связано с ограниченностью информации (например, при отказе пациента отвечать на конкретные вопросы), с логическими или механическими ошибками (например, при неисправности оборудования) или ограниченными ресурсами исследования (например, при отсутствии назначения на исследования) [15].

Для решения проблемы применяются методы удаления и импутации пропусков. Исключение пропущенных значений применятся только в случае большого количества данных и приводит к расширению доверительных интервалов статистических значений, чем в случае использования всех данных [36]. Методы импутации заполняют пропуски на основе известных значений. Частым используется импутация средним значением или медианой. Однако

данный подход может привести к смещению статистических оценок и сужению доверительных интервалов.

Распределение вероятностей времени наступления событий

В анализе выживаемости может наблюдаться «эффект задержки лечения» (delayed treatment effect) [37]. В задачах здравоохранения при госпитализации пациенту назначается схема лечения, однако реакция на лечение зависит от свойств организма. Длительность, интенсивность и эффективность лечения напрямую влияет на распределение времени событий. Аналогично, существуют случаи смещения распределение времени цензурирования. В опубликованных работах случаи раннего и позднего цензурирования связывают с временем входа в исследование [38], используемыми схемами лечения [39] и наличием скрытой статификации наблюдений по риску [28]. При построении моделей выживаемости необходимо учитывать распределения терминальных и цензурированных событий, определяя важность каждого момента в зависимости от исходных данных.

Также, в исследованиях [33, 40] рассматривается соотношение классов (терминальных и цензурированных) событий в данных. В частности, в некоторых наборах возникает проблема дисбаланса классов [41]. Например, при наблюдении за неизлечимо больными пациентами, находящихся на жизнеобеспечении, наблюдается доминирование терминальных событий [42,43]: из 9105 наблюдений были цензурированы 2904 пациента (доля летального исхода – 0.681). Напротив, при исследовании пациентов [15], госпитализированных с атеросклеротическим заболеванием сосудов и выраженными факторами риска атеросклероза, большая часть наблюдений была цензурирована: из 3873 наблюдений цензурировано 3416 (доля летального исхода – 0.119).

Построение моделей на данных с дисбалансом классов может приводить к ложным выводам (например, к искажению оценки эффекта лечения [29, 33]). Несбалансированные наборы данных часто делают прогнозные модели ненадежными, поскольку они имеют тенденцию фокусироваться на доминирующем классе и игнорировать редкий [40]. В частности, при доминировании цензурированных событий модели склонны завышать функцию выживания (приближая к константной 1) для уменьшения ошибки описания данных.

Таким образом, для построения моделей анализа выживаемости необходим механизм обработки категориальных признаков и пропущенных значений. Для обеспечения надежности моделей следует учитывать обе проблемы целевых переменных, анализируя распределения времени и дисбаланс цензурированных и терминальных событий. При различии распределений можно говорить об информативности цензурирования.

1.2 Метрики качества

1.2.1 Точечные метрики

Индекс согласованности

Индекс согласованности (Concordance Index, CI) [44, 45] является наиболее используемой метрикой для задач анализа выживаемости. CI измеряет долю верно упорядоченных пар по ожидаемому времени события среди всех сопоставимых пар в наборе данных. Значение интерпретируется как вероятность того, что прогнозируемое время события двух случайно

выбранных наблюдений сохраняет тот же относительный порядок, что и истинное время событий.

Лучшее значение CI равно 1 (при полностью верном упорядочивании), худшее значение равно 0 (все пары упорядочены неверно), а значение 0.5 отражает случайность отклика модели. Для расчета CI в данной работе используется следующая формула:

$$CI = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\hat{T}_j<\hat{T}_i}}{\sum_{i,j} 1_{T_j<T_i}}, \tag{4}$$

где T_k – истинное время наступления события, а $\hat{T}_k$ – прогноз времени наступления события.

Однако данная метрика основывается на точечном прогнозе времени наступления события и не позволяет оценить прогнозируемую функцию выживания в целом. Значение CI не изменяется при сдвиге функций выживания, хотя само прогнозируемое время будет сильно искажено по сравнению с истинным.

Оценка правдоподобия

В общем виде, функция правдоподобия (Likelihood, LL) представляет собой совместное распределение выборки, рассматриваемое как функция параметра. Пусть X – пространство признаков, T – время события, а θ – параметр, описывающий прогнозную модель. Следовательно, функция правдоподобия для выборки $\{X, T\}$ будет иметь следующий вид:

$$L(\theta \mid X, T) = \prod_{i=0}^{n} P_\theta(T_i \mid X_i)).$$

Уточним определение правдоподобия в терминах анализа выживаемости. Как говорилось ранее, δ – флаг наступления события. В случае использования оценок функции выживания $\hat{S}$ и функции риска $\hat{h}$ прогнозной модели, говорят о «полном правдоподобии» (Full Likelihood) [46]:

$$L(\hat{S}, \hat{h} \mid X, T) = \prod_{i}^{n} \hat{h}(T_i|X_i)^{\delta_i} \hat{S}(T_i|X_i).$$

Для решения задачи оптимизации часто используется логарифмический вид:

$$\log L(\hat{S}, \hat{h} \mid X, T) = \sum_{i}^{n} \delta_i \log(h(T_i|X_i)) + \log(S(T_i|X_i)).$$

В случае использования только оценки функции риска $\hat{h}$ прогнозной модели, говорят о «частичном правдоподобии» (Partial Likelihood) [46]:

$$L(\hat{S}, \hat{h} \mid X, T) = \prod_{i}^{n} \frac{\hat{h}(T_i|X_i)}{\sum_{j \in T_k \geq T_i} \hat{h}(T_i|X_j)}.$$

И соответствующий логарифмический вид:

$$\log L(\hat{S}, \hat{h} \mid X, T) = \sum_{i}^{n} \log(\hat{h}(T_i|X_i)) - \log(\sum_{j \in R_i} \hat{h}(T_i|X_j)).$$

1.2.2 Интегральные метрики

Интегральная площадь под ROC-кривой

Альтернативной метрикой оценки качества ранжирования является интегральная площадь под ROC-кривой (Integrated AUC, $IAUC$), предложенная в работе [47]. В работе представлен метод распространения вычисления ROC-кривой [48] и площади под кривой (AUC) на многоклассовые или временные случаи. Для каждого момента времени t определяются два множества наблюдений, для которых событие произошло до и после момента t. Метрика $\widehat{AUC}(t)$ измеряет взвешенную долю пар наблюдений из каждого множества, имеющих согласованный порядок рисков (наблюдения с наступившими событиями должны иметь больший риск в момент t):

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(T_j > t)I((T_i \le t)\cdot\delta_i)w_i I(\hat{h}(t|X_j) \le \hat{h}(t|X_i))}{(\sum_{j=1}^{n} I(T_j > t))(\sum_{i=1}^{n} I((T_i \le t)\cdot\delta_i)w_i)}, \tag{5}$$

где $\hat{h}(t|X_i)$ – оценка кумулятивного риска для наблюдения X_i в момент времени t, $w_i = G(T_i) = P(C > T_i)$ – вероятность цензурирования наблюдения X_i, построенная на основе непараметрической оценки Каплана–Мейера (раздел 1.3.2). Для агрегации оценок $\widehat{AUC}(t)$ по всем моментам времени используется метрика Integrated AUC:

$$IAUC(t_{min}, t_{max}) = \frac{1}{\hat{S}(t_{min}) - \hat{S}(t_{max})} \int_{t_{min}}^{t_{max}} \widehat{AUC}(t) d\hat{S}(t). \tag{6}$$

Расстояние Кульбака–Лейблера

Классический вид расстояния Кульбака–Лейблера (Kullback-Leibler divergence, KL) был предложен в работе [49] для оценки расстояния между двумя вероятностными распределениями. В работе [50] был представлена модификация расстояния Кульбака–Лейблера функций выживания (Kullback-Leibler divergence of Survival functions, KLS), определяющая метрику KL в терминах анализа выживаемости. Пусть $G_n(t)$ – эмпирическая оценка функция $S(t)$ для n наблюдений. $F(t)$ – истинная $S(t)$, определенная семейством двухпараметрической функции выживания Вейбулла: $F(x) = \exp\left[-\left(\frac{x}{\sigma}\right)^m\right], x \ge 0, m, \sigma > 0$. Тогда модификация KLS определяется по следующей формуле:

$$KLS(G_n||F) = \int_0^\infty G_n(t)\log\left(\frac{G_n(t)}{F(t)}\right) - [G_n(t) - F(t)]dt.$$

Главным недостатком метрики является использование непараметрической оценки $G_n(t)$ для оценки близости к теоретической функции Вейбулла $F(t)$. В таком случае, KLS не позволяет оценивать качество частных прогнозов наблюдений, основанных на признаках X_i. Также, KLS не учитывает информацию о цензурировании наблюдений.

Интегрированная оценка Брайера

Интегрированная оценка Брайера (Integrated brier score, IBS) [51,52] основана на расчете квадратичного отклонения прогнозируемой функции выживания от истинной. Истинная функция выживания равна 1 до момента наступления события и 0 после.

Для оценки качества прогноза в фиксированный момент времени t используется метрика BS (Brier score), основанная на следующей формуле:

$$BS(t) = \frac{1}{N}\sum_i \begin{cases} \frac{(0-S(t|X_i))^2}{G(T_i)}, & \text{if } T_i \le t, \delta_i = 1, \\ \frac{(1-S(t|X_i))^2}{G(t)}, & \text{if } T_i > t, \\ 0, & \text{if } T_i = t, \delta_i = 0, \end{cases} \tag{7}$$

где $S(t|X_i)$ – прогноз функции выживания в момент t для наблюдения x_i с временем наступления события T_i. Таким образом, для фиксированного момента t и наблюдения x_i верны следующие утверждения:

- Если событие произошло до момента t, то ожидается низкая вероятность выживания (близкая к 0);
- Если событие произошло после момента t, то ожидается высокая вероятность выживания (близкая к 1).

Для учета цензурированных данных определяется функция $G(t) = P(C > t)$ – оценка Каплана–Мейера (раздел 1.3.2), построенная на цензурированных наблюдениях (при построении оценки обращается флаг цензурирования).

В модифицированной оценке BS квадраты отклонений взвешиваются исходя из обратной вероятности цензурирования: $1/G(T_i)$, если событие происходит до момента t, и $1/G(t)$, если событие происходит после момента t. Наблюдения, цензурированные до момента t, не учитываются. Для агрегации оценок BS по всем моментам времени используется метрика integrated brier score:

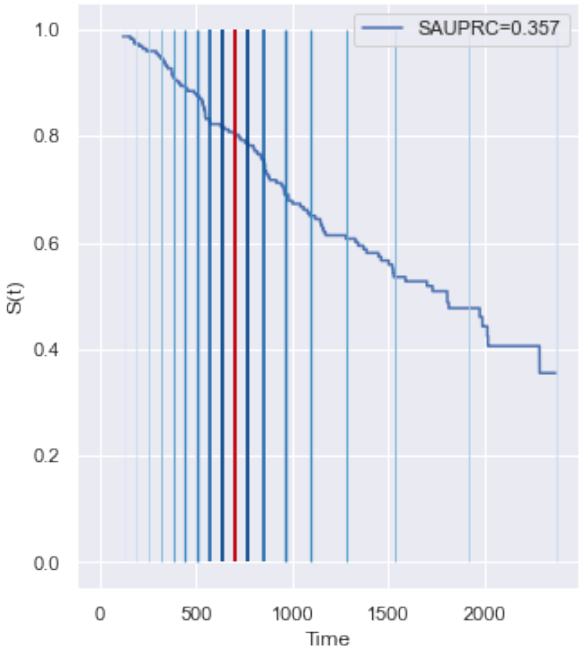
$$IBS = \frac{1}{t_{max}} \int_0^{t_{max}} BS(t)dt. \tag{8}$$

Площадь под кривой выживания точности-полноты

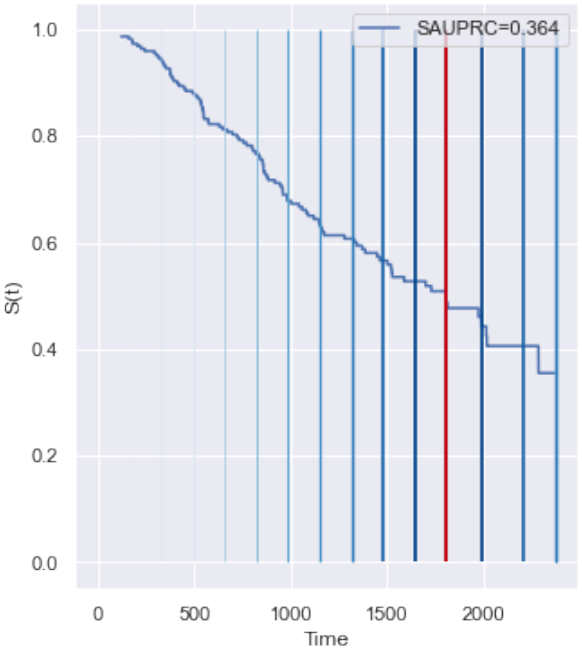
Оценка площади под кривой выживаемости точности-полноты (Survival Area Under Precision-Recall Curve, $AUPRC$) представлена в работе [53] и основана на идее измерения концентрации массы распределения вокруг истинного момента времени. В случае терминальных событий, оценка качества сводится к усреднению качества прогнозирования при разных промежутках $[T_i \cdot \varphi, T_i/\varphi]$, где $\varphi \in [0, 1]$.

$$AUPRC_{\delta=1}(\hat{S}, T_i) = \int_0^1 \hat{S}(T_i \cdot \varphi) - \hat{S}(T_i/\varphi)d\varphi = \int_0^1 P(T_i/\varphi > T > T_i \cdot \varphi)d\varphi. \tag{9}$$

На рисунке 4 представлен пример вычисления $AUPRC$ для двух терминальных наблюдений набора GBSG [17] с временем наступления события $T_i = 698$ (левый рисунок), $T_i = 1807$ (правый рисунок). В качестве прогноза функции выживания рассматривается оценка Каплана–Мейера (раздел 1.3.2), построенная на полных данных. Вертикальная красная линия отражает момент наступления события. Синие вертикальные линии основаны на уровне φ (толщина и цвет линии отражают величину значения) и расположены по обе стороны от момента наступления события T_i для времен $T_i \cdot \varphi$ и T_i/φ. Таким образом, линии отражают сравниваемые значения $\hat{S}(T_i \cdot \varphi), \hat{S}(T_i/\varphi)$ для расчета метрики при разных φ.



(a) Истинное время события $T_i = 698$



(b) Истинное время события $T_i = 1807$

Рис. 4: Пример расчета метрики $AUPRC$ для двух терминальных наблюдений набора GBSG с временем наступления события $T_i = 698$ и $T_i = 1807$.

В случае цензурированных событий, рассматривается только оценка функции выживания до момента наступления события:

$$AUPRC_{\delta=0}(\hat{S}, T_i) = \int_0^1 \hat{S}(T_i \cdot \varphi) d\varphi.$$

Лучшее значение 1 достигается в случае, когда функция выживания представляет собой пороговую функцию, равную 1 до наступления события и 0 после. Наименьшее значение 0 достигается, если функции выживания является константой (для терминальных событий любая константная функция, для цензурированных наблюдений константный 0).

Однако в работе [53] не описан способ агрегации метрики $AUPRC$ для множества наблюдений (формула (9) основана на оценке качества прогнозирования для одного наблюдения). Далее, в качестве метрики агрегации рассматривается среднее значение:

$$AUPRC = \frac{1}{N} \sum_{i=0}^{N} AUPRC_{\delta_i}(\hat{S}(t|X_i), T_i). \tag{10}$$

1.2.3 Мотивация выбора метрик качества

В таблице 1 представлены свойства рассмотренных метрик качества. Во-первых, каждая метрика качества позволяет оценивать одну из прогнозируемых величин: ожидаемое время события T, функцию выживания $S(t)$, функцию риска $h(t)$. В данной работе в качестве целевой функции выживания рассматривается пороговая функция, равная 1 до момента

наступления события и 0 после. Для цензурированных наблюдений, после момента цензурирования значение функции не определено.

Относительно метода сравнения величин, метрики могут быть классифицированы на метрики ранжирования (сравнение относительных значений) и метрики регрессии (сравнение абсолютных значений).

Также, метрики могут быть классифицированы на точечные метрики и интегральные метрики. Точечные метрики основаны на однократной оценке значений функций выживания и риска в момент наступления события. Интегральные метрики основаны на сравнении прогнозируемой функции выживания и риска с целевой функцией.

Наконец, необходимо различать обработку цензурированных и терминальных событий. В частности, из-за неопределенного поведения цензурированных наблюдений после момента выхода, результат сравнения прогнозов может быть отличен от реальной картины.

Таблица 1: Таблица свойств метрик качества.

Метрика	**Оценка**	**Ранжирование**	**Интегральная**	**Учет цензуры**
CI	T	Да	Нет	Да
IAUC	$h(t \mid x)$	Да	Да	Да
LL	$h(T \mid x)$	Нет	Нет	Да
KL	$S(t)$	Нет	Да	Нет
IBS	$S(t \mid x)$	Нет	Да	Да
AUPRC	$S(t \mid x)$	Нет	Да	Да

Таким образом, предлагается использовать метрики качества для оценки прогнозирования следующих величин:

1. T – метрика CI. Для цензурированных наблюдений в метрике рассматриваются только те пары наблюдений, событие которых наступило до момента цензурирования;
2. $h(t)$ – метрика $IAUC$. В отличие от метрики LL, $IAUC$ оценивает функцию выживания для всех моментов времени;
3. $S(t)$ – метрики IBS и $AUPRC$. В отличие от KL, метрики позволяет учитывать особенности цензурированных наблюдений. Также, метрика KL позволяет оценивать только безусловную функцию выживания $S(t)$, не зависящую от признакового пространства наблюдений.

1.3 Статистические методы

1.3.1 Таблицы времен жизни

Исторически первым методом описания данных по выживаемости были таблицы времен жизни (mortality table) [54]. Таблицу можно рассматривать как «расширенную» таблицу частот. Временная шкала событий разбивается на определенное число интервалов.

Для каждого интервала вычисляется число объектов, для которых в начале рассматриваемого интервала ещё не наступило целевое событие (переменная «Число в начале»), и число объектов, для которых событие наступило в данном интервале (переменная «Число

	Начало	Конец	Число в начале	Число изъятых	Число изучаемых	Число умерших	Доля умерших	Доля выживших
0	1.0	4.0	1808.0	167.0	1725.0	1009.0	0.585097	0.414903
1	5.0	8.0	632.0	72.0	596.0	283.0	0.474832	0.525168
2	9.0	12.0	277.0	30.0	262.0	85.0	0.324427	0.675573
3	13.0	16.0	162.0	21.0	152.0	38.0	0.250825	0.749175
4	17.0	20.0	103.0	18.0	94.0	19.0	0.202128	0.797872
5	21.0	24.0	66.0	6.0	63.0	14.0	0.222222	0.777778
6	25.0	28.0	46.0	9.0	42.0	6.0	0.144578	0.855422
7	29.0	32.0	31.0	5.0	29.0	8.0	0.280702	0.719298
8	33.0	36.0	18.0	3.0	17.0	5.0	0.303030	0.696970
9	37.0	40.0	10.0	5.0	8.0	0.0	0.000000	1.000000
10	41.0	44.0	5.0	2.0	4.0	0.0	0.000000	1.000000

Рис. 5: Пример таблицы времен жизни для 12 равномерных интервалов

умерших»). Также вычисляется число цензурированных объектов на каждом интервале – переменная «Число изъятых» На основе проведенных расчетов может быть вычислено число «изучаемых» объектов, которые были живы в начале рассматриваемого временного интервала, за вычетом половины от числа изъятых.

Далее, вычисляется «доля умерших» – отношение числа объектов, умерших в соответствующем интервале, к числу объектов, изучаемых на этом интервале. Наконец, вычисляется «доля выживших», равная единице минус «доля умерших». На рисунке 5 представлен пример таблицы времен жизни для 12 равномерных интервалов.

Для оценки функции выживания рассчитывается кумулятивная доля выживших к началу временного интервала. Поскольку вероятности выживания считаются независимыми на разных интервалах, доля равна произведению долей выживших объектов по всем предыдущим интервалам.

Пример оценки функции выживания с помощью таблиц времен жизни представлен на Рисунке 6. Слева представлена исходная функция выживаемости, а справа 3 оценки таблицы времен жизни для 5, 10 и 30 интервалов. Исходя из графика, видно, что предварительный выбор количества интервалов сильно влияет на приближение функции выживания. Наибольшее качество имеет оценка, построенная на 30 интервалах.

1.3.2 Метод Каплана–Мейера

Наиболее распространенным методом оценки функции выживания является метод Каплана–Мейера [55]. В данном методе для исходной выборки наблюдений рассматривается множество времен наступления события t_i. Для всех моментов времени рассчитывается: число оставшихся наблюдений N_i и число произошедших событий O_i на момент времени t_i.

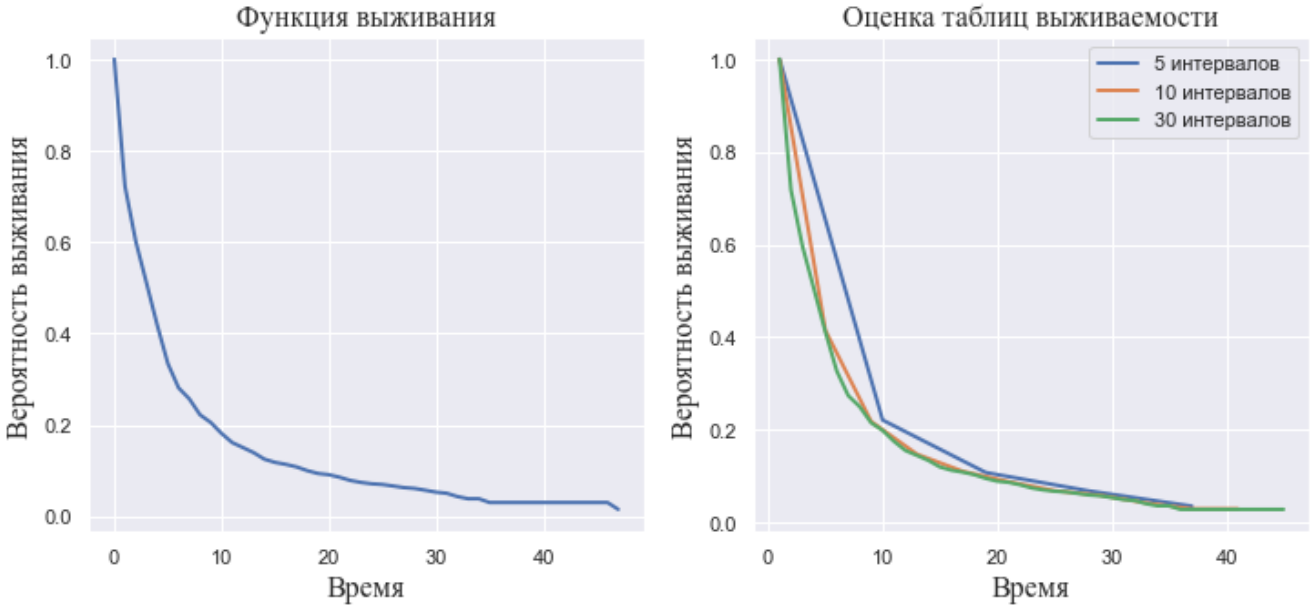


Рис. 6: Пример оценки функции выживания на основе таблиц времен жизни

В таком случае функцию выживания в момент t можно оценить кумулятивным произведением долей выживших по прошедшим моментам времени:

$$S(t) = \prod_{i:t_i \leq t} (1 - P(t_i)) = \prod_{i:t_i \leq t} \left(1 - \frac{O_i}{N_i}\right). \tag{11}$$

На основе оценки функции выживания также можно рассчитать ожидаемое время жизни $t_i : S(t_i) = 0.5$. Также, для оценки Каплана–Мейера может быть расчитано стандартное отклоненение σ_s:

$$\sigma_s = S(t)\sqrt{\sum_{i=0}^{t} \frac{O_i}{N_i(N_i - O_i)}}, \tag{12}$$

и доверительные интервалы для любого момента t и заданного уровня $\alpha : S(t) - \sigma_s Z_{1-\alpha/2} < S(t) < S(t) + \sigma_s Z_{1-\alpha/2}$, где $Z_{1-\alpha/2}$ – это $100(1-\alpha/2)$й перцентиль стандартного нормального отклонения. На Рисунке 7 представлен пример построения оценки функции выживания и её доверительных интервалов с помощью метода Каплана–Мейера.

Стоит отметить, что предположение неинформативности цензурировании может искажать прогноз непараметрической модели. В исследовании [31] отмечается, что метод Каплана–Мейера приводит к завышению оценки функции при положительной корреляции времени события и времени цензурирования, и занижению оценки функцию при отрицательной корреляции.

1.3.3 Метод Нельсона–Аалена

Наиболее распространенным методом оценки функции риска является метод Нельсона–Аалена [56]. Метод Нельсона–Аалена требует, чтобы наблюдения были независимыми. Во-

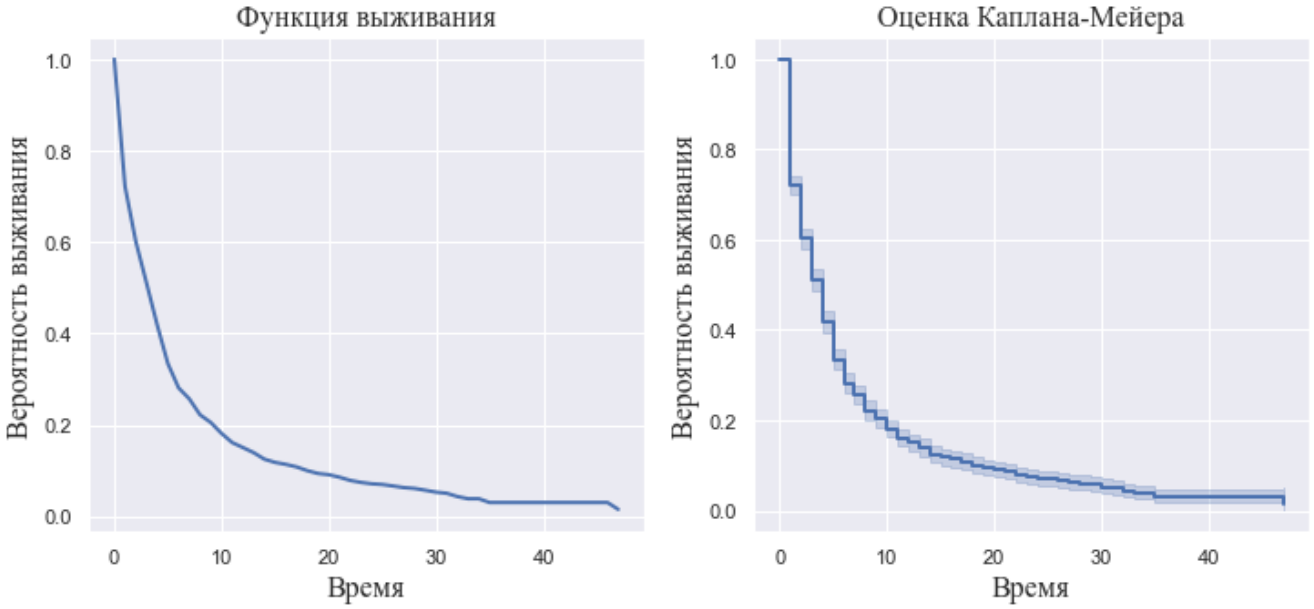


Рис. 7: Пример оценки функции выживания на основе метода Каплана–Мейера

вторых, цензурирование должно быть независимым — если рассматривать два случайных объекта в момент времени $t-1$, и один из них подвергается цензуре в момент времени t, а другой выживает, то оба объекта должны иметь равные шансы выжить в момент времени t.

Пользуясь терминами из раздела 1.3.2, кумулятивную функцию риска в момент t можно оценить кумулятивной суммой долей наступивших событий по прошедшим моментам времени:

$$\hat{H}(t) = \sum_{i:t_i \le t} P(t_i) = \sum_{i:t_i \le t} \left(\frac{O_i}{N_i} \right). \tag{13}$$

Также, можно вычислить ожидаемую вариацию:

$$\hat{V}(t) = \sum_{i:t_i \le t} \left(\frac{O_i}{N_i^2} \right).$$

При условии $\Delta\hat{H}(t_j) = \hat{H}(t_j) - \hat{H}(t_{j-1})$ и $\Delta\hat{V}(t_j) = \hat{V}(t_j) - \hat{V}(t_{j-1})$, оценка функции риска определяется по следующей формуле:

$$\hat{h}(t) = \frac{1}{b} \sum_{j=1}^{D} K_t \left(\frac{t - t_j}{b} \right) \Delta\hat{H}(t_j), \tag{14}$$

где $K_t(\cdot)$ – функция ядра [57], b – параметр пропускной способности (bandwidth), а суммирование происходит по D объектам, для которых событие наступило. Соответственно, можно вычислить ожидаемую вариацию:

$$\hat{\sigma^2}(\hat{h}(t)) = \frac{1}{b^2} \sum_{j=1}^{D} K_t \left(\frac{t - t_j}{b} \right)^2 \Delta\hat{V}(t_j).$$

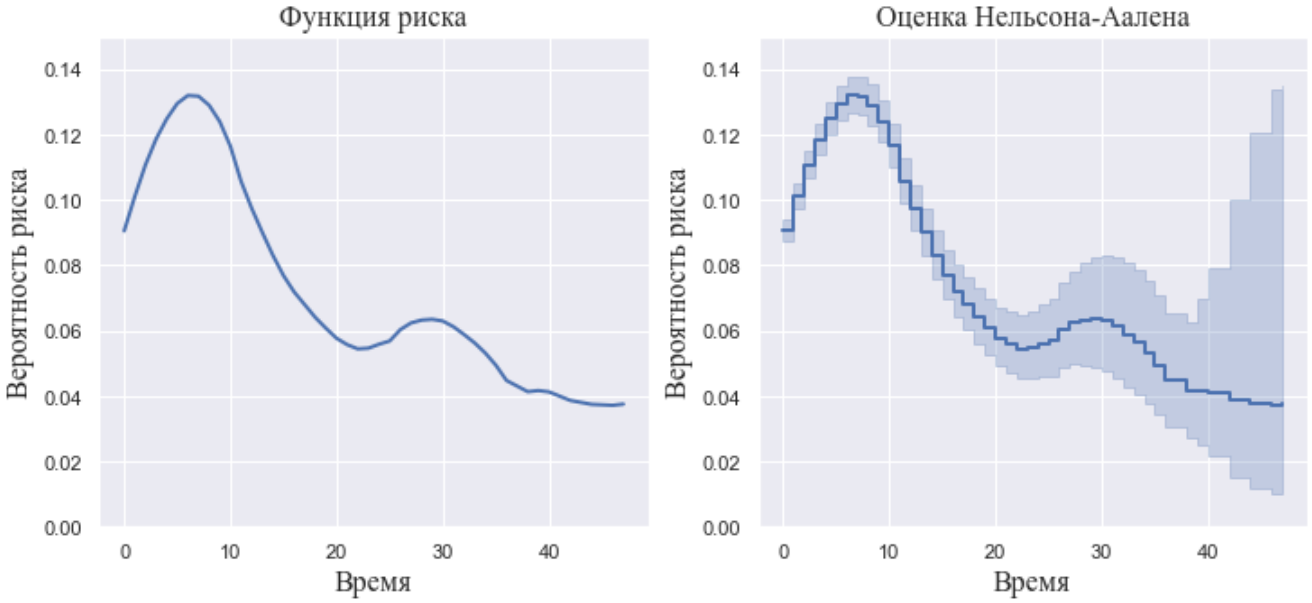


Рис. 8: Пример оценки функции выживания на основе метода Нельсона–Аалена

Точечные доверительные интервалы для функций опасности вычисляются с использованием метода, основанного на логарифмическом преобразовании:

$$\hat{h}(t) = \hat{h}(t) \cdot \exp\left(\pm \frac{Z_{1-\alpha/2}\hat{\sigma}(\hat{h}(t))}{\hat{h}(t)}\right),$$

где $Z_{1-\alpha/2}$ это $100(1-\alpha/2)$й перцентиль стандартного нормального распределения.

Пример оценки функции риска с помощью метода Нельсона–Аалена представлен на Рисунке 8. Для оценки также отмечены доверительные интервалы. Также, в литературе часто используется оценка Breslow [58], представляющая собой экспоненту отрицательной оценки Нельсона–Аалена.

1.3.4 Модель пропорциональных рисков Кокса

Непараметрические методы анализа выживаемости не позволяют оценить влияние признаков наблюдений на их целевые переменные времени и флага цензурирования. Для преодоления данного ограничения, полупараметрические и параметрические методы предполагают теоретическую связь между признаками и целевыми переменными.

Наиболее распространенным методом прогнозирования функции выживания является метод пропорциональных рисков Кокса (Cox Proportional Hazard, CoxPH) [59]. Метод основывается на допущении, что все наблюдения имеют одинаковую форму функции риска $h_0(t)$ и отличаются положительным коэффициентом пропорциональности $\alpha = X^T\beta$:

$$h(t \mid X) = h_0(t) \exp\left(X^T\beta\right), \tag{15}$$

где $h_0(t)$ – базовая функция риска, x — вектор признаков, β – вектор весов линейной модели. Функция $R(X) = \exp(X^T\beta)$ также называется функцией относительного риска. Базовые

функции риска $h_0(t)$ и кумулятивного риска $H_0(t)$ определяются на основе оценок Каплана–Мейера и Нельсона–Аалена (разделы 1.3.2 и 1.3.3 соответственно).

Прогноз индивидуальной функции выживания основан на расчете базовой функции выживания $S_0(t) = \exp(-H_0(t))$ с последующим смещением функции с учетом коэффициента пропорциональности:

$$S(t \mid X) = \exp\left(-H_0(t)\exp(X^T\beta)\right) = S_0(t)^{\exp\left(X^T\beta\right)}. \tag{16}$$

Пусть выборка состоит из N наблюдений и P признаков. Обучение модели проводится через подбор линейных коэффициентов β по методу максимального правдоподобия. Для выражения зависимости функции правдоподобия от линейных коэффициентов используются модифицированные функции частичного правдоподобия $L(\beta)$ и отрицательного логарифмического частичного правдоподобия $LL(\beta)$:

$$\begin{aligned} L(\beta) &= \prod_{j=1}^{N}\left[\frac{\exp(X_j\beta)}{\sum_{t_i\geq t_j}\exp(X_i\beta)}\right]^{\delta_j}, \\ LL(\beta) &= -\sum_{j=1}^{N}\delta_j\left(X_j\beta - \log\sum_{t_i\geq t_j}\exp(X_i\beta)\right). \end{aligned} \tag{17}$$

Вычисление коэффициентов производится по методу Ньютона-Рафсона [60], который итеративно находит оценки параметров, минимизируя $LL(\beta)$ с алгоритмической сложностью $O(NP^2)$. Также, для полученных весов β могут быть рассчитаны коэффициенты отношения рисков (hazard ratio) $\exp(\beta)$ для интерпретации важности признаков.

Применимость предположения пропорциональности рисков

Однако существуют клинические исследования, подвергающие критике применимость предположения о пропорциональности рисков на практике. В статье [61] анализируются результаты трех исследований и делается вывод относительно применимости предположения пропорциональных рисков.

В исследовании Texas Coronary Atherosclerosis Prevention Study пациентам с сердечно-сосудистыми заболеваниями было случайным образом назначено лечение или плацебо, причем отношение рисков составило 0,63 для лечения по сравнению с плацебо. Однако кумулятивная частота наступивших сердечно-сосудистых событий в двух группах была почти одинаковой (коэффициент риска близок к 1) в первых 6 месяцах наблюдения, а затем расходилась (коэффициент меньше 1). Таким образом, влияние терапии на сердечно-сосудистые события проявилось только после 6 месяцев.

Норвежское исследование рака прямой кишки случайно распределяло лиц в возрасте от 50 до 64 лет для скрининга с помощью гибкой сигмоидоскопии или без скрининга. Отношение рисков рака составило 0,80. В группе скрининга кумулятивный риск был выше 1 первые 5 лет наблюдения и ниже 1 после этого времени. Таким образом, скрининг на рак прямой кишки имел как немедленный эффект на выявление недиагностированного рака, так и отсроченный профилактический эффект.

В исследовании Women's Health Initiative женщины в постменопаузе случайным образом распределялись между гормональной терапией или плацебо. Отношение рисков ишеми-

ческой болезни сердца составило 1,24. В течение первого года отношение рисков составило 1,8 и 0,70 через 5 лет наблюдения. Исследователи приходят к выводу, что анамнез пациентов повлиял на восприимчивость к болезни, а риски не являются пропорциональными почти во всех клинических исследованиях. Исключением является случай, когда лечение не дает эффекта и коэффициент риска равен 1 на протяжении всего периода наблюдения.

Вторым ограничением применимости модели Кокса является непересекаемость прогнозируемых функций выживания. В клинической практике, пересекаемость функций выживания двух схем лечения говорит об отсутствии различий между схемами. В то же время, модель Кокса описывает прогноз через масштабирование базовой функции, позволяя прогнозам полностью совпадать или не пересекаться вовсе.

Утверждение 1. *Пусть X_1, X_2 – вектора признаков двух различных наблюдений, а для модели Кокса определен вектор весов β и базовая функция $S_0(t)$. Тогда индивидуальные прогнозы $S(t \mid X_1)$, $S(t \mid X_2)$ модели Кокса либо не пересекаются $\forall t : 1 > S_0(t) > 0$, либо совпадают для всех моментов времени t.*

Доказательство. Обозначим $\alpha_1 = X_1\beta$ и $\alpha_2 = X_2\beta$. Пусть индивидуальные прогнозы функции выживания $S(t \mid X_1)$ и $S(t \mid X_2)$ пересекаются в точке $t = \bar{t} > 0 : 1 > S_0(\bar{t}) > 0$, следовательно:

$$S(\bar{t} \mid X_1) = S(\bar{t} \mid X_2).$$

Прогноз описывается моделью Кокса и равенство имеет следующий вид:

$$S_0(\bar{t})^{\exp(\alpha_1)} = S_0(\bar{t})^{\exp(\alpha_2)}.$$

Основания правой и левой части $1 > S_0(\bar{t}) > 0$ равны и верно равенство степеней $\exp(\alpha_1) = \exp(\alpha_2)$. Аналогично, равенство оснований exp приводит к равенству аргументов $\alpha_1 = \alpha_2$. Следовательно, коэффициенты масштабирования базовой функции $S_0(t)$ для двух наблюдений X_1, X_2 равны. Таким образом, индивидуальные прогнозы функций выживания совпадают для всех моментов времени:

$$S(t \mid X_1) = S_0(t)^{\exp(\alpha_1)} = S_0(t)^{\exp(\alpha_2)} = S(t \mid X_2),\ \forall t.$$

□

Таким образом, метод имеет несколько значительных недостатков:

- Отношение двух функций риска для двух разных векторов признаков не изменяется во времени;
- Независимость значимости признаков от времени. В реальной клинической практике влияние факторов на риск может изменяться во времени. Например, после проведения операции пациент более подвержен риску, а после реабилитации более стабилен;
- Веса модели определяют линейную комбинацию исходных признаков;
- Не поддерживаются категориальные признаки и пропущенные значения.

1.3.5 Модель ускоренного времени отказа

Модель ускоренного времени отказа (Accelerated Failure Time, AFT) [62] основана на предположении масштабируемости функции выживания относительно коэффициента ускорения. В частности, коэффициент ускорения представляет собой константу γ для описания сдвига базовой функции выживания $S_0(t)$ на основе признакового пространства наблюдения A: $S_A(t) = S_0(\gamma t)$. Коэффициент ускорения используется для сравнения времени выживания двух групп и оценки значимости признаков.

Пусть X – матрица ковариат, β — вектор коэффициентов, $\sigma(\sigma > 0)$ – параметр масштабирования, ϵ – переменная ошибки, имеющая распределение величины $\log(T)$. Модель AFT предполагает случайность цензурирования и линейную связь между логарифмом времени выживания T и ковариатами:

$$\ln T = X\beta + \sigma\epsilon.$$

Также, модель предполагает, что известно теоретическое распределение ошибки ϵ. В Таблице 2 представлены возможные теоретические распределения для модели AFT и соответствующие оценки функции плотности, выживания и риска.

Таблица 2: Теоретические распределения модели AFT и соответствующие оценки функции плотности $f_0(t)$, функции выживания $S_0(t)$ и функции риска $h_0(t)$.

Распределение	**Плотность $f_0(t)$**	**Выживаемость $S_0(t)$**	**Риск $h_0(t)$**
Exponential	$\lambda exp(-\lambda t)$	$exp(-\lambda t)$	λ
Weibull	$\lambda p t^{p-1} exp(-\lambda t^p)$	$exp(-\lambda t^p)$	$\lambda p t^{p-1}$
Logistic	$\frac{e^{-(t-\mu)/\sigma}}{\sigma(1+e^{-(t-\mu)/\sigma})^2}$	$\frac{e^{-(t-\mu)/\sigma}}{1+e^{-(t-\mu)/\sigma}}$	$\frac{1}{\sigma(1+e^{-(t-\mu)/\sigma})^2}$
Log-logistic	$\frac{\lambda p t^{p-1}}{(1+\lambda t^p)^2}$	$\frac{1}{(1+\lambda t^p)^2}$	$\frac{\lambda p t^{p-1}}{1+\lambda t^p}$

Экспоненциальное распределение (Exponential) характеризуется только постоянной степенью риска λ. Большее значение λ указывает на более высокий риск и более короткий период выживания. Семейство распределений Вейбулла (Weibull) [63] характеризуется параметрами формы $(k > 0)$ и масштаба $\lambda > 0$. При k = 1 распределение совпадает с экспоненциальным, при k < 1, функция риска убывает с течением времени. В отличие от распределения Вейбулла, логистическое (logistic) и лог-логистическое (log-logistic) распределение допускает немонотонное поведение функции риска. Для логистического распределения определяются 2 параметра: μ – местоположение функции, а σ – параметр масштаба. Для лог-логистического распределения определяется параметр формы $k > 0$. При $k \leq 1$, функция риска монотонно убывает, однако при $k > 1$ является унимодальной. Оценки параметров распределений вычисляются на основе обучающей выборки.

Прогноз времени до события согласно модели AFT определяется по формуле: $T = e^{X\beta}e^{\sigma\epsilon}$. Обозначим: $T_0 = e^{\sigma\epsilon}$, тогда прогноз функции выживания $S(t|X)$, функции плотности $f(t|X)$, функции риска $h(t|X)$ имеют следующий вид:

$$S(t|X) = P(T > t) = P(e^{X\beta}T_0 > t) = P(T_0 > e^{-X\beta}t) = S_0(e^{-X\beta}t),$$

$$f(t|X) = -\frac{d}{dt}S(t|X) = -e^{-X\beta}S_0'(e^{-X\beta}t) = -e^{-X\beta}f_0(e^{-X\beta}t),$$

$$h(t|X) = \frac{f(t|X)}{S(t|X)} = e^{-X\beta}\left(-\frac{S_0'(e^{-X\beta}t)}{S_0(e^{-X\beta}t)}\right) = e^{-X\beta}\left(-\frac{d}{dt}\ln[S_0(e^{-X\beta}t)]\right) = e^{-X\beta}h_0(e^{-X\beta}t).$$

Обучение AFT основано на поиске параметров β и σ с помощью метода максимального правдоподобия. При условии $w_i = \frac{t_i - x_i\beta}{\sigma}$, функция правдоподобия имеет вид:

$$L(\beta, \sigma) = \prod_{i=1}^{n} \left(\sigma^{-1}h_0(w_i)\right)^{\delta_i} S_0(w_i).$$

Утверждение 2. *Пусть X_1, X_2 – вектора признаков двух различных наблюдений, β – вектор весов модели AFT. Тогда индивидуальные прогнозы $S(t \mid X_1)$, $S(t \mid X_2)$ модели AFT либо не пересекаются, либо совпадают для всех моментов времени t.*

Доказательство. Обозначим $\alpha_1 = X_1\beta$ и $\alpha_2 = X_2\beta$. Пусть индивидуальные прогнозы функции выживания $S(t \mid X_1)$ и $S(t \mid X_2)$ пересекаются в точке $t = \bar{t} > 0$, следовательно:

$$S(\bar{t} \mid X_1) = S(\bar{t} \mid X_2).$$

Прогноз функции выживания описываются моделью AFT и равенство принимает вид:

$$S_0(e^{-\alpha_1}\bar{t}) = S_0(e^{-\alpha_2}\bar{t}).$$

Функция S_0 монотонна и можно перейти от равенства значений к равенству аргументов: $e^{-\alpha_1}\bar{t} = e^{-\alpha_2}\bar{t}$. Поскольку $\bar{t} > 0$, сократим обе части на $\bar{t}$. Также, от равенства оснований перейдем к равенству аргументов $\alpha_1 = \alpha_2$. Следовательно, коэффициенты сдвига базовой функции $S_0(t)$ для двух наблюдений X_1, X_2 равны. Таким образом, индивидуальные прогнозы функций выживания совпадают для всех моментов времени:

$$S(t \mid X_1) = S_0(e^{-\alpha_1}t) = S_0(e^{-\alpha_2}t) = S(t \mid X_2),\ \forall t.$$

□

Таким образом, метод AFT имеет несколько значительных недостатков:

- Строгое предположение теоретического распределения ошибки;
- Предположение масштабируемости функций выживания по времени ограничивает возможность пересечения прогнозов функций выживания;
- Веса модели определяют линейную комбинацию исходных признаков;
- Не поддерживаются категориальные признаки и пропущенные значения.

1.4 Методы построения деревьев решений

Для преодоления ограничений строгости теоретических предположений (пропорциональность риска, заданное распределение признаков и целевых переменных) могут быть использованы методы, основанные на построении деревьев решений [64–66]. В основе методов лежит алгоритм рекурсивного разбиения признакового пространства на области с близким значением целевой переменной. Разбиение производится по некоторому статистическому критерию, определяющего близость дочерних выборок или прирост качества при разбиении. В

конечных узлах дерева (листах) строятся оценки целевой переменной. Прогнозом модели является оценка целевой переменной в листовом узле, определяемом на основе значений признаков наблюдения.

Модели могут быть применены к различным задачам: классификация, регрессия, поиск аномалий. Для каждой задачи выбирается свой критерий разбиения выборки для расчета близости разбиений по значениям исходных признаков. В случае анализа выживаемости могут быть использованы статистические log-rank критерии (раздел 1.4.1). В разделе 1.4.2 описан современный подход построения бинарного дерева выживаемости.

1.4.1 Критерии разбиения

На этапе поиска лучшего разбиения выборки с цензурированием необходимо решить одну из следующих задач оптимизации:

1. Минимизация ошибки описания данных. Ошибка рассчитывается для непараметрического прогноза родительской и дочерних выборок и вычисляется прирост качества при использовании разбиения. Лучшее разбиения определяется по максимальному приросту качества;
2. Максимизация различий между выборками. Критерии не вычисляют качество описания данных, а сравнивают характеристики самих дочерних выборок;

Наиболее популярны следующие критерии минимизации ошибки описания данных:

1. Exponential log-likelihood loss [65] вычисляет логарифм правдоподобия при условии постоянства функции риска, а лучшее разбиения максимизирует правдоподобие. Критерий использует строгие предположения и точечную оценку правдоподобия;
2. Relative Risk [67] определяет полное правдоподобие через предположение пропорциональности рисков, а лучшее разбиение максимизирует сумму логарифмических правдоподобий дочерних узлов. Критерий использует предположение Кокса и точечную оценку правдоподобия;
3. Impurity [68] определяет весовую схему смеси ошибки описания терминальных событий (через сумму квадратных отклонений истинного времени в выборке от среднего) и цензурированных событий (через энтропию распределения событий). Лучшее разбиение минимизирует значение критерия. При использовании impurity вычисляется только точечная оценка качества, а также требуется дополнительный подбор весов;
4. Square error of Cox residuals [69, 70] вычисляет сумму квадратных отклонений остатков модели Кокса в каждой выборке. Лучшее разбиение минимизирует сумму ошибок по дочерним выборкам. Критерий основан на предположении Кокса.

Для измерения различий между выборками с цензурированием наибольшее распространение получил L1-Wasserstein [64] и семейство статистик log-rank [71]. Критерий L1-Wasserstein вычисляет площадь между оценками функции выживания, построенных на дочерних выборках, а лучшее разбиение максимизирует площадь. Используя для оценки функции метод Каплана–Мейера, метрика наследует предположения метода.

Наконец, статистический критерий log-rank является модификацией критерия Кокрана–Мантеля–Хензеля (Cochran–Mantel–Haenszel, CMH) [72] для применения к цензурированным данным. Большее значение статистики log-rank определяет большее различие

между двумя выборками. Нулевая гипотеза критерия предполагает H_0, что функции риска двух выборок совпадают $h_1(t) = h_2(t)$.

Рассмотрим две группы с n_1 и n_2 наблюдениями. По двум группам определим упорядоченный набор времен наступления событий: $\tau_1 < \tau_2 < ... < \tau_K$. Пусть $N_{1,j}$ и $N_{2,j}$ — количество наблюдениями на момент τ_j, а $O_{1,j}$ и $O_{2,j}$ — количество событий в момент τ_j.

Тогда общее число наблюдениями и событий на момент τ_j: $N_j = N_{1,j} + N_{2,j}$ и $O_j = O_{1,j}+O_{2,j}$ соответственно. Определим ожидаемое число событий на момент τ_j как $E_{i,j} = \frac{N_{i,j}O_j}{N_j}$. На основе имеющихся данных можно рассчитать статистику log-rank критерия:

$$LR = \frac{\sum_{j=1}^{K} w_j \left(O_{1,j} - E_{1,j}\right)}{\sqrt{\sum_{j=1}^{K} w_j^2 E_{1,j} \left(\frac{N_j - O_j}{N_j}\right) \left(\frac{N_j - N_{1,j}}{N_j - 1}\right)}}, \tag{18}$$

где $w_j = 1$. Квадрат взвешенной статистики log-rank имеет распределение хи-квадрат. Тест log-rank применяется для обнаружения различий в выборках, в которых функции риска пропорциональны друг другу.

В исследованиях [71, 73] высказывается предположение о плохой чувствительности log-rank к особенностям реальных данных, основанных на раннем возникновении событий. Критерий log-rank основывается на предположении, что индикатор цензурирования не связан с прогнозом, вероятности выживания одинаковы для событий на ранних и поздних этапах исследования.

В работе [74] представлено сравнение качества критериев разбиения при построении дерева выживаемости. Рассматривались 3 случая теоретического поведения функции риска: константный риск, снижение риска во времени, повышение риска во времени. В случае константного риска, наилучшие результаты показали критерии: log-rank, exponential log-likelihood loss, relative risk. В случае снижения риска, лучшее качество имеет критерий log-rank и square error of Cox residuals. Наконец, в случае повышения риска рекомендуется использовать exponential log-likelihood loss и impurity.

В силу наличия строгих предположений критериев exponential log-likelihood loss, relative risk и square error of Cox residuals, рекомендуется использовать семейство статистик log-rank для поиска лучшего разбиения выборки с цензурированием. Важно отметить, что чувствительность критерия к определенному интервалу наступления событий может быть повышена с помощью модификации весовой схемы.

1.4.2 Метод построения дерева выживаемости

В статье [66] был представлен метод построения дерева выживаемости, основанный на идее рекурсивного разделения выборки на группы с разной выживаемостью. Дерево строится, начиная с корневого узла, который содержит все данные. Корневой узел разделяется на два дочерних узла на основе критерия разбиения. Далее, разделяется каждый дочерний узел. Процесс повторяется рекурсивным образом для каждого последующего узла.

Предложенный в статье [66] алгоритм основан на построении бинарного дерева. При разбиения узла, в подходе рассматриваются всевозможные промежуточные значения по каждому признаку из множества X. Для каждого промежуточного значения строятся 2 ветви разбиения, в которых, на основе целевых признаков T и E, вычисляется значение крите-

рия. После расчета всевозможных вариантов разбиения, в подходе выбирается разбиение с максимальным значением статистики.

Для измерения различий между функциями выживания двух групп наибольшее распространение получил критерий log-rank [71]. Большее значение статистики log-rank определяет большее различие в функциях выживания двух выборок. Разбиение выбирается по максимальному значению статистики. Нулевая гипотеза критерия предполагает, что различия в выживаемостях двух выборок отсутствуют.

Прогноз функции выживания для наблюдения с вектором признаков X вычисляется на основе данных, которые находятся в том же листе (конечном узле), что и X, с помощью оценки Каплана–Мейера.

Преимуществом метода является сильная интерпретация. Для каждого полученного листа соответствует набор правил, получаемый при проходе от корня к листу. Таким образом, если глубина дерева не слишком велика, экспертом может быть проанализирован набор правил для каждого листа на предмет логичности и корректности.

Однако метод имеет значимые недостатки. Во-первых, построение дерева возможно только на заполненных данных. Во-вторых, при отсутствии ограничений на количество наблюдений в узле дерево решений имеет склонность к переобучению. Наконец, для построения дерева решений, имеющего высокую точность прогнозирования, необходимо достаточное количество данных. В случае ограниченных данных, модель дерева решений часто используют в качестве базовой «слабой» модели при ансамблировании.

1.5 Регрессионные методы машинного обучения

1.5.1 Нейронные сети

Нейронная сеть имитирует функционирование биологической нейронной системы мозга. Единицей обработки информации является нейрон, который получает на вход набор сигналов, проводит преобразования и выдает результат. Поскольку модель нейрона реализует функцию от его входов, нейроны можно объединять в соответствии с правилами суперпозиции функций, получая более сложные модели, называемые персептронами или искусственными нейронными сетями прямого распространения. Стоит отметить, что нейросетевые модели позволяют обрабатывать только непрерывные значения входных и целевых переменных.

Наибольшее распространение получили регрессионные нейросетевые модели [13, 34, 75]. Откликом моделей является точечная оценка вероятности выживания или времени события. Однако функциональность таких моделей ограничена из-за невозможности прогнозирования функций анализа выживаемости.

Нейросетевые модели непрерывного времени [27, 42, 46, 75] основаны на расширении статистических моделей анализа выживаемости. Гибкость архитектуры позволяет описывать нелинейные зависимости между ковариатами и вероятностью события. Однако такие модели наследуют строгие статистические предположения.

Дискретные нейросетевые модели [27, 75, 76] не использует статистические предположения, но имеет ограниченные функциональные возможности прогнозирования из-за дискретной шкалы. В частности, необходим дополнительный подбор количества и размера интервалов временной шкалы, от которых зависит вычислительная сложность модели.

Модели непрерывного времени

В работах [27, 42] предложены методы расширения модели пропорциональных рисков Кокса с помощью многослойного персептрона. Откликом нейронной сети является значение $g(X)$ для расчета относительного риска $\exp(g(X))$. Наконец, модифицированная модель Кокса имеет следующий вид:

$$h(t \mid X) = h_0(t) \exp\left(g(X)\right). \tag{19}$$

При обучении моделей, параметризованных нейронной сетью, функция потерь минимизируется с помощью стохастического градиентного спуска [77] и дополняется штрафом $\sum_{i:\delta_i=1} \sum_{T_j \geq T_i} |g(X_j)|$ на величину $g(X)$ (штраф входит с коэффициентом регуляризации λ). В качестве функции потерь, DeepSurv [42] использует отрицательный логарифм частичного правдоподобия Кокса (раздел 17), а Cox-CC — аппроксимацию правдоподобия для терминальных событий с меньшей вычислительной сложностью.

Модификацией Cox-CC является модель Cox-Time [27], которая включает в функцию $g(\cdot)$ дополнительную зависимость от времени события $g(t, X)$ и частично преодолевает предположение пропорциональности. Однако при обучении модель использует модификацию функции потерь в рамках предположения Кокса. Прогноз кумулятивной функции риска вычисляется через интегрирование по базовому и относительному риску для всех моментов времени.

По результатам исследований, нейросетевые модели показывают рост качества (по сравнению с моделью Кокса) при использовании многослойных архитектур и позволяют преодолеть линейную зависимость коэффициента масштабирования от ковариат. Для построения нейронных сетей применяются современные подходы регуляризации распада веса, пакетной нормализации, слоев dropout и функций активации [78].

Модели дискретного времени

В работе [76] нейронные сети применяются для решения задач дискретного времени анализа выживаемости. Модель DeepHit основана на архитектуре глубокой нейронной сети и масштабируема для случаев нескольких конкурирующих событий.

Откликом модели является вектор вероятностей события в моменты заданной временной шкалы. Архитектура состоит из 2 общих полносвязных слоев, 2 полносвязных слоев для каждого конкурирующего события, 1 выходного слоя с функций активацией softmax и dropout слоями между ними. Для обучения DeepHit используется комбинация функций потерь частичного логарифмического правдоподобия (раздел 1.2.1) и CI (раздел 1.2.1).

Альтернативным подходом [27] является параметризация дискретной функции риска и плотности относительно отклика нейронной сети. Модель LogisticHazard использует идею логистической связи функции риска от вероятности события $\phi(\cdot)$ в момент t_j: $h(t_j|X) = 1/(1+\exp(-\phi_j(X)))$. Модель PMF определяет связь между функцией плотности и вероятностью события через модификацию функции softmax: $f(t_j|X) = \exp(\phi_j(X))/(1+\sum_k \exp(-\phi_k(X)))$.

В качестве функции потерь рассматривается частичное правдоподобие с учетом предполагаемой связи. Откликом нейросетевой модели является вектор вероятностей события $\phi_j(X)$ в t_j моменты времени. Таким образом, модель LogisticHazard имеет выходной логистический слой, а модель PMF — softmax слой.

Метод Piecewise Constant Hazard (PC-Hazard) является модификацией модели PMF и позволяет интерполировать значения функции выживания между дискретными моментами времени. Для интерполяции рассматриваются схемы постоянной плотности (CDI) и постоянного риска (CHI) во временном интервале. Схема CDI предполагает равномерное распределение событий в интервале и приводит к построению кусочно-линейной функции. Схема CHI предполагает, что в начале интервала наступает больше событий и приводит к кусочно-экспоненциальной функции выживания.

При работе с непрерывными данными, PC-Hazard дискретизирует истинное время по равномерной сетке или квантилям. В качестве функции потерь используется дискретная модификация частичного правдоподобия. Для обеспечения неотрицательности откликов выходной слой имеет функцию активации softplus.

Архитектура всех трех моделей LogisticHazard, PMF, PC-Hazard состоит из 8 полносвязных слоев с функциями активации ReLU, пакетной нормализацией и слоями dropout. Прогноз функции выживания определяется через кумулятивное произведение вероятностей ненаступления событий за прошедшие интервалы времени.

1.5.2 Метод опорных векторов

Метод опорных векторов (SVM) [79–82] черезвычайно популярен при решении задач классификации и регрессии. Метод классификации основан на идее поиска гиперплоскости, оптимально разделяющей наблюдения разных классов. Поиск гиперплоскости производится в пространстве признаков высокой размерности, в которое исходные наблюдения неявно отображаются с помощью ядерной функции.

В рамках анализа выживаемости используются три реализации метода опорных векторов: модель регрессии [80, 81], модель ранжирования [13] и гибридный подход [79]. Модель регрессии опорных векторов (SVR) направлена на поиск функции, которая описывает наблюдаемое время события с минимальной ошибкой. Для адаптации модели к цензурированным данным, определяется асимметричная функция потерь, которая определяет штраф в зависимости от типа события [81]. Например, для цензурированных наблюдений ошибка равна нулю, если прогноз времени был завышен.

В модели ранжирования [13] используются дополнительные ограничения для поддержания верного порядка наблюдений. Качество ранжирование определяется на основе метрики согласованности, описанной в разделе 1.2.1. Однако вычислительная сложность алгоритма квадратична относительно размера выборки. Кроме того, метод фокусируется только на упорядочении наблюдений и игнорирует фактические значения отклика.

Гибридный подход объединяет ограничения регрессии и ранжирования для решения задачи квадратичной оптимизации [79]. Для снижения вычислительной нагрузки, каждое наблюдение сравнивается только с ближайшим соседом, а не со всеми наблюдениями выборки. Результаты исследований показали, что качество метода ранжирования значительно ниже методов регресии и гибридного подхода.

Таким образом, существующие реализации метода опорных векторов прогнозируют только точечную оценку времени и вероятности события.

1.5.3 Байесовские методы

Теорема Байеса является одним из наиболее фундаментальных принципов теории вероятностей и математической статистики, обеспечивая связь между апостериорной и априорной вероятностью. В рамках машинного обучения распределение, полученное с учётом данных, называется апостериорным. Переход от априорного распределения к апостериорному отражает обновление нашего представления о параметрах распределения с учётом полученной информации. Байесовские модели используют внешние знания об априорных распределениях параметров моделей для расширения прогноза с точечной оценки на апостериорное распределение целевой переменной при условии наблюдаемых данных. Выбор априорного распределения напрямую влияет на качество итоговых моделей прогнозирования. Также, байесовские модели позволяют делать предположения о вероятности параметров и генерировать новые данные из полученных распределений.

На данный момент, в литературе байесовским моделям анализа выживаемости уделяется гораздо меньше внимания, а существующее количество открытых реализаций очень ограничено [83]. Байесовские реализации непараметрических моделей сводят задачу к поиску оптимального распределения для генерации наблюдений. В частности, используется распределение Дирихле [83, 84] или решается дискретная задача построения гистограммы событий [85]. При этом, в силе остается предположение о неинформативности цензурирования.

Наибольшее распространение получили байесовские расширения параметрических моделей анализа выживаемости: модели пропорциональности Кокса и модели ускоренного времени отказа. Байесовская модель Кокса[1] предполагает априорные распределения линейных коэффициентов β и базового риска $h_0(t)$. Часто рассматривается нормальный закон распределения коэффициентов $\beta \sim N(\mu_\beta, \sigma_\beta^2)$, где $\mu_\beta \sim N(0, 10^2)$ и $\sigma_\beta \sim U(0, 10)$. Для задания априорного распределения базового риска рассматривается дискретный аналог риска $h_j \sim Gamma(10^{-2}, 10^{-2})$ для моментов времени τ_j. Апостериорный риск $h(t \mid X)$ является агрегацией распределений рисков в дискретные моменты времени.

Для байесовских моделей ускоренного времени отказа рассматривается априорное распределение линейных коэффициентов, а также распределения параметров выбранного семества [86]. В отличии от модели Кокса, базовые модели S_0 и h_0 уже имеют теоретическое распределения в модели AFT. Модели байесовского усреднения (BMA) [86] используют формулу полной вероятности для разложении вывода моделей при нескольких различных семействах распределений параметров. Альтернативной реализацией является байесовская AFT модель на основе сплайнов [83].

Таким образом, байесовские модели выживаемости обладают следующими недостатками. Во-первых, модели работают в рамках строгих предположений существующих статистических моделей. Во-вторых, задание априорных распределений накладывает дополнительные теоретические ограничения на использование моделей. Грубая ошибка в спецификации параметров для любого отдельного параметрического семейства снизит качество прогнозирования. Наконец, обучение и вывод байесовских моделей вычислительно сложнее классических моделей анализа выживаемости [83–86].

[1] https://www.pymc.io/projects/examples/en/2021.11.0/survival_analysis/survival_analysis.html

1.6 Методы ансамблирования алгоритмов машинного обучения

Для улучшения точности прогнозирования моделей применяются алгоритмы ансамблирования. Деревья решений хорошо подходят в качестве базовых моделей, так как способны точно описывать обучающую выборку.

Далее рассмотрим наиболее распространенные методы ансамблирования, имеющие аналоги в анализе выживаемости. Метод случайного леса [87] основан на построении ансамбля независимых деревьев решений. Для решения задач анализа выживаемости, необходимо заменить деревья решений на деревья выживания (раздел 1.6.1). Метод градиентого бустинга [88] основан на построении композиции алгоритмов, в которой каждый следующий алгоритм стремится компенсировать ошибки предыдущих (раздел 1.6.2).

Целью данного раздела является обзор и анализ применимости существующих методов и реализаций случайного леса и градиентого бустинга в области анализа выживаемости. Для решения проблемы применимости градиентного бустинга могут использоваться следующие подходы. Во-первых, задача может быть сведена к одной целевой переменной (например, к ожидаемому времени события) с последующим конструированием функций выживания и риска на основе теоретических предположений о распределении времени события.

Во-вторых, может быть поставлена дискретная задача (сведение функции выживания и риска к конечному вектору значений в заданные моменты времени). В таком случае, для каждого момента времени может быть построен свой градиентный бустинг. Альтернативным решением является оптимизация целевого вектора с помощью градиентного бустинга.

1.6.1 Ансамблирование независимых моделей

Наиболее распространенным методом ансамблирования является метод случайного леса. Развитие метода в области анализа выживаемости предложено в статье [87]. Алгоритм Random Survival Forest (RSF) основан на идее построения ансамбля деревьев выживания [66] и агрегации их прогнозов:

1. Строится N бутстреп выборок (с возвращением) из исходной выборки. Каждая бутстреп подвыборка в среднем исключает 37% данных, которые называются out-of-bag (OOB) выборкой;
2. На каждой бутстреп выборке строится дерево выживания [66], в каждом узле дерева выбирается P произвольных признаков для поиска лучшего разбиения. Выбирается разбиение, которое максимизирует разницу между дочерними узлами (в частности, максимизует значение статистики log-rank);
3. Деревья выживания строятся до исчерпания бутстреп выборки (без ограничения на глубину и количество наблюдений в листе).

Прогноз функции выживания для наблюдения с вектором признаков x вычисляется как среднее значение по всем прогнозам деревьев в ансамбле для всех моментов времени. Как было сказано ранее, прогнозом дерева выживания является оценка Каплана–Мейера, построенная на данных, которые находятся в том же листе, что и x. Усреднение прогнозов деревьев решений позволяет улучшить точность и избежать переобучения. При построении ансамбля доступны следующие параметры для вариации: количество деревьев в ансамбле N, размер бутстреп выборки, параметры контроля роста деревьев.

1.6.2 Бустинг ансамблирование моделей

Классический подход ансамблирования моделей градиентный бустинг (Gradient Boosting) был представлен в статье [88] и основывается на итеративном обучении новой модели на ошибках построенного ансамбля.

Целью алгоритма является минимизация заданной функции потерь $loss$. Пусть ансамбль H_0 инициализируется константой. Тогда на m итерации для каждого i наблюдения обучающей выборки производится расчет псевдо-остатков $r_{im} = -\frac{\partial loss(y_i, H_{m-1}(X_i))}{\partial H_{m-1}(X_i)}$, где $H_{m-1}(X_i)$ — индивидуальный прогноз ансамбля с предыдущей итерации для вектора X_i. Новая базовая модель использует псевдо-остатки r_{im} в качестве целевой переменной и строит отображение $h_m : X_i \to r_{im}$. При обновлении ансамбля $H_m(X) = H_{m-1}(X) + \gamma_m h_m(X)$ базовая модель входит с весом $\gamma_m = argmin_\gamma \sum_{i=1}^{N} loss(y_i, H_{m-1}(X_i) + \gamma h_m(X_i))$. Прогноз градиентного бустинга вычисляется как взвешенная сумма прогнозов базовых моделей. Хотя подход обеспечивает высокое качество прогнозирования, построенные модели не имеют строгой интерпретации и склонны к переобучению.

Применение градиентного бустинга к задачам анализа выживаемости ограничего из-за использования одной целевой переменной y_i, по которой производится расчет псевдо-остатков. В то же время, в анализе выживаемости используются две переменные (время события и флаг цензурирования). Также, градиентный бустинг является регрессионной моделью, а основной задачей анализа выживаемости является прогнозирование функции выживаемости и риска. Далее рассмотрим существующие реализации градиентного бустинга, применимых к анализу выживаемости.

Модели непрерывного времени

Метод Gradient Boosting Survival Analysis (GBSA) [80] основан на идее построения ансамбля регрессионных деревьев с функцией потерь, учитывающей цензурированность данных. Вид функции потерь определяет функциональные возможности модели.

При использовании отрицательного логарифмического правдоподобия с преобразованием пропорциональности риска (формула 17), прогнозом ансамбля является точечная оценка относительного риска $f(X)$ события. В таком случае, модель применяется для расширения классической модели пропорциональных рисков Кокса путем замены скалярного произведения $X\beta$ на отклик модели $f(X)$. Следуя обозначениям раздела 1.3.4, определяются прогнозы функции риска и функции выживания:

$$H(t \mid X) = H_0(t) \exp(f(X)),$$

$$S(t \mid X) = S_0(t)^{\exp(f(X))}.$$

При использовании в качестве функции потерь отрицательного логарифма частичного правдоподобия (Negative log-partial Likelihood) $loss = -\sum_i^n \delta_i f(X_i) + \log(\sum_{t_j \geq t_i} \exp(f(X_j)))$, предположение Кокса применяется для представления функций риска и выживания через отклик модели $f(X)$. При использовании эмпирических функций потерь $loss = 0.5 \cdot \sum_i (I((t_i > \hat{T}(X_i) | \delta_i) \cdot (t_i - \hat{T}(X_i)))^2$ и $loss = IPWC$ (обратная вероятность наступления события вычисляется по методу Каплана–Мейера для инвертированного флага цензурирования), модель GBSA позволяет прогнозировать ожидаемое время наступления события без использования

строгих предположений. Однако прогнозирование функций выживания и риска в таком случае не предусмотрено.

Модель GBSA имеет следующие недостатки. Во-первых, ансамбль состоит не из деревьев выживаемости, а из деревьев решений. В таком случае, при поиске лучшего разбиения выборки и построении листового прогноза участвует только одна целевая переменная (время события или флаг цензурирования). Во-вторых, прогнозирование функции выживания и риска возможны только в случае расширения модели Кокса. Такие модели наследуют проблемы оценки $Breslow$ и метода пропорциональных рисков Кокса. Наконец, модель основана на приближении относительного риска и ожидаемого времени события, а не функций выживания и риска.

Метод Component-wise Gradient Boosting (CWGBSA) [89] основан на идее обучения каждой базовой модели не на всем наборе признаков $X = x_1, ...x_p$, а только на одной переменной. Подход также называют бустингом на основе повышения правдоподобия, так как ансамбль максимизирует общую вероятность на каждой итерации, выбирая тот базовый алгоритм, который приводит к наибольшему увеличению вероятности.

Применяя подход к анализу выживаемости, целью градиентного бустинга является оптимизация функции потерь относительно линейных коэффициентов $\beta = w$ модели пропорциональных рисков Кокса. Градиент функции потерь g_m вычисляется по весам w. В таком случае, каждая функция потерь должна быть выражена через веса модели CoxPH (дополнительно накладывая условие пропорциональности рисков).

В качестве базовой модели используется метод наименьших квадратов. Следовательно, на каждой итерации определяется лучшая компонента $\theta_j = (X_j^T X_j)^{-1} X_j^T g_m$ сдвига линейных коэффициентов модели Кокса, минимизируя функцию потерь. Обновление весов производится по формуле $w_m = w_{m-1} + learning\_rate \cdot \theta_k$, где $learning\_rate$ — гиперпараметр скорости обучения. Откликом модели является $\eta = f(x, w) = x^T w$. Построение функций $S(t), h(t)$ осуществляется согласно формулам 16 и 15.

В качестве $loss$ может быть использована одна из следующих функций потерь:

1. Negative log-partial Likelihood $= -\log pl = -\sum_i^n \delta_i f(X_i, w) + \log(\sum_{t_j \geq t_i} \exp(f(X_j, w)))$;
2. Concordance index $= -P(\eta_i > \eta_j \mid t_i < t_j) = -\frac{\sum_{i,j} I(t_i > t_j) \cdot I(\eta_i > \eta_j)}{\sum_{i,j} I(t_i > t_j)}$;
3. Gehan Loss $= -\frac{1}{n} \sum_{i=1}^{n} \left[-\frac{1}{n} \delta_i \sum_{j=1}^{n} (r_i(w) - r_j(w)) I(r_i(w) \leq r_j(w)) \right], r_i(w) = \log t_i - w^T X_i$;
4. Brier Score $= \sum_{i=1}^{n} \left(\delta_i - S_0(t)^{\exp(f(X_i, w))} \right)^2$.

Метод CWGBSA имеет следующие недостатки. Во-первых, целевой переменной для базовых моделей являются только псевдо-остатки. Следовательно, как в случае модели GBSA [80], при обучении базовых алгоритмов не учитывается связь между целевыми переменными времени и флага цензурирования. Во-вторых, целью CWGBSA является подгонка линейных коэффициентов $\beta = w$ для модели CoxPH. При прогнозировании функций $S(t), h(t)$ наследуются проблемы метода пропорциональных рисков Кокса. Наконец, представленные функции потерь являются модификациями теоретических функций с помощью преобразования Кокса. Таким образом, при расчете градиента, на функцию потерь также накладываются ограничения пропорциональных рисков.

Метод CoxBoost [90] основан на идее оптимизации отрицательного частичного логарифмического правдоподобия со штрафом [91]: $pl_{pen}(w) = pl(w) - 0.5\lambda w^\top Pw$, где P — единичная матрица $p \times p$, а λ – коэффициент штрафного члена. Пусть $p < n$ и начальное значение коэффициентов w=(0,...,0). Итерация модели CoxBoost заключается в расчете коэффициентов $\hat{w}$, максимизирующих $pl_{pen}(\hat{w}|w)$. При добавлении компоненты смещения, логарифмическое правдоподобие в пространстве параметров «смещается» к началу координат, (а коэффициент λ влияет на скорость «смещения»).

Для расчета оптимальных коэффициентов рассмотрим функцию потерь для накопленных коэффициентов w и новых коэффициентов $\hat{w}$. Компонент смещения $\eta = X^\top w$ включен в логарифмическую вероятность для итеративного обновления оценки параметра. Таким образом, используется функция вида: $pl_{pen}(\hat{w}|w) = \sum_{i=1}^{n} \delta_i \left[\eta_i + X_i^\top \hat{w} - \log\left(\sum_{t_l \geq t_i} exp(\eta_l + X_l^\top \hat{w})\right)\right] - \frac{\lambda}{2}\hat{w}^\top P\hat{w}$. Оптимальные коэффициенты $\hat{w}$ выражаются на основе градиента и гессиана функции потерь и добавляются к компоненте смещения.

На практике наиболее распространена component-wise реализация, согласно которой процедура применяется только к p признакам (и вычисляются частичные логарифмические правдоподобия $pl(\hat{w}_j)$ соответственно). На каждой итерации ограниченные частичные логарифмические вероятности «смещаются» в сторону $\hat{w}_j$, получая ограниченные штрафные частичные логарифмические вероятности $pl_{pen}(\hat{w}_j|w)$.

Модель CoxBoost обладает двумя недостатками. Во-первых, аналогично модели CWGBSA, целью CoxBoost является подгонка коэффициентов $\beta = w$ для модели CoxPH. При прогнозировании функций $S(t), h(t)$ наследуются проблемы метода пропорциональных рисков Кокса. Во-вторых, используемая функция потерь со штрафом являются модификацией теоретического частичного логарифмического правдоподобия с помощью преобразования Cox PH. Как и в случае с моделью CWGBSA, на функцию потерь накладываются ограничения пропорциональных рисков Кокса.

Модели дискретного времени

Метод Gradient Boosting machine for concordance index (GBMCI) [92] основан на идее последовательной оптимизации метрики CI (раздел 1.2.1) и позволяет прогнозировать точечную оценку ожидаемого времени события. Ансамбль инициализируется моделью пропорциональных рисков Кокса и уточняет прогноз с помощью множества регрессионных деревьев решений. В качестве функции потерь используется сглаженная версия индекса согласованности $SCI = \frac{1}{|P|}\sum_{(i,j)\in P} \frac{1}{1+\exp(\alpha(F(X_i)-F(x_j))}$, где $F(x)$ — прогноз времени события, P — множество верно упорядоченных пар по истинному времени наступления события, α — гиперпараметр сглаженности. Алгоритм построения модели аналогичен градиентному бустингу.

Модель GBMCI обладает следующими недостатками. Во-первых, целевой переменной для базовых регрессионных деревьев являются псевдо-остатки. Как указывалось ранее (для алгоритмов GBSA и CWGBSA), такой подход обучения не позволяет выявить связь между целевыми переменными анализа выживаемости. Во-вторых, из-за использования модели Cox PH в качестве первичного приближения, GBMCI остается в рамках допущения пропорциональных рисков. Для преодоления данной проблемы достаточно использовать альтернативную модель первичного приближения, не основанную на строгих предположениях.

EXtreme Gradient Boosting (XGBoost) [93] — популярная библиотека алгоритмов машинного обучения, ориентированная на скорость вычислений и производительность моделей. Библиотека использует распределенные вычисления, графические процессоры и параллелизацию для ускорения работы. XGBoost также содержит методы построения моделей анализа выживаемости. Целевая переменная события выражается в виде диапазона нижней `y_lower_bound` и верхней границы `y_upper_bound` времени события. В частности, модель обрабатывает 4 вида наблюдений:

1. Терминальные: метка $[a, a]$, где a — истинное время события;
2. Цензурированные справа: метка $[a, \infty)$, где a — является нижней границей;
3. Цензурированные слева: метка $[0, b]$, где b — верхняя граница;
4. Цензурированные в интервале: метка $[a, b]$, где a и b — нижняя и верхняя границы соответственно.

Метод XGBoost основан на расширении модели ускоренного времени отказа (раздел 1.3.5), заменяя скалярное произведение $X\beta$ на отклик ансамбля регрессионных деревьев $F(X)$. XGBoost использует гистограммный подход и квантилизацию точек разбиения для поиска «приблизительного» (approximate) разбиения в деревьях решения. Гистограммный подход применяется в задачах классификации и регрессии и основан на итеративном расчете двух гистограмм для левой и правой ветви разбиения. Для каждой промежуточной точки разбения, частотности целевой переменной «вычитаются» из левой гистограммы и прибавляются к правой.

При построении бустинг ансамблей XGBoost использует алгоритм Ньютона-Рафсона [60]. В частности, если на итерации m был получен отклик ансамбля $H_{m-1}(X_i)$, то базовая модель f_m обучается с целью минимизации функции потерь: $loss_m = \sum_{i=1}^{n}[g_i f_m(X_i) + \frac{1}{2} h_i f_m^2(X_i)]$, где $g_i = \partial_{H_{m-1}(X_i)} loss(y_i, H_{m-1}(X_i))$ и $h_i = \partial^2_{H_{m-1}(X_i)} loss(y_i, H_{m-1}(X_i))$ — градиент и гессиан функции потерь в точке X_i. Расчет значений градиента и гессиана также используются на этапе поиска лучшего разбиения регрессионного дерева.

Модель XGBoost обладает следующими недостатками. Во-первых, бустинг ансамбль расширяет модель AFT, наследуя её недостатки и предположения. Во-вторых, ансамбль основан на агрегации регрессионных деревьев и использует информацию о цензурированности наблюдений только на этапе расчета функции потерь.

Метод Gradient Boosting Survival Tree (GBST) [94] основан на идее построения множества бустинг ансамблей деревьев выживания для каждого момента времени дискретной шкалы. В качестве функции потерь рассматривается отрицательное логарифмическогое правдоподобие с учетом предположения пропорциональности рисков: $loss = \sum_i \sum_j \log(1 + \exp(-y_j(t_i) f(\tau_j|X_i)))$.

Для каждого момента τ_j времени ансамбль решает собственную задачу и прогнозирует вероятность риска $\hat{h}(\tau_j|X)$. Для поиска оптимального веса прогноза базовой модели вычисляется градиент и гессиан для каждого наблюдения и момента времени. Данные значения также используются для поиска лучшего разбиения выборки. Таким образом, прогноз дискретной функции выживания имеет вид:

$$S(\tau_j|X_i) = \prod_j (1 - \hat{h}(\tau_j|X_i)).$$

Модель GBST имеет следующие недостатки. Во-первых, с помощью ансамбля решается задача дискретного времени и необходим дополнительный поиск оптимального числа интервалов для каждого набора данных. Во-вторых, редукция задачи приводит к решению множества бинарных задач для каждого момента времени. В таком случае напрямую не анализируется связь между временем до события и флагом цензурирования. Наконец, построение ансамбля является вычислительно сложной задачей. При масштабировании количества интервалов возрастают затраты по времени и используемой памяти.

Открытые сравнения ансамблей

В исследовании [95] рассматривается задача оценки влияния факторов на прогрессирование деменции. При моделировании высокоразмерных, гетерогенных клинических данных стандартные статистические модели показывают некачественные и нестабильные результаты. Для анализа ретроспективных данных используется анализ выживаемости. В работе оцениваются как статистические алгоритмы (CoxPH, Ridge, Lasso), так и бустинговые модели машинного обучения (CoxBoost, GLMBoost, XGBoost). Согласно выводам авторов, модель CoxBoost показывает лучшие результаты прогнозирования на двух наборах данных как с использованием, так и без отбора признаков.

В рамках статьи [91] проводилось сравнение популярных реализаций бустинг моделей библиотек mboost и CoxBoost языка R. Несмотря на общие принципы ускорения, метод GLMBoost является адаптацией бустинга на основе моделей, в то время как CoxBoost позволяет усилить базовую модель на основе правдоподобия. Согласно комментариям авторов, метод GLMBoost может привести к лучшим результатам, когда для набора определены важные факторы, коэффициенты которых не должны сильно изменяться. Напротив, подход CoxBoost позволяет изменять коэффициенты важных факторов с помощью бустинг ансамблей. Наилучший результат по скорости и качеству достигается с помощью подхода GLMBoost.

1.7 Выводы

Из проведённого обзора современных подходов анализа выживаемости можно сделать следующие выводы:

- Анализ событийных данных широко применяется в здравоохранении, анализе надежности, биоинформатике и маркетинге. Наибольшую популярность получило применение методов анализа выживаемости для анализа медицинских данных. Реальные данные характеризуются наличием цензурированных наблюдений, могут включать категориальные и непрерывные признаки, а также иметь пропущенные значения. Целевые переменные времени и флага цензурирования могут иметь различное распределение вероятностей, а также быть связаны в случае информативности цензурирования.
- Для оценки качества прогнозирования величин анализа выживаемости применяются точечные и интегральные метрики. Метрика CI применяется для оценки качества ранжирования ожидаемого времени события, метрики IBS и AUPRC — для оценки функции выживания, IAUC — для оценки качества ранжирования наблюдений относительно значений кумулятивного риска во времени.
- Наибольшую популярность получили статистические методы анализа выживаемости, основанные на строгих предположениях. Непараметрические методы не описывают за-

висимость между признаками наблюдения и целевыми переменными, а также предполагают неинформативность флага цензурирования. Метод пропорциональных рисков Кокса предполагает единую форму функции выживания и линейную зависимость между коэффициентом масштабирования и признаками наблюдения. Метод ускоренного времени отказа предполагает теоретическое распределение вероятностей времени события. Применение статистических моделей ограничено, поскольку предположения могут не выполняться на практике.

- Методы построения деревьев выживаемости основаны на рекурсивном разбиении пространства признаков на области, содержащие наблюдения с максимально близкой выживаемостью внутри области и максимально различной между областями. В листах деревьев выживаемости строится непараметрическая оценка времени наступления и вероятности события, а также функций выживания и риска. Разбиение производится на основе критерия log-rank, нулевая гипотеза которого предполагает равенство функций риска дочерних выборок. Однако критерий разбиения и непараметрические модели основаны на предположении неинформативности цензурирования. Метод случайного леса позволяет повысить качество прогнозирования, однако наследует недостатки древовидных моделей.
- Нейросетевые модели, байесовские модели, метод опорных векторов и бустинг ансамбли основаны на трех подходах. Точечные методы позволяют прогнозировать только значения времени и вероятности события. Дискретные модели используют заранее заданную временную шкалу, в рамках которой проводится прогнозирование функций выживания и риска. Непрерывные модели сводят задачи прогнозирования функций анализа выживаемости к классификации и регрессии с последующим построением функций на основе статистических моделей.
- Существующие программные системы анализа выживаемости работают с заполненными числовыми данными и используют строгие статистические предположения.

Сформулированные выводы являются обоснованием направлений дальнейших исследований:

- *Исследование и разработка методов построения интерпретируемых деревьев выживаемости, применимых к категориальным и неполным данным с цензурированием и позволяющих прогнозировать время, вероятность события и непрерывные функции выживания и риска без учета статистических предположений;*
- *Исследование влияния особенностей данных на метрики качества анализа выживаемости и разработка метода оценки качества прогнозирования, определяющего равный вклад событий и временных интервалов;*
- *Исследование и разработка методов ансамблирования деревьев выживаемости для повышения качества прогнозирования без использования строгих статистических предположений.*

2 МЕТОД ПОСТРОЕНИЯ ДЕРЕВЬЕВ ВЫЖИВАЕМОСТИ

При работе (при подготовке) над данным разделом диссертации использованы следующие публикации автора, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования:

- *Васильев Ю. А., Петровский М. И., Машечкин И. В. Применение регуляризации при вычислении критериев разбиения в моделях анализа выживаемости // Вычислительные методы и программирование. – 2024. – Т. 25, № 3. – С. 357-377.*

В данной диссертационной работе проводится исследование и разработка методов построения нелинейных моделей анализа выживаемости. Исходя из проведенного обзора существующих решений, можно выделить несколько значимых проблем. Во-первых, в существующих древовидных подходах (разделы 1.4 и 1.6.1) при поиске лучшего разбиения используется критерий log-rank (формула (18)) для оценки различия выборок. Критерий основывается на предположении, что индикатор цензурирования не связан с прогнозом, вероятности выживания одинаковы для событий на ранних и поздних этапах исследования. В открытых исследованиях [71, 73] высказывается предположение о плохой чувствительности критерия log-rank к особенностям реальных данных, основанных на раннем возникновении событий.

Во-вторых, существующие подходы основываются на заполненных данных. На практике, проблема наличия пропусков в данных очень распространена, а происхождение пропусков может быть неизвестно. Для применения моделей к реальным данным необходимо также разработать подход для обработки отсутствующих данных.

Также, актуальна проблема применимости алгоритмов на больших объемах данных. В таких данных могут быть нарушены изначальные предположения моделей, а также возвращается сложность вычисления. Сложность существующих подходов может быть ограничена гиперпараметрами, однако построенные модели дают меньшую точность. Для решения проблемы роста сложности, в модели могут быть заложены дополнительные принципы обработки признаков, имеющие одинаковую сложность для любых объемов данных.

В данной работе проводится исследование и разработка методов анализа выживаемости на основе машинного обучения. Предлагаемые методы должны решать перечень объявленных проблем существующих методов и иметь практическое применение в качестве системы поддержки врачебных решений.

Настоящая глава посвящена разработке методов построения деревьев выживаемости. Формализация задачи представлена в разделе 1.1.3.

2.1 Описание используемых для исследования наборов данных

В данной работе рассматриваются 6 открытых медицинских наборов данных с различными характеристиками (таблица 3) типа события, количества наблюдений, количества признаков, дисбаланса цензурирования и заполненности данных.

Таблица 3: Характеристики открытых медицинских наборов данных (упорядочены по убыванию процента терминальных событий). N – количество наблюдений, d – количество признаков, (Cens, Event) – количество терминальных и цензурированных событий, Event (%) – доля терминальных событий, NaN (feat) – количество признаков с неполными данными.

Название	**Событие**	**N**	**feat**	**Cens, Event**	**Event (%)**	**NaN (feat)**
SUPPORT2 [43]	Смерть	9105	35	(2904, 6201)	0.681	21
WUHAN [19]	Выписка	375	224	(201, 174)	0.464	222
GBSG [17]	Рецидив	686	8	(387, 299)	0.436	0
ROTT2 [18]	Рецидив	2982	11	(1710, 1272)	0.427	0
PBC [96]	Смерть	418	17	(257, 161)	0.385	12
SMARTO [15]	Смерть	3873	26	(3413, 460)	0.119	16

2.1.1 Описание наборов данных

Набор данных Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT2) [43] содержит показатели неизлечимо больных пациентов, находящихся на жизнеобеспечении. В качестве события рассматривается смерть пациента. Набор данных содержит 9105 наблюдений и 35 признаков по анамнезу, классу заболевания пациента, тяжести физиологических отклонений и сопутствующим заболеваниям. Категориальными являются 11 признаков: sex, dzgroup, dzclass, num_co, race, diabetes, dementia, ca, dnr, sfdm2, income. Пропуски содержатся в 21 признаке, максимальное количество пропусков имеет ADL (5641 пропуск). В ходе исследования были цензурированы 2904 пациента.

Набор данных WUHAN, собранный с 10 января по 18 февраля 2020 года, был представлен в работе [19]. В качестве события рассматривается время выписки пациента. Набор данных содержит 375 наблюдений и 76 признаков по анамнезу и результатам клинических исследований за время лечения. Пространство признаков формируется из минимальных, максимальных и средних показателей клинических исследований пациента. Все признаки набора данных могут содержать пропуски, максимальное количество пропусков имеется в показателях антитромбина и продуктах распада фибрина (173 пропуска). В ходе исследования был цензурирован 201 пациент.

Набор данных German Breast Cancer Study Group (GBSG), собранный с 1984 по 1989 год, был представлен в работе [17]. В качестве события рассматривается время рецидива рака. Набор данных содержит 686 наблюдений и 8 признаков по анамнезу, характеристикам опухоли и стратегии лечения. Категориальными являются 3 признака: htreat, menostat, tumgrad. Пропусков набор данных не содержит. В ходе исследования были цензурированы 387 пациентов.

Набор данных Cohort study on breast cancer patients from the Netherlands (ROTT2) [18] содержит информацию о пацентах с раком молочной железы, перенесших операцию на груди. В качестве события рассматривается рецидив рака. Набор данных содержит 2982 наблюдения и 11 признаков по анамнезу, характеристикам опухоли и стратегии лечения. Категориальными являются 6 признаков: meno, tsize, grade, hormon, chemo, recent. Пропусков набор данных не содержит. В ходе исследования были цензурированы 1710 пациентов.

Набор данных Primary Biliary Cirrhosis (PBC), собранный с 1974 по 1984 год, был представлен в работе [96]. В качестве события рассматривается время смерти. Набор данных

содержит 418 наблюдений и 17 признаков по анамнезу, статусу цирроза, стратегии лечения и клиническим показателям. Категориальными являются 5 признаков: trt, sex, ascites, hepato, spiders. Также, 12 признаков набора данных могут содержать пропуски (в частности, стратегии лечения и клинические показатели), максимальное количество пропусков имеется в показателе холестерина (134 пропуска) и в показателе триглицеридов (136 пропусков). В ходе исследования были цензурированы 257 пациентов.

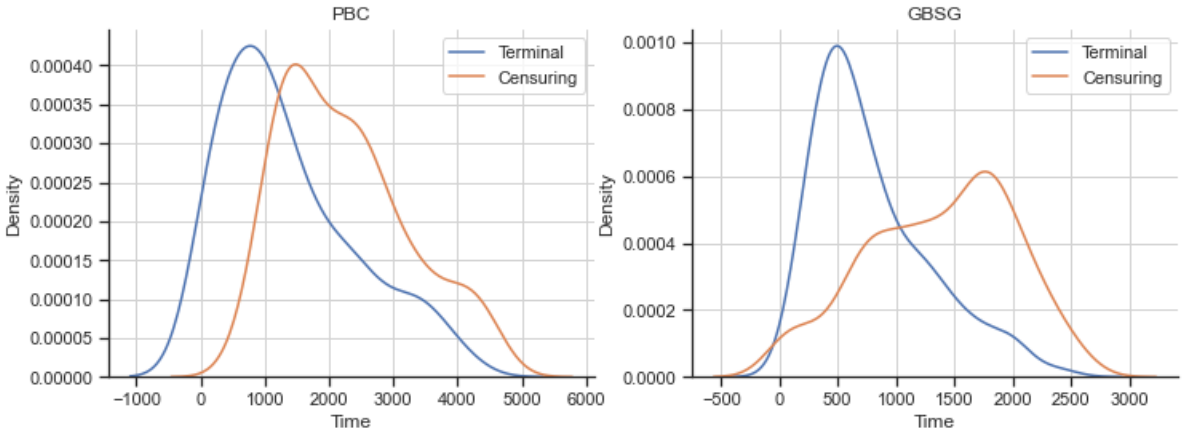
Набор данных Second Manifestations of ARTerial Disease (smarto) [15] содержит сведения о пациентах, госпитализированных с клинически манифестным атеросклеротическим заболеванием сосудов или выраженными факторами риска атеросклероза. В качестве события рассматривается смерть пациента. Набор данных содержит 3873 наблюдения и 26 признаков по анамнезу, клиническим показателям и маркерам атеросклероза. Категориальными являются 9 признаков: sex, diabetes, cerebral, aaa, periph, stenosis, albumin, smoking, alcohol. Пропуски содержатся в 16 признаках, максимальное количество имеется в показателях артериального давления: diastolic by hand (1499 пропусков), systolic by hand (1498 пропусков). В ходе исследования 3413 пациентов были цензурированы.

2.1.2 Выполнимость статистических предположений

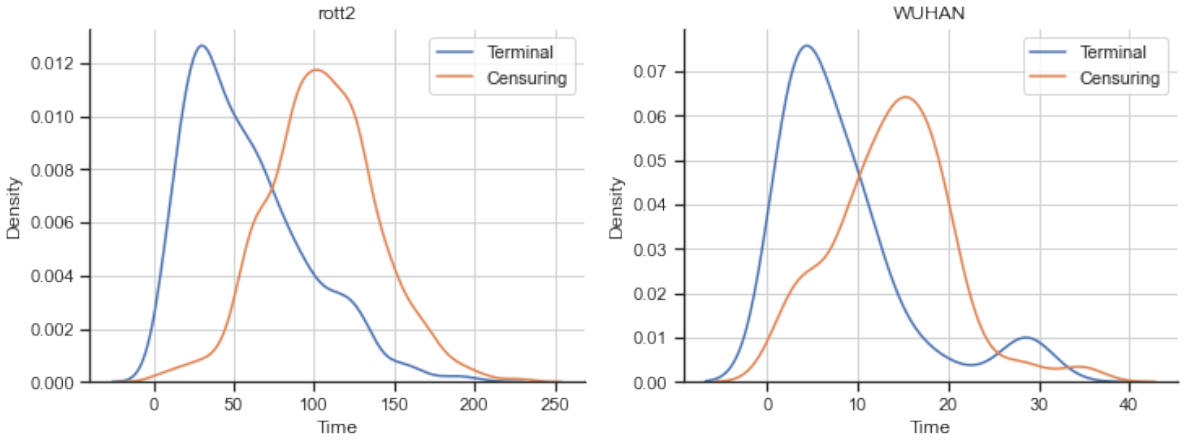
На рисунке 9 представлена ядерная оценка плотности времени событий наборов данных. Распределения времени терминальных и цензурированных событий различаются как относительно параметра локализации распределения, так и относительно формы.

Таблица 4: Результаты проверки выполнимости статистических предположений на открытых наборах данных. Столбцы «KS test value» и «KS test p-value» содержат значения статистики и p-value теста Колмогорова–Смирнова при сравнении распределений времени терминальных и цензурированных событий. Также, для каждого набора определяется топ-5 переменных, не удовлетворяющих условию пропорциональности рисков (по наименьшему p-value).

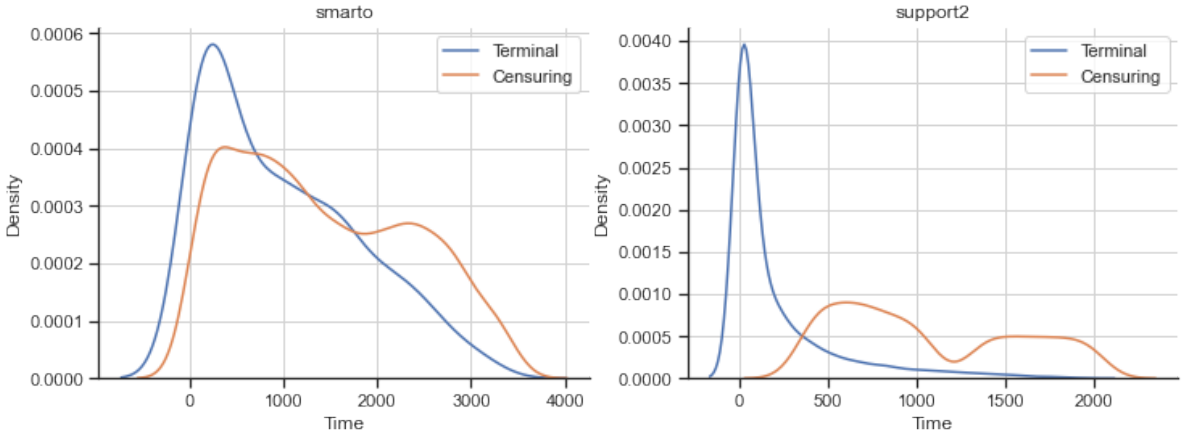
Набор	**KS test value**	**KS test p-value**	**Переменные с непропорциональным риском**		
			Кол-во	**Топ-5 переменных**	**P-value**
GBSG	0.4325	6.89e-29	1	tumgrad	0.0353
PBC	0.4306	4.92e-17	1	protime	0.0327
WUHAN	0.4601	2.27e-18	7	min_neutrophils_count	0.0103
				min_White_blood_cell_count	0.0176
				min_International_standard_ratio	0.0211
				max_Prothrombin_activity	0.0275
				min_Lactate_dehydrogenase	0.0307
SMARTO	0.1847	1.39e-12	1	num_age	0.0353
ROTT2	0.5370	2.25e-193	3	num_estrogen	3.5e-07
				num_progesterone	1.32e-05
				num_age	0.0289
SUPPORT2	0.8035	0.0	15	fac_sfdm2	8.92e-184
				num_sps	1.24e-52
				num_scoma	7.44e-39
				num_surv6m	2.35e-11
				fac_sex	9.32e-07



(a) Распределение времени событий наборов GBSG, PBC



(b) Распределение времени событий наборов WUHAN, ROTT2



(c) Распределение времени событий наборов SMARTO, SUPPORT2

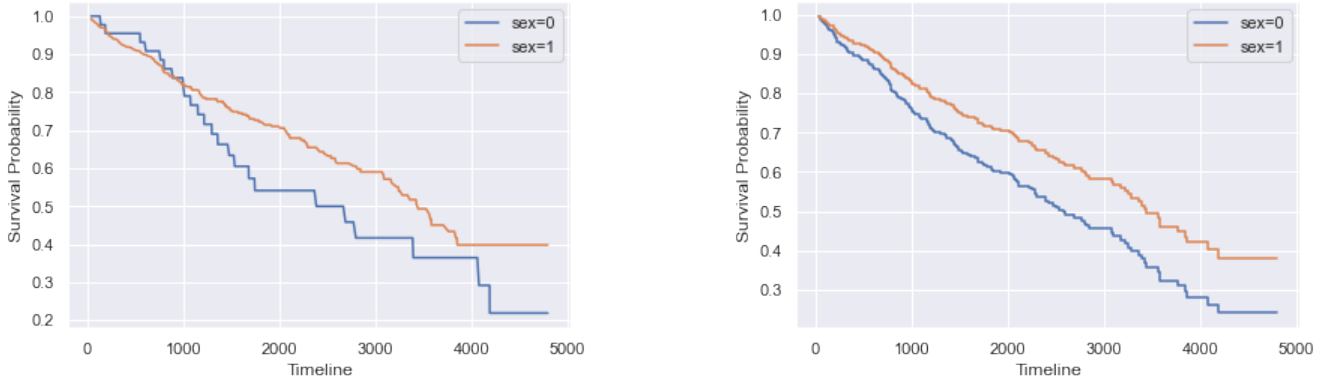
Рис. 9: Распределение времени терминальных (отмечено синим) и цензурированных наблюдений (отмечено оранжевым) для наборов данных GBSG, PBC (рисунок a), WUHAN, ROTT2 (рисунок b), SMARTO, SUPPORT2 (рисунок c). Для наборов PBC, WUHAN, ROTT2 наблюдается смещение распределений во времени при сохранении формы. Для наборов GBSG, SMARTO, SUPPORT2 наблюдается смещение и изменение формы распределений.

В таблице 4 представлены результаты проверки наборов данных на выполнимость статистических предположений пропорциональных рисков Кокса и информативности цензури-

рования. Для проверки информативности цензурирования выполнялось сравнение распределений времени терминальных и цензурированных событий по тесту Колмогорова–Смирнова. Значение статистики определяет максимальное расстояние между двумя эмпирическими функциями распределения. Таким образом, для всех наборов данных наблюдаются значимые различия между распределениями, что говорит об информативности индикатора цензурирования.

Также, таблица 4 содержит результаты проверки статистического теста на пропорциональности рисков. Для каждого набора данных определено количество непропорциональных переменных и отражены 5 переменных с наибольшим значением p-value. Таким образом, для каждого набора данных определяется хотя бы одна переменная, нарушающая предположение пропорциональности.

Для повышения наглядности ограничения предположения пропорциональных рисков рассмотрим следующие примеры на наборах PBC и GBSG. В случае категориальных признаков, оценим функции выживания для каждой страты на основе метода Каплана-Майера и пропорциональных рисков Кокса. На рисунке 10 представлен пример влияния фактора «sex» на функцию выживания для набора PBC. В случае оценки KM, функции выживания для каждой страты пересекаются: вероятность выживания страты «sex=0» выше вероятности выживания страты «sex=1» для ранних моментов времени (до 1000 дней). Пересечение кривых говорит о том, что эффект фактора «sex» на выживаемость не является постоянным во времени и может меняться в зависимости от длительности наблюдения. Согласно клинической медицинской практике, нельзя сделать однозначный вывод о влиянии фактора и две группы считаются неразличимыми. В то же время, модель пропорциональных рисков Кокса (CoxPH) не учитывает изменение влияния фактора во времени и определяет выживаемость страты «sex=1» выше «sex=0», чем ложно вызывает постоянный эффект значимости фактора «sex» на различие в функциях выживаемости.



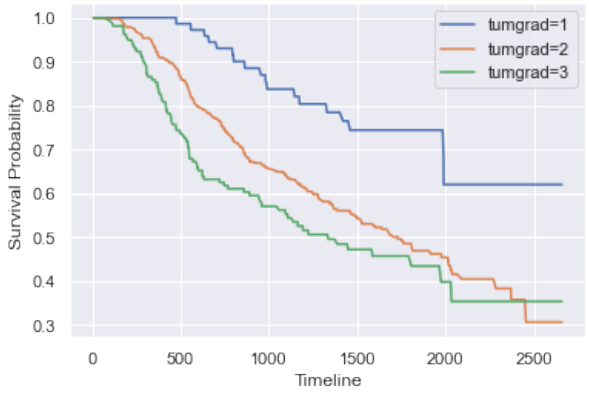
(a) Оценка Каплана-Майера

(b) Оценка по методу Кокса

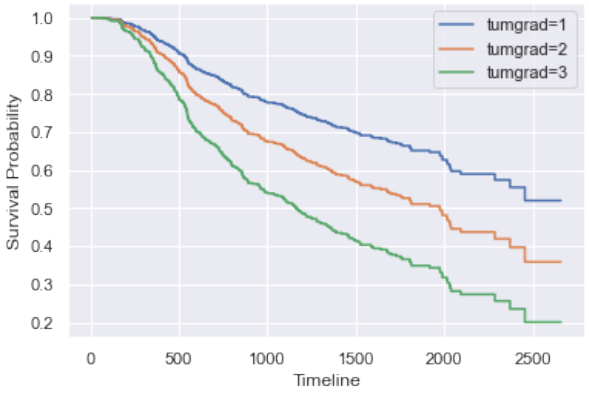
Рис. 10: Пример влияния фактора «sex» набора PBC на оценку функции выживания по методам KM и CoxPH

На рисунке 11 представлен пример влияния фактора «tumgrad» на функцию выживания для набора GBSG. В данном случае, обратим внимание на тренд оцениваемых функций. Для оценки KM функция выживания страты «tumgrad=1» расположена выше остальных, однако на момент 2000 дня обладает резким снижением вероятности выживания (из-за малого

количества наблюдений, доживших до текущего момента времени). В случае оценки CoxPH, функция выживания страты «tumgrad=1» имеет почти линейный тренд. Аналогично примеру 10 , по оценке KM страты «tumgrad=2» и «tumgrad=3» являются неразличимыми, в то время как оценка CoxPH определяет строгий порядок оценок вероятностей, постоянный во времени.



(a) Оценка Каплана-Майера



(b) Оценка по методу Кокса

Рис. 11: Пример влияния фактора «tumgrad» набора GBSG на оценку функции выживания по методам KM и CoxPH

Таким образом, выделяются следующие особенности наборов данных:

1. Для всех наборов данных нарушаются статистические предположения неинформативности цензурирования и пропорциональности рисков;
2. Ранние события имеют наибольшую важность в наборе SUPPORT2, так как набор посвящен исследованию неизлечимых пациентов на жизнеобеспечении. Дисбаланс цензурирования событий смещен в сторону терминальных событий.
3. Наборы данных WUHAN, GBSG, ROTT2, PBC сбалансированы относительно классов событий и наибольшую важность вносят ранние и средние события.
4. Набор SMARTO имеет высокий дисбаланс цензурированных событий. Важно отметить, что формы функций плотности распределения событий набора SMARTO близки, однако статистически различимы по тесту Колмогорова–Смирнова.

2.2 Поиск лучшего бинарного разбиения выборки

Одним из важнейших факторов качества древовидных моделей анализа выживаемости является метод разделения выборки на группы с максимальными различиями по выживаемости. Существующие подходы (раздел 1.4) основаны на переборе всевозможных промежуточных значений признаков с расчетом близости между соответствующими подвыборками.

Стоит отметить, что данный подход позволяет обрабатывать только непрерывные признаки. В случае категориальных признаков сравнение значений некорректно (например, для признаков диагноза или типа опухоли). Также, не обрабатываются пропущенные значения признаков. Наконец, поиск лучшего разбиения среди множества всех возможных промежуточных значений является вычислительно сложным процессом для реальных данных.

В разделе 2.2.1 предлагается гистограммный метод поиска лучшего разбиения выборки с цензурированием по непрерывному признаку. Сравнение ветвей разбиения проводится с использованием взвешенных критериев log-rank (раздел 2.2.2). В разделе 2.2.4 представлено расширение метода для обработки категориальных признаков, а в разделе 2.2.3 — для обработки пропущенных значений.

2.2.1 Гистограммный метод поиска разбиения

Рассмотрим алгоритм поиска лучшего бинарного разбиения по непрерывному признаку f. Пусть дана выборка $\{(v_k, T_k, \delta_k)\}$, в которой каждое наблюдение представляется в виде тройки значений, где T_k — время события, δ_k — флаг цензурирования, v_k — значение признака f. Определим $\{\bar{v}_i\}$ как упорядоченное множество значений признака f.

Промежуточные точки $s_1, s_2, ...s_k$ удовлетворяют цепочке неравенств $\bar{v}_1 < s_1 < \bar{v}_2 < s_2 < ... < s_{n-1} < \bar{v}_n$ и равны $s_i = (\bar{v}_i + \bar{v}_{i-1})/2$. Для каждой промежуточной точки s_i определяются две ветви разбиения выборки: левая содержит наблюдения с $v \leq s_i$, правая — с значениями $v > s_i$. Наконец, проведем дискретизацию признака: истинные значения v_i заменим на значение ближайшей правой промежуточной точки.

Отметим, что вычислительная сложность поиска линейно зависит от количества уникальных значений признака. При n уникальных значений признака необходимо рассмотреть $n-1$ промежуточную точку для поиска разбиения. Для контроля вычислительной сложности, предлагается квантилизовать точки разбиения (точка разбиения s_i равна i процентилю выборки v) при $n > 100$.

Пусть $\{\tau_j\}$ — упорядоченный набор времени наступления событий $\tau_1 < \tau_2 < ... < \tau_K$ в выборке. Тогда, для каждого момента времени τ_j, количество наблюдений n_j и количество событий O_j определяется по следующим формулам:

$$n_j = \sum_k I(T_k = \tau_j),$$

$$O_j = \sum_k I(T_k = \tau_j) \cdot I(\delta_k = 1).$$

Также, количество наблюдений $n_{j|v=s}$ и событий $O_{j|v=s}$ в момент τ_j при условии $v = s$ (условие на значения модифицированного признака f) равно:

$$n_{j|v=s} = \sum_k I(T_k = \tau_j) \cdot I(v_k = s),$$

$$O_{j|v=s} = \sum_k I(T_k = \tau_j) \cdot I(v_k = s) \cdot I(\delta_k = 1).$$

Таким образом, для каждой промежуточной точки s_i определяются гистограммы левой и правой ветви: n^l и n^r (гистограммы наблюдений), O^l и O^r (гистограммы событий) по следующим формулам:

$$n^l_{s_i,j} = \sum_{k:s_k \leq s_i} n_{s_k,j}, \; n^r_{s_i,j} = \sum_{k:s_k > s_i} n_{s_k,j}, \tag{20}$$

$$O^l_{s_i,j} = \sum_{k:s_k \le s_i} n_{s_k,j},\ O^r_{s_i,j} = \sum_{k:s_k > s_i} n_{s_k,j}. \tag{21}$$

Классический алгоритм поиска лучшего разбиения (раздел 1.4) вычисляет значения $n^l_{s_i,j}, n^r_{s_i,j}, O^l_{s_i,j}, O^r_{s_i,j}$ для каждой промежуточной точки s_i. Следовательно, метод избыточно вычисляет $n_{s_k,j}$ на каждой итерации, причем кратность равна количеству промежуточных точек. Для оптимизации времени вычисления, можно предварительно рассчитать гистограммы $n_{s_k,j}$. В таком случае, объем используемой памяти линейно зависит от количества промежуточных точек.

Гистограммный метод позволяет одновременно оптимизировать время вычисления и объем используемой памяти. Метод основан на идее итеративного обновления гистограмм левой и правой ветви и выполняется по следующим правилам:

$$n^l_{0,j} = 0,\ n^r_{0,j} = n_j,$$

$$n^l_{s_i,j} = n^l_{s_{i-1},j} + n_{j|v=s_i},\ n^r_{s_i,j} = n^r_{s_{i-1},j} - n_{j|v=s_i},$$

$$O^l_{0,j} = 0,\ O^r_{0,j} = O_j,$$

$$O^l_{s_i,j} = O^l_{s_{i-1},j} + O_{j|v=s_i},\ O^r_{s_i,j} = O^r_{s_{i-1},j} - O_{j|v=s_i}.$$

Левые гистограммы $(n^l_{0,j}, O^l_{0,j})$ инициализируются 0, а правые $(n^r_{0,j}, O^r_{0,j})$ — количеством наблюдений и событий для всех моментов времени. Итеративно исчерпывая упорядоченное множество точек s_i, для каждой промежуточной точки s_i вычисляются 2 гистограммы $(n_{j|v=s_i}, O_{j|v=s_i})$, которые векторно прибавляются к гистограмме левой ветви ($n^l_{s_i,j}$ и $O^l_{s_i,j}$ соответственно) и вычитаются из правой ($n^r_{s_i,j}$ и $O^r_{s_i,j}$ соответственно). Таким образом, для каждой точки s_i остаются верными равенства (20) и (21).

Наконец, на основе $n^l_{s_i,j}$ и $n^r_{s_i,j}$ вычисляются гистограммы оставшихся наблюдений $N^l_{s_i,j}, N^r_{s_i,j}$ и гистограмма ожидаемого количества событий для левой ветви $E^l_{s_i,j}$:

$$N^l_{s_i,j} = \sum_{\tau_k \ge \tau_j} n^l_{s_i,k},\ N^r_{s_i,j} = \sum_{\tau_k \ge \tau_j} n^r_{s_i,k},$$

$$E^l_{s_i,j} = \frac{N^l_{s_i,j} \cdot O_j}{N_j}.$$

Используя гистограммы оставшихся наблюдений и наступивших событий, для промежуточной точки s_j вычисляется значение взвешенной статистики log-rank (модификация формулы (18)):

$$LR_{s_i} = \frac{\sum_{j=1}^K w_j \left(O^l_{s_i,j} - E^l_{s_i,j}\right)}{\sqrt{\sum_{j=1}^K w_j^2 \cdot E^l_{s_i,j} \left(\frac{N_j - O_j}{N_j}\right) \left(\frac{N_j - N^l_{s_i,j}}{N_j - 1}\right)}}. \tag{22}$$

Лучшее разбиение для признака f определяется по максимальному значению статистики log-rank (или минимальному p-value) и соответствует паре правил $v \le s_{best,f}$ и $v > s_{best,f}$, где $s_{best,f} = argmax_i\ LR_{s_i}$.

На Рисунке 12 представлен пример разбиения выборки с цензурированием по признаку

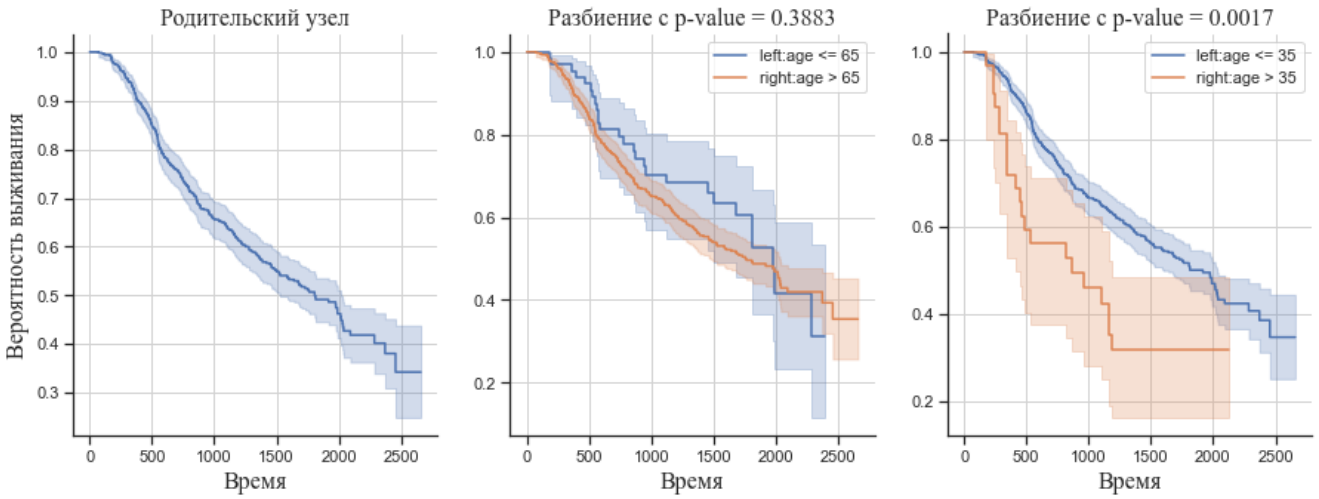


Рис. 12: Пример разбиения выборки по возрасту с промежуточными точками 35 и 65.

«age» (возраст). На левом графике представлена функция выживаемости исходной выборки (родительский узел). На центральном и правом графике представлены два возможных разбиения выборки по промежуточным значениям признака 65 и 35. При разбиении по значению 65, p-value равно 0.3883 и функции выживаемости пересекаются. При разбиении по значению 35, p-value меньше 0.01 и функции выживаемости имеют сильные различия (ветвь с правилом «$age \leq 35$» расположена значительно выше второй ветви). Для максимизации различия между ветвями разбиения, алгоритмом будет выбрано разбиение с меньшим p-value (разбиение с промежуточной точкой 35).

2.2.2 Взвешенный критерий log-rank

Как было сказано ранее, для оценки близости между ветвями разбиения используется статистика log-rank, большее значение которой определяет большее различие между функциями риска. Для решения проблемы малой чувствительность критерия log-rank, при оценке близости ветвей разбиения (формула (22)) предлагается использовать различные весовые схемы для учета особенностей распределения времени событий. В частности, гистограммный метод поддерживает 4 взвешенные схемы log-rank критерия: log-rank (с единичным весом $w_j = 1$), wilcoxon [97], tarone-ware [98], peto-peto (peto) [99]. Аналогичного исследования применимости взвешенных критериев log-rank для построения деревьев выживаемости в литературе найдено не было.

Взвешенные критерии основаны на определении величины w_j в формуле (22):

1. Критерий wilcoxon: $w_j = N_j$
   Значение статистики вычисляется путем взвешивания вклада по числу наблюдений, подверженных риску. Это придает больший вес более ранним временам событий, когда число подверженных риску выше. Однако данный критерий сильно зависит от различий в структуре цензуры групп.
2. Критерий peto-peto (peto): $w_j = \hat{S}(\tau_j)$, где $\hat{S}(t)$ — оценка Каплана–Мейера, построенная на полной выборке.

Критерий подходит для случаев, когда считается, что функции риска различаются не пропорционально. В отличие от теста wilcoxon, на него не влияют различия в структуре цензуры групп.

3. Критерий Tarone-Ware: $w_j = \sqrt{N_j}$
   Значение статистики вычисляется путем взвешивания вклада квадратным корнем по числу наблюдений, подверженных риску. Как и критерий wilcoxon, он дает более высокие веса (хотя и не такие большие) для более ранних периодов возникновения событий. В исследовании [100] отмечается, что критерий является «золотой серединой» между критериями wilcoxon и peto.

2.2.3 Обработка пропущенных значений

Как говорилось ранее, существующие реализации деревьев выживания не позволяют работать с пропусками и требуют проводить предобработку данных. Наиболее распространен подход предварительного заполнения (импутации) пропусков. Заполнение пропусков константой корректно только в случае определенности семантики признака. Заполнение пропусков типичным значением признака (средним значением, медианой или модой) искажает распределение значений. Также, применяются подходы прогнозирования пропущенных значений с помощью вспомогательных моделей классификации или регрессии. Однако при таком подходе теряется смысл пропущенных значений, а также сглаживаются данные всего признакового пространства.

Также, пропущенные значения могут рассматриваться в виде отдельной категории признака. Так как значение пропусков неопределенно, то и положение категории на числовой шкале определяется индивидуально для каждой задачи. В таком случае, необходима дополнительная обработка категорий внутри модели прогнозирования.

В данной работе предлагается следующий подход обработки отсутствующих значений. На этапе поиска лучшего разбиения наблюдения разделяются на два множества: с пропусками v_{nan} и без пропусков $v_{\overline{nan}}$. Для множества v_{nan} вычисляется гистограмма наблюдений $n_{nan,j}$ и событий $O_{nan,j}$ в момент времени τ_j, а также гистограмма оставшихся наблюдений $N_{nan,j}$.

Для множества $v_{\overline{nan}}$ применим предложенный алгоритм поиска лучшего разбиения (раздел 2.2.1). При этом, на каждой итерации, после вычисления гистограмм левой и правой ветви, правила разбиения дополняются пропущенными значениями:

$$N^l_{s_i|nan,j} = N^l_{s_i,j} + N_{nan,j},\ O^l_{s_i|nan,j} = O^l_{s_i,j} + O_{nan,j},$$

$$N^r_{s_i|nan,j} = N^r_{s_i,j} + N_{nan,j},\ O^r_{s_i|nan,j} = O^r_{s_i,j} + O_{nan,j}.$$

Для пары правил $v \le s_i \mid nan$ и $v > s_i$ вычисляется статистическое значение $LR_{s_i|left\ nan}$ критерия log-rank (по формуле (22)). Аналогично, для пары правил $v \le s_i$ и $v > s_i \mid nan$ вычисляется статистическое значение $LR_{s_i|right\ nan}$. Пропущенные значения определяются для ветви с наибольшим статистическим значением разбиения. Следовательно, одно из правил $v \le s_i$ и $v > s_i$ дополняется условием наличия пропусков. Например, при $LR_{s_i|right\ nan} > LR_{s_i|left\ nan}$ пропущенные значения назначаются правой ветви и рассматриваются правила $v \le s_i$ и $v > s_i \mid nan$.

2.2.4 Обработка категориальных признаков

Как говорилось ранее, модели существующих библиотек не позволяют работать с категориями и требуют проводить предобработку данных. Классический метод поиска разбиения для категориальных признаков основан на переборе всевозможных пар непересекающихся множеств l, r уникальных значений признака f. Следовательно, для каждой пары множеств l, r определяются 2 правила разбиения признака f: $v \in [l]$ (левая ветвь) и $v \in [r]$ (правая ветвь). Заметим, что при использовании k различных значений категориального признака f порождается $2^{k-1} - 1$ различных пар множеств (рассматриваются непустые множества). Таким образом, вычислительная сложность алгоритма экспоненциально зависит от количества уникальных значений.

Предлагаемая реализация использует метод Weight of Evidence (WOE) [101] для отображения категорий признаков на непрерывную числовую шкалу. Метод основан на построении бинарной модели, описывающую связь между категориями признака и вероятностью наступления целевого события. Пусть $C(T)$ — функция подсчета наблюдений признака f со значением T, B — входной признак, D — целевой бинарный признак. В терминах поставленной задачи, B определяет фиксированную категорию признака f, а D — флаг цензурирования.

Тогда $C(B)$ — число наблюдений с категорией B, а $C(D)$ — число наблюдений, для которых событие наступило. Соответственно, $C(\overline{B})$ — число наблюдений **не** категории B, $C(\overline{D})$ — число наблюдений, для которых событие **не** наступило. Тогда, определим 4 вероятностные величины по следующим формулам:

1. $P(D|B) = \frac{C(B \cap D)}{C(B)}$ — отношение количества терминальных событий с категорией B к общему числу событий с категорией B;
2. $P(\overline{D}|B) = \frac{C(B \cap \overline{D})}{C(B)}$ — отношение количества цензурированных событий с категорией B к общему числу событий с категорией B;
3. $P(D|\overline{B}) = \frac{C(\overline{B} \cap D)}{C(\overline{B})}$ — отношение количества терминальных событий не категории B к общему числу событий не категории B;
4. $P(\overline{D}|\overline{B}) = \frac{C(\overline{B} \cap \overline{D})}{C(\overline{B})}$ — отношение количества цензурированных событий не категории B к общему числу событий не категории B.

Значения WOE вычисляются по следующим формулам (под обозначением ln подразумевается натуральный логарифм):

$$WOE^{+}(B) = \ln \frac{P(D|B)}{P(\overline{D}|B)},\ WOE^{-}(B) = \ln \frac{P(D|\overline{B})}{P(\overline{D}|\overline{B})}. \tag{23}$$

Веса оценивают меру пространственной связи между исходным и целевым признаками. Положительный вес $WOE^{+}(B)$ означает, что число наблюдений с категорией B выше, чем при независимом распределении значений признака и флага цензурирования. Отрицательный вес $WOE^{-}(B)$ означает, что число наблюдений с категорией B ниже, чем в случае независимости значений признака и флага цензурирования.

Предлагаемая реализация сопоставляет каждой категории B признака f значение $WOE^{+}(B)$. Категории с одинаковыми значениями объединяются в общую группу. Стоит отметить, что предложенный подход отличается от существующих подходов из-за учета факта цензурирования наблюдений вместо ожидаемого времени события. Предложенный подход

учитывает информативность цензурирования путем расположения категорий на основе схожести подвыборок по цензурируемости.

При поиске лучшего разбиения, предложенный гистограммный метод (раздел 2.2.1) рассматривает непрерывное отображение категориального признака. После нахождения разбиения с промежуточной точкой $\hat{s}$, правила разбиения формулируются в терминах исходных категорий. Разбиение признака f определяется правилами: $v \in [l]$ (левая ветвь) и $v \in [r]$ (правая ветвь), где множества l, r имеют следующий вид:

$$l = \{B \mid \hat{s} \leq WOE^{+}(B)\},\ r = \{B \mid \hat{s} > WOE^{+}(B)\}.$$

2.3 Модель дерева выживания

В данном разделе представляется собственный метод построения деревьев выживания на основе предложенного гистограммного метода поиска лучшего разбиения выборки. В разделе 2.3.1 описывается алгоритм построения дерева выживания и его основные отличия от существующих методов.

Главными недостатками модели дерева решений является большой размер итоговой модели и склонность к переобучению. Переобучение возникает в случае, когда модель прогнозирования вырабатывает предсказания, которые слишком близко или точно соответствуют конкретному набору данных и поэтому не подходят для применения алгоритма к дополнительным данным или будущим наблюдениям. В случае переобучения модели ошибка прогнозирования на тестовых данных сильно превышает ошибку на тренировочных данных.

Для борьбы с переобучением в деревьях решений применяется обрезка (prunning). Обрезка — это метод сжатия данных в алгоритмах машинного обучения и поиска, который уменьшает размер деревьев решений за счет удаления некритичных и избыточных участков дерева. Одновременно, решается проблема большого размера итоговой модели.

Подход предварительной обрезки (pre-pruning) состоит в остановке процесса дерева при достижении заданных параметров (раздел 2.3.2). Альтернативным подходом является обрезка построенного дерева (post-pruning). Идея подхода заключается в первоначальном построении дерева решений, после чего лишние ветии дерева отсекаются на основе минимизации или максимизации метрики качества по отложенной выборке. На практике, подход post-pruning занимает большее время, однако обеспечивает лучшее качество, так как позволяет оценить качество всех возможных вариантов структуры дерева решений. Реализация обрезки построенного дерева описана в разделе 2.3.3.

2.3.1 Построение и прогноз дерева выживания

В данном разделе предлагается новый метод построения моделей дерева выживания. Как и в рассмотренных ранее моделях 1.4, 1.6.1, построение дерева начинается с корневого узла, содержащего все данные.

На этапе разбиения узла, для каждого допустимого признака производится поиск лучшего разбиения по предложенному гистограммному методу поиска лучшего разбиения (раздел 2.2.1). Важно отметить, что поиск может выполняться параллельно (по множеству признаков), поскольку алгоритм использует только значения одного признака. Для каждого

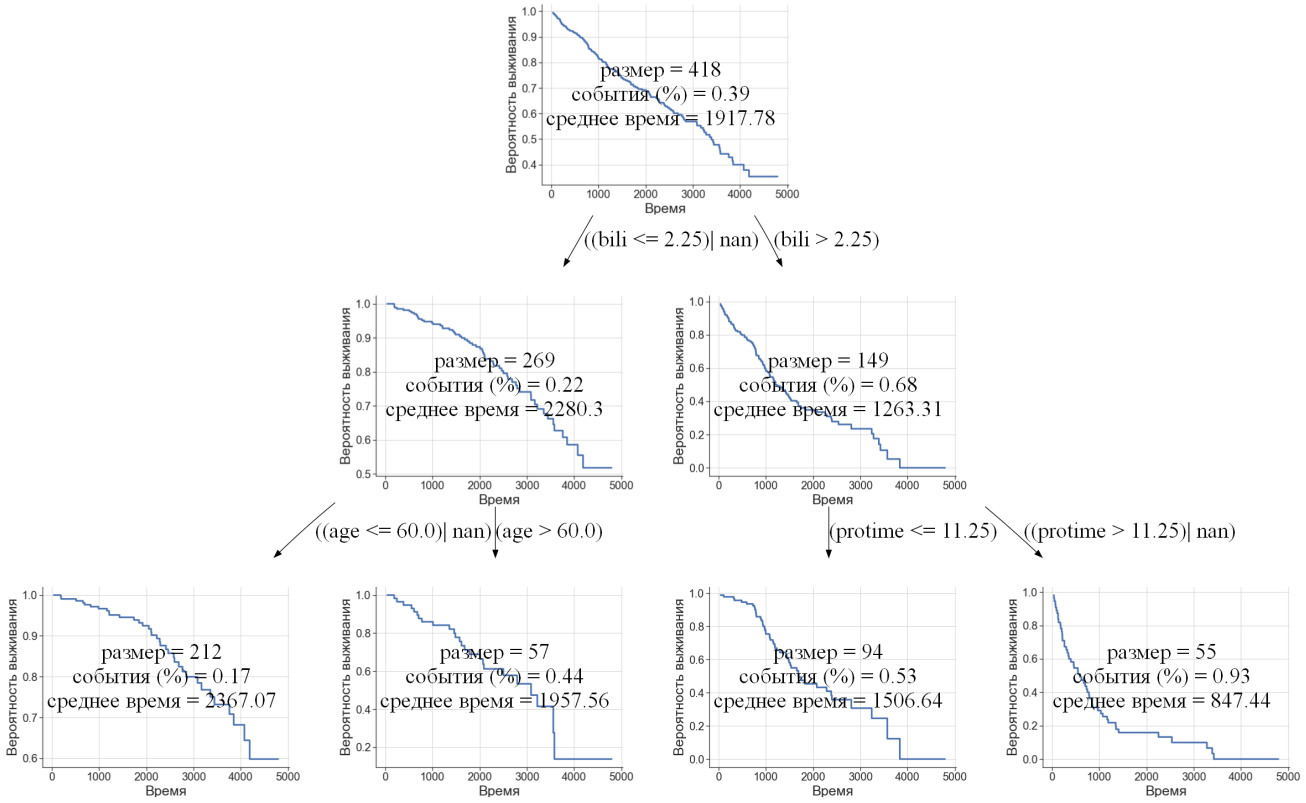


Рис. 13: Пример построенного дерева выживаемости глубины 2 на наборе PBC. Каждый узел дерева содержит визуализацию функции выживания — изменение вероятности выживания относительно времени, а также содержит информацию о размере листа (размер), доле терминальных событий (события (%)) и среднем времени события (среднее время).

признака f определяется тройка значений $(R_f, P\text{-}value_{R,f}, N_{sign,f})$, где R_f — лучшее правило разбиения выборки по признаку f, $P\text{-}value_{R,f}$ — значение $p\text{-}value$ критерия log-rank для правила R, $N_{sign,f}$ — количество «значимых» разбиений по признаку f. Разбиение по правилу R будем называть «значимым», если $P\text{-}value_{R,f}$ не превосходит заранее заданный порог значимости.

Перед выбором лучшего признака разбиения применяется поправка Бонферрони [102] на множественную проверку гипотез. Поправка заключается в расчете скорректированного значения $\overline{P\text{-}value}_{R,f} = P\text{-}value_{R,f} \cdot N_{sign,f}$. Поправка Бонферрони уменьшает значимость более общих признаков, давая предпочтение редким значимым разбиениям. Итоговое правило R разбиения выборки минимизирует скорректированное значение $p\text{-}value$: $R = argmin_f \overline{P\text{-}value}_{R,f}$.

Описанный алгоритм приводит к разбиению корневого узла на две дочерние выборки. Далее, алгоритм разбиения выборок рекурсивно применятся для каждого дочернего узла. Пример построенного дерева глубины 2 на наборе PBC представлен на Рисунке 13. Дерево основано на признаках bili (показатель билирубина), protime (стандартизированное время свертывания крови), age (возраст). Значения «nan» указывают на ветвь определения пропущенных значений.

Получение прогноза

На этапе применения модели к наблюдению с вектором признаков, наблюдению сопоставляется конечный узел (листа) дерева на основе системы правил разбиения. Для непрерывных признаков применяются правила на основе интервалов, для категориальных признаков — правила вхождение в множество допустимых категорий. Для пропущенных наблюдений выбирается ветвь, допускающая наличие пропусков.

На основе листовой выборки определяются следующие прогнозы дерева выживания. Во-первых, точечные прогнозы вероятности и времени наступления события вычисляются как агрегация исходов и времен наблюдений в листе. В качестве функции агрегации используется медиана, среднее значение или средневзвешенная сумма.

Для прогнозирования функции выживания и риска, в листах дерева вычисляются непараметрические статистические оценки (раздел 5.3.2). Оценка функции выживания строится по методу Каплана–Мейера, а оценка функции кумулятивного риска — по методу Нельсона–Аалена.

Наконец, модель дерева выживания позволяет прогнозировать показатели описательной статистики входных признаков по листовой выборке. Например, среднее значение возраста или результатов клинических анализов. Например, для врачебной экспертизы полезна информация о наиболее характерных схемах лечениях в листе дерева. В таком случае, кроме функций выживания и риска, прогноз пациента дополняется полным описанием возможных схем лечений с вероятностью и временем наступления события.

2.3.2 Pre-pruning: контроль роста дерева

Подход предварительной обрезки позволяет ограничить структуру дерева на основе заранее заданных характеристик. Например, при ограничении глубины дерева числом n, узлы глубины n объявляются конечными (листовыми) и не участвуют в процессе поиска лучшего правила разбиения. Таким образом, pre-pruning помогает снизить риск переобучения модели, сделать дерево более интерпретируемым для пользователя и ускорить процесс обучения.

Стоит отметить, что для достижения баланса между простотой модели и ее предсказательной способностью необходимо обладать знаниями о поставленных задачах и характеристиках данных.

Предложенный гистограммный метод (алгоритм) поиска лучшего бинарного разбиения выборки (раздел 2.2.1) поддерживает следующие параметры:

1. Максимальное число промежуточных точек признака для поиска разбиений;
2. Критерий разбиения: log-rank, wilcoxon, peto, tarone-ware;
3. Уровень значимости разбиений (максимальное допустимое значение p-value);
4. Минимальное число наблюдений в ветви разбиения;
5. Флаг обработки категориальных признаков: полный перебор непересекающихся множеств категорий или расчет значений Weight Of Evidence.

Предложенный метод построения дерева выживания (раздел 2.3.1) поддерживает следующие параметры для контроля роста модели:

1. Максимальная глубина дерева;
2. Максимальное количество признаков при поиске лучшего разбиения;
3. Максимальное количество наблюдений в узле.

2.3.3 Post-pruning: обрезка дерева

Вторым методом борьбы с перебучением является обрезка построенного дерева выживаемости (post-pruning). Основный принцип обрезки дерева заключается в удалении некоторых узлов и их ветвей, чтобы улучшить обобщающую способность построенной модели. Стоит отметить, что существующие реализации деревьев выживания не поддерживают операцию post-pruning. До начала обучения, входная выборка разделяется в пропорции 80%/20% на обучающую и валидационную выборки соответственно.

После построения дерева решений на обучающей выборке, применяется итерационный алгоритм обрезки на валидационной выборке:

1. Вычисляется качество дерева решений на валидационной выборке;
2. Определяется множество поддеревьев без одного листового разбиения. Листовым разбиением будем называть разбиение узла, оба дочерних узла которого являются листами;
3. Для всех поддеревьев пункта 2 вычисляется качество на валидационной выборке. Выбирается поддерево с лучшим качеством;
4. Если поддерево не является корнем, для него применяется пункт (1) алгоритма.

Результатом алгоритма является множество деревьев из пункта (1) с количеством листовых узлов от 1 (корневой узел) до n (число листов в изначальном дереве). Наконец, среди множества деревьев выбирается модель с лучшим качеством. Также, возможна визуализация множества деревьев для обрезки на основе экспертного решения.

Заметим, что для оценки качества могут быть использованы как метрики анализа выживаемости (раздел 1.2), так и метрики отклонения прогноза (например, сумма квадратов разности или сумма модулей разности). В частности, при использовании метрик отклонения, обрезка может проводиться по любому признаку обучающей выборки. В частности, обрезка позволит строить деревья выживания, оптимизирующие затраты по дополнительному признаку.

Например, в области медицинского страхования может быть решена проблема определения лучшей схемы лечения пациента с учетом минимальной стоимости. В данном примере метрикой качества будет служить сумма квадратов разности, а обрезка будет производиться по признаку «стоимость лечения», который при построении дерева не является целевым.

2.4 Обработка информативности цензурирования

В реальных данных анализа выживаемости наблюдаются случаи мультимодального распределения времени наступления событий (рисунок 9). Одной из причин является ограничение длительности исследования. В таком случае, ненаступившие события становятся обрезанными (truncated) и представляются как цензурированные в последний наблюдаемый момент времени. Другой причиной является информативность цензурирования (например, принятие решения о выписки пациента по истечении определенного времени).

При разделении исходной выборки возникает проблема появления мультимодальных распределений времени в подмножествах. На рисунке 14 представлен пример построения дерева решений на наборе данных PBC. Для узлов дерева представлена ядерная оценка плотности времени событий, а также сводная информация о выборке: размер узла (размер),

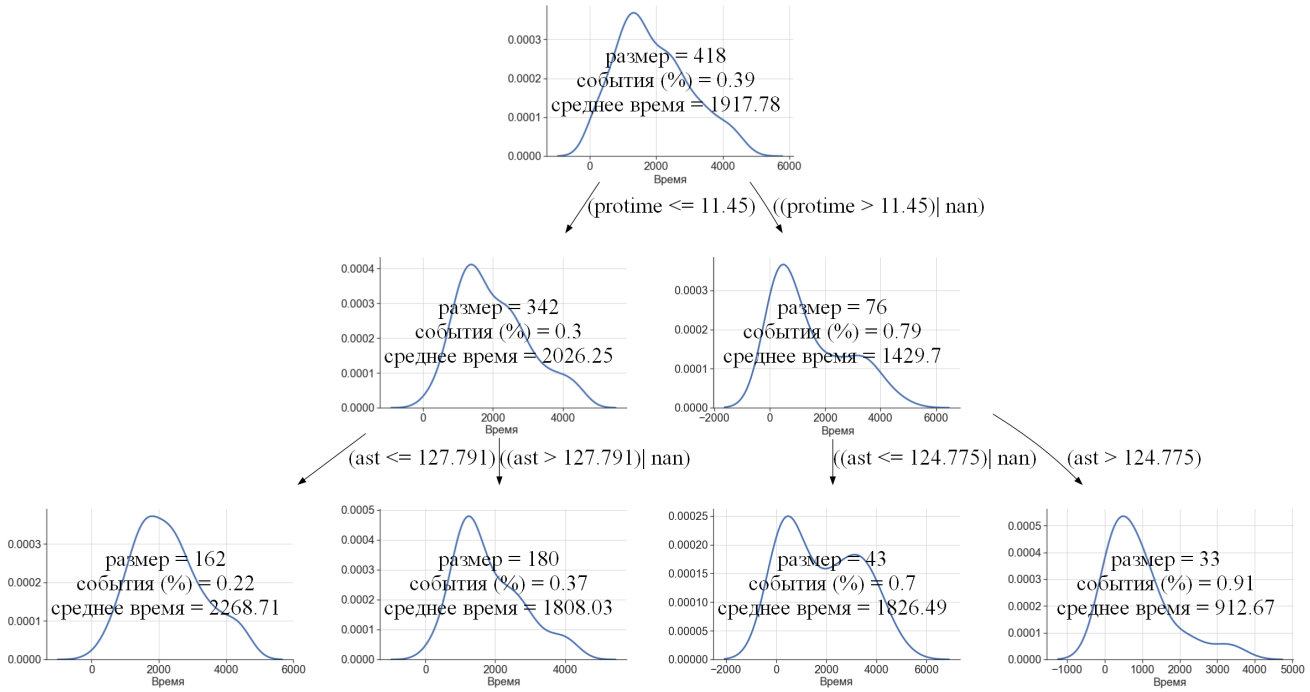


Рис. 14: Пример построения дерева выживаемости на наборе данных PBC. Построение проводилось с гиперпараметрами: глубина ($depth = 2$), размер листа ($min\_samples\_leaf = 0.5$), критерий разбиения ($criterion = logrank$). Дерево основано на признаках: protime (стандартизированное время свертывания крови), ast (аспартат-аминотрансфераза, U/мл).

доля терминальных событий (события (%)), среднее время наступления события (среднее время).

Стоит отметить, что априорное время событий набора PBC распределено одномодально (корневой узел), однако, некоторые листы дерева выживания имеют мультмодальное распределение времени. В частности, лист дерева с правилом $((protime > 11,45)|nan)\&((ast \leq 124.775)|nan)$ имеет мультимодальное распределение времени.

Одной из причин появления подвыборок с мультимодальным распределением времени являются признаки, связанные со схемой лечения. Например, летальный исход для некоторых пациентов наступает на начальном этапе после операции, а для других — намного позже этапа реабилитации. Признаки диагностики заболеваний также неявно указывают на схему лечения. Попадание пациента в зону риска (например, на основе результатов клинических анализов или анамнеза) требует врачебного вмешательства и приводит к стратификации пациентов относительно успешности схемы лечения.

Важно отметить, что критерий разбиения может приводить к повышению значимости разбиений с мультимодальностью. В разделе 2.4.1 мы рассмотрим случаи малой чувствительности критерия log-rank к выборкам с мультимодальным распределением времени. В частности, значение статистика для некоторых разбиений включает информацию о различиях выборок только на ранних моментах времени.

Наконец, в разделе 2.4.2 мы рассмотрим недостатки непараметрических оценок, построенных на выборках с мультимодальным распределением времени.

2.4.1 Чувствительность критерия log-rank

Классичический log-rank критерий предполагает независимость и случайность обеих выборок, а также неинформативность индикатора цензурирования. При работе с малыми выборками, критерий также ограничен временной шкалой и не чувствителен к поздним событиями, наблюдаемых после исчерпания одной из выборок. Для демонстрации недостатков log-rank критерия вновь рассмотрим формулу расчета статистики (18). Далее все обозначения согласованы с разделом 1.4.1.

Пусть максимальное время события в первой группе равно τ_{K1}, а во второй τ_{K2}, причем $K1 \leq K2$ (в противном случае поменяем группы местами). Тогда для $\tau_k \in (\tau_{K1}, \tau_{K2}]$ первая группа не содержит событий и $O_{1,k} = 0, N_{1,k} = 0$. Следовательно, ожидаемое количество событий $E_{i,k} = O_{i,k}$ и вклад отклонений в числитель и знаменатель равен 0. Таким образом, после наступления всех событий одной из групп, вклад **поздних событий** другой группы не учитывается.

Данный эффект влияет на появление подвыборок с мультимодальным распределением времени, поскольку критерий использует ограниченную информацию о наблюдаемых данных. Рассмотрим следующий пример. Пусть дана выборка терминальных событий с временем от 0 до 150 с шагом 10. Пусть после разбиения выборки, в первую группу вошли события с временем: $[0, 10, 20, 30, 80, 90, 100]$, а во вторую с временем $[40, 50, 60, 70, 110, 120, 130, 140, 150]$. Таким образом, исходная выборка имеет равномерное распределение времени, в то время как подгруппы имеют мультимодальное распределение. При расчете log-rank стастистики, количество наблюдений $N_{1,j}$ и $N_{2,j}$ и ожидаемое количество событий имеют следующие значения (значения округлены для удобства восприятия):

$$\overrightarrow{N}_{1,j} = [7, 6, 5, 4, 3, 3, 3, 3, 3, 2, 1, \mathbf{0,0,0,0,0}],$$

$$\overrightarrow{N}_{2,j} = [9, 9, 9, 9, 9, 8, 7, 6, 5, 5, 5, \mathbf{5,4,3,2,1}],$$

$$\overrightarrow{E}_{1,j} = [0.44, 0.4, 0.36, 0.31, 0.25, 0.27, 0.3, 0.33, 0.38, 0.29, 0.16, \mathbf{0,0,0,0,0}].$$

Жирным выделены моменты времени с поздними событиями второй группы, которые имеют нулевой вклад из-за наступления всех событий первой группы ($N_{2,j} = N_j, E_{2,j} = O_{1,j}$). С точки зрения критерия log-rank, разбиение является значимым (уровень значимости 0.05): статистика критерия равна 5.313, а p-value — 0.0211.

Однако, значимость разбиений с мультимодальными распределением времени групп не устойчива относительно изменения данных. Добавим к каждой группе по одному цензурированному наблюдению в момент времени $\tau = 160$. Обновленные значения $N_{1,j}$ и $N_{2,j}$ имеют следующий вид (жирным выделены моменты времени, которые ранее имели нулевой вклад):

$$\overrightarrow{N}_{1,j} = [8, 7, 6, 5, 4, 4, 4, 4, 4, 3, 2, \mathbf{1,\ 1,\ 1,\ 1,\ 1}, 0],$$

$$\overrightarrow{N}_{2,j} = [10, 10, 10, 10, 10, 9, 8, 7, 6, 6, 6, \mathbf{6,\ 5,\ 4,\ 3,\ 2}, 1].$$

Данное разбиение является незначимым: статистика критерия равна 1.298, а p-value 0.2545. Разительное отличие статистик связано с тем, что в обновленных данных **поздние события** имеют ненулевой вклад при расчете статистики.

Таким образом, по критерию log-rank существуют значимые разбиения, которые приводят к появлению двух групп с мультимодальным распределением времени. При добавлении новых событий, значимость разбиений неустойчива по критерию log-rank. Для использования полной информации о наблюдаемых данных, необходимо учитывать вклад поздних событий даже при исчерпании одной из выборок.

2.4.2 Недостатки непараметрических оценок

Классическая модель Каплана–Мейера также предполагает неинформативность флага цензурирования. В данном разделе мы рассмотрим влияние мультимодальности распределения времени в выборке на качество прогноза непараметрической оценки.

На рисунке 15 представлено сравнение непараметрической оценки функции выживания (красная пунктирная линия) по сравнению с целевой функцией (красная сплошная линия). Функция выживания построена на мультимодальной выборке. Синим цветом помечены отклонения до момента наступления события (занижение прогноза). Оранжевым цветом изображены отклонения после момента наступления события (завышение прогноза).

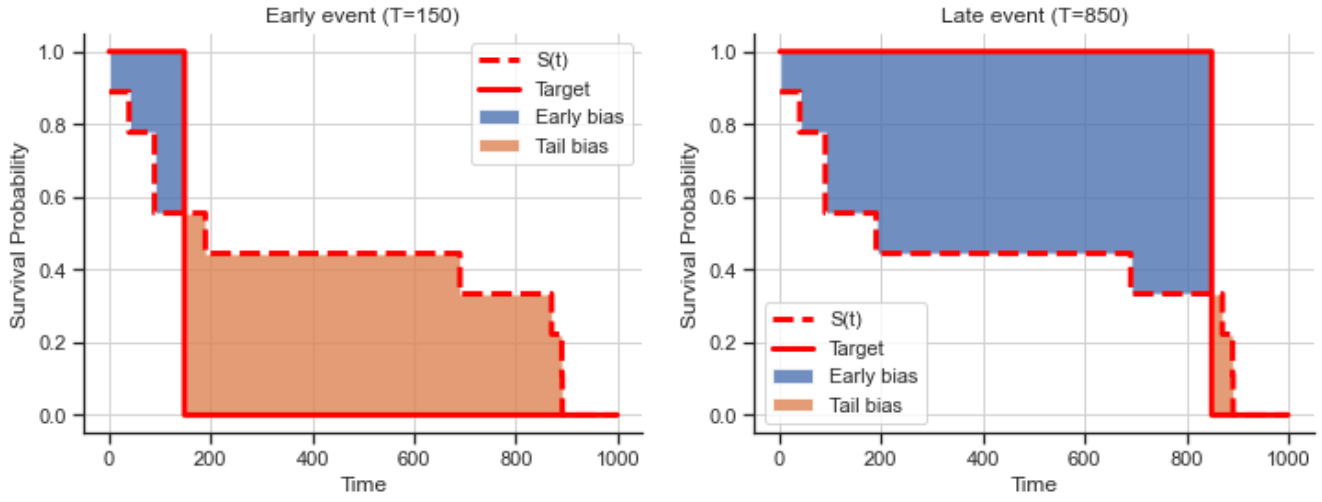


Рис. 15: Демонстрация ошибки прогноза для выборки с мультимодальным распределением. Ошибка прогноза для раннего наблюдения (T=150) основана на позднем отклонении (оранжевый цвет), для позднего наблюдения (T=850) — на раннем отклонении (синий цвет).

При построении классической модели Каплана–Мейера на выборке с ранними и поздними событиями, для ранних событий наблюдается большой вклад ошибки тяжелых хвостов — высокое значение выживаемости после истинного наступления события. Для поздних событий наблюдается большой вклад ошибки ранних отклонений — низкое значение выживаемости до момента наступления события. Таким образом, ранние события имеют завышенный прогноз функции выживания, а поздние события — заниженный прогноз.

Отметим следующие недостатки представленной непараметрической оценки. Во-первых, прогноз имеет слабую интерпретацию для целевого пользователя. Наличие стратификации по времени приводит к низкой пороговости функции. Если исходные данные репрезентативны, то новые наблюдения должны принадлежать к одной из полученных мод.

Однако, при большом расстоянии между модами, пользователь не может установить ожидаемое время события (событие ожидается либо в начале исследования, либо в конце)

Во-вторых, оценка функции выживания имеет низкое качество описания исходных данных. Непараметрическая оценка Каплана–Мейера равна константе между модальностями (из-за отсутствия наблюдений) и вносит дополнительную ошибку при оценке как ранних, так и поздних событий.

Наконец, снижается качество ранжирования исходных данных. В случае попадания ранних и поздних наблюдений в одну выборку, их прогноз вероятности и времени события, а также функции выживания и риска совпадает. Следовательно, в метриках CI и IAUC возникают ошибки неотличимости ранних и поздних событий.

Таким образом, проблема мультимодальности распределения времени возникает и в исходных данных, и при построении древовидных моделей. Причина возникновения лежит в наличии скрытых стратифицирующих признаков стратегии лечения, а также в малой чувствительности критерия log-rank. При описании данных с помощью непараметрической оценки Каплана–Мейера, наблюдается слабая интерпретация прогноза, низкое качество описания данных и повышение ошибки ранжирования.

Далее мы рассмотрим комплекс подходов для преодоления проблемы мультимодальности распределения времени как на этапе построения древовидных моделей, так и при работе с листовыми выборками.

2.4.3 Регуляризация критерия разбиения

Как говорилось ранее, современные способы борьбы с переобучением деревьев решений позволяют ограничить сложность дерева на этапе построения модели или обрезки избыточных листов. Однако подходы не используют информацию о распределении вероятностей времени событий в выборке и не позволяют преодолеть проблему обнуления вклада поздних событий (раздел 2.4.1).

В данном разделе предлагается подход регуляризации критерия log-rank для повышения чувствительности критерия к распределению вероятностей времени событий. Метод позволяет учитывать информацию об априорном распределении при сравнении выборок с цензурированием.

Формально, подход основан на добавлении информации об априорном распределении времени оставшихся и наступивших событий для всех моментов времени. Априорная информация добавляется с коэффициентом регуляризации к наблюдаемым распределениям вероятностей в выборках и определяет ненулевой вклад всех интервалов времени.

Обозначим через N_j^A априорное количество оставшихся наблюдений к моменту τ_j, O_j^A — априорное количество событий в момент τ_j. Введем дополнительный параметр регуляризации λ. Аналогично формуле критерия log-rank (раздел 2.4.1) рассмотрим две группы с количеством оставшихся наблюдений $N_{1,j}, N_{2,j}$, количеством событий $O_{1,j}, O_{2,j}$. Введем обновленные значения $\hat{N}_{1,j}, \hat{N}_{2,j}, \hat{O}_{1,j}, \hat{O}_{2,j}$ согласно следующим формулам:

$$\hat{N}_{i,j} \leftarrow N_{i,j} + \frac{\lambda}{2} \cdot N_j^A,$$

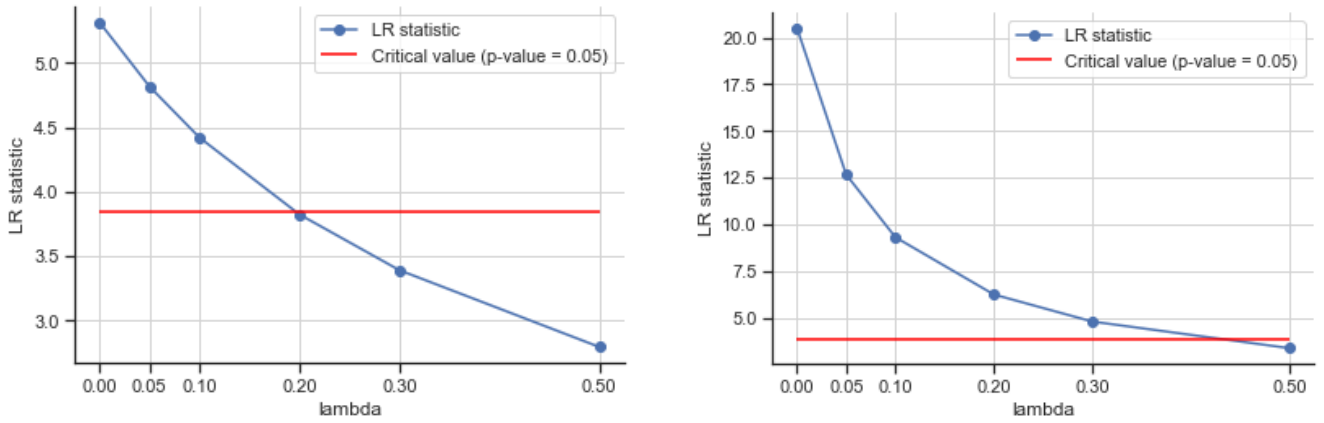
$$\hat{N}_j \leftarrow \hat{N}_{1,j} + \hat{N}_{2,j},$$

$$\hat{O}_{i,j} \leftarrow O_{i,j} + \frac{\lambda}{2} \cdot O_j^A,$$

$$\hat{O}_j \leftarrow \hat{O}_{1,j} + \hat{O}_{2,j}.$$

Определим математическое ожидание и дисперсию числа событий на момент τ_j как $\hat{E}_{i,j} = \frac{\hat{N}_{i,j}\hat{O}_j}{\hat{N}_j}$ и $\hat{V}_{i,j} = \hat{E}_{i,j}\left(\frac{\hat{N}_j - \hat{O}_j}{\hat{N}_j}\right)\left(\frac{\hat{N}_j - \hat{N}_{i,j}}{\hat{N}_j - 1}\right)$. Для оценки значимости разбиения с регуляризацией используем следующий вид статистики log-rank, основанный на нулевой гипотезе $\hat{h}_1(t) = \hat{h}_2(t)$ (подстановка полученных значений в формулу (22)):

$$LR = \frac{\sum_{j=1}^{K} w_j \left(\hat{O}_{1,j} - \hat{E}_{1,j}\right)}{\sqrt{\sum_{j=1}^{K} w_j^2 \hat{V}_{1,j}}}.$$



(a) Мультимодальное разбиение

(b) Одномодальное разбиение

Рис. 16: Оценка изменения значимости разбиения с увеличением коэффициента регуляризации. На левом графике представлено разбиение с мультимодальным распределением времени групп, на правом — с одномодальным распределением.

Приведенная модификация log-rank критерия позволяет учитывать априорное распределения вероятностей времени при $\lambda > 0$. На рисунке 16 представлен пример изменения значимости двух разбиений. На левом графике представлено разбиение с мультимодальным распределением времени групп из примера раздела 2.4.1. На правом графике представлено разбиение с одномодальным распределением групп (первая группа содержит события со временем $[0, 10, 20, 30]$, а вторая — $[40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150]$). Красная линия определяет уровень значимости статистики log-rank. Видно, что значимость разбиений убывает, однако разбиение с одномодальностью групп имеет большую устойчивость к эффекту регуляризации.

Подход регуляризации критерия log-rank позволяет учесть все интервалы времени на этапе поиска лучшего разбиения. Недостатком подхода является необходимость подбора гиперпараметра λ. При высоких значениях параметра значимость разбиения уменьшается и группы становятся неразличимы.

Псевдокод предложенного алгоритма поиска лучшего бинарного разбиения выборки с помощью взвешенных регуляризованных log-rank критериев (раздел 2.2) представлен на

Алгоритм 1 HistogramFindBestSplit – поиск лучшего разбиения выборки

Вход:

T ▷ Время наступления событий выборки
δ ▷ Флаг цензурирования наблюдений выборки
v ▷ Значение признака наблюдений выборки
T_A ▷ Время наступления событий корневой выборки
δ_A ▷ Флаг цензурирования наблюдений корневой выборки
λ ▷ Коэффициент регуляризации
$FeatureType$ ▷ Тип признака: непрерывный или категориальный
$Crit$ ▷ Весовая схема критерия log-rank

Выход: BestSplit ▷ Лучшее бинарное разбиение выборки по признаку

T_{nan}, δ_{nan} ← getNanValues(T, δ, v)
$T_{\overline{nan}}$, $\delta_{\overline{nan}}$, $v_{\overline{nan}}$ ← getNotNanValues(T, δ, v)
if FeatureType = Categorical **then**
    $v_{\overline{nan}}$, map_categ ← WeightOfEvidence($T_{\overline{nan}}$, $\delta_{\overline{nan}}$, $v_{\overline{nan}}$)
s_i ← GetQuantile($v_{\overline{nan}}$)

n_j^l, O_j^l ← 0
n_j^r, O_j^r ← CountHistogram($T_{\overline{nan}}$, $\delta_{\overline{nan}}$)
$n_{nan,j}$, $O_{nan,j}$ ← CountHistogram($T_{\overline{nan}}$, $\delta_{\overline{nan}}$)
n_j^A, O_j^A ← CountHistogram(T_A, δ_A)
(n_j^l, O_j^l) += $(\lambda * n_j^A, \lambda * O_j^A)$
(n_j^r, O_j^r) += $(\lambda * n_j^A, \lambda * O_j^A)$

Splits ← []
for s in s_i **do**
    Mask ← $v_{\overline{nan}} = s$
    $n_{j|v=s}$, $O_{j|v=s}$ ← CountHistogram($T_{\overline{nan}}$[Mask], $\delta_{\overline{nan}}$[Mask])
    (n_j^l, O_j^l) += $(n_{j|v=s}, O_{j|v=s})$
    (n_j^r, O_j^r) −= $(n_{j|v=s}, O_{j|v=s})$
    Splits ← Splits ∪ LogRankCriterion(n_j^l, O_j^l, $n_j^r + n_{nan,j}$, $O_j^r + O_{nan,j}$, $Crit$)
    Splits ← Splits ∪ LogRankCriterion($n_j^l + n_{nan,j}$, $O_j^l + O_{nan,j}$, n_j^r, O_j^r, $Crit$)
BestSplit ← SelectBest(Splits)
return BestSplit

Рис. 17: Псевдокод алгоритма поиска лучшего разбиения выборки.

Рисунке 17. На рисунке 18 представлен псевдокод алгоритма разбиения вершины дерева выживания на две дочерние выборки (раздел 2.3.1), а на рисунке 19 представлен псевдокод алгоритма построения дерева выживания. Наконец, на рисунке 20 представлен псевдокод алгоритма обрезки дерева post-pruning (раздел 2.3.3).

2.4.4 Модификация листовых оценок

Представленные выше подходы позволяют уменьшить вероятность появления выборок с мультимодальным распределением времени. Но применение подходов ограничено предположением, что существует значимое разбиение выборки с одномодальным распределением двух подвыборок.

Алгоритм 2 Split – разбиение вершины на две дочерние выборки

Вход:
Node ▷ Вершина дерева выживания
Выход: SubNodes ▷ Набор дочерних вершин
$X \leftarrow$ GetFeatures(Node)
$T, \delta \leftarrow$ GetTarget(Node)
$R_f \leftarrow []$
for parallel(f in X) **do**
    $R_f \leftarrow R_f \cup$ HistogramFindBestSplit(T, δ, f)
$R \leftarrow$ SelectSplitByFeature(R_f)
SubNodes $\leftarrow []$
if $R[N_{sign}] = 0$ **then**
    return SubNodes
Samples $\leftarrow$ SplitByRule(Node, $R[\hat{s}]$)
for Sample in Samples **do**
    SubNodes $\leftarrow$ SubNodes $\cup$ CreateNode(Sample)
return SubNodes

Рис. 18: Псевдокод алгоритма разбиения вершины дерева выживания на две дочерние выборки.

Алгоритм 3 TreeConstruction – построение дерева выживания

Вход:
X, ▷ Признаковое пространство наблюдений
T, ▷ Время наступления событий
δ, ▷ Флаг цензурирования наблюдений
MaxDepth ▷ Максимальная глубина дерева
Выход: Tree ▷ Корень дерева выживания
InitNode $\leftarrow$ CreateNode(X, T, δ)
StackNodes $\leftarrow$ [InitNode]
while length(StackNodes) > 0 **do**
    Node $\leftarrow$ StackNodes.pop()
    if Node.depth $\geq$ MaxDepth **then**
        continue
    SubNodes $\leftarrow$ Split(Node)
    StackNodes $\leftarrow$ StackNodes $\cup$ SubNode
return InitNode

Рис. 19: Псевдокод алгоритма построения дерева выживания.

На практике, такого разбиения может не существовать. Например, в случае появления одного мультимодального распределения времени для всех точек разбиения, или в случае мультимодального распределения времени исходной выборки. Для преодоления данных недостатков предлагается модифицировать непараметрические оценки функции выживаемости и риска.

Главным недостатком прогноза является константность функции между модами распределения времени. Мультимодальное распределение ставит перед экспертом выбор между

Алгоритм 4 TreePostPrunning – обрезка дерева выживания

Вход:

$Tree$, ▷ Дерево выживания
X, ▷ Признаковое пространство наблюдений валидационной выборки
T, ▷ Время наступления событий валидационной выборки
δ, ▷ Флаг цензурирования наблюдений валидационной выборки
Q ▷ Метрика качества

Выход: BestTree ▷ Дерево выживания после обрезки

```
BestTrees ← [Tree]
TreesQuality ← [Q(T, δ, Tree.predict(X))]
while Tree.NumLeaves() > 1 do
    BestSubTrees ← []
    SubTreesQuality ← []
    for l in Tree.GetLeaves() do
        SubTree ← Tree.remove(l)
        BestSubTrees ← BestSubTrees ∪ SubTree
        SubTreesQuality ← BestSubTrees ∪ Q(T, δ, SubTree.predict(X))
    BestInd ← SelectBestValue(SubTreesQuality, Q)
    BestTrees ← BestSubTrees[BestInd]
    TreesQuality ← SubTreesQuality[BestInd]
BestInd ← SelectBestValue(TreesQuality, Q)
BestTree ← BestTrees[BestInd]
return BestTree
```

Рис. 20: Псевдокод алгоритма обрезки дерева выживания.

несколькими наиболее возможными исходами. Также, необходимо обрабатывать случаи информативного цензурирования.

Древовидные модели предполагают, что наблюдения в листе обладают схожей теоретической функцией выживания. Для повышения качества описания данных предлагатся заменить мультимодальное распределение вероятностей времени на одномодальное распределение, а также избавиться от информативности цензурирования. Функция выживания, построенная на одномодальных данных, имеет лучшую интерпретацию для пользователя за счет улучшения пороговости.

Для достижения данных целей, в работе предлагается подход генерации виртуальных событий. На основе характеристик выборки определим теоретическое распределение, которое наилучшим образом описывает время событий. Данное распределение используется для генерации времени виртуальных событий, которые не наблюдались в исходной выборке. Для поддержки предположения о неинформативности цензурирования, каждому виртуальному событию сопоставляется индикатор исхода с вероятностью исхода в выборке.

Рассмотрим семейство нормальных распределений $\mathcal{N}(\mu, \sigma^2)$ со средним значением μ и дисперсией σ^2. Теоретическое распределение определяется на основе выборочного среднего $\overline{T} = (\sum_{i=1}^{n} T_i)/n$ и выборочной дисперсии $S^2 = (\sum_{i=1}^{n}(T_i - \overline{T})^2)/(n-1)$, где n — размер выборки, $\{T_i\}$ — значения времени в выборке. Виртуальные события описываются временем $\tilde{T} \sim \mathcal{N}(\overline{T}, S^2)$ и флагом цензурирования $\tilde{\delta} \sim Bernoulli((\sum_{i=1}^{n} \delta_i)/n)$, где $\{\delta_i\}$ — флаг цензурирования в выборке. Наконец, на основе множеств $\{\tilde{T}\}, \{\tilde{\delta}\}$ строится непараметрическая

оценка Каплана–Мейера.

Таким образом, предложенная модификация непараметрической оценки Каплана–Мейера (KMWV) не требует выполнения предположения о неинформативности индикатора наступления события и применима к данным с мультимодальным распределением вероятностей времени.

2.5 Выводы

В данной главе проводились исследование и разработка метода построения деревьев выживаемости. По итогам проведенного исследования были достигнуты следующие результаты:

- Определены особенности 6 реальных медицинских наборов данных. К особенностям признакового пространства относится наличие категориальных признаков и пропущенных значений. К особенностям целевых переменных относятся: различное распределение вероятностей времени событий, информативность цензурирования, дисбаланс цензурирования.
- Предложен алгоритм поиска лучшего бинарного разбиения выборки по непрерывным и категориальным признакам, который позволяет сократить время работы и используемую память по сравнению с подходом независимого перебора промежуточных точек. Подход отображения категориальных значений на числовую прямую по методу Weight Of Evidence располагает категории на основе схожести подвыборок по цензурируемости. Для контроля сложности вычислений, промежуточные точки непрерывных признаков дискретизируются на основе квантилей. Для обработки пропущенных значений применяется подход выбора ветви с наибольшей значимостью разбиения. Сравнение выборок с цензурированием производится по взвешенному критерию log-rank, обеспечивая большую гибкость при учете распределения времени событий.
- Предложен метод построения деревьев выживания. Построение моделей основано на поиске лучших разбиений по предложенному алгоритму. Лучший признак для разбиения выборки выбирается с учетом поправки Бонферрони, отдавая предпочтение редким значимым разбиениям. Для борьбы с переобучением модели разработаны подходы контроля роста дерева и обрезки модели. Построенная модель позволяет прогнозировать вероятность и время наступления события, функции выживания и риска;
- Исследована проблема влияния информативности цензурирования на обучение древовидной модели. Информативность цензурирования может приводить к появлению мультимодального распределения вероятностей времени событий в узлах дерева выживания. Определены недостатки критерия log-rank и непараметрической оценки Каплана–Мейера при работе с мультимодальными распределениями времени. Предложен подход регуляризации критериев log-rank и модификация оценки Каплана–Мейера с виртуальными событиями. Регуляризация критерия позволяет учитывать информацию об априорном распределении времени событий при поиске разбиений выборок. Непараметрическая модель $KMWV$ основана на генерации виртуальных событий по одномодальному распределению времени с последующим построением оценки Каплана–Мейера.

3 ОЦЕНКА И СРАВНЕНИЕ МОДЕЛЕЙ АНАЛИЗА ВЫЖИВАЕМОСТИ

При работе (при подготовке) над данным разделом диссертации использованы следующие публикации автора, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования:

- *Vasilev I., Petrovskiy M., Mashechkin I. Sensitivity of Survival Analysis Metrics // Mathematics. – 2023. – Т. 11. – №. 20. – С. 4246.*

В предыдущей главе данной диссертационной работы были выделены особенности рассматриваемых наборов данных и предложен метод построения деревьев выживаемости, учитывающий особенности на этапе обучения модели.

В частности, распределение вероятности времени событий влияет на значимость правил разбиения выборок и построение непараметрических оценок. Для учета распределения событий используется подход регуляризации деревьев и взвешенные схемы критериев разбиения. Также, дизбаланс цензурированных и терминальных событий может смещать прогноз моделей. Классическим подходом борьбы с дисбалансом является балансировка данных (увеличение минорного класса или уменьшение доминирующего класса). Однако такой подход может приводить к искажению априорных вероятностей типов событий.

Влияние выделенных особенностей распространяется не только на сами прогнозные модели, но и на методы оценки их качества. В классическом машинном обучении метрики позволяют проводить оценку качества моделей, а также используются в качестве функции потерь для решения задач оптимизации на этапе построения модели. Наконец, метрики влияют на отбор гиперпараметров и выбор оптимальной модели для введения в эксплуатацию.

Критический анализ преимуществ и недостатков существующих метрик позволит достоверно оценивать качество прогнозирования моделей и мотивировать исследователей проводить дополнительную проверку устойчивости метрик к особенностям данных. В дальнейшем, информация об устойчивости функций потерь может быть использована для разработки бустинговых моделей анализа выживаемости.

Задачей данной главы является исследование корректности выбранных метрик анализа выживаемости и выбор итогового набора метрик для оценки качества моделей. На основе выбранных метрик проводится экспериментальное исследование качества предложенного метода построения деревьев выживаемости, а также сравнение с существующими методами построения древовидных моделей, описанных в разделе 1.4.

3.1 Анализ чувствительности метрик качества

В разделе 1.2.3 была представлена мотивация выбора метрик качества для оценки прогнозирования ожидаемого времени события, функции выживания и функции риска. Для обеспечения корректной оценки моделей и верной интерпретации качества прогноза, необходимо исследовать применимость метрик к особенностям данных.

В данном разделе проводится исследование избыточной чувствительности метрик к особенностям данных. Под избыточной чувствительностью будем понимать наличие неравного вклада отдельных величин при равенстве условий оценки. Для обеспечения равных

условий оценки событий рассмотрим случай неопределенности времени наступления события при $S(t) = 0.5$. Такой прогноз предполагает, что все наблюдения имеют одинаковый риск события в любой момент времени. Прогноз не имеет преимущества для определенных индивидуальных событий, интервалов временой шкалы или типов события.

В то же время, при использовании единичного $S(t) = 1$ прогноза, качество описания цензурированных наблюдений выше качества терминальных (учитывается прогноз до момента цензурирования). При использовании нулевого прогноза $S(t) = 0$, качество описания ранних событий по метрике IBS выше качества поздних (отклонения от эталонной функции после момента наступления события равны 0).

Для всестороннего исследования метрик качества предлагается рассмотреть следующие случаи избыточной чувствительности:

1. **Значимость вклада отдельных событий.** Метрика может иметь неявную зависимость вклада событий от истинного времени, влияющую на достоверность оценки. Например, повышенный вклад поздних событий может быть не обоснован для данных с доминирующим числом ранних событий и искажать качество прогнозирования.
2. **Зависимость интегральных сумм от времени.** Значения метрики могут иметь неявную зависимость от временной шкалы. В таком случае, повышенная значимость определенного промежутка времени может быть необоснована для различных данных.
3. **Влияние времени при расчете интегральной метрики.** Метод интегрирования непосредственно влияет на агрегацию во времени и может иметь завышенную значимость определенного промежутка времени.
4. **Устойчивость к дисбалансу цензурирования.** Доминирование цензурированных наблюдений может привести к ложному завышению или занижению метрики.

Устойчивость метрики к каждому случаю определяется равенством вкладов: (1) отдельных наблюдений, (2) интервалов временной шкалы, (3) частичных отрезков разбиения и (4) типов событий. Стоит отметить, что случаи (2)–(3) корректны только в терминах интегральных метрик. Далее подробнее рассмотрим каждый случай избыточной чувствительности.

3.1.1 Значимость вклада отдельных событий

Исследуем метрики IBS, $IAUC$, $AUPRC$ на наличие зависимости вклада одиночных событий. Для проверки зависимости визуализируем значения метрики каждого наблюдения относительно истинного времени наступления события. Необходимым требованием для проверки является возможность представления метрики в виде суммы оценок на отдельных наблюдениях (расчет качества по каждому наблюдению).

IBS

Утверждение 3. *Вклад отдельных наблюдений в метрике IBS монотонно неубывает относительно истинного времени события в случае неопределенности времени наступления события.*

Доказательство. Приведем альтернативный вид формулы (8). В частности, внесем операцию интегрирования для каждой итерации суммирования. Так как суммирование и константа

N не зависят от времени наблюдения, формула имеет следующий вид:

$$IBS = \frac{1}{N}\sum_i \frac{1}{t_{max}} \int_0^{t_{max}} \left(\begin{cases} \frac{(0-S(t|X_i))^2}{G(T_i)}, & \text{if } T_i \leq t, \delta_i = 1, \\ \frac{(1-S(t|X_i))^2}{G(t)}, & \text{if } T_i > t, \\ 0, & \text{if } T_i = t, \delta_i = 0, \end{cases} \right) dt.$$

Раскроем значение интеграла для каждого из условий системы:

$$IBS = \frac{1}{N}\sum_i \frac{1}{t_{max}} \left(\int_0^{T_i} \frac{(1-S(t|X_i))^2}{G(t)} dt + \int_{T_i}^{t_{max}} \delta_i \cdot \frac{(0-S(t|X_i))^2}{G(T_i)} dt \right).$$

Следовательно, IBS представляется в виде суммы значений IBS для конкретного наблюдения i с соотвествующим временем события T_i, признаками X_i и флагом цензурирования δ_i:

$$IBS^i = \frac{1}{t_{max}} \left(\int_0^{T_i} \frac{(1-S(t|X_i))^2}{G(t)} dt + \int_{T_i}^{t_{max}} \delta_i \cdot \frac{(0-S(t|X_i))^2}{G(T_i)} dt \right), i = 1, 2, ..., N. \tag{24}$$

Тогда альтернативный вид метрики IBS является усреднением отдельных IBS_i для каждого из наблюдений выборки. Приведенный вид формулы IBS позволяет расчитывать значение метрики для каждого отдельного наблюдения:

$$IBS = \frac{1}{N}\sum_i IBS^i. \tag{25}$$

Рассмотрим значения отдельных вкладов для константного прогноза $S(t) = 0.5$. Таким образом, формула (24) имеет вид:

$$IBS^i = \frac{1}{t_{max}} \left(\int_0^{T_i} \frac{(0.5)^2}{G(t)} dt + \int_{T_i}^{t_{max}} \delta_i \cdot \frac{(0.5)^2}{G(T_i)} dt \right), i = 1, 2, ..., N. \tag{26}$$

Рассмотрим два терминальных события i, j наступивших в моменты времени $T_i \leq T_j$. Следовательно, $G(T_i) \geq G(T_j)$ и разность вкладов равна:

$$IBS^j - IBS^i = \frac{1}{t_{max}} \left(\int_{T_i}^{T_j} \frac{0.25}{G(t)} dt + \int_{T_j}^{t_{max}} \frac{0.25}{G(T_j)} dt - \int_{T_i}^{t_{max}} \frac{0.25}{G(T_i)} dt \right).$$

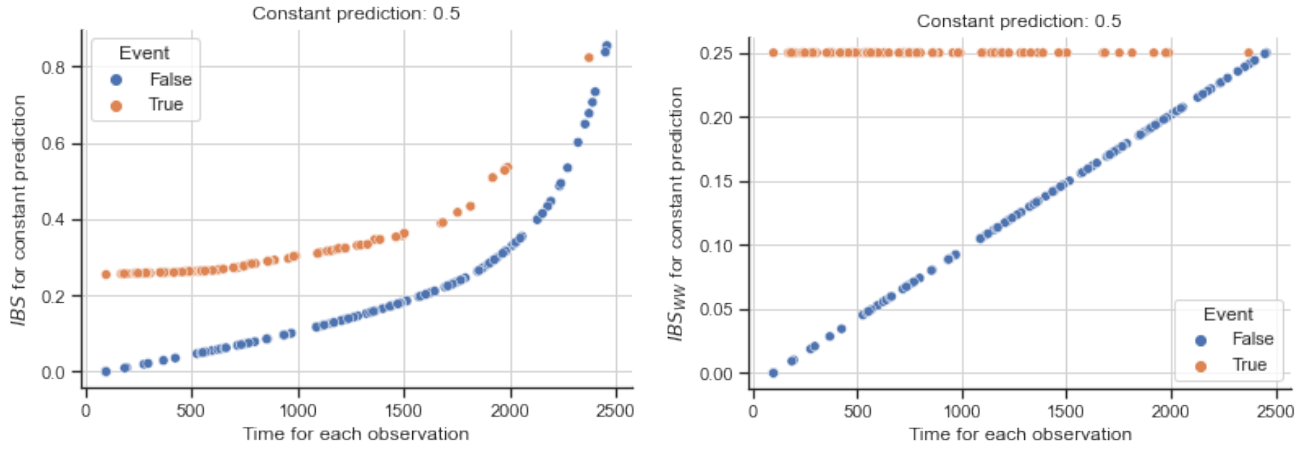
Согласно неравенству $\frac{1}{G(T_i)} \leq \frac{1}{G(t)} \leq \frac{1}{G(T_j)}$, при $t \in [T_i, T_j]$, имеем:

$$IBS^j - IBS^i \geq \frac{1}{t_{max}} \left(\int_{T_i}^{T_j} \frac{0.25}{G(t)} dt - \int_{T_i}^{T_j} \frac{0.25}{G(T_i)} dt \right) \geq 0.$$

Таким образом, $IBS^j \geq IBS^i$ при $T_i \leq T_j$, причем равенство достигается в случае $G(T_i) = G(T_j)$.

□

Таким образом, одним из недостатков IBS является увеличение значений метрики относительного истинного времени события. Рассмотрим пример зависимости вкладов наблюдений от времени при прогнозировании константной функции выживания $S(t) = 0.5$. На левом рисунке 21 представлены значения метрики IBS для наблюдений набора GBSG. По горизонтальной оси отмечено истинное время события, по вертикальной оси — полученное значение метрики IBS^i. Цвет точек определяет тип события: синие точки определяют цензурированные события, а оранжевые определяют терминальные события. Исходя из левого графика видно, что при увеличении истинного времени события увеличивается и значение IBS для отдельного наблюдения. Следовательно, при расчете общей метрики для всех наблюдений по формуле (25), больший вклад вносят поздние редкие наблюдения.



(a) IBS (b) IBS_{WW}

Рис. 21: Пример зависимости значений IBS^i и IBS^i_{WW} для каждого наблюдения набора данных GBSG относительно истинного времени события. IBS увеличивается в зависимости от времени события и определяет больший вклад для поздних наблюдений. IBS_{WW} постоянен во времени для терминальных событий и определяет равные вклады.

Отметим, что существующие исследования нацелены на анализ ранних событий, которые имеют меньший вклад при расчете метрики IBS. Например, использование взвешенных схем критерия log-rank приводит к увеличению значимости ранних событий при поиске разбиения. Также отметим, что терминальные и цензурированные наблюдения определяют две кривые, причем кривая цензурированных событий расположена ниже. Такой эффект наблюдается из-за расчета отклонений только до момента цензурирования.

Для преодоления недостатков существующей метрики IBS предлагается использовать метрику $IBS_{WW}(t)$ без весовой схемы (предполагая $G(t) = 1$):

$$BS_{WW}(t) = \frac{1}{N}\sum_i \begin{cases} (0 - S(t|X_i))^2, & \text{if } T_i \leq t, \delta_i = 1, \\ (1 - S(t|X_i))^2, & \text{if } T_i > t, \\ 0, & \text{if } T_i = t, \delta_i = 0, \end{cases} \tag{27}$$

$$IBS_{WW} = \frac{1}{t_{max}} \int_0^{t_{max}} BS_{WW}(t)dt. \tag{28}$$

На правом рисунке 21 представлены значения IBS_{WW} для каждого наблюдения набора данных GBSG. Обозначения эквивалентны левому рисунку 21. Согласно рисунку, значение метрики для терминальных событий равно константе 0.25 и устраняет возрастающую зависимость от истинного времени. Для цензурированных наблюдений сохраняется возрастающая зависимость, однако она принимает линейный вид. Интерпретация линейной зависимости основана на расчете отклонений только до момента цензурирования и ограничена значением 0.25 (в случае последнего подвергнутого цензуре наблюдения).

Таким образом, использование весовой схемы $1/G(t)$ приводит к избыточной чувствительности к поздним наблюдениям даже в случае константного прогноза функции выживания. При переходе от весовой схемы к константному вкладу отклонений возрастающая зависимость исчезает. Альтеративным недостатком системы весов $1/G(t)$ является зависимость метрики от оценки генеральной функции выживания цензурированных наблюдений $G(t)$. Оценка $G(t)$ изменяется с увеличением объема данных, что приводит к искажению предыдущих значений метрики.

IAUC и CI

Метрики $IAUC$ и CI (формулы (6) и (4)) не представимы в виде суммы $IAUC^i$ и CI^i, которые зависят только от прогноза для i-го наблюдения. В частности, расчет отдельных оценок предполагает сравнение истинной и прогнозируемой функции риска, а принцип расчета качества $IAUC$ и CI основан на сравнении значений (функций риска для $IAUC$ и ожидаемого времени для CI) для пар наблюдений.

AUPRC

Утверждение 4. *Вклад отдельных наблюдений в метрике $AUPRC$ не зависит от истинного времени события в случае неопределенности времени наступления события.*

Доказательство. Представление $AUPRC$ в виде суммы отдельных оценок определен в формуле (9). Следовательно, при условии $S(t) = 0.5$, значение метрики для терминальных наблюдений имеет следующий вид:

$$AUPRC_{\delta=1}(\hat{S}, T_i) = \int_0^1 \hat{S}(T_i \cdot \varphi) - \hat{S}(T_i/\varphi)d\varphi = \int_0^1 0.5 - 0.5 \; d\varphi = 0.$$

При условии $S(t) = 0.5$, значение метрики для цензурированных наблюдений имеет следующий вид:

$$AUPRC_{\delta=0}(\hat{S}, T_i) = \int_0^1 \hat{S}(T_i \cdot \varphi)d\varphi = \int_0^1 0.5 \; d\varphi = 0.5.$$

Следовательно, для каждого типа события метрика определяет равный вклад отдельных наблюдений, не зависящий от времени. □

3.1.2 Зависимость метрики от времени

В данном разделе проводится исследование поведения временных компонент метрик $AUC(t)$, $BS(t)$ и $AUPRC(\varphi)$ относительно временной шкалы. В отличии от раздела 3.1.1, в данном разделе рассматривается общее качество прогноза по выборке в фиксированный момент времени. Для расчета интегральных значений временные компоненты агрегируются по всей временной шкале.

IBS

Метрики $BS(t)$ (формула (7)) и $BS_{WW}(t)$ (формула (27)) основаны на усреднении отклонений всех наблюдений (общее число наблюдений N) для момента времени t. Рассмотрим пример оценки качества прогноза в случае неопределенности времени события и докажем следующие утверждения.

Утверждение 5. *Метрика IBS определяет различные значения компонент $BS(t)$ в случае неопределенности времени наступления события.*

Доказательство. Для фиксированного момента времени t значение $BS(t)$ определяется по формуле (7). С учетом непрерывности временной шкалы, определим момент времени $t < \hat{t}$, такой, что за промежуток времени $(t, \hat{t}]$ произошло ровно одно событие k с истинным временем T_k и индикатором цензурирования δ_k.

Следовательно, разность между $BS(\hat{t})$ и $BS(t)$ отличается только ошибкой на событии k: в $BS(t)$ событие входит с ошибкой $\frac{0.25}{N \cdot G(t)}$ (поскольку событие ещё не произошло), а в $BS(\hat{t})$ событие входит с ошибкой $\frac{0.25}{N \cdot G(T_k)}$ (в случае терминального события) или с 0 (в случае цензурированного события).

Таким образом, в случае терминального события разность равна: $BS(\hat{t}) - BS(t) = \frac{0.25}{N \cdot G(T_k)} - \frac{0.25}{N \cdot G(t)} \geq 0$ (поскольку $G(t) \geq G(T_k)$). В случае цензурированного события разность равна: $BS(\hat{t}) - BS(t) = 0 - \frac{0.25}{N \cdot G(t)} < 0$. Следовательно, $\forall\, t\ \exists\, \hat{t} : \hat{t} > t$ верно $BS(\hat{t}) \neq BS(t)$ (при $G(\hat{t}) \neq G(t)$). □

Утверждение 6. *Значения временных компонент $BS_{WW}(t)$ монотонно невозрастают в случае неопределенности времени наступления события.*

Доказательство. Для фиксированного момента времени t значение $BS_{WW}(t)$ определяется по формуле (27). Аналогично утверждению 5, определим момент времени $\hat{t}$.

Следовательно, разность между $BS_{WW}(\hat{t})$ и $BS_{WW}(t)$ отличается только ошибкой на событии k: в $BS_{WW}(t)$ событие входит с ошибкой $\frac{0.25}{N}$ (поскольку событие ещё не произошло), а в $BS_{WW}(\hat{t})$ событие входит с ошибкой $\frac{0.25}{N}$ (в случае терминального события) или с 0 (в случае цензурированного события).

Если событие терминальное, то разность $BS(\hat{t})$ и $BS(t)$ равна 0. Если событие цензурировано, то разность отрицательна и равна $-\frac{0.25}{N}$. Таким образом, $\forall\ t, \hat{t} : \hat{t} > t$ верно $BS_{WW}(\hat{t}) \leq BS_{WW}(t)$. Следовательно, нулевая ошибка цензурированных наблюдений приводит к ложному повышению качества прогнозирования. □

На рисунке 22 представлено изменение квадратичных отклонений во времени. По горизонтальной оси отмечена временная шкалы, по вертикальной оси — полученное значение метрик во времени.

Синия линия определяет значения метрики $BS(t)$. Наибольшие значения метрики достигаются в временном интервале от 1500 до 2500, а ранние наблюдения (наступившие до 1000 момента времени) определяют меньший вклад в итоговое значение IBS. Оранжевая линия определяет значения метрики $BS_{WW}(t)$, которые, в отличие от метрики $BS(t)$, монотонно убывают во времени и определяют повышенный вклад для ранних событий.

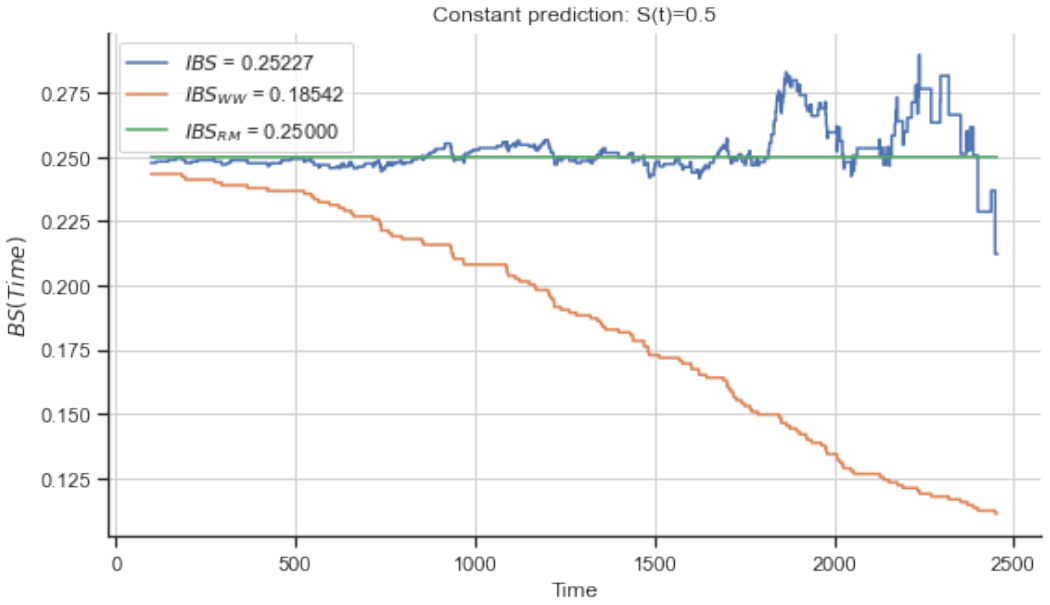


Рис. 22: Пример изменения метрик семейства $BS(t)$ во времени при константном прогнозе $S(t) = 0.5$. Метрика IBS отмечена синим, IBS_{WW} оранжевым, а IBS_{RM} зеленым. Видно, что метрики IBS и IBS_{WW} изменяются во времени и имеют ложную чувствительность. Предложенная метрика IBS_{RM} определяет равный вклад во времени.

Для преодоления проблем метрик $BS(t)$ и $BS_{WW}(t)$ рассмотрим модификацию $BS_{RM}(t)$, основаннную на контролируемом усреднении наблюдаемых событий к моменту t. В таком случае, константа общего числа событий N заменяется на переменную $N(t) = N_{event} + N_{cens}(t) = N_{event} + \sum_{i:\delta_i=0} I(T_i > t)$. Следовательно, следующая модификация метрики не учитывает вклад цензурированных наблюдений после момента выхода:

$$BS_{RM}(t) = \frac{1}{N(t)} \sum_i \begin{cases} (0 - S(t|X_i))^2, & \text{if } T_i \leq t, \delta_i = 1, \\ (1 - S(t|X_i))^2, & \text{if } T_i > t, \\ 0, & \text{if } T_i = t, \delta_i = 0, \end{cases} \tag{29}$$

$$IBS_{RM} = \frac{1}{t_{max}} \int_0^{t_{max}} BS_{RM}(t)dt. \tag{30}$$

На рисунке 22 зеленая линия определяет значения метрики $BS_{RM}(t)$. Метрика на содержит ложную чувствительность относительно времени так как вклад каждого времени равен 0.25. Важно отметить, что исследование раздела 3.1.1 основано на расчете единичных IBS^i, предполагая $N = N(t) = 1$. Следовательно, данная особенность чувствительности от

времени влияет только на расчет интегральной метрики по множеству наблюдений. В случае единичных наблюдений, $IBS^i_{RM} = IBS^i_{WW}$.

Таким образом, весовая схема $1/G(t)$ и подход полного усреднения отклонений вносит излишнюю чувствительность к поздним и ранним событиям соответственно. При использовании модификации с константным вкладом отклонений и контролируемым усреднением наблюдаемых событий, ложная чувствительность исчезает. Также, докажем следующее свойство метрик семейства IBS.

Утверждение 7. *Метрика IBS вогнута вниз относительно истинного времени наступления событий при равном прогнозе функции выживания.*

Доказательство. При $\delta = 1$ функция $IBS(T, \hat{S}, \delta = 1)$ имеет следующий вид:

$$IBS(T, \hat{S}) = \frac{1}{t_{max}} \left(\int_0^T (1 - S(t))^2 dt + \int_T^{t_{max}} S^2(t) dt \right).$$

Так как $\frac{d}{dT} S(t) = 0$ ($S(t)$ не зависит от T), по формуле Лейбница имеем:

$$\frac{dIBS(T, \hat{S})}{dT} = \frac{1}{t_{max}} \left((1 - S(T))^2 \cdot 1 - S^2(T) \cdot 1 \right) = \frac{1}{t_{max}} \left(1 - 2 \cdot S(T) \right).$$

По определению функции риска: $\frac{d}{dT} S(T) = -f(T)$ (где $f(T) \geq 0$), имеем:

$$\frac{d^2 IBS(T, \hat{S})}{dT^2} = \frac{2 \cdot f(T)}{t_{max}} \geq 0.$$

По достаточному свойству вогнутости функций, $IBS(T, \hat{S}, \delta = 1)$ вогнута на множестве $\{T\}$, причем минимальное значение функции достигается при $S(T) = 0.5$.

Аналогично, при $\delta = 0$ функция $IBS(T, \hat{S}, \delta = 0)$ имеет следующий вид:

$$IBS(T, \hat{S}) = \frac{1}{t_{max}} \left(\int_0^T (1 - S(t))^2 dt \right).$$

Так как $\frac{d}{dT} S(t) = 0$ ($S(t)$ не зависит от T), по формуле Лейбница имеем:

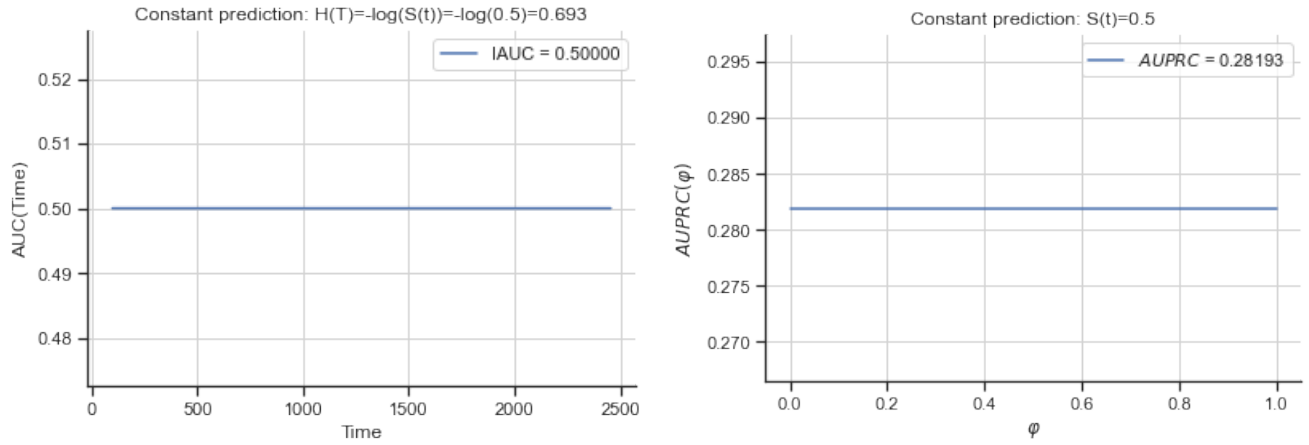
$$\frac{dIBS(T, \hat{S})}{dT} = \frac{1}{t_{max}} \left((1 - S(T))^2 \cdot 1 \right).$$

По определению функции риска: $\frac{d}{dT} S(T) = -f(T)$ (где $f(T) \geq 0, 1 \geq S \geq 0$), имеем:

$$\frac{d^2 IBS(T, \hat{S})}{dT^2} = \frac{2}{t_{max}} \cdot (1 - S(T)) \cdot f(T) \geq 0.$$

По достаточному свойству вогнутости функций, $IBS(T, \hat{S}, \delta = 0)$ вогнута на множестве $\{T\}$, причем функция монотонно возрастает. □

IAUC



(a) $AUC(t)$ (b) $AUPRC(\varphi)$

Рис. 23: Пример изменения метрик $AUPRC(\varphi)$ и $AUC(t)$ во времени при константном прогнозе $S(t) = 0.5$ и $H(t) = -\log(S(t)) = -\log(0.5)$. Видно, что обе метрики определяют равный вклад во времени.

Аналогично, рассмотрим поведение значений метрики AUC во времени (левый график рисунка 23). Так как метрика позволяет оценивать качество прогнозирования функции кумулятивного риска, воспользуемся эквивалентным преобразованием: $H(t) = -\log(S(t))$. Для константного прогноза $S(t) = 0.5$ получаем $H(t) = -\log(0.5) = \log(2) = 0.693$. На основе визуализации можно сделать вывод, что вклад каждого времени равен 0.5 и метрика не имеет ложной чувствительности относительно времени.

AUPRC

На правом графике рисунка 23 представлен пример изменения метрики $AUPRC$ по временной шкале (определяется параметром φ) для константного прогноза $S(t) = 0.5$ набора данных GBSG. На основе визуализации можно сделать вывод, что метрика $AUPRC$ определяет равный вклад во времени.

3.1.3 Влияние времени при расчете интеграла

В разделе 3.1.2 была рассмотрена чувствительность временных компонент значений метрик относительно момента t. Перед получением интегральных значений необходимо исследовать наличие чувствительности метрик к переменной интегрирования.

Отличительной особенностью интегральных метрик (IBS, IAUC, AUPRC) является агрегация качества для всех моментов времени путем вычисления интеграла по временной шкале. На практике, временная шкала может быть задана пользователем для прогнозирования функции выживания и функции риска. В данной работе, временная шкала является множеством всех времен между наступлением первого и последнего события тренировочной выборки (далее, множество бинов $\{t_i\} : t_{min} \leq T \leq t_{max}$).

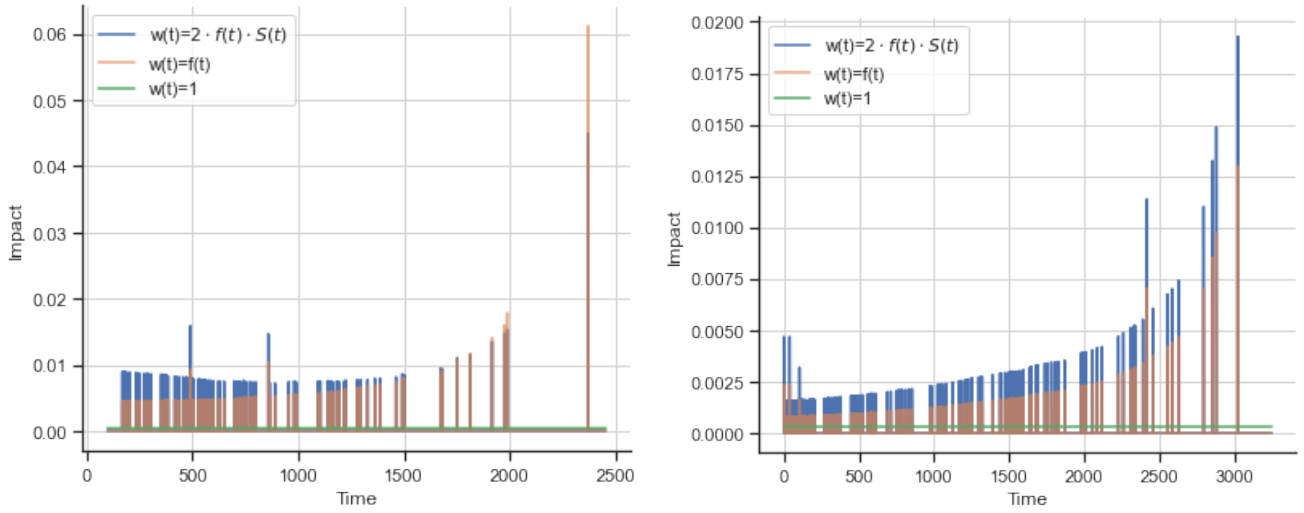
Интегрирование по определенной временной шкале приводит к равному вкладу всех моментов времени, поскольку $dt = 1$. В терминах метрик семейства IBS (IBS, IBS_{WW}, IBS_{RM}) интеграл вычисляется напрямую от времени t и имеет равные вклады. Аналогично,

при расчете интеграла по переменной φ в метрике $AUPRC$ равны все вклады моментов времени.

Для метрики $IAUC$ (раздел 1.2.2) существует несколько определений дифференциала для вычисления интеграла. Все весовые схемы [47, 103–105] обобщаются формулой $IAUC = \frac{1}{\int w(t)dt} \int AUC(t)w(t)dt$. В исследованиях [103–105] рассматривается взвешивание на основе функции плотности $w(t) = \hat{f}(t)$. Следовательно $w(t)dt = \hat{f}(t)dt = -d\hat{S}(t)$, где оценка $\hat{S}(t)$ построена на основе модели Каплана–Мейера (раздел 1.3.2).

В работе [47] рассматривается весовая схема $w(t) = 2 \cdot \hat{f}(t) \cdot \hat{S}(t)$. Следовательно $w(t)dt = 2 \cdot \hat{f}(t) \cdot \hat{S}(t)dt = -2 \cdot \hat{S}(t)d\hat{S}(t) = -d\hat{S}^2(t)$, где $\hat{S}(t)$ — оценка Каплана–Мейера.

На рисунке 24 представлен пример изменения вклада $w(t)$ относительно временной шкалы для наборов данных $GBSG$ и $SMARTO$. Зеленым цветом обозначена весовая схема с равным вкладом каждого момента времени (используется при расчете IBS и $AUPRC$). Синим цветом представлена весовая схема $w(t) = 2 \cdot \hat{f}(t) \cdot \hat{S}(t)$, оранжевым — весовая схема $w(t) = \hat{f}(t)$ (используются при расчете $IAUC$).



(a) $GBSG$ (b) $SMARTO$

Рис. 24: Пример зависимости вклада $w(t)$ от времени для весовых схем $w(t) = 1$, $w(t) = \hat{f}(t)$ и $w(t) = 2 \cdot \hat{f}(t) \cdot \hat{S}(t)$ для наборов данных $GBSG$ и $SMARTO$.

Весовые схемы $w(t) = \hat{f}(t)$ и $w(t) = 2 \cdot \hat{f}(t) \cdot \hat{S}(t)$ обладают схожим поведением и имеют два недостатка. Во-первых, не все моменты времени влияют на значение интегральной метрики. Наличие нулевого вклада обусловлено совпадением функций выживания для начального t_1 и конечного t_2 значений интервала $S(t_1) - S(t_2) = 0$. Согласно модели Каплана–Мейера, такие случаи возникают, если в интервале $[t_1, t_2]$ не наблюдалось ни одного события. В таком случае, цензурирование наблюдений в интервале $[t_1, t_2]$ влияет на значение $AUC(t)$ (так как не будут учитываться пары с цензурированным наблюдением), но не учитываются при расчете интегральной метрики $IAUC$.

Вторым недостатком является повышение значимости вклада $AUC(t)$ для поздних моментов времени. Согласно графикам плотности времени наступления события 9, наборы данных $GBSG$ и $SMARTO$ содержат большое количество ранних событий. Однако на рисунке

24 отмечается высокая важность качества ранжирования для поздних моментов времени. Для преодоления недостатков метрики $IAUC$ рекомендуется использовать единичную весовую схему $w(t) = 1$.

Таким образом, метрики IBS и $AUPRC$ основаны на агрегации $BS(t)$ и $AUPRC(\varphi)$ с равными вкладами во времени (без ложной чувствительности). В то же время, при вычислении $IAUC$ обе весовые схемы основаны на повышенном вкладе $AUC(t)$ поздних моментов времени и игнорировании значений $AUC(t)$ в моменты ненаступления терминальных событий (не учитывая факт возможной цензурируемости).

3.1.4 Влияние дисбаланса цензурирования

На практике, реальные наборы данных имеют различное соотношение цензурированных и терминальных событий. Для терминального наблюдения целевая функция выживания равна 1 до наступления события и 0 после. Для цензурированного наблюдения целевая функция выживания равна 1 до момента цензурирования и имеет неопределенное значение после.

IBS

Согласно формуле (8), для цензурированного наблюдения эталонной функцией выживания может быть любая функция, равная 1 до момента цензурирования. Следовательно, для цензурированных наблюдений наименьшее значение $IBS = 0$ будет достигаться при константном прогнозе $S(t) = 1$. Для терминальных событий рассматривается отклонение до и после момента T_i. Следовательно, для фиксированного прогноза S(t) значение $IBS_{\delta=1} > IBS_{\delta=0}$.

Для демонстрации чувствительности метрики IBS к дисбалансу цензурирования приведем альтернативный вид формулы (8). В частности, представим метрику $BS(t)$ в виде суммы отклонений для каждого типа события $BS(t) = BS_{\delta=1}(t) + BS_{\delta=0}(t)$, где $BS_{\delta=1}(t)$ — доля отклонений терминальных событий, $BS_{\delta=0}(t)$ — доля отклонений цензурированных наблюдений:

$$BS_{\delta=1}(t \mid N) = \frac{1}{N} \sum_{i:\delta_i=1} \begin{cases} \frac{(0-S(t|X_i))^2}{G(T_i)}, & \text{if } T_i \le t, \\ \frac{(1-S(t|X_i))^2}{G(t)}, & \text{if } T_i > t, \end{cases} \tag{31}$$

$$BS_{\delta=0}(t \mid N) = \frac{1}{N} \sum_{i:\delta_i=0} \begin{cases} \frac{(1-S(t|X_i))^2}{G(t)}, & \text{if } T_i > t, \\ 0, & \text{if } T_i \le t. \end{cases} \tag{32}$$

Следовательно, IBS может быть представлена в следующем виде:

$$IBS = \frac{1}{t_{max}} \int_0^{t_{max}} BS_{\delta=1}(t \mid N) + BS_{\delta=0}(t \mid N) dt. \tag{33}$$

При работе с данными с преобладанием цензурированных наблюдений $N_{\delta=1} << N_{\delta=0}$ прогноз $S(t)$ смещается к 1, обеспечивая меньшую ошибку для доминирующих цензурированных наблюдений.

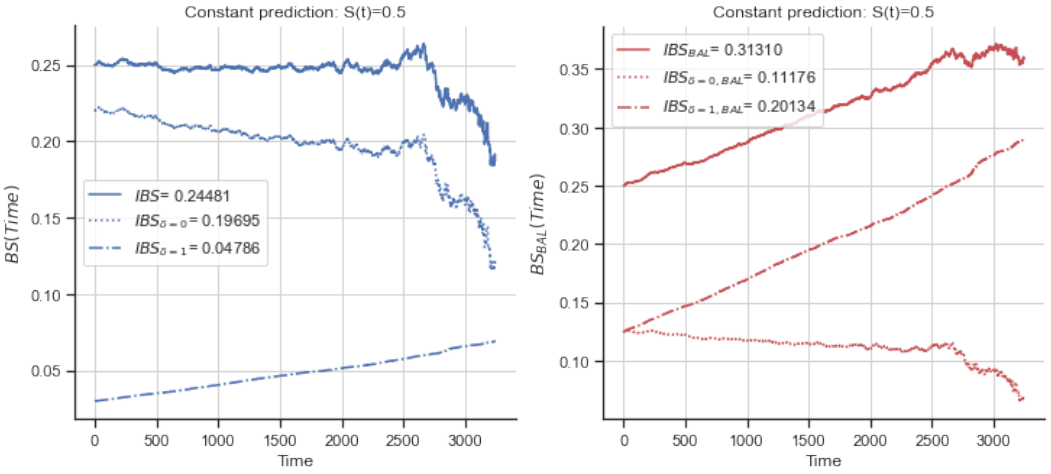


Рис. 25: Пример изменения метрик $BS(t)$ и $BS_{BAL}(t)$ во времени для набора SMARTO при прогнозе $S(t) = 0.5$. Кривая $BS(t)$ близка к $BS_{\delta=0}(t)$ из-за дисбаланса цензурирования. Синие линии отражают значения IBS (событие, цензура и итог) с исходным соотношением цензурированных и терминальных событий. Красные линии отражают значения IBS_{BAL} с равными долями типов событий.

На левом рисунке 25 представлено изменение метрики $BS(t)$ для набора данных $SMARTO$ с дисбалансом цензурирования (12% терминальных событий, 88% цензурированных наблюдений). Точечная и пунктирная линии определяют значения метрики $BS(t)$ для цензурированных $BS_{\delta=0}(t \mid N)$ и терминальных событий $BS_{\delta=1}(t \mid N)$ соответственно. Сплошная линия определяет сумму значений $BS(t)$. Исходя из графика, возникает занижение значимости ошибки терминальных событий и кривая $BS(t)$ близка к $BS_{\delta=0}(t)$. Следовательно, отклонения цензурированных наблюдений вносят больший вклад в метрику IBS.

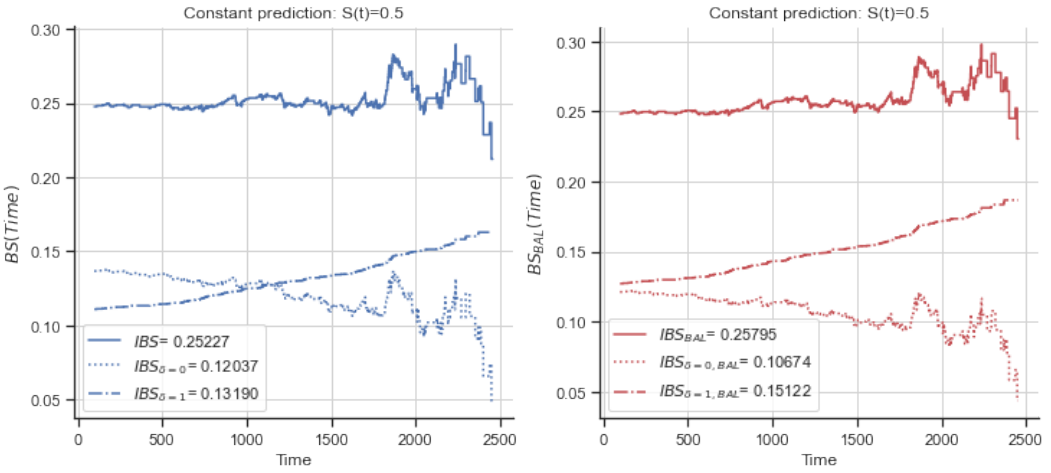


Рис. 26: Пример изменения метрик $BS(t)$ и $BS_{BAL}(t)$ во времени для набора GBSG при прогнозе $S(t) = 0.5$. Значения метрик близки, так как набор GBSG сбалансирован относительно типов событий.

На левом рисунке 26 представлен график изменения метрики $BS(t)$ для набора данных $GBSG$ с малым дисбалансом цензурирования (44% терминальных событий, 56% цензурированных наблюдений). Обозначения эквивалентны графику 25. Исходя из графика, нельзя

однозначно утверждать о доминировании вклада определенного типа события. В таком случае, отклонения для обоих типов событий оказывают значимое влияние на значение метрики IBS.

Для преодоления недостатка различия вкладов событий, предлагается определить равный вклад цензурированных и терминальных событий в значении $BS(t)$. Для этого, заменим в формуле (31) общее число событий N на число терминальных событий $N_{\delta=1}$. Аналогично, в формуле (32) заменим N на число цензурированных событий $N_{\delta=0}$. Определим метрику $BS_{BAL}(t)$ как сбалансированное среднее значение двух типов событий:

$$BS_{BAL}(t) = \frac{1}{2}\left(BS_{\delta=1}(t \mid N_{\delta=1}) + BS_{\delta=0}(t \mid N_{\delta=0})\right),$$

$$IBS_{BAL} = \frac{1}{t_{max}} \int_0^{t_{max}} BS_{BAL}(t)dt. \tag{34}$$

Отметим, что метрики $BS_{\delta=1}(t \mid N_{\delta=1}), BS_{\delta=0}(t \mid N_{\delta=0}), BS_{BAL}(t)$ могут быть определены для ранее описанных модификаций: $BS_{WW}(t)$, $BS_{RM}(t)$ аналогичным образом.

На правом рисунке 25 представлен график изменения метрики $BS_{BAL}(t)$ для набора данных $SMARTO$. По сравнению с метрикой $BS(t)$ на левом рисунке, вклад отклонений терминальных и цензурированных событий равнозначно влияет на значения $BS_{BAL}(t)$. На правом рисунке 26 представлен график изменения метрики $BS_{BAL}(t)$ для набора данных $GBSG$. Значения метрик близки и вклад отклонений терминальных и цензурированных событий значимо влияет на значения $BS(t)$ и $BS_{BAL}(t)$.

Таким образом, при дисбалансе определенного типа наблюдений среднее отклонение смещается в сторону преобладающего класса. При использовании сбалансированной модификации возможно преодолеть избыточную чувствительность к преобладающему классу.

AUPRC

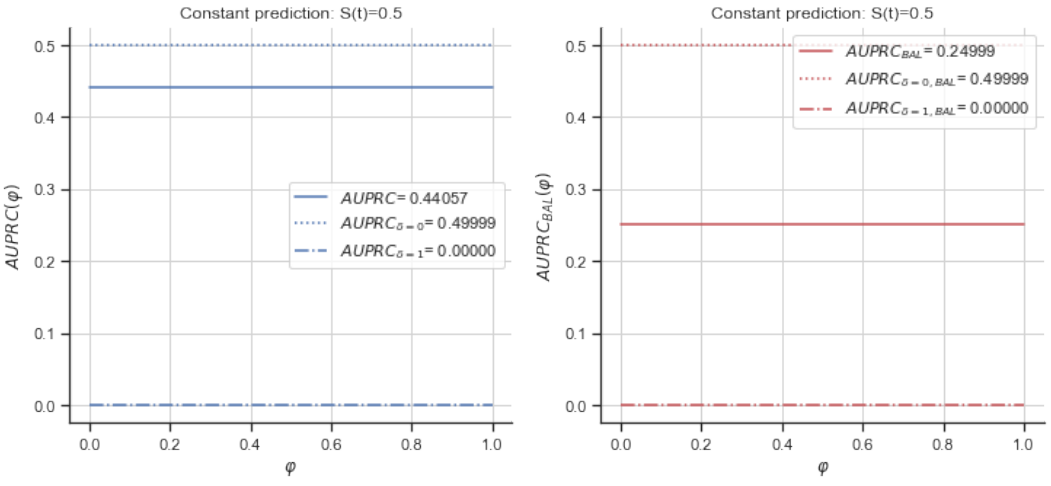


Рис. 27: Пример изменения метрики $AUPRC(\varphi)$ от φ для набора SMARTO при прогнозе $S(t) = 0.5$. На левом графике значение $AUPRC$ смещено в сторону доминирующего класса цензурированных событий. $AUPRC_{BAL}$ определяет равный вклад классов.

Аналогичным образом можно провести рассуждения для метрики $AUPRC$. На левом рисунке 27 представлен график изменения метрики $AUPRC(\varphi)$ для набора данных $SMARTO$. Обозначения повторяют рисунок 25. Исходя из графика, возникает занижение значимости ошибки терминальных событий и кривая $AUPRC(\varphi)$ близка к $AUPRC_{\delta=0}(t)$. Следовательно, отклонения цензурированных наблюдений вносят больший вклад в метрику $AUPRC$.

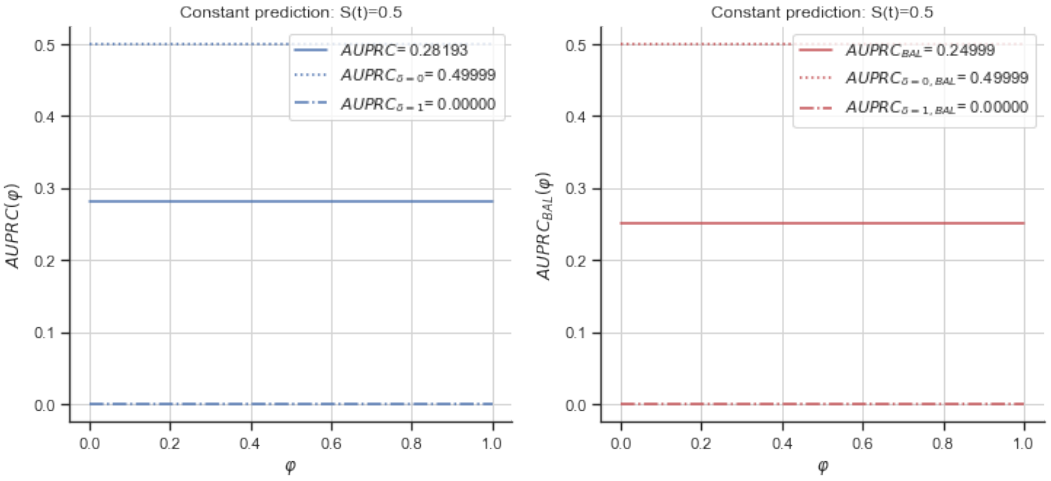


Рис. 28: Пример изменения метрики $AUPRC(\varphi)$ во времени для набора GBSG при прогнозе $S(t) = 0.5$. $AUPRC_{\delta=0}$ и $AUPRC_{\delta=1}$ близки, поскольку GBSG сбалансирован относительно типов событий. $AUPRC_{BAL}$ определяет равное соотношение вкладов.

На левом рисунке 28 представлен график изменения метрики $AUPRC(\varphi)$ для набора данных $GBSG$. В отличии от метрики IBS, итоговой метрикой является среднее двух значений. Остальные обозначения повторяют рисунок 26. Исходя из графика, нельзя точно сказать о доминировании определенного класса событий.

Следовательно, для повышения устойчивости метрики $AUPRC$ предлагается следующая сбалансированная модификация (аналогично метрике (34)):

$$AUPRC_{BAL}(\varphi) = \frac{1}{2}\left(AUPRC_{\delta=1}(\varphi \mid N_{\delta=1}) + AUPRC_{\delta=0}(\varphi \mid N_{\delta=0})\right),$$

$$AUPRC_{BAL} = \int_0^1 AUPRC_{BAL}(\varphi)d\varphi. \tag{35}$$

Важно отметить, что использование сбалансированных метрик приводит к нарушению требований проверки первого случая избыточной чувствительности 3.1.1, поскольку индивидуальный вклад наблюдений заменяется на усредненный вклад во времени.

IAUC

Среди рассмотренных ранее метрик (таблица 1), $IAUC$ можно считать устойчивой к дисбалансу цензурирования. В частности, в работах [48, 106] авторы советуют использовать метрику AUC в случае несбалансированных данных. Также, в работах [107, 108] отмечается, что метрика AUC инвариантна относительно априорных вероятностей классов. В терминах

анализа выживаемости, аналогичными свойствами обладает метрика $AUC(t)$ и интегральное значение относительно времени $IAUC$.

CI

В исследованиях [44, 45] отмечаются случаи завышения и занижения метрики при дисбалансе между цензурированными и терминальными наблюдениями. Проблемы проявляются при доминировании цензурированных наблюдений, поскольку количество упорядоченных пар изменяется после выхода наблюдений из исследования.

3.1.5 Сравнение чувствительности метрик

На основе проведенных исследований (разделы 3.1.1 — 3.1.4) построена сводная таблица 5. В таблице представлены все рассмотренные интегральные метрики и их модификации. Также, сводная таблица содержит информацию об использовании генеральной оценки функции выживания цензурированных наблюдений $G(t)$. На практике, расчет и использование $G(t)$ требует дополнительных данных и вычислительных ресурсов. Кроме того, оценка функции выживания изменяется с увеличением объема данных, приводя к искажению предыдущих значений метрики.

Таблица 5: Наличие избыточной чувствительности метрик качества. Названиями столбцов служат номера разделов, в которых исследовалась особенность метрик: 3.1.1 — зависимость вклада от наблюдения, 3.1.2 — зависимость вклада от времени, 3.1.3 — влияние бинов на интеграл, 3.1.4 — влияние дисбаланса. Столбец «$G(T)$» отражает зависимость метрики от генеральной функции выживания. Значения «?» указывают на отсутствие сведений. Лучшие метрики отмечены серым цветом.

Название	3.1.1	3.1.2	3.1.3	3.1.4	**G(T)**
IBS	Да	Да	Нет	Да	Да
IBS_{WW}	Нет	Да	Нет	Да	Нет
IBS_{RM}	Нет	Нет	Нет	Да	Нет
IBS_{BAL}	?	Да	Нет	Нет	Да
$IBS_{WW,BAL}$	?	Да	Нет	Нет	Нет
$IBS_{RM,BAL}$	?	Нет	Нет	Нет	Нет
$IAUC$ $w(t) = 2 \cdot f(t) \cdot S(t)$	?	Нет	Да	Нет?	Да
$IAUC$ $w(t) = f(t)$	?	Нет	Да	Нет?	Да
$IAUC$ $w(t) = 1$	?	Нет	Нет	Нет?	Нет
$AUPRC$	Нет	Нет	Нет	Да	Нет
$AUPRC_{BAL}$	?	Нет	Нет	Нет	Нет
CI	?	Нет	Нет	Да?	Нет

Лучшие метрики обладают наибольшей устойчивостью к случаям избыточной чувствительности. Невозможность проверки или отсутствие сведений помечены как «?». Например, для семейства $IAUC$ невозможно провести проверку вклада отдельных наблюдений, поскольку метрика не представима в виде суммы отдельных оценок. Также, устойчивость семейства

$IAUC$ к дисбалансу цензурирования определена из исследований [107, 108], однако не было проверена на практике. Лучшая метрика из каждого семейства помечена серым цветом.

Таким образом, наиболее устойчивыми метриками являются $IBS_{RM}, IBS_{RM,BAL}$, $IAUC(w(t) = 1)$ и $AUPRC, AUPRC_{BAL}$. Как было сказано ранее, сбалансированные метрики не позволяют вычислить качество отдельных событий и не применимы для анализа случая 3.1.1. Неизвестное поведение $IAUC(w(t) = 1)$ в отношении дисбаланса и отдельных событий приводит к меньшей наглядности и достоверности метрики по сравнению с остальными семействами. При сравнении метрик IBS_{RM} и $AUPRC$ отметим важное свойство последней. При оценке вероятности $P(T_i/\varphi > T > T_i \cdot \varphi)$ попадания события в интервал $[T_i \cdot \varphi, T_i/\varphi]$ определяется качество функции выживания до и после момента наступления события. Следовательно, ранние (до события) и поздние (после события) интервалы вносят равный вклад в метрику $AUPRC$. В случае метрики IBS вклад ранних и поздних интервалов пропорционален длине временного интервала.

Также, метрика $AUPRC$ позволяет избежать ложной чувствительности при расчете интеграла по времени. В частности, расстояние между бинами не влияет на интеграл и позволяет избежать повышения вклада редких поздних событий. В случае метрики IBS наличие выбросов приводит к появлению длинных интервалов, повышающих вклад поздних событий в интегральную метрику. Таким образом, по результатам проведенных исследований, наиболее устойчивой к случаям избыточной чувствительности метрикой качества является $AUPRC$.

3.2 Экспериментальное исследование

По результатам аналитического исследования проблемы мультимодального распределения вероятностей времени событий, мы предложили несколько модификаций существующей модели 1.4 дерева выживаемости. Первая модификация основана на регуляризации дерева на этапе выбора лучшего разбиения выборки.

Вторая модификация основана на построении непараметрической модели $KMWV$ в листах дерева выживаемости. Для преодоления мультимодальности распределения времени, мы генерируем виртуальные события по нормальному распределению с вероятностью цензурирования в листовой выборке.

В данном разделе мы провдем экспериментальное исследование качества моделей анализа выживаемости на 6 медицинских наборах данных (раздел 2.1). Первой целью экспериментального исследования является оценка влияения коэффициента регуляризации на качество предложенных моделей. Также, мы оценим влияние листовой оценки KM и $KMWV$ на качество моделей. Наконец, мы сравним предложенный алгоритм построения деревьев выживания с существующей реализацией ST и методом пропорциональных рисков Кокса $CoxPH$.

3.2.1 Постановка эксперимента

Первоначально проводится предобработка набора данных, формирование признакового пространства и целевых переменных (время до наступления события, индикатор цензурирования).

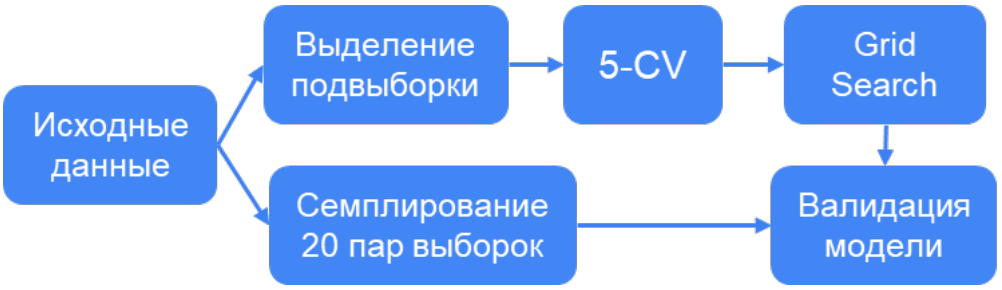


Рис. 29: Схема экспериментального исследования. Основные этапы: выделение обучающей выборки, поиск лучших гиперпараметров по сетке (кросс-валидация) и валидация на 20 независимых подмножествах.

Алгоритм эксперимента разделяется на 3 этапа (Рисунок 29). На первом этапе, исходные данные разделяются на тренировочную и тестовую выборку (66% и 33% соответственно) со стратификацией по индикатору цензурирования и времени.

На втором этапе, проводится пятикратная кросс-валидация [109] по сетке гиперпараметров на тренировочной выборке. При кросс-валидации исходная выборка разделяется на 5 непересекающихся частей, 4 из которых используются для обучения модели, а 1 часть используется для тестирования модели и вычислении метрик качества. Всего проводится 5 итераций обучения/тестирования модели, причем каждая часть единожды используется для оценки качества на тестовой выборки. Качество модели по кросс-валидации оценивается как среднее значение метрики по всем итерациям. Данный этап включает выбор лучших гиперпараметров для каждой модели.

На третьем этапе, исходные данные 20-кратно разбиваются на тренировочные и тестовые данные (66% и 33% соответственно). Лучшие модели (выбранные в ходе кросс-валидации) обучаются на тренировочных данных и применяются к тестовым. Итоговое качество модели вычисляется как среднее качество для 20 тестовых выборок.

Таблица 6: Гиперпараметры моделей прогнозирования

Название модели	**Гиперпараметры**	**Сетка значений**
Survival Tree (ST)	split strategy	best, random
	max depth	from 10 to 30 step 5
	min sample leaf	from 1 to 20 step 1
	max features	sqrt, log2, None
TREE	max depth	from 10 to 30 step 5
	min sample leaf	0.05, 0.01, 0.001
	signif	0.05, 0.1, 1.0
	lambda	0.0, 0.01, 0.1, 0.5, 0.9
	leaf model	KM, KMWV
	criterion	peto, tarone-ware, wilcoxon, logrank

Реализация модели Survival Tree взята из открытой библиотеки *Scikit-survival*. Реализация модели $TREE$ была представлена в разделе 2.3.1 данной работы. Модель $TREE_{KMWV}$ является модификацией модели $TREE$ с заменой непараметрических листовых моделей

Каплана–Мейера на предложенную модель $KMWV$. Обе модели поддерживают аппарат регуляризации, описанный в разделе 2.4.3 (коэффициент регуляризации λ используется в качестве гиперпараметра). Полный список гиперпараметров используемых моделей представлен в таблице 6.

3.2.2 Оценка качества непараметрических моделей

На рисунке 30 представлено сравнение непараметрических оценок на наборе GBSG. На правом графике представлено распределение вероятностей времени исходных данных (мультимодально) и виртуальных событий (унимодально). На левом графике изображены прогнозы функций выживания по классическом модели KM и предложенной модели $KMWV$. Качество KM достигло $IBS_{RM} = 0.19145, AUPRC = 0.58663$, а $KMWV$ достигло $IBS_{RM} = 0.17374, AUPRC = 0.63810$. Предложенная непараметрическая оценка с виртуальными событиями позволила построить более точную функцию выживания на полных данных набора GBSG.

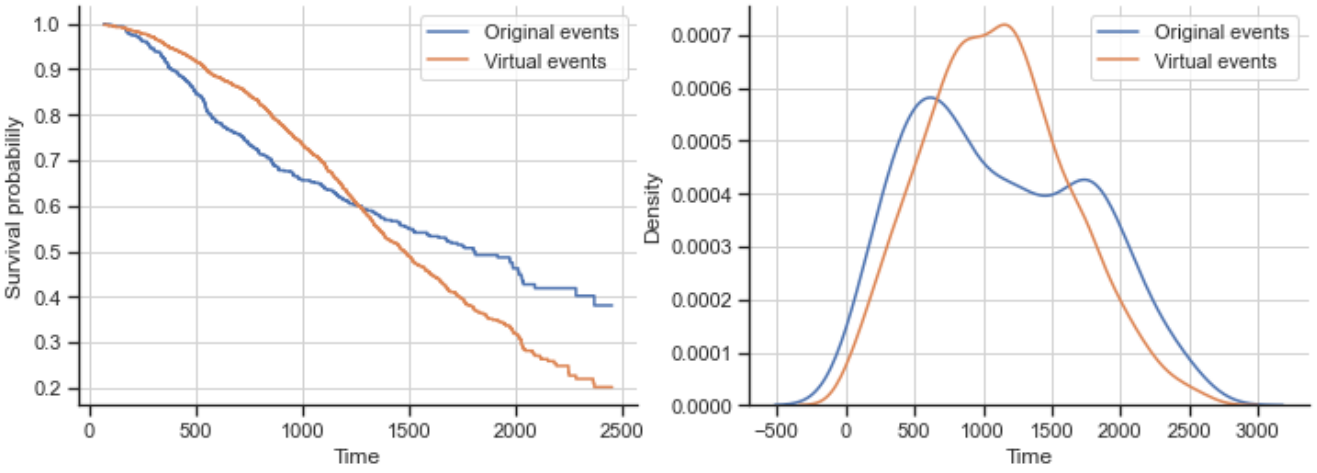


Рис. 30: Сравнение оценки функций выживания (набор GBSG) по методу Каплана–Мейера для исходных событий («Original events») и виртуальных событий («Virtual events») на наборе GBSG. На правом графике представлено распределение времени событий (виртуальные события распределены унимодально).

В древовидных моделях анализа выживаемости, непараметрические модели используются для построения оценки на подмножестве данных. На рисунках 31 и 32 представлено сравнение качества прогнозов моделей KM и $KMWV$ на 100 случайных подмножествах набора GBSG размером 5% и 30% от общего объема. На рисунке 31 размер подмножества составляет 30 наблюдений (менее 5% от общего объема), а на рисунке 32 размер равен 200 (30% от общего объема). На левых графиках представлено качество описания подмножества по метрике IBS_{RM}, на правых графиках — по метрике $AUPRC$. Горизонтальные линии определяют среднее значение метрик для каждой модели. Для оценки различия между прогнозами, определено p-value равенства средних по статистическому тесту Уэлча.

В обоих случаях качество модели $KMWV$ значимо превосходит KM. Для меньших подмножеств разница в качестве менее значима. Также, по метрике $AUPRC$ качество модели $KMWV$ имеет больший прирост, чем по метрике IBS_{RM}.

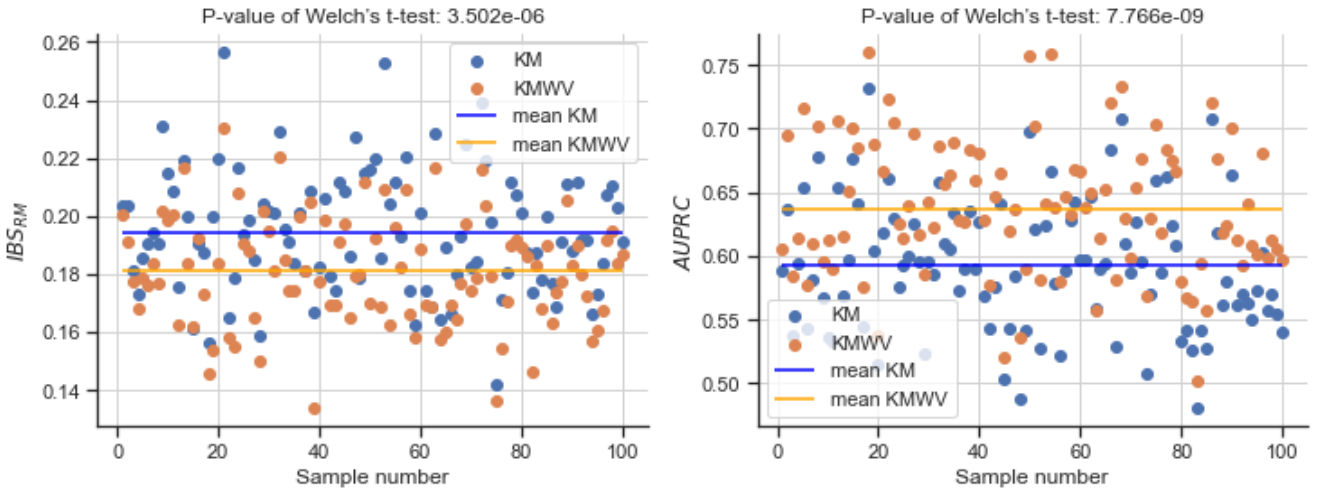


Рис. 31: Сравнение качества прогнозов моделей KM и $KMWV$ на подмножествах набора данных GBSG размера 30. Прогнозы модели $KMWV$ имеют лучшее качество по метрикам $AUPRC$ и IBS_{RM}, причем различие по качеству значимо по тесту Уэлча (для уровня значимости 0.01).

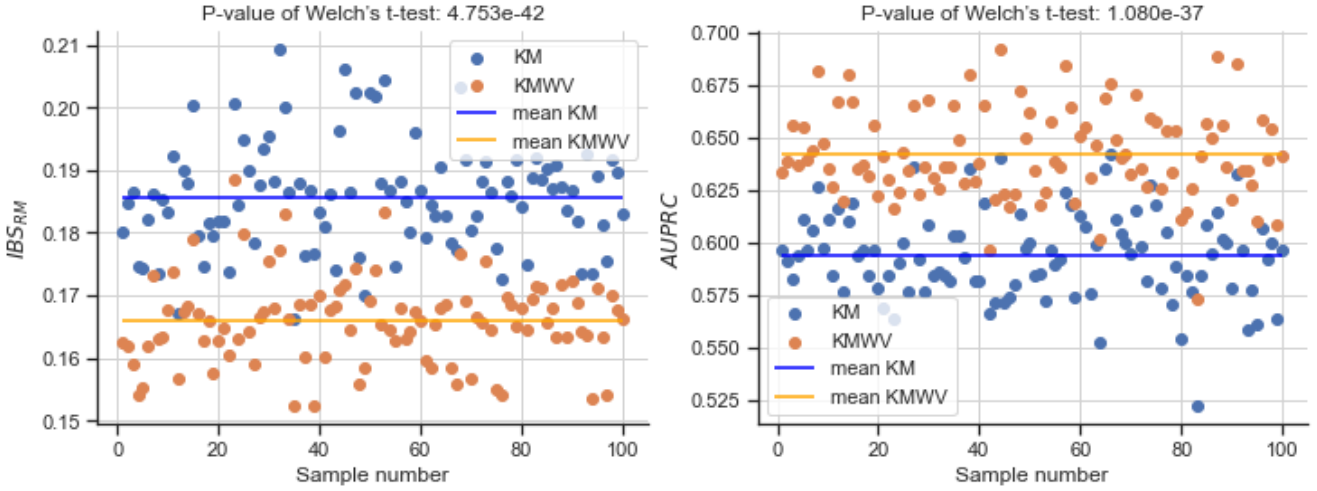


Рис. 32: Сравнение качества прогнозов моделей KM и $KMWV$ на подмножествах набора данных GBSG размера 200. Прогнозы модели $KMWV$ имеют лучшее качество по метрикам $AUPRC$ и IBS_{RM}, причем различие по качеству значимо по тесту Уэлча (для уровня значимости 0.01).

Наконец, проведем аналогичные эксперименты для остальных наборов данных. В таблице 7 приведено сравнение качества моделей KM и $KMWV$ на 100 случайных подмножествах каждого набора данных для подвыборок размером 30% и 5% от общего числа. Для каждого размера и модели определялось качество прогноза по метрикам $AUPRC$ и IBS_{RM}, а в ячейках таблицы указано среднее значение качества. Для оценки разброса между значениями метрик, таблица содержит p-value статистического теста Уэлча на равенство средних.

Таблица 7: Сравнение модели Каплана–Мейера и предложенной модификации KMWV на 100 подвыборках медицинских наборов данных размером 5% и 30%.

Model	**KM**	**KMWV**	**Welch p-value**	**KM**	**KMWV**	**Welch p-value**
	GBSG			PBC		
$IBS(30\%)$	0.1856	**0.1659**	4.753e-42	0.1961	**0.1732**	7.081e-32
$IBS(5\%)$	0.1944	**0.1809**	3.502e-06	0.2095	**0.1923**	3.123e-06
$AUPRC(30\%)$	0.5937	**0.6416**	1.080e-37	0.6056	**0.6512**	1.723e-15
$AUPRC(5\%)$	0.5926	**0.6374**	7.766e-09	0.6301	**0.6685**	1.741e-03
	WUHAN			ROTT2		
$IBS(30\%)$	0.1931	**0.1569**	1.282e-42	0.1708	**0.1391**	8.159e-68
$IBS(5\%)$	0.2116	**0.1890**	1.084e-08	0.1865	**0.1582**	2.252e-28
$AUPRC(30\%)$	0.5694	**0.6074**	1.011e-16	0.5975	**0.6592**	7.023e-75
$AUPRC(5\%)$	0.5651	**0.6064**	5.797e-05	0.5896	**0.6537**	5.490e-27
	SMARTO			SUPPORT2		
$IBS(30\%)$	0.2053	**0.1835**	7.973e-51	0.1723	**0.1643**	1.454e-42
$IBS(5\%)$	0.1996	**0.1791**	3.393e-10	0.1726	**0.1649**	2.799e-12
$AUPRC(30\%)$	0.8431	**0.8546**	4.652e-11	0.2819	**0.3137**	5.191e-107
$AUPRC(5\%)$	0.8476	**0.8579**	1.543e-02	0.2831	**0.3141**	6.831e-50

Во всех случаях модель $KMWV$ значительно превосходит модель KM. Для меньших подмножеств разница в качестве менее значительна, чем для больших подмножеств. Таким образом, предложенная непараметрическая оценка показывает выигрыш в качестве по метрикам IBS_{RM}, $AUPRC$. Подход не требует предположения о неинформативности индикатора наступления события и применим к данным с мультимодальным распределением времени.

3.2.3 Влияние весовых схем log-rank

Таблицы 8-13 содержат результаты экспериментального исследования влияния весовой схемы критерия разбиения на качество дерева выживаемости.

Таблица 8: Сравнение весовых схем критерия на наборе PBC

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC}$	$\boldsymbol{AUPRC}$
$TREE(logrank)$	**0.6602**	0.1330	0.8366	0.7615
$TREE(peto)$	0.6586	0.1282	0.8399	0.7608
$TREE(tarone\text{-}ware)$	0.6559	0.1281	0.8385	0.7605
$TREE(wilcoxon)$	0.6588	**0.1245**	**0.8473**	**0.7634**

По совокупности метрик, для набора PBC лучшее качество показала весовая схема *wilcoxon*, для набора $GBSG$ — схема *tarone-ware*, для наборов $WUHAN, SUPPORT2$ — схема *peto*. На наборе $ROTT2$ лучшие результаты показали схемы *logrank* и *tarone-ware*, а на наборе $SMARTO$ — схемы *tarone-ware* и *wilcoxon*. Таким образом, использование весовых схем вместо классического *logrank* критерия приводит к повышению качества на 5

Таблица 9: Сравнение весовых схем критерия на наборе GBSG

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC}$	$\boldsymbol{AUPRC}$
$TREE(logrank)$	0.5944	0.1678	0.7035	0.7092
$TREE(peto)$	0.5935	0.1690	0.7019	**0.7095**
$TREE(tarone\text{-}ware)$	**0.6015**	**0.1646**	**0.7117**	**0.7095**
$TREE(wilcoxon)$	0.5992	0.1655	0.7083	0.7085

Таблица 10: Сравнение весовых схем критерия на наборе WUHAN

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC}$	$\boldsymbol{AUPRC}$
$TREE(logrank)$	0.6926	0.1010	0.8217	0.757
$TREE(peto)$	0.7073	**0.0954**	**0.8400**	**0.7601**
$TREE(tarone\text{-}ware)$	0.7078	0.0996	0.8349	**0.7601**
$TREE(wilcoxon)$	**0.7130**	0.097	0.8328	0.7577

Таблица 11: Сравнение весовых схем критерия на наборе ROTT2

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC}$	$\boldsymbol{AUPRC}$
$TREE(logrank)$	**0.6297**	0.1282	**0.7359**	0.7226
$TREE(peto)$	0.6203	0.1281	0.7356	**0.7229**
$TREE(tarone\text{-}ware)$	0.6200	**0.1280**	0.7330	**0.7229**
$TREE(wilcoxon)$	0.6188	0.1282	0.7354	0.7226

Таблица 12: Сравнение весовых схем критерия на наборе SUPPORT2

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC}$	$\boldsymbol{AUPRC}$
$TREE(logrank)$	0.7995	0.1034	0.8888	0.5457
$TREE(peto)$	**0.8029**	**0.1030**	**0.8903**	**0.5570**
$TREE(tarone\text{-}ware)$	0.8021	0.1032	**0.8903**	0.5562
$TREE(wilcoxon)$	0.8027	0.1032	0.8897	**0.5570**

Таблица 13: Сравнение весовых схем критерия на наборе SMARTO

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC}$	$\boldsymbol{AUPRC}$
$TREE(logrank)$	0.5272	0.1712	0.61948	0.8719
$TREE(peto)$	0.5267	0.1710	0.61734	0.8719
$TREE(tarone\text{-}ware)$	0.5394	0.1698	**0.62255**	**0.8720**
$TREE(wilcoxon)$	**0.5435**	**0.1695**	0.6212	0.8718

из 6 наборов. Стоит отметить, что выбор подходящей весовой схемы должен производиться экспериментальным путем на основе характеристик исходных данных.

3.2.4 Сравнение методов построения деревьев выживаемости

На рисунках 33-36 представлены результаты экспериментального сравнения классической модели ST и предложенных $TREE$, $TREE_{KMWV}$ деревьев выживаемости. Для описания листовой выборки, $TREE$ использует оценку Каплана–Мейера, а $TREE_{KMWV}$ — пред-

ложенную непараметрическую оценку $KMWV$. Также, обе модели поддерживают предложенный подход регуляризации критерия при поиске лучшего разбиения (величина коэффициента λ задана в названии модели). Проверка качества проводилась по метрикам: $CI \uparrow$, $IBS_{RM} \downarrow$, $IAUC_{WW,TI} \uparrow$, $AUPRC \uparrow$. Для каждого метода мы составляем *boxplot*-график по показателям метрики на 20 тестовых выборках. Кроме того, медиана значений отмечена серой линией.

Для набора данных GBSG 33 при повышении коэффициента регуляризации λ наблюдается улучшение качества по метрикам $CI, IBS_{RM}, IAUC_{WW,TI}$. По метрике $AUPRC$ качество модели $TREE$ падает с ростом λ, качество модели $TREE_{KMWV}$ повышается для $\lambda = 0.1$. При попарном сравнении качества моделей, модель $TREE_{KMWV}$ показывает улучшение качества по метрикам $IBS_{RM}, IAUC_{WW,TI}, AUPRC$.

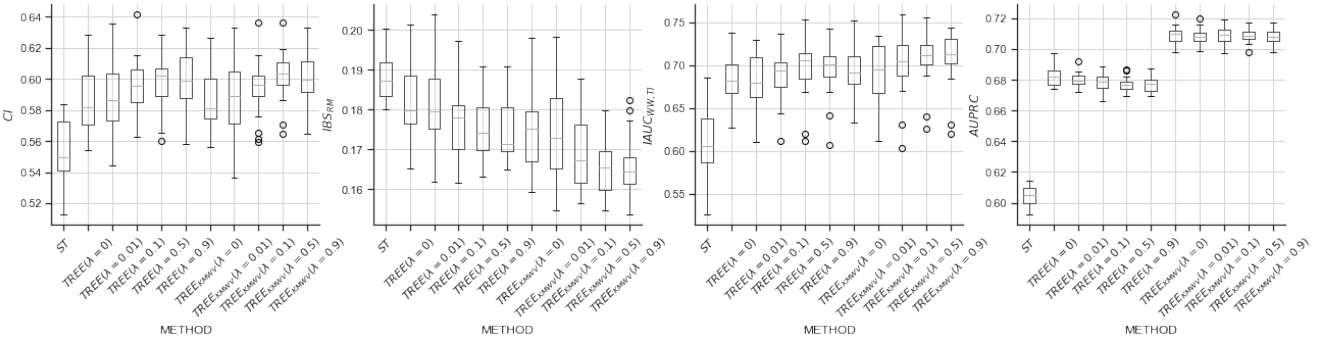


Рис. 33: Сравнение качества древовидных методов на наборе GBSG

Для набора данных WUHAN 34 при повышении коэффициента регуляризации λ наблюдается улучшение качества по метрикам $CI, AUPRC$. При попарном сравнении качества моделей, модель $TREE_{KMWV}$ показывает улучшение качества по метрикам $IBS_{RM}, AUPRC$.

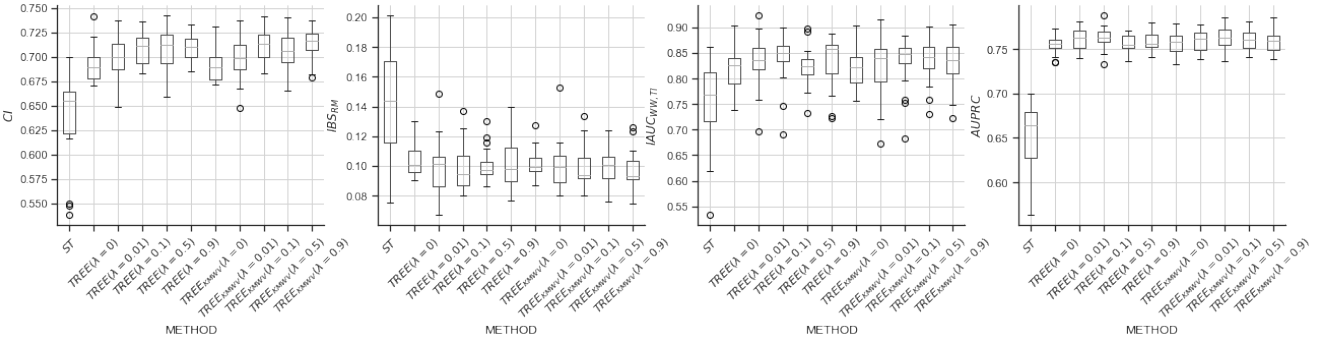


Рис. 34: Сравнение качества древовидных методов на наборе WUHAN

Для набора данных PBC 35 при повышении коэффициента регуляризации λ наблюдается улучшение качества по всем метрикам. При попарном сравнении качества моделей, модель $TREE_{KMWV}$ показывает улучшение качества по всем метрикам.

Для набора данных ROTT2 36 при повышении коэффициента регуляризации λ наблюдается улучшение качества по метрикам CI, IBS_{RM}. При попарном сравнении качества моделей, модель $TREE_{KMWV}$ показывает улучшение качества по всем метрикам.

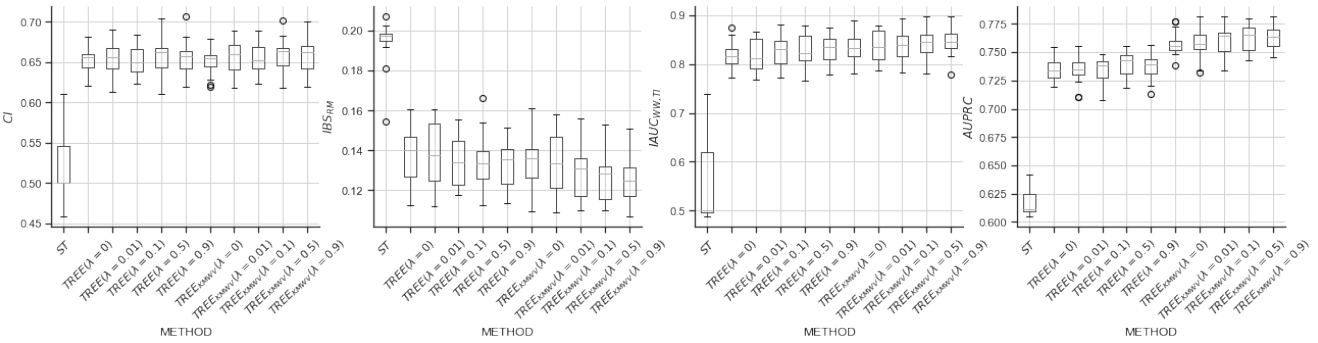


Рис. 35: Сравнение качества древовидных методов на наборе PBC

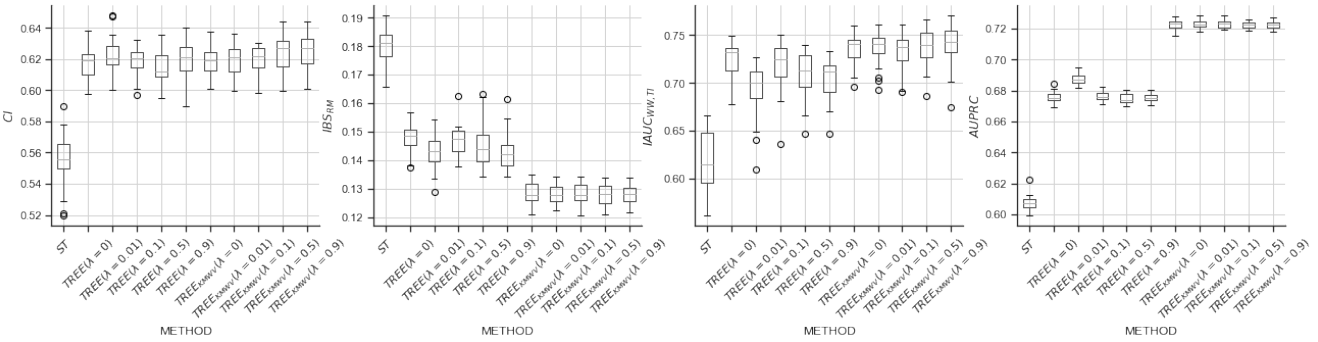


Рис. 36: Сравнение качества древовидных методов на наборе ROTT2

Для набора данных SUPPORT2 37 при повышении коэффициента регуляризации λ наблюдается улучшение качества по метрикам $CI, IAUC_{WW,TI}, AUPRC$. При попарном сравнении качества моделей, модель $TREE_{KMWV}$ показывает улучшение качества по метрикам $CI, IAUC_{WW,TI}, AUPRC$.

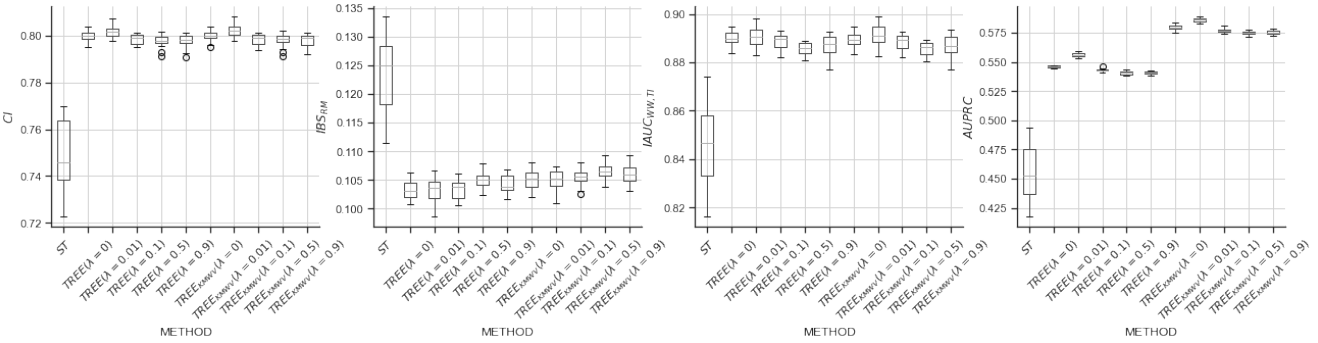


Рис. 37: Сравнение качества древовидных методов на наборе SUPPORT2

Для набора данных SMARTO 38 при повышении коэффициента регуляризации λ наблюдается улучшение качества по всем метрикам. При попарном сравнении качества моделей, модель $TREE_{KMWV}$ показывает улучшение качества по метрикам $CI, IBS_{RM}, AUPRC$.

На основе результатов можно сделать следующие выводы. Внедрение регуляризации в этап поиска лучших разбений позволил повысить гибкость моделей. Для всех наборов наблюдается рост качества при использовании регуляризации. При сравнении классической и предложенной непараметрической оценки, модель $TREE_{KMWV}$ достигает лучшего качества

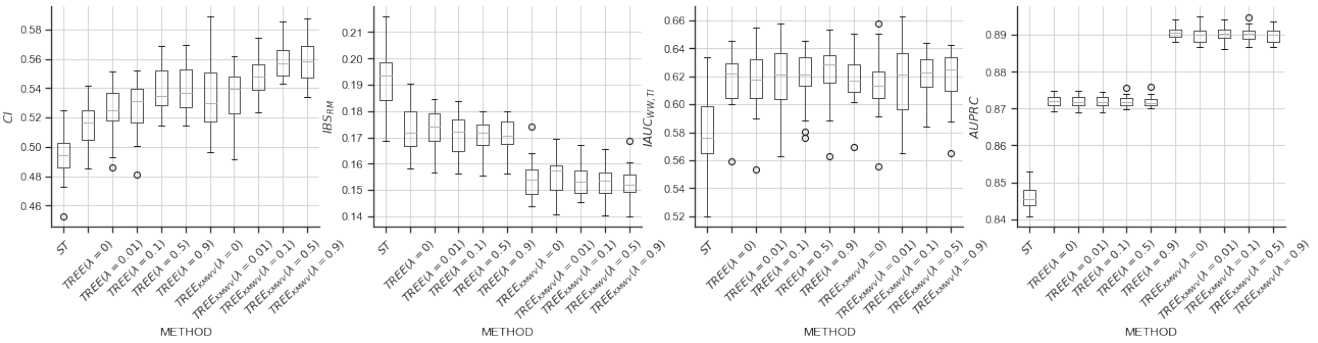


Рис. 38: Сравнение качества древовидных методов на наборе SMARTO

по сравнению с $TREE$. Наконец, оба предложенных подхода $TREE, TREE_{KMWV}$ превзошли существующий метод построения деревьев выживания ST на всех наборах данных.

3.3 Выводы

В данной главе проводились исследование избыточной чувствительности метрик анализа выживаемости и разработка модификаций метрик. По итогам проведенного исследования были достигнуты следующие результаты:

- Определены 4 случая избыточной чувствительности метрик анализа выживаемости к вкладу (1) отдельных событий, (2) временных компонент, (3) временных интервалов, (4) дисбалансу цензурирования. Исследование чувствительности должно проводится при константной функции выживания $S(t) = 0.5$, поскольку такой прогноз не дает преимуществ для определенных индивидуальных событий, интервалов временой шкалы или типов события.
- Разработаны модификации для преодоления избыточной чувствительности метрик к выявленным случаям. Для обеспечения равного вклада отдельных событий и временных компонент предлагается определить равные веса наблюдений и использовать контролируемое усреднение вкладов во времени. Для учета всех временных интервалов при интегрировании метрики рекомендуется проводить интегрирование напрямую по времени, без использования весовых схем $w(t) = f(t), w(t) = 2 \cdot f(t) \cdot S(t)$. Для определения равного вклада цензурированных и терминальных событий предложена модификация сбалансированного усреднения значений во времени.
- Проведено экспериментальное исследование качества предложенного метода построения дерева выживания на модифицированных метриках. Взвешенные схемы критерия разбиения привели к улучшению качества на 5 наборах данных. Предложенный подход регуляризации (на этапе поиска разбиения и построения непараметрической модели) позволил улучшить качество моделей и превзойти существующий метод построения деревьев выживания на всех наборах данных.

4 АНСАМБЛИ ДЕРЕВЬЕВ ВЫЖИВАЕМОСТИ

При работе (при подготовке) над данным разделом диссертации использованы следующие публикации автора, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования:

- *Vasilev I., Petrovskiy M., Mashechkin I. Adaptive Sampling for Weighted Log-Rank Survival Trees Boosting //International Conference on Pattern Recognition Applications and Methods. – Cham : Springer International Publishing, 2023. – С. 98-115.*

В предыдущих главах данной диссертации был предложен подход построения деревьев выживаемости (раздел 2.3). Использование взвешенных критериев и подхода регуляризации дерева позволило достичь высокого качества прогнозирования функций выживания и риска, а также превзойти качество существующего подхода построения деревьев выживаемости.

В данной главе представляется развитие предложенного метода для построения ансамблей деревьев выживаемости. Основная задача базовой модели — точно описать соответствующую обучающую подвыборку. Агрегация прогнозов нескольких базовых моделей позволяет улучшить качество прогнозирования и предотвратить переобучение. В отличие от деревьев выживания, ансамблевые модели не имеют интерпретации зависимостей и нацелены на построение точных прогнозов, хотя и позволяют выявить наиболее важные признаки.

В данной главе рассматривается два метода ансамблирования. Бутстреп метод (раздел 4.1) основан на построении множества независимых деревьев выживания. По сравнению с существующим методом Random Survival Forest (раздел 1.6.1), предлагаемый метод оснащен функцией выбора размера ансамбля, а также имеет высокое качество описания данных с помощью базовых моделей.

Существующие бустинг ансамбли анализа выживаемости имеют несколько значительных недостатков. Модели дискретного времени не позволяют прогнозировать непрерывные функции и имеют высокую вычислительную сложность при масштабировании. Точечные модели ограничены прогнозированием вероятности и времени наступления события. Наконец, модели, основанные на расширении статистических подходов, наследуют строгие предположения и недостатки подходов.

Для преодоления недостатков существующих моделей предлагается исследовать и разработать методы построения адаптивных бустингов анализа выживаемости. Адаптивные бустинги основаны на идее итеративного построения базовых моделей, обученных на подмножествах наблюдений с высокой ошибкой прогнозирования ансамбля. Ранее, адаптивные модели применялись только для задач классификации и регресии.

Задачей данного раздела является исследование и разработка методов построения ансамблей деревьев выживаемости, экспериментальное исследование качества существующих и предложенных моделей анализа выживаемости.

4.1 Бутстреп ансамбль независимых деревьев выживаемости

Бутстреп ансамблирование (bootstrap), также известное как беггинг (bagging), основано на обучении каждой модели в ансамбле на случайном подмножестве данных. Из исходной выборки данных генерируется множество подвыборок с заменой (бутстрэп), для каждой подвыборки строится отдельная модель. Предсказание ансамбля производится путем голо-

сования или усреднения предсказаний всех моделей. Метод позволяет уменьшить дисперсию модели и повысить ее стабильность, особенно при работе с данными, содержащими шум или выбросы. Также, бутстреп ансамблирование эффективно для снижения переобучения и улучшения качества отдельных моделей [110].

4.1.1 Обучение и прогноз ансамбля

Рассмотрим алгоритм построения ансамбля независимых деревьев выживания с фиксированным размером M. Алгоритм принимает на вход следующие переменные: $Sample_{TR} = \{(X_i, T_i, \delta_i)\}$ — обучающая выборка, M — размер ансамбля, N — размер подвыборки. На этапе инициализации алгоритма определяется пустые списки базовых моделей $Models$ и out-of-bag выборок OOB.

Обучение ансамбля основано на итеративном алгоритме из 3 шагов:

1. **Семплирование данных.** На основе алгоритма семплинга с возвращением и равной вероятностью попадания наблюдений, определяется обучающая выборка $Sample_{bstr,i} = (X_i, T_i, \delta_i)_{n=1}^{N}$. Множество наблюдений, не вошедших в выборку $Sample_{bstr,i}$, называется out-of-bag (OOB) выборкой и равно $Sample_{OOB,i} = Sample_{TR} \setminus Sample_{bstr,i}$, где $\setminus$ — операция вычитания множеств. Каждая бутстреп подвыборка в среднем исключает $N \cdot 37\%$ данных [87].
2. **Обучение базовой модели.** Модель дерева выживаемости $Model_i$ (предложенная в разделе 2.3) строится на основе выборки $Sample_{bstr,i}$.
3. **Сохранение базовой модели.** Построенная модель $Model_i$ добавляется в список $Models$, а выборка $Sample_{OOB,i}$ — в список OOB.

Стоит отметить, что алгоритм обучения поддерживает параллелизацию вычислений иза. В таком случае, для каждой итерации $i = 1..M$ запускается параллельный процесс обучения i базовой модели. Результатом вычислений является пара значений: $Sample_{OOB,i}$ и $Model_i$. При условии сохранения порядка запуска процессов, формируются списки $Models$ и OOB, эквивалентные результатам итеративного алгоритма.

Для обеспечения баланса между скоростью и вычислительными затратами, параллелизация алгоритма может производиться по $k \leq M$ процессам, где при $k = 1$ выполняется итеративный алгоритм, а при $k = M$ — параллельный алгоритм по всем моделям.

На этапе применения модели, для входного наблюдения вычисляются прогнозы по всем базовым моделям из $Models$. В частности, для каждого дерева $Model_i$ определяется листовой узел и вычисляется точечный прогноз, функция выживания или функция риска. Прогноз анасамбля вычисляется путем агрегации прогнозов базовых моделей. В качестве функции агрегации может использоваться медиана, среднее или взвешенное среднее. Для прогнозирования функции выживания и кумулятивного риска расчитывается агрегация прогнозов для каждой точки временной шкалы.

Для управления сложности модели и контроля вычислительной нагрузки доступны следующие гиперпараметры: количество деревьев в ансамбле M, размер бутстреп выборки N, количество параллельных процессов k и параметры контроля роста базовых моделей (раздел 2.3.2). Дополнительными параметрами являются: доля случайных признаков для поиска разбиений в базовых моделях, метод агрегации прогнозов.

4.1.2 Определение размера ансамбля

Классические ансамбли деревьев выживаемости (раздел 5.3.4) имеют фиксированное количество моделей M. Выбор размера перекладывается на пользователя и, в то же время, сильно влияет на итоговое качество модели. При малом размере ансамбля возникают случаи недообучения из-за недостаточной сложности модели. Ансамбль высокого размера требует большого количества вычислительных ресурсов и часто избыточен для построения качественных прогнозов из-за схожести базовых моделей.

Для автоматического подбора размера ансамбля предлагается дополнить алгоритм построения модели Bootstrap. Алгоритм итеративного поиска размера основан на идее контролируемого добавления новых базовых моделей при росте качества ансамбля. Этап инициализации алгоритма построения ансамбля дополняется входным параметром метрики качества Q и созданием пустого списка показателей качества $Qualities$. Вместо фиксированного размера ансамбля пользователь указывает максимально допустимый размер M. К описанному итеративному алгоритму обучения добавляются следующие 2 шага (после шага (3) «Сохранение базовой модели»):

4. **Оценка качества.** После добавления модели в ансамбль вычисляется качество ансамбля q_i на OOB выборке по метрике Q. Важно отметить, что прогноз $pred_i$ базовой модели i вычисляется на соответствующей OOB_i выборке и для наблюдения с вектором x прогнозом ансамбля является агрегация $pred_i(x)$ среди всех моделей с $x \in OOB_i$. При вычислении прогноза по всем моделям возникает смещение прогноза из-за появления наблюдений в бутстреп-выборке некоторых моделей (следовательно, повышается риск переобучения). Далее, значение l_i добавляется в список $Qualities$.
5. **Проверка роста качества.** Если добавленная модель ухудшает качество ансамбля $q_i < q_{i-1}$, то она удаляется и построение ансамбля прекращается. В противном случае алгоритм возвращается к пункту (1).

Представленный алгоритм итеративного поиска размера позволяет ограничить вычислительную сложность, обучая только подмножество базовых моделей, которые приводят к росту качества ансамбля. Однако для оценки прироста качества определяется зависимость от предыдущей итерации. Следовательно, невозможна параллелизация алгоритма. Таким образом, алгоритм позволяет получить локальный максимум качества, склонен к ранней остановке и может приводить к недообучению моделей.

Альтернативный алгоритм основан на идее толерантности. В таком случае, проводится обучение всех базовых моделей (по пунктам 1-3) без проведения оценки качества на каждой итерации. На основе обученного ансамбля размера M итеративно вычисляется список $Qualities$, аналогичный пункту 4. Лучший размер определяет максимальное качество ансамбля на OOB выборке. Таким образом, толерантный алгоритм позволяет проводить параллельное обучение ансамбля и определяет размер с глобальным максимумом качества с учетом верхней границы M.

На рисунке 39 представлен пример выбора размера ансамбля при гиперпараметрах: критерий разбиения — «peto», максимальное количество деревьев — 50, метрика качества — «CI», глубина дерева — 5, минимальный размер листа — 0.01 (доля от обучающей выборки), размер подвыборки — 0.6 (доля от обучающей выборки). Синия линия определяет значения массива $Qualities$ относительно итерации i. Вертикальные линии определяют вы-

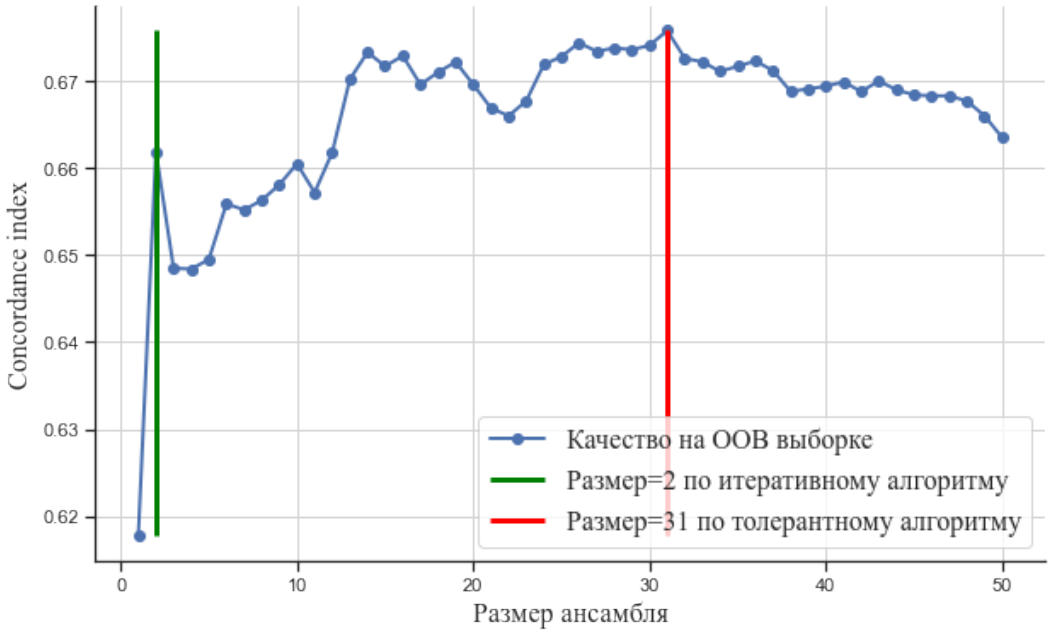


Рис. 39: Пример выбора размера бутстреп ансамбля независимых деревьев выживаемости по итеративному алгоритму (равен 2) и по толерантному алгоритму (равен 31).

бранный размер ансамбля по итеративному (равен 2) и толератному алгоритму (равен 31). Размер ансамбля по толерантному алгоритму имеет лучшее качество по метрике CI на OOB выборке.

4.2 Адаптивный бустинг с перевыборкой

4.2.1 Взвешенный бустинг

Альтернативным подходом построения бустинг ансамбля деревьев решений является алгоритм адаптивного бустинга AdaBoost [111]. В данном подходе каждая последующая базовая модель строится по выборке с наблюдениями, имеющими низкое качество прогноза на предыдущих итерациях ансамбля.

После добавления в ансамбль базовой модели, корректируются веса наблюдений, которые отвечают важности каждого из наблюдений обучающей выборки. Для наблюдений с низкой оценкой качества прогноза веса возрастают. Используя нормированные веса в качестве вероятностей попадания наблюдений в подвыборку, новое дерево решений строится на более «сложных» для прогноза наблюдениях.

В статье [112] предлагается модификация алгоритма AdaBoost. Метод основан на взвешивании не «важности вкладов» самих наблюдений, а их вероятностей попадания в последующие обучающие подвыборки (adaptive sampling). В таком случае каждое последующее дерево обучается на более «сложном» для прогнозирования подмножестве наблюдений.

Подробно рассмотрим алгоритм построения модели. На этапе инициализации каждому наблюдению обучающей выборки i сопоставляется вес $w_i = 1$. Обучение проводится итеративно, до тех пор пока средняя потеря $\bar{L}$ не превосходит 0.5:

1. Из обучающей выборки выбирается N наблюдений (с возвращением) с вероятностью $p_i = \frac{w_i}{\sum w_i}$, где p_i определяет вероятность попадания i наблюдения в выборку.
2. На полученной подвыборке строится регрессионная модель $h_t : x \to y$, на основе которой для каждого наблюдения обучающей выборки вычисляется прогноз $y_i^{(t)}(x_i)$.
3. Для каждого наблюдения из обучающей выборки вычисляется ошибка прогнозирования L_i. Пусть $D = \sup \left| y_i^{(p)}(x_i) - y_i \right|$, тогда в модели определяются линейная, квадратичная и экспоненциальная функции потерь:

$$L_i^{linear} = \frac{\left| y_i^{(p)}(x_i) - y_i \right|}{D}, \tag{36}$$

$$L_i^{sqr} = \frac{\left| y_i^{(p)}(x_i) - y_i \right|^2}{D^2}, \tag{37}$$

$$L_i^{exp} = 1 - \exp\left[\frac{-\left| y_i^{(p)}(x_i) - y_i \right|}{D} \right]. \tag{38}$$

4. Вычисляется значение средней потери $\bar{L} = \sum_{i=1}^{N_1} L_i p_i$ и меры уверенности модели t в прогнозе $\beta_t = \frac{\bar{L}}{1-\bar{L}}$. Меньшее значение отражает большую уверенность в прогнозе.
5. Веса наблюдений в обучающей выборке обновляются согласно правилу $w_i \leftarrow w_i \beta_t^{(1-L_i)}$. Чем меньше потеря для наблюдения, тем сильнее уменьшается вес и вероятность того, что наблюдение попадет в следующую обучающую выборку.

Прогнозом для наблюдения с вектором признаков x будет служить взвешенная сумма прогнозов $y^{(t)}(x)$ базовых регрессионных моделей ансамбля с нормированными весами $\log \frac{1}{\beta_t}$. Для обучения модели адаптивного бустинга необходимо определить функцию потерь L.

Данная модель применяется только для решения задач регресии, а базовые модели ансамбля h_t не способны учитывать связь между двумя целевыми переменными анализа выживаемости. В частности, расчет ошибки прогнозирования основан на вычислении абсолютной разницы между ожидаемым и наблюдаемым значением. Также, в открытых источниках не исследована сходимость алгоритма (в частности, выполнимость свойства остановки $\bar{L} \geq 0.5$). В то же время, идея алгоритма позволяет не вычислять градиент напрямую (раздел 1.6), работая только с вероятностями попадания наблюдений в обучающую выборку.

4.2.2 Предлагаемый метод

В данной диссертационной работе предлагается метод адаптивного бустинга деревьев выживаемости с перевыборкой (boosting), расширяющий рассмотренный подход адаптивного бустинга регрессионных моделей к задачам анализа выживаемости. Для обеспечения гибкости, предлагаемый метод использует большое количество гиперпараметров, задаваемых пользователем.

Рассмотрим алгоритм построения адаптивного бустинга деревьев выживаемости с фиксированным размером M. Аналогично описанию бутстреп модели (раздел 4.1), алгоритм принимает на вход следующие переменные: $Sample_{TR} = \{(X_i, T_i, \delta_i)\}$ — обучающая выборка, M — размер ансамбля, N — размер подвыборки, L — функция потерь.

При инициализации алгоритма определяются пустые списки базовых моделей $Models$, весов моделей $bettas$, out-of-bag выборок OOB и назначаются веса $w_i = 1$ для всех наблюдений обучающей выборки. Обучение модели основано на следующем итеративном алгоритме (рассмотритм итерацию t):

1. **Семплирование данных.** Строится подвыборка $Sample_{bstr,t} = (X_i, T_i, \delta_i)_{n=1}^{N}$ размера N с помощью семплинга с возвращением и вероятностью попадания $p_i = \frac{w_i}{\sum w_i}$ наблюдения i обучающей выборки. Оставшиеся наблюдения составляют выборку $Sample_{OOB,t}$.
2. **Обучение базовой модели.** Модель дерева выживаемости $Model_t$ строится на построенной подвыборке. Для наблюдения с вектором признаков X_i вычисляется прогноз $y_i^{(t)}(X_i)$. Прогноз согласуется с заданной функции потерь L (например, для L = IBS вычисляется прогноз функции выживания).
3. **Обновление весов наблюдений.** Для каждого наблюдения подвыборки с временем наступления события T_i и флагом цензурирования δ_i вычисляются ошибки прогнозирования $E_i = L(T_i, \delta_i, y_i^{(t)}(X_i))$ и значение $D = \max E_i$. Для i наблюдения потеря L_i вычисляется по модифицированным формулам (36)-(38):

$$L_i^{linear} = \frac{E_i}{D},\ L_i^{sqr} = \frac{E_i^2}{D},\ L_i^{exp} = 1 - \exp\left(\frac{-E_i}{D}\right).$$

   Также, предлагается новая сигмоидальная потеря:

$$L_i^{sigmoid} = 1/\left(1 + \exp\left(\frac{-E_i}{D}\right)\right)$$

   Формулы расчета средней потери $\bar{L} = \sum_{i=1}^{N_1} L_i p_i$, меры уверенности модели t: $\beta_t = \frac{\bar{L}}{1-\bar{L}}$ и обновления весов наблюдений $w_i \leftarrow w_i \beta_t^{(1-L_i)}$ эквивалентны разделу 4.2.1.
4. **Сохранение базовой модели.** Построенная модель $Model_t$ добавляется в список $Models$, вес модели β_t — в список $bettas$, а выборка $Sample_{OOB,t}$ — в список OOB.

Прогноз boosting модели для наблюдения с вектором признаков X вычисляется через взвешенную сумму прогнозов базовых моделей с весами $wei_t = \log(\frac{1}{\beta_t})$):

$$y(X) = \sum_{t=1}^{M} \frac{wei_j}{\sum wei_k} y^{(j)}(X). \tag{39}$$

В частности, функция выживания и кумулятивного риска вычисляются как агрегация по каждому моменту времени. Таким образом, модель boosting объединяет алгоритмы построения адаптивного бустинга регрессионных моделей и бутстреп ансамбля деревьев выживаемости для повышения качества прогнозирования. Стоит отметить, что предложенный метод может быть дополнен итеративным и толерантным алгоритмом поиска размера ансамбля (раздел 4.1.2). Итеративный алгоритм (с максимальным числом итераций M) позволяет преодолеть проблемы со сходимостью и выполнимостью эмпирического правила $\bar{L} \geq 0.5$ остановки обучения.

Необходимым требованием для функции потерь L является возможность представления функции в виде суммы отдельных оценок (раздел 3.1.1) для расчета индивидульных

Рис. 40: Пример прогнозируемых функций выживания для каждого дерева решений адаптивного бустинга с перевыборкой (размер ансамбля равен 3)

потерь по отдельному наблюдению. Как было отмечено в разделе 3.1.1, таким свойством не обладают метрики ранжирования (например, $CI, IAUC, IAUC_{WW,TI}$) и сбалансированные метрики. Далее необходимо провести дополнительное исследование влияния функции потерь на качество моделей.

На рисунках 40 и 41 представлен пример прогнозирования для модели адаптивного бустинга с перевыборкой с размером ансамбля $M = 3$. На рисунке 40 изображены прогнозируемые функции выживания каждого из деревьев ансамбля (Tree 1, Tree 2, Tree 3). На рисунке 41 изображен процесс получения итогового прогноза: на левом рисунке изображены прогнозы функции выживания по каждой модели, в центре прогнозы взвешены относительно веса дерева wei_t, на правом рисунке изображена итоговая функция выживания (взвешенная сумма прогнозов моделей ансамбля).

На Рисунке 42 представлен псевдокод предложенного алгоритма построения адаптивного бустинга с перевыборкой. Для контроля вычислительной сложности доступны следующие параметры для вариации: максимальное количество деревьев в ансамбле, размер бутстреп выборки, флаг режима толерантности, метод агрегации прогнозов моделей ансамбля, метрика расчета OOB ошибки, метрика расчета весов наблюдений, параметры контроля роста деревьев выживаемости.

4.2.3 Стратегии локализации обновления весов

В отличии от регрессионного адаптивного бустинга, предложенный адаптивный бустинг с перевыборкой корректирует только веса наблюдений бутстреп-подвыборки на t итерации. Данная модификация позволяет повысить качество прогнозирования моделей путем увеличения различий между базовыми моделями ансамбля.

Далее подробно рассмотрим различные стратегии локализации обновления весов. Будем называть стратегию глобальной, если формула обновления весов применяется для всей обучающей выборки ансамбля. Будем называть стратегию локальной, если формула обновления весов применяется только для наблюдений, попавших в последнюю выборку $Sample_{bstr,t}$.

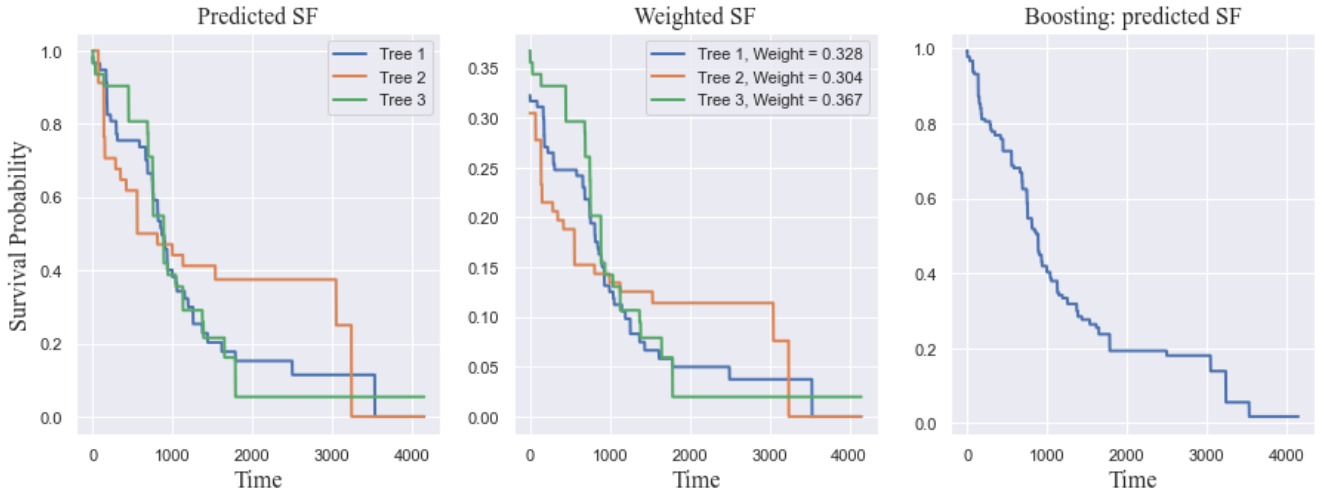


Рис. 41: Пример построения итогового прогноза адаптивного бустинга с перевыборкой. На левом рисунке изображены прогнозируемые функции выживания по всем моделям ансамбля. В центре прогнозы взвешиваются по весам деревьев решений. На правом рисунке изображен прогноз адаптивного бустинга с перевыборкой (взвешенная сумма прогнозируемых функций базовых моделей)

Алгоритм 5 BoostingConstruction – построение адаптивного бустинга с перевыборкой

Вход:

X, ▷ Признаковое пространство наблюдений
T, ▷ Время наступления событий выборки
δ, ▷ Флаг цензурирования наблюдений выборки
M, ▷ Размер ансамбля
N, ▷ Размер подвыборки
L ▷ Функция потерь

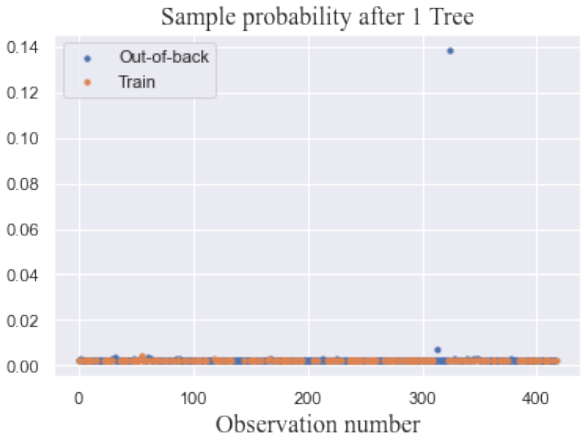
Выход: BoostingEnsemble ▷ Модель адаптивного бустинга с перевыборкой

$w_i \leftarrow 1$
Ensemble $\leftarrow$ []
Bettas $\leftarrow$ []
EnsembleScore $\leftarrow$ []
for t in range(M) **do**
    $X_{bstr}, T_{bstr}, \delta_{bstr} \leftarrow$ GetBootstrapSample(X, T, δ, w_i, N)
    $X_{OOB}, T_{OOB}, \delta_{OOB} \leftarrow (X, T, \delta) \setminus (X_{bstr}, T_{bstr}, \delta_{bstr})$
    Model $\leftarrow$ TreeConstruction($X_{bstr}, T_{bstr}, \delta_{bstr}$)
    wei, betta $\leftarrow$ CountModelWeights(T_{bstr}, δ_{bstr}, Model.predict(X_{bstr}))
    $E_i \leftarrow$ L(T_{bstr}, δ_{bstr}, Model.predict(X_{bstr}))
    $L_i \leftarrow$ WeightScheme(E_i)
    Betta $\leftarrow \frac{Mean(L_i)}{1-Mean(L_i)}$
    $w_i[X_{bstr}.\text{Index}] *= \text{Betta}^{(1-L_i)}$

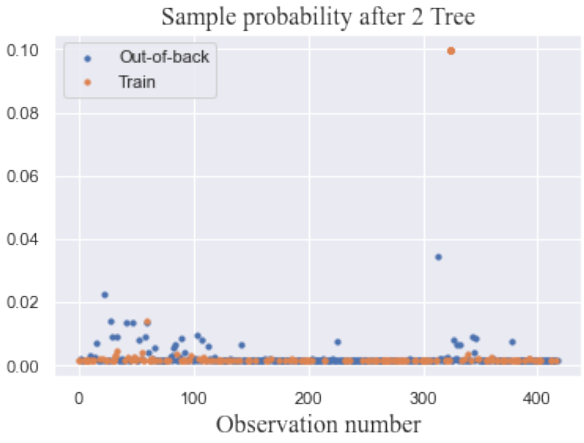
    Ensemble $\leftarrow$ Ensemble $\cup$ Model
    Bettas $\leftarrow$ Bettas $\cup$ Betta
    EnsembleScore $\leftarrow$ EnsembleScore $\cup$ Ensemble.OOBScore(L)
Ensemble $\leftarrow$ SelectBestSize(Ensemble, EnsembleScore)
return Ensemble

Рис. 42: Псевдокод алгоритма построения адаптивного бустинга с перевыборкой.

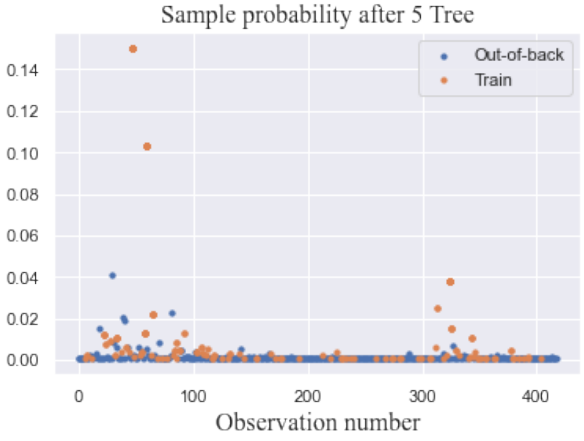
Рассмотрим 2 ансамбля с глобальной и локальной статегией обновления весов, для которых остальные гиперпараметры совпадают. Для каждого дерева ансамбля визуализируем на графиках распределение весов наблюдений после их обновления на основе нового дерева. Таким образом, по оси x отметим истинное время наблюдения, а по оси y — обновленную вероятность попадания в следующую выборку $P(x \in Sample_{bstr,t})$. Синим отметим наблюдения OOB выборки (не используются при обучении дерева), оранжевым отмечены наблюдения бутстреп-подвыборки.



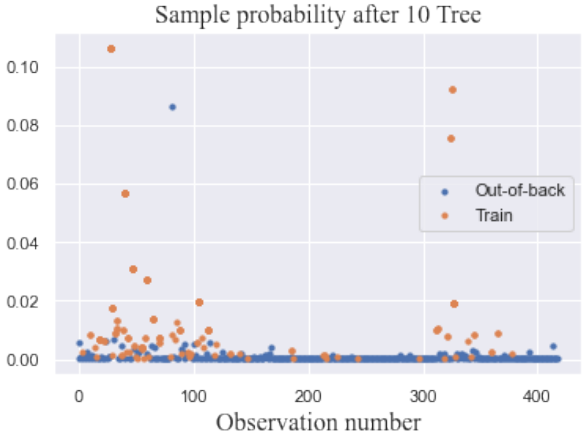
(a) $P(x \in Sample_{bstr,1})$



(b) $P(x \in Sample_{bstr,2})$



(c) $P(x \in Sample_{bstr,5})$



(d) $P(x \in Sample_{bstr,10})$

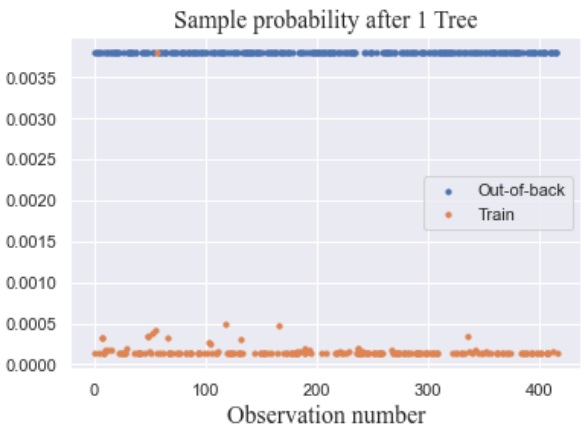
Рис. 43: Пример глобальной стратегии обновления весов в модели адаптивного бустинга с перевыборкой. Для всех деревьев ансамбля определяется небольшое количество вырожденных повторяющихся точек, вес которых значимо отличается от нуля.

На рисунке 43 представлена глобальная стратегия обновления весов для всех данных ансамбля. Видно, что для всех деревьев ансамбля существует только небольшое количество вырожденных повторяющихся точек, значимо отличающихся от нуля. Так как данные точки обладают намного большей вероятностью, то про построении бутстреп-выборки (с перевыборкой) они будут встречаться чаще остальных наблюдений. В таком случае, деревья выживаемости будут получаться схожими.

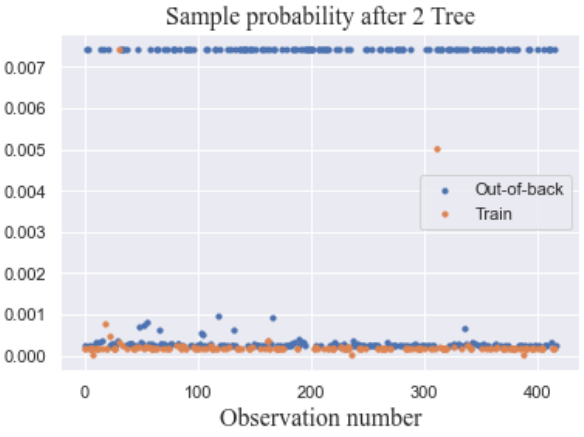
Появление таких вырожденных точек можно объяснить наличием выбросов в данных (наблюдений, которые не могут быть описаны на основе общих зависимостей в наборе). В

таком случае, после построения дерева, прогноз для таких наблюдений не приближается к истинному значению и при обновлении весов наблюдения сохраняют высокую вероятность попадания в выборку.

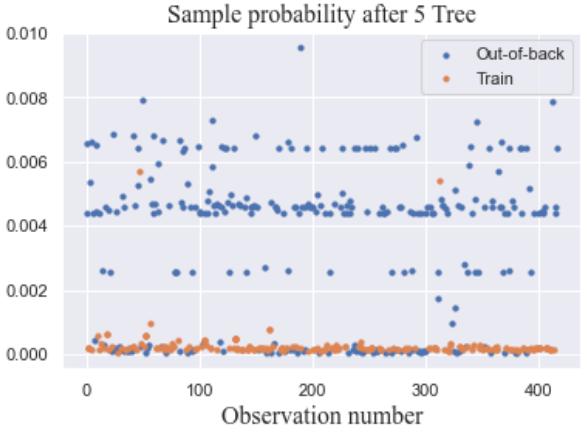
Также, рассмотрим результаты поиска размера ансамбля по метрике IBS. При построении ансамбля, на каждом этапе добавления дерева были получены следующие значения метрики: `[0.168 0.1591 0.1688 0.1648 0.1696 0.1651 0.165 0.1655 0.1646 0.1645]`. С точки зрения минимизации IBS, лучший размер ансамбля равен 2 (минимальный IBS=0.1591).



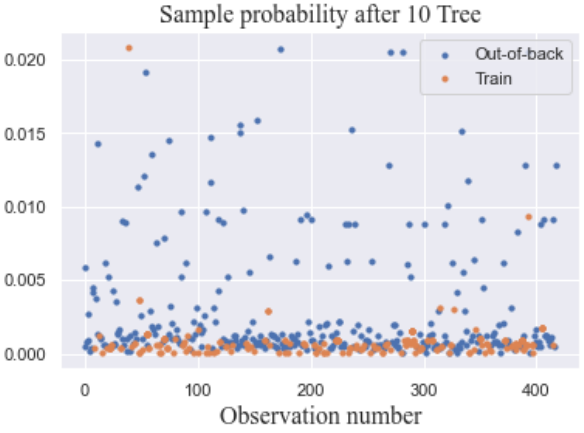
(a) $P(x \in Sample_{bstr,1})$



(b) $P(x \in Sample_{bstr,2})$



(c) $P(x \in Sample_{bstr,5})$



(d) $P(x \in Sample_{bstr,10})$

Рис. 44: Пример локальной стратегии обновления весов в модели адаптивного бустинга с перевыборкой. При обновлении весов не сохраняется подгруппа вырожденных наблюдений с большим весом.

На рисунке 44 представлена локальная стратегия обновления весов только для наблюдений выборки обучения дерева S_{n-1}. Видно, что для всех деревьев ансамбля не сохраняется вырожденность конкретных точек выборки. Вместо это, для множества точек обучающей выборки S_{n-1} снижается вероятность попадания в следующую выборку и точки OOB «преобладают» при построении последующих выборок. Таким образом, последующие деревья выживания не имеют сильного сходства с предыдущими, что позволяет точнее описывать различные подвыборки.

Также, рассмотрим результаты отбора размера ансамбля по метрике IBS. При построении ансамбля, на каждом этапе были получены следующие значения метрики: `[0.168, 0.1564, 0.1551, 0.1547, 0.1513, 0.1478, 0.1438, 0.143, 0.1414, 0.1422]`. Относительно метрики IBS, размер ансамбля равен 9 (минимальный IBS=0.1414). Отметим, что минимальное значение IBS по глобальной стратегии взвешивания превышает значение IBS по локальной стратегии. Также, в обоих случаях, для выбранного подмножества деревьев достигается лучшее качество, чем при полном ансамбле размера 10.

4.3 Экспериментальное исследование

В результате проведенных исследований были предложены два метода ансамблирования моделей анализа выживаемости. Бутстреп модель основана на построении независимых деревьев выживания и агрегации их прогнозов. Модель адаптивного бустинга с перевыборкой основана на идее итеративного расчета вероятностей наблюдений относительно качества прогноза базовых моделей с последующем построением базовых моделей на сложных подмножествах наблюдений.

Оба метода используют функцию потерь для поиска оптимального размера ансамбля и оценки качества прогнозирования на этапе обучения. В разделе 3.1 было проведено сравнение характеристик существующих метрик качества и сделан вывод об устойчивости метрик $CI, IBS_{RM}, IAUC_{WW,TI}, AUPRC$ к рассматриваемым случаям избыточной чувствительности. Однако при построении предложенных ансамблевых моделей необходимо зафиксировать функцию потерь.

Первой целью экспериментального исследования является выбор оптимальной функции потерь и анализ влияния модификаций существующих метрик на качество модели. В частности, исследование проводится на предложенном бутстреп ансамбле древовидных моделей.

Второй целью экспериментального исследования является оценка и сравнение качества существующих и предложенных моделей анализа выживаемости. Рассматриваются следующие классы существующих моделей: статистические модели (непараметрические, полупараметрические, параметрические), модели машинного обучения (древовидные, ансамблевые).

4.3.1 Постановка эксперимента

Схема исследовния качества функций потерь представлена на Рисунке 45. В разделе 2.1 были описаны характеристики 6 открытых медицинских наборов данных и выявлены особенности целевых переменных: информативность цензурирования, различие форм распределений времени событий, дисбаланс цензурирования. Как было показано в разделе 3.1, данные особенности влияют не только на построение прогнозных моделей, но и на метрики оценки качества. Были представлены 4 случая искажения значимости вкладов: отдельных событий, временных компонент, временных интервалов и классов событий. Для преодоления существующих недостатков были предложены модифицированные метрики качества. По результатам сравнения устойчивости 3.1.5, метрики $CI, IBS_{RM}, IAUC_{WW,TI}, AUPRC$ были рекомендованы для оценки качества прогнозирования 3х величин анализа выживаемости: ожидаемого времени события, функции выживания, функции риска.

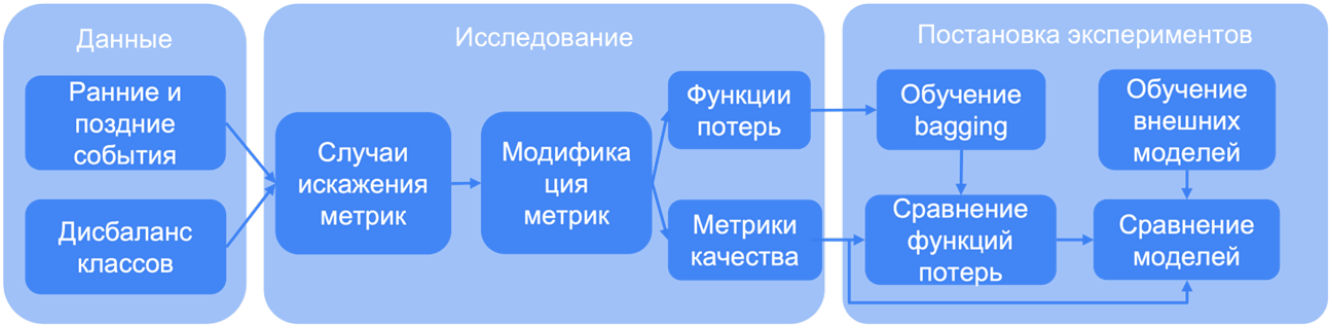


Рис. 45: Схема исследования качества функций потерь.

В данном разделе проводится экспериментальное сравнение качества прогнозирования бутстреп ансамбля при использовании различных функций потерь. По результатам сравнения фиксируется функция потерь с лучшим качеством моделей.

Далее проводится сравнение качества предложенных моделей (с выбранной функцией потерь) с существующими методами анализа выживаемости. Алгоритм оценки качества моделей аналогичен алгоритму в разделе 3.2.1: на подмножестве исходных данных проводится поиск лучших гиперпараметров по сетке (используется 5-кратная кросс-валидация), с которыми лучшие модели 20-кратно обучаются на подмножествах исходных данных и применяются к оставшимся данным (на каждой итерации).

Таблица 14: Гиперпараметры существующих моделей

Модель	**Параметр**	**Сетка**
Cox Proportional Hazard (CoxPH)	penalty	0.1, 0.01, 0.001
	ties	breslow, efron
Accelerated failure time (AFT)	penalizer	0, 0.01, 0.1, 0.5, 1.0
	l1 ratio	10^x for x from -3 to 3
	distribution	Weibull, LogNormal, LogLogistic
Survival Tree (ST)	split strategy	best, random
	max depth	from 10 to 30 step 5
	min sample leaf	from 1 to 20 step 1
	max features	sqrt, log2, None
Random Survival Forest (RSF)	num estimators	from 10 to 100 step 10
	max depth	from 10 to 30 step 5
	min sample leaf	from 1 to 20 step 1
	max features	sqrt, log2, None
Component-wise Gradient Boosting (CWGBSA)	num estimators	from 10 to 100 step 10
	loss function	coxph
	learning rate	from 0.01 to 0.5 step 0.01
	subsample	from 0.5 to 1.0 step 0.1
	dropout rate	from 0.0 to 0.5 step 0.1
Gradient Boosting SA (GBSA)	num estimators	from 10 to 100 step 10
	max depth	from 10 to 30 step 5
	min sample leaf	from 1 to 20 step 1
	max features	sqrt, log2, None
	loss function	coxph
	learning rate	from 0.01 to 0.5 step 0.01
DeepSurv & CoxTime	batch size	from 0.01 to 0.1 step 0.01
	batch norm	true, false
	dropout	from 0.0 to 0.5 step 0.1
	num nodes	$[2^x, 2^x]$ for x from 5 to 9

Таблица 15: Гиперпараметры предложенных моделей

Модель	**Параметр**	**Сетка**
TREE	max depth	from 10 to 30 step 5
	min sample leaf	0.05, 0.01, 0.001
	signif	0.05, 0.1, 1.0
	lambda	0.0, 0.01, 0.1, 0.5, 0.9
	leaf model	KM, KMWV
	criterion	peto, tarone-ware, wilcoxon, logrank
Bootstrap	subsample	from 0.5 to 1.0 step 0.1
	num estimators	from 10 to 50 step 10
	max depth	from 10 to 30 step 5
	min sample leaf	0.05, 0.001
	max features	0.3, sqrt
	leaf model	KM, KMWV
	criterion	peto, tarone-ware, wilcoxon, logrank
Boosting	subsample	from 0.3 to 0.9 step 0.1
	num estimators	from 10 to 50 step 5
	max depth	from 10 to 30 step 5
	min sample leaf	from 1 to 20 step 1
	weight scheme	linear, exp, sigmoid
	ensemble metric	IBS_{RM}
	weighted sum	true, false

Реализации моделей Cox Proportional Hazard, Survival Tree, Random Survival Forest, Component-wise Gradient Boosting, Gradient Boosting SA были взяты из открытой библиотеки Scikit-survival, реализации WeibullAFT, LogLogisticAFT, LogNormalAFT — из библиотеки Lifelines, а реализации DeepSurv, CoxTime — из библиотеки PyCox. Гиперпараметры моделей представлены в таблице 14. Гиперпараметры предложенных моделей Bootstrap (раздел 4.1) и Boosting (раздел 4.2) представлены в таблице 15.

4.3.2 Сравнение функций потерь

В таблицах 16-21 представлены результаты экспериментального сравнения функций потерь при построении бутстреп ансамбля. Рассматривались 11 функций потерь: CI, LL, IBS, IBS_{WW}, IBS_{RM}, IBS_{BAL}, $IBS_{WW,BAL}$, $IBS_{RM,BAL}$, $IAUC_{WW,TI}$, $AUPRC$, $AUPRC_{BAL}$.

Оценка качества прогнозирования проводилась по метрикам: $CI \uparrow$, $IBS_{RM} \downarrow$, $IAUC_{WW,TI} \uparrow$, $AUPRC \uparrow$. Ячейки содержат среднее качество по 5 моделям с лучшими гиперпараметрами и фиксированной функцией потерь. Жирным отмечены 3 лучших значения по каждой метрики качества без учета оптимизируемой метрики (подчеркнута).

Таблица 16: Сравнение функций потерь на наборе GBSG

Функция потерь	$\boldsymbol{CI}$	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
CI	0.617	**0.1517**	**0.7344**	0.6963
LL	0.6154	0.1539	0.7293	0.6887
$IAUC_{WW,TI}$	0.6175	0.152	0.7399	**0.7004**
$AUPRC$	0.5879	0.1759	0.6808	0.7026
$AUPRC_{BAL}$	0.5919	0.1718	0.6856	**0.7038**
IBS	0.6164	0.1536	0.7298	0.6876
IBS_{RM}	**0.6182**	0.1516	**0.7423**	**0.7016**
IBS_{WW}	**0.6177**	**0.1514**	**0.7336**	0.6945
IBS_{BAL}	0.6165	0.153	0.7302	0.6877
$IBS_{RM,BAL}$	0.616	0.154	0.7285	0.6876
$IBS_{WW,BAL}$	**0.6181**	**0.1511**	0.7314	0.691

Таблица 17: Сравнение функций потерь на наборе PBC

Функция потерь	$\boldsymbol{CI}$	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
CI	0.6993	**0.113**	**0.8724**	0.7458
LL	0.6932	**0.113**	0.8650	0.739
$IAUC_{WW,TI}$	**0.6963**	0.1138	0.8735	**0.7464**
$AUPRC$	0.666	0.1392	0.8097	0.7469
$AUPRC_{BAL}$	0.6684	0.1332	0.8214	**0.7476**
IBS	0.6921	0.114	0.8599	0.7415
IBS_{RM}	**0.6934**	0.1126	**0.8678**	**0.7449**
IBS_{WW}	0.6895	**0.1135**	0.8601	0.7421
IBS_{BAL}	0.6916	0.1137	0.8597	0.7419
$IBS_{RM,BAL}$	0.692	0.114	0.8606	0.7399
$IBS_{WW,BAL}$	**0.6937**	0.1138	**0.8652**	0.7393

Таблица 18: Сравнение функций потерь на наборе ROTT2

Функция потерь	$\boldsymbol{CI}$	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
CI	0.668	0.1192	**0.7621**	**0.7182**
LL	0.6687	**0.1173**	0.738	0.7106
$IAUC_{WW,TI}$	0.6622	0.1212	0.7633	0.7131
$AUPRC$	0.6435	0.1256	0.7434	0.7185
$AUPRC_{BAL}$	0.6473	0.1246	0.7451	**0.7191**
IBS	0.6664	0.1186	0.7404	0.7119
IBS_{RM}	**0.6695**	0.1173	**0.7491**	**0.717**
IBS_{WW}	**0.669**	**0.1173**	**0.7452**	0.7123
IBS_{BAL}	0.6666	0.1185	0.7408	0.7118
$IBS_{RM,BAL}$	0.6641	0.1192	0.7341	0.7112
$IBS_{WW,BAL}$	**0.6688**	**0.1173**	0.745	0.7122

Таблица 19: Сравнение функций потерь на наборе WUHAN

Функция потерь	$\boldsymbol{CI}$	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
CI	<u>0.7634</u>	**0.0735**	**0.8586**	0.7491
LL	0.7505	0.0753	0.8425	0.7414
$IAUC_{WW,TI}$	0.747	0.0767	<u>0.8505</u>	0.7498
$AUPRC$	0.7055	0.0968	0.824	<u>0.7552</u>
$AUPRC_{BAL}$	0.7017	0.0962	0.8163	**0.7546**
IBS	0.749	0.0758	0.8429	0.7491
IBS_{RM}	**0.7562**	<u>0.071</u>	**0.8508**	**0.7534**
IBS_{WW}	**0.7561**	**0.0711**	**0.8508**	**0.7533**
IBS_{BAL}	0.7463	0.0761	0.8403	0.7497
$IBS_{RM,BAL}$	0.7378	0.0785	0.8422	0.7525
$IBS_{WW,BAL}$	**0.756**	**0.0711**	0.8501	0.753

Таблица 20: Сравнение функций потерь на наборе SMARTO

Функция потерь	$\boldsymbol{CI}$	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
CI	<u>0.7313</u>	0.1434	0.6485	0.8846
LL	**0.7245**	0.1441	0.6403	0.8846
$IAUC_{WW,TI}$	0.714	**0.1395**	<u>0.6689</u>	**0.8896**
$AUPRC$	0.6305	0.1551	0.6033	<u>0.8889</u>
$AUPRC_{BAL}$	0.644	0.1524	0.6221	0.8888
IBS	0.709	0.1458	0.652	0.8795
IBS_{RM}	**0.7169**	<u>0.1395</u>	**0.6667**	**0.8895**
IBS_{WW}	**0.7227**	**0.1408**	**0.6627**	0.8873
IBS_{BAL}	0.7144	**0.1394**	**0.6684**	**0.8889**
$IBS_{RM,BAL}$	0.7152	0.1418	0.6613	0.8852
$IBS_{WW,BAL}$	0.7151	0.141	0.6623	0.8872

Таблица 21: Сравнение функций потерь на наборе SUPPORT2

Функция потерь	$\boldsymbol{CI}$	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
CI	<u>0.807</u>	0.0986	0.8982	0.557
LL	0.805	0.0992	0.8942	0.545
$IAUC_{WW,TI}$	**0.8062**	0.0978	<u>0.9004</u>	0.5592
$AUPRC$	0.7914	0.1123	0.8721	<u>0.5797</u>
$AUPRC_{BAL}$	0.7923	0.1114	0.8764	**0.5803**
IBS	0.8052	**0.0974**	**0.8993**	0.5608
IBS_{RM}	**0.8055**	<u>0.0973</u>	**0.8985**	0.5592
IBS_{WW}	**0.8055**	**0.0973**	**0.8985**	0.5593
IBS_{BAL}	0.805	**0.0977**	0.8983	**0.5618**
$IBS_{RM,BAL}$	0.8045	0.0986	0.898	0.5603
$IBS_{WW,BAL}$	**0.8062**	0.0979	**0.8989**	**0.5632**

Таблица 22: Сводная таблица лучших функций потерь для каждого набора данных по 4м метрикам качества (порядок убывания наборов данных по доле терминальных событий). Столбец «Итог» определяет лучшую функцию потерь для каждого набора данных.

Набор	$\boldsymbol{CI}$	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$	**Итог**
$SUPPORT2$	$IAUC_{WW,TI}$ $IBS_{WW,BAL}$ IBS_{RM} IBS_{WW}	IBS IBS_{WW} IBS_{BAL}	IBS $IBS_{WW,BAL}$ IBS_{RM} IBS_{WW}	$AUPRC_{BAL}$ $IBS_{WW,BAL}$ IBS_{BAL}	$IBS_{WW,BAL}$ IBS_{RM} IBS_{WW}
$WUHAN$	IBS_{RM} IBS_{WW} $IBS_{WW,BAL}$	IBS_{WW} $IBS_{WW,BAL}$ CI	CI IBS_{RM} IBS_{WW}	IBS_{RM} IBS_{WW} $AUPRC_{BAL}$	IBS_{RM} IBS_{WW}
$GBSG$	IBS_{RM} IBS_{WW} $IBS_{WW,BAL}$	$IBS_{WW,BAL}$ IBS_{WW} CI	IBS_{RM} IBS_{WW} CI	$AUPRC_{BAL}$ $IAUC_{WW,TI}$ IBS_{RM}	IBS_{RM} IBS_{WW}
$ROTT2$	IBS_{RM} IBS_{WW} $IBS_{WW,BAL}$	LL IBS_{WW} $IBS_{WW,BAL}$	CI IBS_{RM} IBS_{WW}	$AUPRC_{BAL}$ CI IBS_{RM}	IBS_{RM} IBS_{WW}
PBC	IBS_{RM} $IAUC_{WW,TI}$ $IBS_{WW,BAL}$	LL CI IBS_{WW}	CI IBS_{RM} $IBS_{WW,BAL}$	$AUPRC_{BAL}$ $IAUC_{WW,TI}$ IBS_{RM}	IBS_{RM}
$SMARTO$	IBS_{RM} IBS_{WW} LL	$IAUC_{WW,TI}$ IBS_{WW} IBS_{BAL}	IBS_{RM} IBS_{WW} IBS_{BAL}	$IAUC_{WW,TI}$ IBS_{RM} IBS_{BAL}	IBS_{RM} IBS_{WW} IBS_{BAL}

В таблице 22 представлены сводные результаты по выбору топ-3 функций потерь по каждой метрике качества. Столбец «Итог» определяет лучшую функцию потерь для каждого набора данных. По результатам экспериментов, метрика IBS_{RM} стала лучшей для шести наборов данных, метрика IBS_{WW} для пяти наборов данных, метрики $IBS_{WW,BAL}$ и IBS_{BAL} для одного набора данных.

Таким образом, при использовании функции потерь IBS_{RM} достигается лучшее качество по совокупности метрик на всех наборах данных. Далее функция потерь IBS_{RM} фиксируется для построения предложенных бутстреп и бустинг ансамблей, а также используется на этапе отбора лучших гиперпараметров по кросс-валидации.

4.3.3 Сравнение методов анализа выживаемости

В таблицах 23-28 представлены результаты экспериментального сравнения прогнозных моделей анализа выживаемости. Оценка качества прогнозирования проводилась по метрикам: $CI \uparrow$, $IBS_{RM} \downarrow$, $IAUC_{WW,TI} \uparrow$, $AUPRC \uparrow$. Для предложенной модели *Boosting* определены 3 вариации весовой схемы: *exp*, *linear*, *sigmoid*. Жирным отмечены 3 лучших значения по каждой метрики качества.

По результатам экспериментального исследования, предложенная модель дерева выживаемости $TREE_{KMWV}$ показала лучшее качество по метрике $AUPRC$ на всех набо-

Таблица 23: Сравнение качества моделей на наборе GBSG

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
$KaplanMeier$	0.500±0.000	0.194±0.002	0.495±0.008	0.591±0.003
$CoxPH$	0.609±0.016	0.171±0.004	0.723±0.030	0.622±0.004
$LogLogisticAFT$	0.612±0.015	0.155±0.004	**0.737±0.026**	0.616±0.002
$LogNormalAFT$	0.610±0.015	0.155±0.005	0.734±0.028	0.626±0.003
$WeibullAFT$	0.612±0.014	0.158±0.003	**0.737±0.024**	0.603±0.002
ST	0.553±0.021	0.188±0.006	0.609±0.038	0.605±0.007
RSF	0.616±0.013	0.172±0.003	0.733±0.024	0.610±0.003
$CWGBSA$	0.581±0.019	0.184±0.002	0.690±0.025	0.599±0.003
$GBSA$	0.613±0.013	0.173±0.003	0.728±0.025	0.608±0.003
$DeepSurv$	0.606±0.020	0.159±0.005	0.726±0.030	0.669±0.005
$CoxTime$	0.607±0.021	0.170±0.005	0.723±0.033	0.675±0.005
$TREE_{KMWV}$	0.599±0.019	0.155±0.008	0.708±0.033	**0.709±0.004**
$Bootstrap$	**0.619±0.018**	**0.151±0.005**	0.733±0.032	**0.696±0.005**
$Boosting_{linear}$	0.616±0.019	**0.147±0.005**	0.734±0.030	0.688±0.005
$Boosting_{exp}$	**0.618±0.019**	**0.147±0.005**	0.735±0.031	0.692±0.004
$Boosting_{sigmoid}$	**0.619±0.015**	0.152±0.005	**0.736±0.032**	**0.697±0.004**

Таблица 24: Сравнение качества моделей на наборе PBC

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
$KaplanMeier$	0.500±0.000	0.198±0.002	0.498±0.004	0.609±0.002
$CoxPH$	**0.678±0.019**	0.120±0.009	**0.862±0.026**	0.698±0.006
$LogLogisticAFT$	**0.680±0.020**	0.125±0.010	**0.860±0.027**	0.684±0.006
$LogNormalAFT$	**0.681±0.019**	0.129±0.010	0.857±0.026	0.678±0.008
$WeibullAFT$	0.675±0.022	0.126±0.010	0.854±0.032	0.670±0.005
ST	0.520±0.038	0.195±0.011	0.553±0.080	0.617±0.011
RSF	0.677±0.018	0.133±0.006	**0.868±0.027**	0.661±0.005
$CWGBSA$	0.648±0.031	0.190±0.003	0.820±0.063	0.612±0.002
$GBSA$	0.646±0.023	0.188±0.002	0.820±0.031	0.614±0.002
$DeepSurv$	0.658±0.039	0.135±0.020	0.829±0.075	0.709±0.025
$CoxTime$	0.675±0.013	0.124±0.012	0.859±0.029	0.726±0.009
$TREE_{KMWV}$	0.659±0.019	0.125±0.011	0.847±0.028	**0.763±0.010**
$Bootstrap$	0.672±0.015	**0.113±0.007**	0.852±0.024	**0.738±0.008**
$Boosting_{linear}$	0.667±0.017	**0.113±0.008**	0.858±0.026	0.732±0.006
$Boosting_{exp}$	0.670±0.015	**0.113±0.007**	0.854±0.023	0.730±0.007
$Boosting_{sigmoid}$	0.672±0.015	0.115±0.008	0.843±0.027	**0.738±0.007**

рах данных. Модель бутстреп ансамбля *Bootstrap* показала лучшее качество на наборах $GBSG$, $ROTT2$, $WUHAN$, $SUPPORT2$ и заняла второе место на наборах $SMARTO$ и PBC. Модель адаптивного бустинга со схемой *sigmoid* показала лучшее качество на наборах $GBSG$, $SMARTO$ и заняла второе место на наборах $ROTT2$, $WUHAN$, $SUPPORT2$. Модель адаптивного бустинга со схемой *linear* показала лучшее качество на наборах $ROTT2$,

Таблица 25: Сравнение качества моделей на наборе ROTT2

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
$KaplanMeier$	0.500±0.000	0.179±0.002	0.496±0.004	0.595±0.003
$CoxPH$	0.596±0.009	0.476±0.010	0.710±0.018	0.000±0.000
$LogLogisticAFT$	0.606±0.009	0.153±0.003	0.728±0.017	0.629±0.002
$LogNormalAFT$	0.596±0.008	0.157±0.003	0.711±0.017	0.625±0.002
$WeibullAFT$	0.602±0.009	0.153±0.004	0.719±0.017	0.631±0.002
ST	0.555±0.018	0.180±0.007	0.619±0.031	0.607±0.005
RSF	0.599±0.008	0.154±0.002	0.727±0.017	0.608±0.004
$CWGBSA$	0.609±0.010	0.154±0.003	0.730±0.018	0.629±0.003
$GBSA$	0.619±0.010	0.146±0.004	**0.747±0.021**	0.638±0.004
$DeepSurv$	0.603±0.011	0.153±0.004	0.726±0.020	0.663±0.002
$CoxTime$	0.602±0.010	0.155±0.005	0.720±0.022	0.670±0.003
$TREE_{KMWV}$	0.619±0.010	0.128±0.004	0.735±0.017	**0.723±0.003**
$Bootstrap$	**0.649±0.010**	**0.117±0.003**	0.742±0.021	**0.713±0.003**
$Boosting_{linear}$	**0.652±0.009**	**0.114±0.003**	**0.764±0.019**	0.702±0.002
$Boosting_{exp}$	**0.652±0.009**	**0.115±0.003**	**0.763±0.018**	0.707±0.002
$Boosting_{sigmoid}$	0.646±0.007	0.119±0.003	0.733±0.015	**0.716±0.002**

Таблица 26: Сравнение качества моделей на наборе WUHAN

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
$KaplanMeier$	0.500±0.000	0.191±0.002	0.487±0.013	0.551±0.003
$CoxPH$	0.709±0.028	0.126±0.014	0.820±0.061	0.728±0.018
$LogLogisticAFT$	0.666±0.018	0.188±0.002	0.487±0.013	0.510±0.002
$LogNormalAFT$	0.653±0.019	0.240±0.031	0.735±0.048	0.631±0.018
$WeibullAFT$	0.500±0.000	0.177±0.002	0.487±0.013	0.557±0.001
ST	0.636±0.045	0.144±0.034	0.751±0.088	0.651±0.038
RSF	0.683±0.012	0.088±0.004	0.823±0.046	0.679±0.005
$CWGBSA$	0.692±0.012	0.094±0.005	0.830±0.049	0.675±0.005
$GBSA$	0.681±0.011	0.099±0.005	0.802±0.042	0.662±0.003
$DeepSurv$	0.671±0.026	0.101±0.024	0.827±0.056	0.741±0.039
$CoxTime$	0.665±0.013	0.102±0.013	0.819±0.052	0.733±0.008
$TREE_{KMWV}$	0.693±0.019	0.101±0.010	0.822±0.041	**0.757±0.011**
$Bootstrap$	**0.731±0.019**	**0.071±0.004**	**0.855±0.045**	**0.752±0.006**
$Boosting_{linear}$	**0.734±0.015**	0.074±0.005	0.838±0.050	0.741±0.006
$Boosting_{exp}$	**0.741±0.015**	**0.073±0.006**	**0.847±0.049**	0.742±0.006
$Boosting_{sigmoid}$	0.728±0.015	**0.070±0.004**	**0.853±0.044**	**0.754±0.006**

$SUPPORT2$. Модель адаптивного бустинга со схемой *exp* показала лучшее качество на наборах $ROTT2$, $SUPPORT2$ и заняла второе место на наборе $WUHAN$, $SMARTO$.

Предложенные модели бутстреп ансамбля и адаптивного бустинга с перевыборкой достигли лучшего качества прогнозирования на наборах $GBSG$, $ROTT2$, $WUHAN$, $SMARTO$, $SUPPORT2$ и показали сравнимые результаты на наборе PBC по сравнению с существующими методами. Использование адаптивной бустинга позволяет достичь лучшего качества

Таблица 27: Сравнение качества моделей на наборе SMARTO

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
$KaplanMeier$	0.500±0.000	0.187±0.007	0.500±0.000	0.841±0.000
$CoxPH$	0.473±0.014	0.159±0.007	0.646±0.022	0.849±0.001
$LogLogisticAFT$	0.470±0.018	0.173±0.008	0.645±0.022	0.846±0.001
$LogNormalAFT$	0.452±0.020	0.180±0.009	0.629±0.026	0.840±0.001
$WeibullAFT$	0.502±0.007	0.159±0.007	0.661±0.019	0.848±0.001
ST	0.471±0.022	0.184±0.009	0.572±0.018	0.846±0.001
RSF	0.491±0.010	0.183±0.007	0.638±0.018	0.842±0.000
$CWGBSA$	0.514±0.009	0.179±0.007	0.641±0.024	0.843±0.001
$GBSA$	0.571±0.008	0.171±0.006	0.663±0.021	0.869±0.001
$DeepSurv$	0.499±0.004	0.158±0.007	0.653±0.020	0.873±0.001
$CoxTime$	0.503±0.003	0.174±0.008	0.642±0.018	0.873±0.004
$TREE_{KMWV}$	0.549±0.013	0.153±0.006	0.619±0.027	**0.890±0.002**
$Bootstrap$	**0.598±0.026**	0.140±0.005	**0.666±0.031**	**0.889±0.001**
$Boosting_{linear}$	0.591±0.022	**0.132±0.003**	0.656±0.025	0.856±0.002
$Boosting_{exp}$	**0.597±0.014**	**0.132±0.003**	**0.666±0.022**	0.871±0.001
$Boosting_{sigmoid}$	**0.601±0.019**	**0.140±0.003**	**0.670±0.021**	**0.890±0.001**

Таблица 28: Сравнение качества моделей на наборе SUPPORT2

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
$KaplanMeier$	0.500±0.000	0.178±0.001	0.500±0.000	0.278±0.000
$CoxPH$	0.786±0.003	0.126±0.002	0.862±0.005	0.464±0.002
$LogLogisticAFT$	0.789±0.003	0.109±0.002	0.868±0.005	0.469±0.002
$LogNormalAFT$	0.789±0.003	0.109±0.002	0.870±0.005	0.470±0.002
$WeibullAFT$	0.789±0.003	0.117±0.002	0.871±0.005	0.456±0.002
ST	0.749±0.015	0.124±0.007	0.844±0.017	0.455±0.023
RSF	0.791±0.003	0.129±0.001	0.881±0.004	0.359±0.002
$CWGBSA$	0.748±0.002	0.128±0.002	0.818±0.004	0.396±0.001
$GBSA$	0.803±0.003	0.099±0.002	0.893±0.004	0.520±0.002
$DeepSurv$	0.801±0.003	0.100±0.002	0.897±0.004	0.541±0.002
$CoxTime$	0.799±0.003	0.102±0.002	0.893±0.004	0.519±0.004
$TREE_{KMWV}$	0.802±0.003	0.105±0.002	0.891±0.004	**0.586±0.002**
$Bootstrap$	0.802±0.003	**0.098±0.001**	**0.899±0.003**	**0.560±0.003**
$Boosting_{linear}$	**0.806±0.003**	**0.095±0.001**	**0.902±0.004**	0.549±0.003
$Boosting_{exp}$	**0.805±0.003**	**0.096±0.001**	**0.901±0.003**	0.552±0.003
$Boosting_{sigmoid}$	**0.805±0.003**	0.098±0.001	0.898±0.003	**0.568±0.003**

на наборах $GBSG$, $ROTT2$, $SMARTO$, $SUPPORT2$.

В таблице 29 представлены средние ранги метрик качества прогнозирования моделей выживаемости по всем наборам данных. Ранжирование проводилось исходя из свойств метрик качества, а наилучшее качество определяется наименьшим рангом. Модель дерева выживаемости $TREE_{KMWV}$ достигла лучшего качества по метрике $AUPRC$. Модель бутстреп ансамбля $Bootstrap$ вошла в тройку лучших моделей по всем метрикам. Модель адаптивно-

Таблица 29: Сравнение средних рангов качества по всем наборам данных

Модель	**CI**	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
$KaplanMeier$	15.0	15.5	16.0	15.5
$CoxPH$	9.5	10.5	9.67	10.17
$LogLogisticAFT$	9.17	8.67	8.0	11.33
$LogNormalAFT$	10.67	11.17	10.33	11.33
$WeibullAFT$	9.67	9.5	8.33	11.67
ST	14.5	14.0	14.33	12.67
RSF	8.33	11.67	7.67	12.33
$CWGBSA$	10.5	11.67	10.83	12.83
$GBSA$	7.33	8.83	8.0	9.83
$DeepSurv$	10.17	8.5	8.33	6.17
$CoxTime$	9.17	9.17	8.67	6.17
$TREE_{KMWV}$	7.83	6.83	9.67	**1.0**
$Bootstrap$	**3.67**	**2.67**	**4.33**	**2.83**
$Boosting_{linear}$	3.83	**1.83**	**3.67**	5.5
$Boosting_{exp}$	**3.33**	**2.17**	**3.67**	4.5
$Boosting_{sigmoid}$	**3.33**	3.33	4.5	**2.17**

го бустинга со схемой *sigmoid* достигла лучшего среднего качества по метрике CI и заняла второе место по метрике $AUPRC$. Модель адаптивного бустинга со схемой *linear* показала лучшее среднее качество по метрикам IBS_{RM}, $IAUC_{WW,TI}$. Наконец, модель адаптивного бустинга со схемой *exp* показала лучшее качество по метрикам CI, IBS_{RM}, $IAUC_{WW,TI}$.

Таким образом, предложенные модели бутстреп ансамбля и адаптивного бустинга с перевыборкой превзошли по среднему рангу существующие методы. Модель бутстреп ансамбля демонстрирует лучшее качество по совокупности метрик, в то время как модель адаптивного бустинга позволяет достичь лучшего качества прогнозирования по всем метрикам при различных весовых схемах.

4.4 Выводы

В данной главе проводилось исследование и разработка новых методов ансамблирования древовидных моделей анализа выживаемости. Агрегация прогнозов нескольких моделей используется для повышения качества прогнозирования и снижения переобучения. По итогам проведенного исследования были достигнуты следующие результаты:

- Предложен метод построения бутстреп ансамбля независимых деревьев выживаемости, предложенных в главе 2. Предложена как реализация с определением размера ансамбля по первому локальному максимуму качества на OOB выборке, так и параллельная реализация с толерантным поиском размера по глобальному максимуму качества прогнозирования. Прогнозом модели является агрегация прогнозов базовых моделей.
- Предложен метод построения адаптивного бустинг ансамбля деревьев выживаемости с перевыборкой. Метод основан на идее итеративного построения ансамбля моделей, в котором каждая последующая базовая модель строится по выборке с наблюдениями, имеющими низкое качество прогноза на предыдущих итерациях ансамбля. Предложен

метод расчета ошибки и корректировки весов наблюдений, которые определяют вероятность попадания наблюдений в следующую обучающую выборку дерева выживаемости.

- По результатам экспериментального исследования, среди функций потерь наибольшее качество достигается на предложенной модификации IBS_{RM}. Функция потерь используется для определения размера ансамбля, а также для расчета ошибок прогнозов в модели адаптивного бустинга с перевыборкой. Предложенные методы позволили улучшить качество предложенных базовых моделей и превзойти существующие методы анализа выживаемости на всех наборах данных.

5 ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ОТКРЫТОЙ БИБЛИОТЕКИ АНАЛИЗА ВЫЖИВАЕМОСТИ

При работе (при подготовке) над данным разделом диссертации использованы следующие публикации автора, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования:

- *Васильев Ю. А. РАЗРАБОТКА БИБЛИОТЕКИ ДРЕВОВИДНЫХ МОДЕЛЕЙ АНАЛИЗА ВЫЖИВАЕМОСТИ // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. — 2024. — № 3. — С. 60–72.*

В предыдущих разделах данной диссертационной работы был представлен комплекс алгоритмов построения древовидных моделей анализа выживаемости. Метод построения деревьев выживаемости позволяет строить интерпретированные модели, позволяющие обрабатывать непрерывные и категориальные признаки, пропущенные значения, данные с различной формой распределения времени событий и случаи информативности цензурирования. Методы бутстреп и бустинг ансамблирования деревьев выживаемости позволяют повысить качество прогнозирования моделей, однако лишают моделей строгой интерпретации.

Данный раздел посвящен задачам разработки и реализации открытой библиотеки анализа выживаемости `survivors`, в основе работы которой лежат предложенные в данной работе алгоритмы. Решение поставленных на данном этапе задач состоит из следующих шагов:

- разработка архитектуры библиотеки;
- разработка программных модулей сбора данных, построения древовидных моделей и их ансамблей, оценки качества, проведения экспериментального исследования предложенных и внешних методов анализа выживаемости и визуализации результатов исследования;
- разработка сценариев функционирования библиотеки.

5.1 Обзор альтернативных программных реализаций

Для анализа событийных данных на языке Python, наибольшее распространение получили библиотеки Lifelines, PyCox, Scikit-survival.

Пакет Lifelines [113] включает реализации популярных статистических моделей анализа выживаемости. Непараметрические модели KaplanMeierFitter и NelsonAalenFitter (раздел 5.3.2) позволяют строить оценки функции выживания и риска без использования информации о признаковом пространстве. Полупараметрическая модель CoxPHFitter [114] (раздел 1.3.4) основывается на теоретическом предположении пропорциональности функций риска для различных наблюдений. Параметрическая модель Accelerated Failure Time (раздел 1.3.5) определяет линейную связь между логарифмом времени события и исходными признаками, а также предполагает теоретическое распределение ошибки. В частности, в библиотеке Lifelines реализованы модели WeibullAFTFitter, LogLogisticAFTFitter и LogNormalAFTFitter. Однако, как отмечалось в разделе 2.1, строгие теоретические предположения статистических моделей могут не выполняться на реальных данных.

Пакет PyCox [46] использует нейронные сети для повышения качества статистических моделей выживаемости (раздел 1.5.1). В частности, модели DeepSurv, PCHazard и CoxCC

заменяют линейную зависимость коэффициента масштабирования риска от признаков на отклик нейронной сети. Таким образом, модели остаются в рамках строгих предположений. С другой стороны, модель Cox-Time преодолевает предположение пропорциональности, добавляя коэффициенту масштабирования зависимость от времени. Однако, модель является дискретной и позволяет прогнозировать вероятность выживания только на заранее заданной временной шкале.

Пакет Scikit-survival [80] содержит множество моделей машинного обучения, наследуя интерфейс библиотеки Scikit-learn. Согласно сравнительным таблицам [80] полноты существующих библиотек (Scikit-survival, Lifelines, Statsmodels, PyCox), Scikit-survival содержит наиболее широкий функционал. В частности, в библиотеке реализованы древовидные методы. Модель Survival Tree (раздел 5.3.3) строит дерево выживаемости, используя классический критерий log-rank для поиска разбиения. Однако, как было отмечено в разделе 2.4.1, критерий log-rank имеет низкую чувствительность к особенностям реальных данных.

Для агрегации независимых деревьев выживания используется модель Random Survival Forest (раздел 1.6.1). Модель градиентного бустинга Gradient Boosting Survival Analysis (раздел 1.6.2) наследует предположение Кокса и строит ансамбль регрессионных деревьев решения. Наконец, модель Component-wise Gradient Boosting Survival Analysis (раздел 1.6.2) использует метод наименьших квадратов в качестве базовой модели и позволяет уточнить линейные коэффициенты в модели Кокса. Библиотека содержит большое количество вспомогательных функций для расчета метрик качества и загрузки данных.

Важно отметить, что описанные библиотеки работают только с заполненными данными и непрерывными признаками. В таком случае, требуется дополнительная предобработка данных (например, заполнение пропусков и обработка категориальных признаков). Также, открытые реализации имеют ограниченную применимость на реальных данных. Дискретные модели используют фиксированную временную шкалу, статистические модели основаны на строгих предположениях, а древовидные модели используют критерий log-rank.

5.2 Архитектура

В библиотеку `survivors` встроены существующие и предложенные статистические модели и модели машинного обучения. Исходный код библиотеки Survivors доступен на платформе GitHub[2].

Реализация основана на языке программирования Python 3.10 и поддерживается на операционных системах семейства Linux, macOS, и Windows. Программное обеспечение распространяется под лицензией BSD-3-Clause license[3]. BSD-3-Clause обеспечивает разработчикам и пользователям свободу, гибкость, простоту в использовании и распространении библиотеки, что способствует ее успешному прикладному применению.

[2] `https://github.com/iuliivasilev/dev-survivors`

[3] `https://opensource.org/license/bsd-3-clause/`

5.2.1 Требования к реализации

При проектировании библиотеки `survivors` преследовались следующие цели: широкая функциональность моделей, применимость к реальным данным, простота внедрения, экспериментальный уклон.

Широта функциональности моделей достигается за счет непрерывности прогнозируемых функций. Как говорилось ранее, дискретные модели анализа выживаемости фиксируют временную шкалу до обучения и рассчитывают вероятности события в рамках заданной шкалы. Непрерывные модели не зависят от шкалы и прогнозируют значения функции для любых моментов времени. Также, модели библиотеки должны прогнозировать как точечные величины (вероятность и время события), так и функции выживаемости и риска.

Библиотека предоставляет интерфейс для построения современных древовидных моделей анализа выживаемости: деревьев выживаемости и их ансамблей. Модели не требуют предобработки данных, способны обрабатывать категориальные признаки, пропущенные значения, случаи информативного цензурирования и мультимодальности распределения времени. Также, модели имеют повышенную чувствительность к особенностям данных: взвешенные критерии log-rank оценивают значимость разбиений при различном распределении времени, подходы регуляризации позволяют снизить вероятность переобучения. Таким образом, достигается применимость моделей к реальным данным.

Модели реализованы на основе NumPy [115], Numba [116], Pandas [117] и не требуют установки сложных зависимостей и тяжеловесных библиотек. Для ускорения вычислений на этапе поиска разбиения используется параллелизация процессов на CPU и JIT-компиляция (Just-In-Time). В `survivors` встроены исходные данные 9 открытых медицинских наборов данных, упрощая процесс ознакомления с функционалом на реальных примерах.

Отметим, что `survivors` поддерживает возможность проведения экспериментальных исследований. Модуль «Experiments Layer» включает множество стратегий обучения и валидации качества моделей с учетом классических метрик анализа выживаемости и их модификаций, представленных в разделе 3.1.

В отличии от существующих библиотек, `survivors` позволяет работать с реальными данными «из коробки», имеет повышенную чувствительность к особенностям данных (учитывает распределение времени и информативность цензурирования). Реализованные древовидные модели: деревья выживаемости и их ансамбли продемонстрировали лучшую прогнозную способность по сравнению с существующими реализациями (раздел 4.3.3). Экспериментальный модуль совместим с моделями библиотеки Scikit-survival и обладает интерфейсом для внедрения собственных моделей прогнозирования (в частности, для расширения функционала моделей библиотеки Lifelines).

5.2.2 Описание программных компонентов

На рисунке 46 представлена UML-диаграмма (Unified Modeling Language) [118] пакетов библиотеки `survivors`. Блоки отражают модули библиотеки, а пунктирные стрелки — зависимости между модулями (в частности, импортирование модуля). Для удобства восприятия архитектура представлена в виде иерархической структуры, в которой более низкие пакеты импортируются более высокими.

Для модулей «Experiments Layer», «Ensemble Layer», «Data Layer», «Tree Layer», «External Layer» отображена внутренняя структура дочерних пакетов. При этом, пользователем могут использоваться как все разработанные модули, так и отдельные из них.

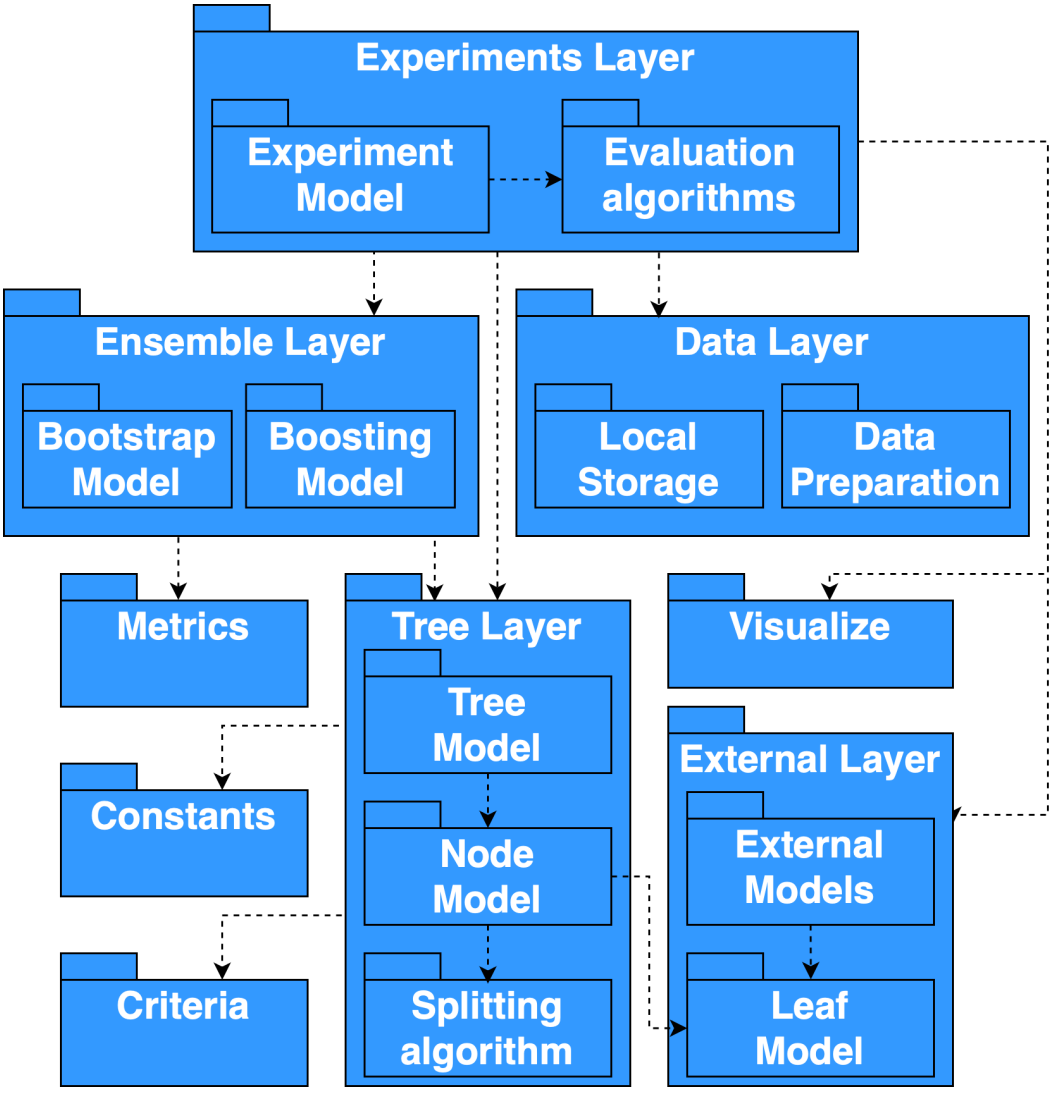


Рис. 46: Диаграмма пакетов библиотеки `survivors` на универсальном языке моделирования (Unified Modeling Language, UML).

Реализованная библиотека содержит следующие модули.

1. *Модуль внутреннего представления* ***Constants***.
   Модуль содержит константы (имена целевых переменных и словари отображения входных параметров) и функции (разбиение входных данных по типу переменных, построение целевой переменной, бинаризация временной переменной), необходимые для работы с внутренним представлением данных и прогнозными моделями.
   Используются библиотеки: Os, NumPy, объем — 200 строк.
2. *Модуль метрик качества* ***Metrics***.
   Модуль содержит реализации классических метрик анализа выживаемости Concordance Index, Likelihood, Integrated Brier Score, Integrated AUC, AUPRC, а также их модификации (например, IBS_WW, IBS_REMAIN, IBS_BAL), разработан-

ные при исследовании случаев избыточной чувствительности метрик.
Используются библиотеки: Pandas, Lifelines, объем — 300 строк.
Для параллелизации и векторизации вычислений использовались библиотеки Numba и Joblib.

3. *Модуль статистических критериев* ***Criteria***.
Модуль содержит векторизованные реализации статистических критериев: log-rank, wilcoxon, tarone-ware, peto-peto и др. Ускорение вычислений осуществляется за счет использования JIT-компиляции исходного кода в байт-код средствами библиотеки Numba [116].
Используются библиотеки: NumPy, Numba, объем — 600 строк.
4. *Модуль визуализации* ***Visualize***.
Модуль содержит функции демонстрации результатов работы алгоритмов в удобном для пользователя формате. Выделяются следующие группы функций: визуализация прогнозов анализа выживаемости, анализ поведения метрик качества, визуализация характеристик данных.
Используются библиотеки: Pandas, NumPy, Matplotlib, Seaborn, объем — 1000 строк.
5. *Модуль внешних моделей* ***External Layer***.
Модуль содержит реализации существующих моделей анализа выживаемости с унифицированным интерфейсом LeafModel. Модели с интерфейсом LeafModel могут использоваться как для проведения экспериментов, так и для описания листовой выборки в дереве выживаемости. В рамках модуля реализованы непараметрические модели Каплана–Мейера, Нельсона–Аалена и их модификации, а также обертки для моделей библиотеки Lifelines: KaplanMeierFitter, NelsonAalenFitter, WeibullAFT, LogNormalAFT, LogLogisticAFT.
Используются библиотеки: NumPy, Lifelines, объем — 400 строк.
6. *Модуль построения дерева выживаемости* ***Tree Layer***.
Модуль содержит описание основных сущностей (класс вершины Node, класс дерева выживаемости CRAID), реализацию предложенных методов выбора лучшего разбиения выборки и построения древовидных моделей. Класс Node отвечает за отдельные структурные элементы дерева выживаемости (используя класс листовой модели LeafModel), а также за поиск оптимального правила разбиения выборки в узле. Класс CRAID определяет структуру дерева выживаемости (с контролем роста дерева на основе гиперпараметров), а также содержит методы обрезки дерева и агрегации прогнозов листовых моделей.
Используются библиотеки: Pandas, Scikit-learn, Numba и Joblib, объем — 1800 строк.
7. *Модуль построения ансамблей деревьев решений* ***Ensemble Layer***.
Модуль содержит базовый класс ансамблирования деревьев выживаемости BaseEnsemble, а также реализацию интерфейса в виде модели ансамбля независимых деревьев выживаемости BootstrapCRAID и модели адаптивного бустинга с перевыборкой BoostingCRAID. Модели содержат методы расчета ошибки ансамбля, агрегации откликов базовых моделей, подбора лучшего размера ансамбля.
Используются библиотеки: Pandas, Scikit-learn, объем — 1000 строк.
8. *Модуль сбора наборов данных* ***Data Layer***.

Модуль предназначен для загрузки и предобработки 9 открытых наборов данных: GBSG [17], PBC [96], WUHAN [19], ACTG [119], FLCHAIN [120], SMARTO [15], ROTT2 [18], SUPPORT2 [43], FRAMINGHAM [121]. На этапе предобработки исходные данные приводятся к унифицированной структуре: X (пространство признаков), y (упорядоченный массив с двумя целевыми переменными), features (исходные названия признаков наблюдений), categ (подмножество категориальных признаков).
Используются библиотеки: Pandas, Scikit-learn, Re, Math, Lxml, объем — 1000 строк.

9. *Модуль проведения экспериментов* ***Experiments Layer***.
Модуль содержит класс Experiments, для постановки и запуска экспериментов с различными характеристиками. Класс предоставляет гибкий интерфейс для работы с встроенными и внешними моделями выживаемости, их гиперпараметрами, а также стратегиями проведения экспериментов [122]: валидация на отложенной выборке (hold-out), кросс-валидация (CV), Grid-Search кросс-валидация с семплингом (CV+sampling), кросс-валидация во времени (Time-Aware Cross-Validation [123]). Средства модуля включают построение сводных таблиц результатов и визуализации качества на основе точечных графиков и диаграмм описания распределений.
Используются библиотеки: Pandas, Scikit-survival, Lifelines, объем — 2000 строк.

5.3 Сценарии использования

В данной главе рассматриваются примеры использования библиотеки survivors для работы с реальными данными анализа выживаемости. Библиотека содержит 9 встроенных открытых наборов данных, для которых определен интерфейс загрузки и предобработки данных (раздел 5.3.1).

На основе загруженных данных могут быть построены непараметрические модели (раздел 5.3.2), использующие только информацию о целевых переменных: времени и флаге цензурирования. Для построения индивидуальных прогнозов survivors предоставляет комплекс древовидных моделей с повышенной чувствительностью к реальным данным. Деревья выживаемости (раздел 5.3.3) позволяют строить модели со строгой интерпретацией зависимостей. Ансамбли деревьев выживаемости (раздел 5.3.4) повышают качество прогнозирования за счет построения нескольких базовых моделей на разных подмножествах данных.

Также, библиотека предоставляет интерфейс для оценки качества прогнозирования встроенных моделей выживаемости (раздел 5.3.5). Библиотека содержит как классические метрики качества анализа выживаемости, так и модификации, предложенные в главе 3.

5.3.1 Подготовка данных

Для удобства работы с реальными данными, в библиотеку встроены функции загрузки и обработки 9 медицинских наборов данных. Встроенные наборы имеют различные характеристики типа события, количества наблюдений, количества признаков, дисбаланса цензурирования, заполненности данных и распределения времени событий. Каждому набору сопоставляется функция загрузки вида `load_<NAME>_dataset` из модуля «datasets».

```
1   import survivors.constants as cnt
2   import survivors.datasets as ds
3   import survivors.visualize as vis
4
5   # Загрузка данных
6   X, y, features, categ, _ = ds.load_pbc_dataset()
7
8   # Формирование временной шкалы
9   bins = cnt.get_bins(time=y[cnt.TIME_NAME], cens=y[cnt.CENS_NAME])
10  bins_short = [10, 100, 1000, 2000, 3000]
```

Рис. 47: Пример программного кода для работы с встроенными данными библиотеки `survivors`. Представлены этапы загрузки модулей (строки 1-3), загрузки набора данных PBC (строка 6), и два подхода формирования временной шкалы (полной шкалы в строке 9, пользовательской шкалы в строке 10).

На рисунке 47 представлен фрагмент программного кода для работы с набором данных PBC (раздел 2.1). Строки 1-3 содержат импорт модулей «constants», «datasets», «visualize» библиотеки `survivors`. В частности, из «constants» используются внутренние названия целевых переменных: время события `TIME_NAME` и индикатор цензурирования `CENS_NAME`.

Строка 6 описывает интерфейс загрузки набора PBC. Функция `load_pbc_dataset` преобразует исходные данные набора PBC к кортежу значений `X, y, features, categ, sch_nan`. Значение `X` представляет собой признаковое пространство наблюдений в виде Pandas DataFrame. Значение `y` определяет целевые переменные наблюдения в виде именнованного массива NumPy. Список `features` содержит названия признаков, а `categ` — подмножество категориальных признаков. По результатам загрузки, набор содержит 418 наблюдений с 17 признаками (категориальными являются 5 признаков: trt, sex, ascites, hepato, spiders), из которых 12 могут содержать пропуски.

Библиотека `survivors` обеспечивает возможность работы с непрерывными функциями анализа выживаемости. Для работы с конкретными значениями функций необходимо определить временную шкалу, в рамках которой фиксируется наступление событий. В строках 9 и 10 представлены 2 способа определения временной шкалы. Встроенная функция `get_bins` возвращает «полную временную шкалу» — упорядоченное множество моментов между минимальным и максимальным временем наступления события с шагом 1 (в данном случае, равной `[41, 42, ... 4190, 4191]` с размером 4151). Также, поддерживается и «пользовательская временная шкала», состоящая, например, из 5 моментов времени: `[10, 100, 1000, 2000, 3000]`.

Полученный формат данных и временной шкалы применим для дальнейшего построения непараметрических, древовидных и ансамблевых моделей, а также для оценки качества моделей относительно каждой временной шкалы.

5.3.2 Построение непараметрических моделей

В библиотеку встроены собственные реализации непараметрических оценок Каплана–Мейера, Нельсона–Аалена [56], а также предложенная модификация $KWMW$ модели

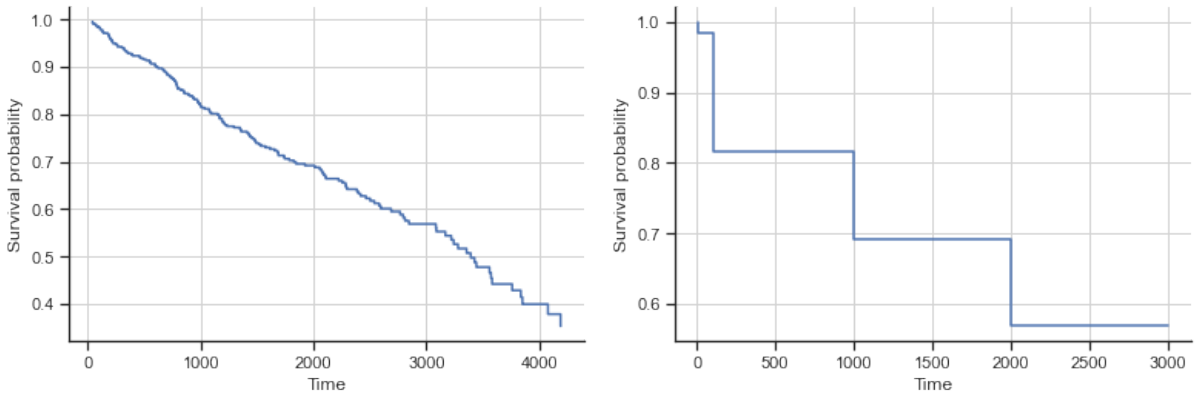
Каплана–Мейера с генерацией виртуальных событий (раздел 2.4.4). Для увеличения производительности, при построении модели используются векторные вычисления библиотеки NumPy.

```
1   from survivors.external import KaplanMeier
2
3   # Построение непараметрической модели Каплана-Мейера
4   km = KaplanMeier()
5   km.fit(durations=y[cnt.TIME_NAME], right_censor=y[cnt.CENS_NAME])
6
7   # Прогнозирование функции выживания для всех моментов времени
8   sf_km = km.survival_function_at_times(times=bins)
9   vis.plot_survival_function(sf_km, bins)
10
11  # Прогнозирование функции выживания для частных моментов времени
12  sf_km_short = km.survival_function_at_times(times=bins_short)
13  vis.plot_survival_function(sf_km_short, bins_short)
```

Рис. 48: Пример программного кода для работы с моделью Каплана–Мейера. В примере представлены этапы инициализации и построения непараметрической оценки (строки 4-5), прогнозирования и визуализации функции выживания для полной (строки 8-9) и пользовательской (строки 12-13) временной шкалы.

На рисунке 48 представлен фрагмент программного кода для построения непараметрической оценки функции выживания. Встроенная реализация метода Каплана–Мейера импортируется из модуля `tree.external` в строке 1. В строках 4-5 представлен интерфейс инициализации и построения модели Каплана–Мейера. Модель использует информацию о времени наступления событий `durations` (соответствует целевой переменной `y[cnt.TIME_NAME]`) и флаге цензурирования (соответствует целевой переменной `y[cnt.CENS_NAME]`).

Обученная модель содержит внутреннее представление непрерывной функции выживания. Для демонстрации гибкости модели к временной шкале, в строках 8 и 12 формируются два прогноза функции выживания: `sf_km` для шкалы `bins` (размером 4151), `sf_km_short` для шкалы `bins_short` (размером 5). С помощью функции `plot_survival_function` модуля «visualize» (строки 9 и 13) визуализируется прогноз функции выживания.



(a) Оценка функции для шкалы `bins` (b) Оценка функции для шкалы `bins_short`

Рис. 49: Пример двух оценок функции выживания по непараметрическому методу Каплана–Мейера (результат представлен в примере 48). Оценка может быть получена как для полной временной шкалы (рисунок a), так и для пользовательских моментов времени (рисунок b).

На рисунке 49 представлены результаты визуализации двух прогнозов функции выживания. Горизонтальная ось соответствует временной шкале, а вертикальная ось отражает вероятность выживаемости для каждого момента времени. На рисунке a изображена функция выживания `sf_km`, на рисунке b — функция выживания `sf_km_short`. Прогноз функции по полной шкале позволяет оценить вероятность выживания для всех моментов времени, однако может содержать избыточную информацию для пользователя.

5.3.3 Построение деревьев выживаемости и интерпретация зависимостей

В библиотеку встроен предложенный метод построения деревьев выживаемости *CRAID* (раздел 2.3), поддерживающий использование взвешенных log-rank критериев для поиска лучшего разбиения выборки, методов обрезки pre-pruning (раздел 2.3.2) и post-pruning (раздел 2.3.3), а также предложенных подходов регуляризации дерева (раздел 2.4.3) и модификации листовых оценок (раздел 2.4.4).

На рисунке 50 представлен фрагмент программного кода для работы с моделью дерева выживаемости CRAID модуля «tree». Строки 4-5 описывают процесс инициализации модели с указанием гиперпараметров для ограничения роста дерева: критерий разбиения `criterion`, глубина дерева `depth`, минимальный размер листа `min_samples_leaf`, уровень значимости разбиения (p-value) `signif`, список категориальных признаков `categ`. В строке 6 инициализированная модель обучается на признаковом пространстве `X` и целевых переменных `y`.

В строке 9 структура дерева визуализируется с помощью метода `visualize` (результат визуализации представлен на рисунке 13). Структура состоит из вершин, отражающих функцию выживания в узле и характеристики выборки (размер, доля терминальных событий и среднее время), и ребер (правил разбиения выборки).

На основе метода `predict_at_times`, в строках 12-14 вычисляется прогноз функции выживания `sf_cr` и риска `chf_cr`. В методе указываются параметры: признаки наблюдений, временная шкала и тип прогноза («surv» для функции выживания, «hazard» для функции риска). Откликом модели является NumPy массив индивидуальных прогнозов функции для всех моментов времени шкалы. В частности, оба прогноза имеют размерность (418, 4151).

```
1  from survivors.tree import CRAID
2
3  # Инициализация и обучение дерева выживаемости
4  cr = CRAID(criterion="logrank", depth=2,
5             min_samples_leaf=0.1, signif=0.05, categ=categ)
6  cr.fit(X, y)
7
8  # Визуализация структуры дерева
9  cr.visualize(target=cnt.TIME_NAME, mode="surv")
10
11 # Прогноз функции выживания для всех наблюдений выборки
12 sf_cr = cr.predict_at_times(X, bins=bins, mode="surv")
13 # Прогноз функции риска
14 chf_cr = cr.predict_at_times(X, bins=bins, mode="hazard")
15
16 # Индивидуальный прогноз для 0 наблюдения
17 print(y[0])  # Целевые переменные: (True, 400.)
18 # Прогноз времени события
19 pred_time = cr.predict(X, target=cnt.TIME_NAME)
20 print(pred_time[0])  # Ожидаемое время: 847.43
21 # Прогноз вероятности события
22 pred_prob = cr.predict(X, target=cnt.CENS_NAME)
23 print(pred_prob[0])  # Вероятность события: 0.9272
24 # Прогноз глубины листа
25 pred_depth = cr.predict(X, target="depth")
26 print(pred_depth[0])  # Глубина листа: 2
```

Рис. 50: Пример программного кода для работы с моделью дерева выживаемости CRAID. Представлены этапы инициализации и обучения модели (строки 4-6), визуализации структуры (строка 9), прогнозирования функций выживания и риска (строки 12-14), а также времени и вероятности события, глубины листа для 0 наблюдения (строки 17-26).

Наконец, в строках 17-26 демонстрируется процесс получения индивидуальных прогнозов для нулевого наблюдения с помощью метода `predict`. Терминальное событие наступило на 400 день исследования. Получены прогнозы: ожидаемое время события равно 847.43 (вызов с параметром `target=cnt.TIME_NAME`), вероятность события равна 0.9272 (вызов с параметром `target=cnt.CENS_NAME`), наблюдение находится в листе глубиной 2 (вызов с параметром `target="depth"`).

5.3.4 Построение ансамблей деревьев выживаемости

В библиотеку встроены предложенные методы ансамблирования деревьев выживаемости $CRAID$: бутстреп ансамбль независимых деревьев $BootstrapCRAID$ (раздел 4.1) и адаптивный бустинг с перевыборкой $BoostingCRAID$ (раздел 4.2).

На рисунке 51 представлен фрагмент программного кода построения модели BootstrapCRAID. В строках 4-7 модель инициализируется с гиперпараметрами: размер ансамбля `n_estimators`, размер бутстреп подмножества `size_sample`, максимальная доля признаков при поиске разбиения `max_features` и гиперпараметры базовых моделей (описаны в разделе 5.3.3). Интерфейс обучения ансамбля (строка 9) и построения прогноза `sf_bstr` функции выживаемости (строка 11) аналогичен модели CRAID.

```
1   from survivors.ensemble import BootstrapCRAID
2
3   # Инициализация ансамбля деревьев выживаемости
4   bstr = BootstrapCRAID(n_estimators=10, size_sample=0.7, depth=3,
5                         max_features=0.3, criterion="peto",
6                         min_samples_leaf=0.01, categ=categ)
7
8   # Обучение ансамбля
9   bstr.fit(X, y)
10  # Прогноз функции выживания для всех наблюдений выборки
11  sf_bstr = bstr.predict_at_times(X, bins=bins, mode="surv")
```

Рис. 51: Пример программного кода для работы с ансамблем независимых деревьев выживаемости `BootstrapCRAID`. Представлены следующие этапы: инициализация модели (строки 4-6) с фиксированными гиперпараметрами, обучение модели (строка 10), прогнозирование функции выживания (строка 11).

```
1   from survivors.ensemble import BoostingCRAID
2
3   # Создание дерева выживаемости
4   boost = BoostingCRAID(mode_wei="linear", n_estimators=10, depth=3,
5                         size_sample=0.5, ens_metric_name="IBS_REMAIN",
6                         max_features=0.3, criterion="peto",
7                         min_samples_leaf=0.01, categ=categ)
8
9   # Обучение дерева
10  boost.fit(X, y)
11  # Прогноз функции выживания для всех наблюдений выборки X
12  sf_boost = boost.predict_at_times(X, bins=bins, mode="surv")
```

Рис. 52: Пример программного кода для работы с адаптивным бустингом деревьев выживаемости с перевыборкой `BoostingCRAID`. Этапы повторяют фрагмент кода 51.

Аналогично, на рисунке 52 представлен фрагмент программного кода построения модели BoostingCRAID. В строках 4-7 модель инициализируется с гиперпараметрами: весовая схема `mode_wei`, размер ансамбля `n_estimators`, размер бутстреп подмножества `size_sample`, метрика расчета весов наблюдений `ens_metric_name`, максимальный размер подмножества признаков при поиске разбиения `max_features` и гиперпараметры базовых моделей (описаны в разделе 5.3.3). Интерфейс обучения ансамбля (строка 10) и построения прогноза `sf_boost` функции выживаемости (строка 12) аналогичен модели CRAID.

5.3.5 Оценка качества прогнозирования

Также, библиотека содержит реализации существующих (раздел 1.2) и предложенных (раздел 3.1.5) методов оценки качества прогнозирования величин анализа выживаемости. Модуль «metrics» содержит 5 семейств метрик: Concordance Index (CI), Likelihood (LL), Integrated brier score (IBS), Integrated Area Under roc-Curve (IAUC), Area Under Precision Recall Curve (AUPRC).

По результатам проведенных сценариев 5.3.2-5.3.4, были получены следующие прогнозы функции выживания: непараметрической модели Каплана–Мейера (`sf_km`), дерева выживаемости (`sf_cr`), бутстреп ансамбля (`sf_bstr`), адаптивного ансамбля с перевыборкой (`sf_boost`).

```
1   # Усредненный интегральный brier score
2   mean_ibs_rm = metr.ibs_remain(y, y, sf_bstr, bins, axis=-1)
3   print(mean_ibs_rm)  # 0.116
4   # Интегральный brier score для отдельных наблюдений
5   ibs_rm_by_obs = metr.ibs_remain(y, y, sf_bstr, bins, axis=0)
6   print(ibs_rm_by_obs)  # [0.0237, 0.0379, ..., 0.0000, 0.0008]
7   # Усредненный brier score во времени
8   ibs_rm_by_time = metr.ibs_remain(y, y, sf_bstr, bins, axis=1)
9   print(ibs_rm_by_time)  # [0.0000, 0.0042, ..., 0.1006, 0.0857]
10
11  # Прогноз функции выживания по трем моделям
12  vis.plot_func_comparison(y[0],
13        [sf_km, sf_cr[0], sf_bstr[0], sf_boost[0]],
14        ["KM", "CRAID", "BootstrapCRAID", "BoostingCRAID"])
15  # Изменение IBS_RM во времени
16  vis.plot_metric_comparison(y[0],
17        [sf_km, sf_cr[0], sf_bstr[0], sf_boost[0]],
18        ["KM", "CRAID", "BootstrapCRAID", "BoostingCRAID"],
19        bins, metr.ibs_remain)
20  # Изменение auprc во времени
21  vis.plot_metric_comparison(y[0],
22        [sf_km, sf_cr[0], sf_bstr[0], sf_boost[0]],
23        ["KM", "CRAID", "BootstrapCRAID", "BoostingCRAID"],
24        bins, metr.auprc)
```

Рис. 53: Пример программного кода для оценки качества прогнозирования. В строках 1-9 проводится оценка качества по выборке, в строках 11-24 — оценка индивидуального прогноза для нулевого наблюдения.

На рисунке 53 представлен фрагмент программного кода для оценки качества прогнозирования моделей. Строки 1-9 содержат пример вычисления метрики качества IBS_{RM} (формула (30)) для всех наблюдений выборки X. Параметрами функции `ibs_remain` служат: целевые переменные обучающей и тестовой выборок, прогноз функции выживания, временная шкала, параметр оси (опционален). В данном случае оценивается качество прогноза модели BootstrapCRAID на обучающей выборке по временной шкале `bins`.

Подробнее рассмотрим параметр `axis` функции `ibs_remain`. Метрики IBS и AUPRC представимы в интегральном виде, а также могут быть стратифицированы относительно наблюдений и моментов времени. При `axis=-1` ответом функции будет скалярное значение усредненного IBS_{RM} для всех наблюдений (размерность 1). При `axis=0` вычисляется ошибка для отдельных наблюдений набора PBC (рассматривая индивидуальную функцию выживания) с размерностью 418 (количество наблюдений). При `axis=1` вычисляется ошибка для всех моментов временной шкалы `bins` (размерность 4151). Таким образом, может быть оценена

как интегральная ошибка по выборке, так и зависимость ошибки от отдельных наблюдений и моментов времени.

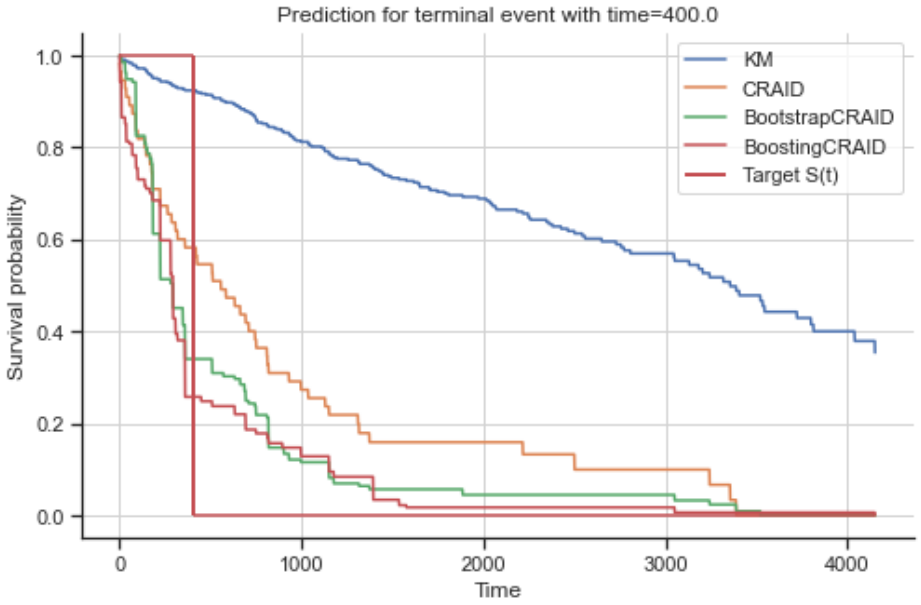


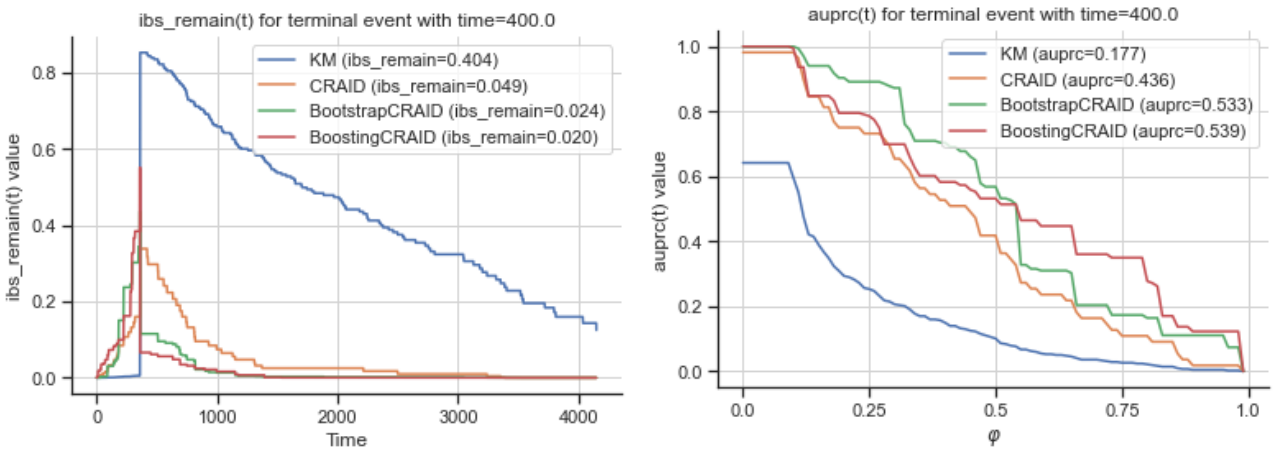
Рис. 54: Визуализация индивидуальных прогнозов функции выживания для 0 наблюдения. На горизонтальной оси отмечена временная шкала, на вертикальной оси — вероятность выживания. Красный график отражает истинную функцию выживания для терминального события, наступившего в момент времени 400. Прогнозы получены с помощью оценки Каплана–Мейера, дерева выживаемости, бутстреп и бустинг ансамблей.

В строках 11–24 оценивается качество прогноза для нулевого наблюдения набора PBC. Для визуализации прогнозов разных моделей, в строках 12-14 используется функция `plot_func_comparison` (результат представлен на рисунке 54). В параметрах функции указывается целевое наблюдение, список индивидуальных прогнозов и список названий моделей. Красная линия отражает эталонную функцию выживания, синяя линия — прогноз модели Каплана–Мейера, оранжевая линия — прогноз модели CRAID, зеленая линия — прогноз модели BootstrapCRAID, розовая линия – прогноз модели BoostingCRAID.

В строках 16-24 визуализируется изменение качества IBS_RM и AURPC в зависимости от временной шкалы (с помощью функции `plot_metric_comparison`). Функция имеет следующие параметры: целевое наблюдение, список индивидуальных прогнозов, список названий моделей, временная шкала и функция метрики качества.

На рисунке 55 представлен результат визуализации изменения метрик. На рисунке a оценивается метрика IBS_RM, цвет линий аналогичен рисунку 54. В легенде графика отмечены интегральные значения метрик. Прогноз ансамбля BoostingCRAID обладает лучшим качеством, а прогноз метода Каплана–Мейера — худшим. Стоит отметить, что наибольшая ошибка прогнозирования достигается в окрестности момента наступления события для всех прогнозов.

На рисунке b оценивается метрика AUPRC, обозначения аналогичны рисунку a. По горизонтальной оси отмечено значение параметра $\varphi \in [0, 1]$. Таким образом, при приближении параметра к значению 1, оценивается качество в окрестности момента наступления события,



(a) Изменение метрики `IBS_RM` во времени

(b) Изменение метрики `auprc` во времени

Рис. 55: Визуализация качества прогнозирования функции выживания по метрикам `IBS_RM` (рисунок a) и `auprc` (рисунок b) во времени. Прогнозы получены с помощью оценки Каплана–Мейера, дерева выживаемости, бутстреп и бустинг ансамблей.

а при приближении к значению 0 — качество в ранних и поздних моментах времени. Прогноз ансамбля BoostingCRAID обладает лучшим качеством, а прогноз метода Каплана–Мейера — худшим. Стоит отметить, что качество AUPRC понижается по мере приближения к моменту события.

5.4 Оценка производительности

Для оценки производительности разработанной библиотеки была проведена серия экспериментальных исследований. Рассматривалось среднее время выполнения обучения и прогнозирования предложенных моделей, а также потребляемый объем оперативной памяти при использовании библиотеки *survivors*. Оценка производительности проводилась на 6 наборах данных: $GBSG$, PBC, $ROTT2$, $WUHAN$, $SMARTO$, $SUPPORT$, рассматривались 3 модели прогнозирования: дерево выживаемости $CRAID$, бутстреп ансамбль $BootstrapCRAID$, адаптивный бустинг с перевыборкой $BoostingCRAID$. Постановка экспериментов аналогична третьему этапу постановки экспериментов при сравнения качества моделей прогнозирования (раздел 4.3.3) с фиксированием лучших гиперпараметров, отобранных на втором этапе.

Эксперименты проходили на рабочей станции со следующими программными характеристиками: ОС Windows Server 2016 Standard (64bit), Python 3.10.4; и аппаратными характеристиками: процессор Intel(R) Core(TM) i7-10700, частота 2.9 ГГц, 8 ядер; оперативная память 128 Гбайт; дисковый накопитель HDD, 697 Гбайт.

Результаты оценки среднего времени работы моделей представлены на рисунке 56. Для каждого набора данных отражено среднее время работы по трём предложенным моделям. Показатели времени отображаются на логарифмической шкале, а для каждого метода указано среднее время работы. Исходя из результатов оценки, построение и применение деревьев выживаемости не превышает 1.8 секунд (в худшем случае, на наборе данных WUHAN), бутстреп ансамбля – 43 секунд (в худшем случае, на наборе данных SUPPORT2), адаптивного ансамбля с перевыборкой – 104 секунд (в худшем случае, на наборе данных SUPPORT2).

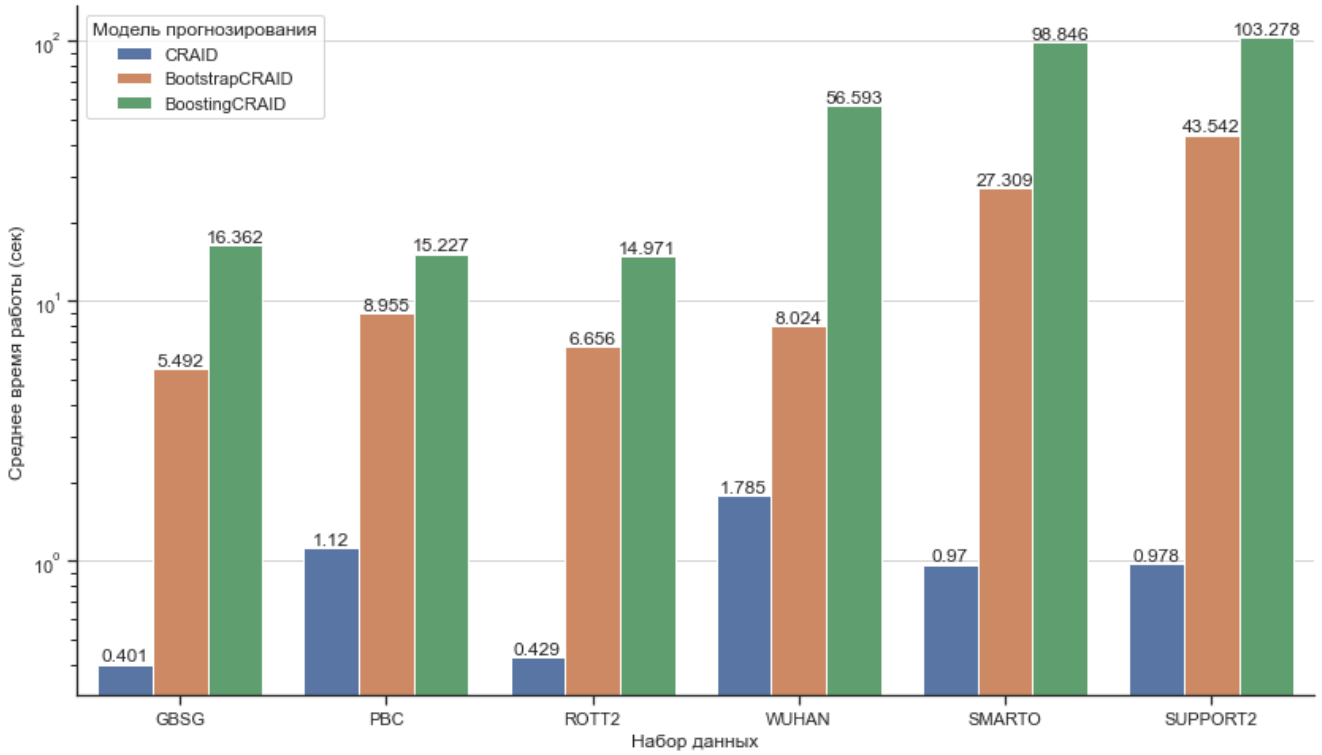


Рис. 56: Показатели среднего времени (сек) обучения и прогнозирования для предложенных моделей

Стоит отметить, что скорость работы зависит от количества наблюдений и признаков в наборе данных.

Результаты оценки средней потребляемой памяти (МБайт) во время работы моделей представлены на рисунке 57. Обозначения аналогичны рисунку 56, а для каждого метода указан средний объем потребляемой памяти. Исходя из результатов оценки, затрачиваемая память при использовании дерева выживаемости не превышает 1.5 МБайт. Для бутстреп ансамбля используемая память варьируется от 38 до 164 МБайт (в худшем случае, на наборе данных SUPPORT2), а для адаптивного бустинга с перевыборкой от 57 по 325 МБайт (в худшем случае, на наборе данных SUPPORT2). Стоит отметить, что потребляемый объем памяти напрямую зависит от объема набора данных, а также от количества уникальных значений и величины временной шкалы.

5.5 Выводы

В данном разделе была проведена разработка библиотеки анализа выживаемости, в основе работы которой лежат предложенные в данной работе алгоритмы. Были достигнуты следующие результаты:

- Разработана программная библиотека анализа выживаемости с открытым исходным кодом, использующая предложенный комплекс алгоритмов. Модели разработанной библиотеки работают с реальными данными анализа выживаемости, включая непрерывные и категориальные признаками, пропущенные значения, различные формы распределений времени событий и случаи информативности цензурирования. Реализованные

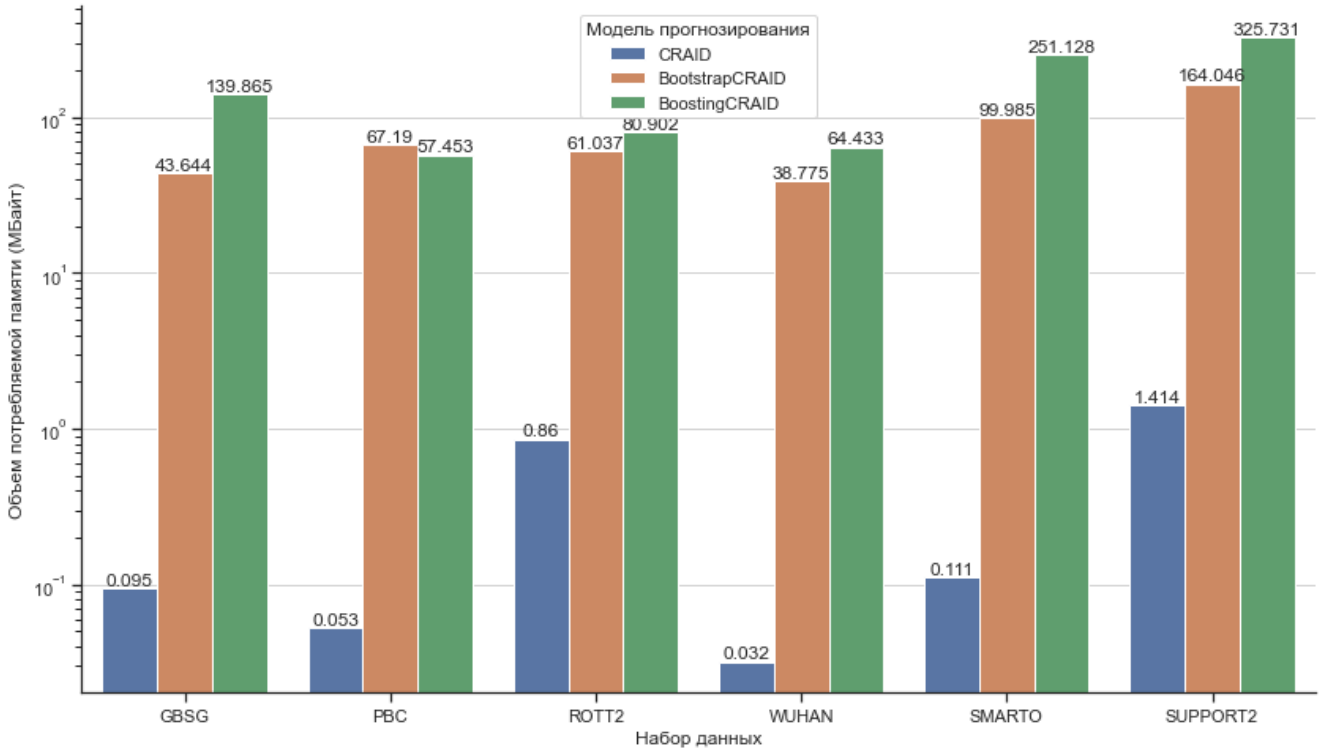


Рис. 57: Показатели потребляемой памяти (МБайт) предложенных моделей

древовидные методы основаны на предложенном гистограммном методе поиска разбиений в цензурированных данных. Алгоритмы используют векторизацию и параллелизацию вычислений для увеличения производительности.

- Представлены базовые сценарии функционирования библиотеки:
  - Сбор и предобработка открытых медицинских данных;
  - Построение статистических моделей анализа выживаемости;
  - Построение предложенных древовидных моделей анализа выживаемости, включая модель дерева выживаемости CRAID, бутстреп ансамбль независимых деревьев BootstrapCRAID и адаптивный бустинг с перевыборкой BoostingCRAID;
  - Оценка качества прогнозирования построенных моделей.
- Проведена экспериментальная оценка времени работы и потребляемой памяти предложенной программной реализации. Полученные оценки производительности свидетельствует о возможности применения разработанной библиотеки анализа выживаемости на практике.

ЗАКЛЮЧЕНИЕ

Основные результаты диссертационной работы заключаются в следующем:

1. Предложен метод построения деревьев выживаемости, основанный на алгоритме поиска лучшего разбиения с учетом взвешенных регуляризованных log-rank критериев. Взвешенные критерии позволяют придавать разный приоритет ранним и поздним событиям, повышая чувствительность к распределению вероятностей времени наступления событий. Регуляризация критерия разбиения позволяет учитывать информацию об априорном распределении времени наступления событий. Метод способен обрабатывать категориальные признаки и пропуски в данных, а также применим к случаям информативного цензурирования.
2. Предложены методы ансамблирования деревьев выживаемости. Бутстреп метод основан на построении ансамбля независимых деревьев на бустреп-выборках с выбором размера по минимальной ошибке вне бустреп-выборки. Модель бустинга основана на построении адаптивного ансамбля деревьев с перевыборкой, в котором каждая последующая модель обучается на наиболее сложных для прогнозирования ансамблем наблюдениях. Выбор функции потерь проводился на основе исследования чувствительности метрик качества к особенностям данных.
3. Реализована открытая программная библиотека анализа выживаемости, использующая предложенный комплекс алгоритмов. Библиотека позволяет проводить сценарии сбора и предобработка данных, построения и применения моделей выживаемости, оценки качества прогнозирования. По результатам проведенных на основе библиотеки экспериментальных исследований на медицинских данных, разработанные методы превзошли по качеству прогнозирования существующие методы анализа выживаемости.

Данные результаты опубликованы в 4 печатных работах [9–12]. Разработанная программная библиотека методов построения моделей анализа выживаемости прошла апробацию в рамках НИР «Выполнение части работ по развитию прикладного программного обеспечения государственной информационной системы обязательного медицинского страхования», 2021–2022 гг., НИР «Выполнение работ в области разработки и внедрения методов искусственного интеллекта и анализа больших данных в сфере здравоохранения», 2020–2021 гг. и НИР «Исследование, разработка и применение инновационных технологий построения интеллектуальных программных систем» 2018–2027 гг. Полученные результаты могут послужить основой для построения перспективных современных систем анализа событийных данных, которые будут включать в себя модели анализа выживаемости. При этом, могут использоваться как все разработанные модули, так и отдельные из них.

Список литературы

[1] Vasilev Iulii, Petrovskiy Mikhail, Mashechkin Igor. Survival Analysis Algorithms based on Decision Trees with Weighted Log-rank Criteria // In Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods. — 2022. — P. 132–140.

[2] Васильев Ю. А., Петровский М. И., Машечкин И. В. Новые алгоритмы анализа выживаемости на основе деревьев решений с взвешенными logrank критериями // Тихоновские чтения: научная конференция: 25–30 октября 2021 г. : тезисы докладов. — Москва : ООО "МАКС Пресс". — 2021. — Т. 46. — С. 90–90.

[3] Васильев Юлий Алексеевич. Исследование и разработка древовидных моделей для задачи анализа выживаемости // Материалы Международного молодежного научного форума «Ломоносов-2022» / под ред. Алешковский Иван Андреевич, Андриянов Андрей Владимирович, Антипов Евгений Александрович, Зимакова Екатерина Игоревна. — Москва : ООО "МАКС Пресс". — 2022. — Т. 39 из Гидрометеорология.

[4] Машечкин И. В., Петровский М. И., Васильев Ю. А. Исследование и разработка нелинейных моделей выживаемости на основе деревьев решений и их ансамблей // Ломоносовские чтения-2022: научная конференция, факультет ВМК МГУ имени М.В.Ломоносова. Тезисы докладов. — Москва : ООО "МАКС Пресс". — 2022. — Т. 2022 из СЕКЦИЯ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ. — С. 26–27.

[5] Машечкин И. В., Петровский М. И., Васильев Ю. А. Исследование и разработка вероятностного бустинг ансамбля анализа выживаемости // Ломоносовские чтения-2023: научная конференция, факультет ВМК МГУ имени М.В.Ломоносова. Тезисы докладов. — Москва : ООО "МАКС Пресс". — 2023. — «Вычислительная математика и кибернетика». — С. 107–109.

[6] Васильев Юлий Алексеевич. Критический обзор методов анализа выживаемости на основе бустинг ансамблей // Материалы Международного молодежного научного форума «Ломоносов-2023» / под ред. Алешковский Иван Андреевич, Андриянов Андрей Владимирович, Антипов Евгений Александрович, Зимакова Екатерина Игоревна. — Москва : ООО "МАКС Пресс". — 2023. — География. — С. 3.

[7] Васильев Юлий Алексеевич. Обзор функциональных возможностей библиотеки Survivors для анализа выживаемости в Python // Ломоносовские чтения-2024: научная конференция, факультет ВМК МГУ имени М.В.Ломоносова. Тезисы докладов. — Москва : ООО "МАКС Пресс". — 2024. — «Вычислительная математика и кибернетика». — С. 147–149.

[8] Васильев Ю. А., Петровский М. И., Машечкин И. В. Библиотека методов машинного обучения для построения моделей анализа выживаемости. – Свидетельство о гос. регистрации программы для ЭВМ; № 2024681935; заявл. 03.09.2024 ; опубл. 16.09.2024 (Рос. Федерация).

[9] Vasilev Iulii, Petrovskiy Mikhail, Mashechkin Igor. Adaptive Sampling for Weighted Log-Rank Survival Trees Boosting // Lecture Notes in Computer Science. — 2023. — Vol. 13822. — P. 98–115.

[10] Vasilev Iulii, Petrovskiy Mikhail, Mashechkin Igor. Sensitivity of Survival Analysis Metrics // Mathematics. — 2023. — Vol. 11, no. 20. — P. 4246.

[11] Васильев Ю. А., Петровский М. И., Машечкин И. В. Применение регуляризации при вычислении критериев разбиения в моделях анализа выживаемости // Вычислительные методы и программирование. — 2024. — Т. 25, № 3. — С. 9.

[12] Васильев Ю. А. РАЗРАБОТКА БИБЛИОТЕКИ ДРЕВОВИДНЫХ МОДЕЛЕЙ АНАЛИЗА ВЫЖИВАЕМОСТИ // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. — 2024. — № 3. — С. 60–72.

[13] Wang Ping, Li Yan, Reddy Chandan K. Machine learning for survival analysis: A survey // ACM Computing Surveys (CSUR). — 2019. — Vol. 51, no. 6. — P. 1–36.

[14] Salerno Stephen, Li Yi. High-dimensional survival analysis: Methods and applications // Annual review of statistics and its application. — 2023. — Vol. 10. — P. 25–49.

[15] Cohort profile: the Utrecht Cardiovascular Cohort–Second Manifestations of Arterial Disease (UCC-SMART) Study–an ongoing prospective cohort study of patients at high cardiovascular risk in the Netherlands / Castelijns Maria C, Helmink Marga AG, Hageman Steven HJ, Asselbergs Folkert W, de Borst Gert J, Bots Michiel L, Cramer Maarten J, Dorresteijn Jannick AN, Emmelot-Vonk Marielle H, Geerlings Mirjam I, et al. // BMJ open. — 2023. — Vol. 13, no. 2. — P. e066952.

[16] Trivella Juan, John Binu V, Levy Cynthia. Primary biliary cholangitis: Epidemiology, prognosis, and treatment // Hepatology communications. — 2023. — Vol. 7, no. 6. — P. e0179.

[17] Schumacher M. Rauschecker for the german breast cancer study group, randomized 2× 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive lbreast cancer patients // Journal of Clinical Oncology. — 1994. — Vol. 12. — P. 2086–2093.

[18] Royston Patrick, Lambert Paul C et al. Flexible parametric survival analysis using Stata: beyond the Cox model. — Stata press College Station, TX, 2011. — Vol. 347.

[19] An interpretable mortality prediction model for COVID-19 patients / Yan Li, Zhang Hai-Tao, Goncalves Jorge, Xiao Yang, Wang Maolin, Guo Yuqi, Sun Chuan, Tang Xiuchuan, Jing Liang, Zhang Mingyang, et al. // Nature machine intelligence. — 2020. — Vol. 2, no. 5. — P. 283–288.

[20] S.M.A.R.T site attributes. — `https://smartlinux.sourceforge.net/smart/attributes.php`. — Accessed: 2024-03-10.

[21] Hard Disk Drive Failure Analysis and Prediction: An Industry View / Miller Zach, Medaiyese Olusiji, Ravi Madhavan, Beatty Alex, and Lin Fred // 2023 53rd Annual

IEEE/IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S) / IEEE. — 2023. — P. 21–27.

[22] Role of Epstein-Barr Virus in Breast Cancer: Correlation with Clinical Outcome and Survival Analysis / Hsu Yi-Chiung, Tsai Ming-Han, Wu Guani, Liu Chien-Liang, Chang Yuan-Ching, Lam Hung-Bun, Su Pei-Yu, Lung Chun-Fan, and Yang Po-Sheng // Journal of Cancer. — 2024. — Vol. 15, no. 8. — P. 2403–2411.

[23] Nagy Ádám, Munkácsy Gyöngyi, Győrffy Balázs. Pancancer survival analysis of cancer hallmark genes // Scientific reports. — 2021. — Vol. 11, no. 1. — P. 6047.

[24] Győrffy Balázs. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer // Computational and structural biotechnology journal. — 2021. — Vol. 19. — P. 4101–4109.

[25] Awit Neil T, Marticio Ramon M. Customer Churn Prediction using Predictive Analytics: Basis for the Formulation of Customer Retention Strategy in the Context of Web-based Collaboration Platform // Proceedings of the International Conference on Industrial Engineering and Operations Management. — 2023.

[26] Customer churn prediction in telecom sector using machine learning techniques / Wagh Sharmila K, Andhale Aishwarya A, Wagh Kishor S, Pansare Jayshree R, Ambadekar Sarita P, and Gawande SH // Results in Control and Optimization. — 2024. — Vol. 14. — P. 100342.

[27] Kvamme Håvard, Borgan Ørnulf. Continuous and discrete-time survival prediction with neural networks // Lifetime data analysis. — 2021. — Vol. 27. — P. 710–736.

[28] Informative Censoring—A Cause of Bias in Estimating COVID-19 Mortality Using Hospital Data / Lin Hung-Mo, Liu Sean TH, Levin Matthew A, Williamson John, Bouvier Nicole M, Aberg Judith A, Reich David, and Egorova Natalia // Life. — 2023. — Vol. 13, no. 1. — P. 210.

[29] Templeton Arnoud J, Amir Eitan, Tannock Ian F. Informative censoring—a neglected cause of bias in oncology trials // Nature Reviews Clinical Oncology. — 2020. — Vol. 17, no. 6. — P. 327–328.

[30] Handbook of survival analysis / Klein John P, Van Houwelingen Hans C, Ibrahim Joseph George, and Scheike Thomas H. — CRC Press Boca Raton, FL:, 2014.

[31] Turkson Anthony Joe, Ayiah-Mensah Francis, Nimoh Vivian. Handling censoring and censored data in survival analysis: A standalone systematic literature review // International journal of mathematics and mathematical sciences. — 2021. — Vol. 2021. — P. 1–16.

[32] Candès Emmanuel, Lei Lihua, Ren Zhimei. Conformalized survival analysis // Journal of the Royal Statistical Society Series B: Statistical Methodology. — 2023. — Vol. 85, no. 1. — P. 24–45.

[33] Informative censoring of surrogate end-point data in phase 3 oncology trials / Gilboa Shai, Pras Yarden, Mataraso Aviv, Bomze David, Markel Gal, and Meirson Tomer // European Journal of Cancer. — 2021. — Vol. 153. — P. 190–202.

[34] Predicting COVID-19-induced lung damage based on machine learning methods / Vasilev IA, Petrovskiy MI, Mashechkin Igor V, and Pankratyeva Liudmila L // Programming and Computer Software. — 2022. — Vol. 48, no. 4. — P. 243–255.

[35] Kosaraju Nishoak, Sankepally Sainath Reddy, Mallikharjuna Rao K. Categorical data: Need, encoding, selection of encoding method and its emergence in machine learning models—a practical review study on heart disease prediction dataset using pearson correlation // Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 1 / Springer. — 2023. — P. 369–382.

[36] Missing data in clinical research: a tutorial on multiple imputation / Austin Peter C, White Ian R, Lee Douglas S, and van Buuren Stef // Canadian Journal of Cardiology. — 2021. — Vol. 37, no. 9. — P. 1322–1331.

[37] Efird Jimmy T. The Inverse Log-Rank Test: A Versatile Procedure for Late Separating Survival Curves // International Journal of Environmental Research and Public Health. — 2023. — Vol. 20, no. 24. — P. 7164.

[38] Rosen Kate, Prasad Vinay, Chen Emerson Y. Censored patients in Kaplan–Meier plots of cancer drugs: An empirical analysis of data sharing // European Journal of Cancer. — 2020. — Vol. 141. — P. 152–161.

[39] Olivier Timothée, Haslam Alyson, Prasad Vinay. Sotorasib in KRASG12C mutated lung cancer: Can we rule out cracking KRAS led to worse overall survival? // Translational Oncology. — 2023. — Vol. 28. — P. 101591.

[40] AutoScore-Imbalance: An interpretable machine learning tool for development of clinical scores with rare events data / Yuan Han, Xie Feng, Ong Marcus Eng Hock, Ning Yilin, Chee Marcel Lucas, Saffari Seyed Ehsan, Abdullah Hairil Rizal, Goldstein Benjamin Alan, Chakraborty Bibhas, and Liu Nan // Journal of Biomedical Informatics. — 2022. — Vol. 129. — P. 104072.

[41] Drysdale Erik. SurvSet: An open-source time-to-event dataset repository // arXiv preprint arXiv:2203.03094. — 2022.

[42] DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network / Katzman Jared L, Shaham Uri, Cloninger Alexander, Bates Jonathan, Jiang Tingting, and Kluger Yuval // BMC medical research methodology. — 2018. — Vol. 18, no. 1. — P. 1–12.

[43] The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults / Knaus William A, Harrell Frank E, Lynn Joanne, Goldman Lee, Phillips Russell S, Connors Alfred F, Dawson Neal V, Fulkerson William J, Califf Robert M, Desbiens Norman, et al. // Annals of internal medicine. — 1995. — Vol. 122, no. 3. — P. 191–203.

[44] Pitfalls of the concordance index for survival outcomes / Hartman Nicholas, Kim Sehee, He Kevin, and Kalbfleisch John D // Statistics in Medicine. — 2023. — Vol. 42, no. 13. — P. 2179–2190.

[45] The Concordance Index decomposition: A measure for a deeper understanding of survival prediction models / Alabdallah Abdallah, Ohlsson Mattias, Pashami Sepideh, and Rögnvaldsson Thorsteinn // Artificial Intelligence in Medicine. — 2024. — Vol. 148. — P. 102781.

[46] Kvamme Håvard, Borgan Ørnulf, Scheel Ida. Time-to-event prediction with neural networks and Cox regression // arXiv preprint arXiv:1907.00825. — 2019.

[47] Heagerty Patrick J, Zheng Yingye. Survival model predictive accuracy and ROC curves // Biometrics. — 2005. — Vol. 61, no. 1. — P. 92–105.

[48] Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation / Carrington André M, Manuel Douglas G, Fieguth Paul W, Ramsay Tim, Osmani Venet, Wernly Bernhard, Bennett Carol, Hawken Steven, Magwood Olivia, Sheikh Yusuf, et al. // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2022. — Vol. 45, no. 1. — P. 329–341.

[49] Kullback Solomon, Leibler Richard A. On information and sufficiency // The annals of mathematical statistics. — 1951. — Vol. 22, no. 1. — P. 79–86.

[50] Yari Gholamhossein, Mirhabibi Alireza, Saghafi Abolfazl. Estimation of the Weibull parameters by Kullback-Leibler divergence of Survival functions // Appl. Math. Inf. Sci. — 2013. — Vol. 7, no. 1. — P. 187–192.

[51] Murphy Allan H. A new vector partition of the probability score // Journal of Applied Meteorology and Climatology. — 1973. — Vol. 12, no. 4. — P. 595–600.

[52] Effective Ways to Build and Evaluate Individual Survival Distributions. / Haider Humza, Hoehn Bret, Davis Sarah, and Greiner Russell // J. Mach. Learn. Res. — 2020. — Vol. 21. — P. 85–1.

[53] Countdown regression: sharp and calibrated survival predictions / Avati Anand, Duan Tony, Zhou Sharon, Jung Kenneth, Shah Nigam H, and Ng Andrew Y // Uncertainty in Artificial Intelligence / PMLR. — 2020. — P. 145–155.

[54] Chiang Chin Long, Organization World Health et al. Life table and mortality analysis. — 1979.

[55] Kaplan Edward L, Meier Paul. Nonparametric estimation from incomplete observations // Journal of the American statistical association. — 1958. — Vol. 53, no. 282. — P. 457–481.

[56] Aalen Odd, Borgan Ornulf, Gjessing Hakon. Survival and event history analysis: a process point of view. — Springer Science & Business Media, 2008.

[57] Muller Hans-Georg, Wang Jane-Ling. Hazard rate estimation under random censoring with varying kernels and bandwidths // Biometrics. — 1994. — P. 61–76.

[58] Lin DY. On the Breslow estimator // Lifetime data analysis. — 2007. — Vol. 13, no. 4. — P. 471–480.

[59] Cox David R. Regression models and life-tables // Journal of the Royal Statistical Society: Series B (Methodological). — 1972. — Vol. 34, no. 2. — P. 187–202.

[60] Akram Saba, Ann Quarrat Ul. Newton raphson method // International Journal of Scientific & Engineering Research. — 2015. — Vol. 6, no. 7. — P. 1748–1752.

[61] Stensrud Mats J, Hernán Miguel A. Why test for proportional hazards? // Jama. — 2020. — Vol. 323, no. 14. — P. 1401–1402.

[62] Wei Lee-Jen. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis // Statistics in medicine. — 1992. — Vol. 11, no. 14-15. — P. 1871–1879.

[63] Hallinan Jr Arthur J. A review of the Weibull distribution // Journal of Quality Technology. — 1993. — Vol. 25, no. 2. — P. 85–93.

[64] Gordon L, Olshen RA. Tree-structured survival analysis // Cancer treatment reports. — 1985. — October. — Vol. 69, no. 10. — P. 1065—1069. — Access mode: `http://europepmc.org/abstract/MED/4042086`.

[65] Davis Roger B, Anderson James R. Exponential survival trees // Statistics in medicine. — 1989. — Vol. 8, no. 8. — P. 947–961.

[66] LeBlanc Michael, Crowley John. Survival trees by goodness of split // Journal of the American Statistical Association. — 1993. — Vol. 88, no. 422. — P. 457–467.

[67] LEBRANC M. Relative risk trees for censored survival data // Biometrics. — 1992. — Vol. 55. — P. 204–213.

[68] Zhang Heping. Splitting criteria in survival trees // Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modelling Innsbruck, Austria, 10–14 July, 1995 / Springer. — 1995. — P. 305–313.

[69] Therneau Terry M, Grambsch Patricia M, Fleming Thomas R. Martingale-based residuals for survival models // Biometrika. — 1990. — Vol. 77, no. 1. — P. 147–160.

[70] Keleş Sündüz, Segal Mark R. Residual-based tree-structured survival analysis // Statistics in medicine. — 2002. — Vol. 21, no. 2. — P. 313–326.

[71] Lee Seung-Hwan. Weighted Log-Rank Statistics for Accelerated Failure Time Model // Stats. — 2021. — Vol. 4, no. 2. — P. 348–358.

[72] Rayner John CW, Rippon Paul. An overview of new results in Cochran–Mantel–Haenszel testing. — 2018.

[73] Buyske Steven, Fagerstrom Richard, Ying Zhiliang. A class of weighted log-rank tests for survival data when the event is rare // Journal of the American Statistical Association. — 2000. — Vol. 95, no. 449. — P. 249–258.

[74] Shimokawa Asanao, Kawasaki Yohei, Miyaoka Etsuo. Comparison of splitting methods on survival tree // The international journal of biostatistics. — 2015. — Vol. 11, no. 1. — P. 175–188.

[75] Deep learning for survival analysis: a review / Wiegrebe Simon, Kopper Philipp, Sonabend Raphael, Bischl Bernd, and Bender Andreas // Artificial Intelligence Review. — 2024. — Vol. 57, no. 3. — P. 65.

[76] Deephit: A deep learning approach to survival analysis with competing risks / Lee Changhee, Zame William, Yoon Jinsung, and Van Der Schaar Mihaela // Proceedings of the AAAI conference on artificial intelligence. — 2018. — Vol. 32.

[77] Haji Saad Hikmat, Abdulazeez Adnan Mohsin. Comparison of optimization techniques based on gradient descent algorithm: A review // PalArch's Journal of Archaeology of Egypt/Egyptology. — 2021. — Vol. 18, no. 4. — P. 2715–2743.

[78] Salehin Imrus, Kang Dae-Ki. A review on dropout regularization approaches for deep neural networks within the scholarly domain // Electronics. — 2023. — Vol. 12, no. 14. — P. 3106.

[79] Support Vector Machines for Survival Analysis with R. / Fouodo Césaire JK, König Inke R, Weihs Claus, Ziegler Andreas, and Wright Marvin N // R Journal. — 2018. — Vol. 10, no. 1.

[80] Pölsterl Sebastian. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. // J. Mach. Learn. Res. — 2020. — Vol. 21, no. 212. — P. 1–6.

[81] Ma Guangzhi, Zhao Xuejing. Regression of survival data via twin support vector regression // Communications in Statistics-Simulation and Computation. — 2022. — Vol. 51, no. 9. — P. 5126–5138.

[82] Roy Atin, Chakraborty Subrata. Support vector machine in structural reliability analysis: A review // Reliability Engineering & System Safety. — 2023. — Vol. 233. — P. 109126.

[83] Bayesian survival analysis using the rstanarm R package / Brilleman Samuel L, Elci Eren M, Novik Jacqueline Buros, and Wolfe Rory // arXiv preprint arXiv:2002.09633. — 2020.

[84] Zhang Chenyang, Yin Guosheng. Bayesian nonparametric analysis of restricted mean survival time // Biometrics. — 2023. — Vol. 79, no. 2. — P. 1383–1396.

[85] Neuenschwander Beat. A Note on the Berliner-Hill Predictive Survival Distribution // Available at SSRN 4380225. — 2023.

[86] Bartoš František, Aust Frederik, Haaf Julia M. Informed Bayesian survival analysis // BMC Medical Research Methodology. — 2022. — Vol. 22, no. 1. — P. 238.

[87] Random survival forests / Ishwaran Hemant, Kogalur Udaya B, Blackstone Eugene H, and Lauer Michael S // The annals of applied statistics. — 2008. — Vol. 2, no. 3. — P. 841–860.

[88] Friedman Jerome H. Greedy function approximation: a gradient boosting machine // Annals of statistics. — 2001. — P. 1189–1232.

[89] Nguyen Nam Phuong. Gradient Boosting for Survival Analysis with Applications in Oncology. — University of South Florida, 2019.

[90] Binder Harald, Binder Maintainer Harald. Package 'CoxBoost'. — 2015.

[91] De Bin Riccardo. Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost // Computational Statistics. — 2016. — Vol. 31. — P. 513–531.

[92] A gradient boosting algorithm for survival analysis via direct optimization of concordance index / Chen Yifei, Jia Zhenyu, Mercola Dan, and Xie Xiaohui // Computational and mathematical methods in medicine. — 2013. — Vol. 2013.

[93] Chen Tianqi, Guestrin Carlos. Xgboost: A scalable tree boosting system // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. — 2016. — P. 785–794.

[94] Bai Miaojun, Zheng Yan, Shen Yun. Gradient boosting survival tree with applications in credit scoring // Journal of the Operational Research Society. — 2022. — Vol. 73, no. 1. — P. 39–55.

[95] A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction / Spooner Annette, Chen Emily, Sowmya Arcot, Sachdev Perminder, Kochan Nicole A, Trollor Julian, and Brodaty Henry // Scientific reports. — 2020. — Vol. 10, no. 1. — P. 1–10.

[96] Kaplan Marshall M. Primary biliary cirrhosis // New England Journal of Medicine. — 1996. — Vol. 335, no. 21. — P. 1570–1580.

[97] Breslow Norman. A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship // Biometrika. — 1970. — Vol. 57, no. 3. — P. 579–594.

[98] Tarone Robert E, Ware James. On distribution-free tests for equality of survival distributions // Biometrika. — 1977. — Vol. 64, no. 1. — P. 156–160.

[99] Peto Richard, Peto Julian. Asymptotically efficient rank invariant test procedures // Journal of the Royal Statistical Society: Series A (General). — 1972. — Vol. 135, no. 2. — P. 185–198.

[100] Klein John P, Moeschberger Melvin L. Statistics for biology and health // Stat. Biol. Health, New York. — 1997. — Vol. 27238.

[101] Weed Douglas L. Weight of evidence: a review of concept and methods // Risk Analysis: An International Journal. — 2005. — Vol. 25, no. 6. — P. 1545–1557.

[102] Benjamini Yoav, Hochberg Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing // Journal of the Royal statistical society: series B (Methodological). — 1995. — Vol. 57, no. 1. — P. 289–300.

[103] Hung Hung, Chiang Chin-Tsang. Estimation methods for time-dependent AUC models with survival data // Canadian Journal of Statistics. — 2010. — Vol. 38, no. 1. — P. 8–26.

[104] Evaluating prediction rules for t-year survivors with censored regression models / Uno Hajime, Cai Tianxi, Tian Lu, and Wei Lee-Jen // Journal of the American Statistical Association. — 2007. — Vol. 102, no. 478. — P. 527–537.

[105] Lambert Jérôme, Chevret Sylvie. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves // Statistical methods in medical research. — 2016. — Vol. 25, no. 5. — P. 2088–2102.

[106] Chawla Nitesh V. Data mining for imbalanced datasets: An overview // Data mining and knowledge discovery handbook. — 2010. — P. 875–886.

[107] Bradley Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms // Pattern recognition. — 1997. — Vol. 30, no. 7. — P. 1145–1159.

[108] Fawcett Tom. An introduction to ROC analysis // Pattern recognition letters. — 2006. — Vol. 27, no. 8. — P. 861–874.

[109] Refaeilzadeh Payam, Tang Lei, Liu Huan. Cross-validation. // Encyclopedia of database systems. — 2009. — Vol. 5. — P. 532–538.

[110] Mienye Ibomoiye Domor, Sun Yanxia. A survey of ensemble learning: Concepts, algorithms, applications, and prospects // IEEE Access. — 2022. — Vol. 10. — P. 99129–99149.

[111] Beja-Battais Perceval. AdaBoost: A theoritical review. — 2023.

[112] Drucker Harris. Improving regressors using boosting techniques // ICML / Citeseer. — 1997. — Vol. 97. — P. 107–115.

[113] Davidson-Pilon Cameron. lifelines: survival analysis in Python // Journal of Open Source Software. — 2019. — Vol. 4, no. 40. — P. 1317.

[114] Methods to analyze time-to-event data: the Cox regression analysis / Abd ElHafeez Samar, D'Arrigo Graziella, Leonardis Daniela, Fusaro Maria, Tripepi Giovanni, and Roumeliotis Stefanos // Oxidative Medicine and Cellular Longevity. — 2021. — Vol. 2021. — P. 1–6.

[115] Array programming with NumPy / Harris Charles R, Millman K Jarrod, Van Der Walt Stéfan J, Gommers Ralf, Virtanen Pauli, Cournapeau David, Wieser Eric, Taylor Julian, Berg Sebastian, Smith Nathaniel J, et al. // Nature. — 2020. — Vol. 585, no. 7825. — P. 357–362.

[116] Lam Siu Kwan, Pitrou Antoine, Seibert Stanley. Numba: A llvm-based python jit compiler // Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC. — 2015. — P. 1–6.

[117] pandas development team The. pandas-dev/pandas: Pandas. — 2020. — Feb. — Access mode: https://doi.org/10.5281/zenodo.3509134.

[118] Fowler Martin. UML distilled: a brief guide to the standard object modeling language. — Addison-Wesley Professional, 2018.

[119] A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less / Hammer Scott M, Squires Kathleen E, Hughes Michael D, Grimes Janet M, Demeter Lisa M, Currier Judith S, Eron Jr Joseph J, Feinberg Judith E, Balfour Jr Henry H, Deyton Lawrence R, et al. // New England Journal of Medicine. — 1997. — Vol. 337, no. 11. — P. 725–733.

[120] Prevalence of monoclonal gammopathy of undetermined significance / Kyle Robert A, Therneau Terry M, Rajkumar S Vincent, Larson Dirk R, Plevak Matthew F, Offord Janice R, Dispenzieri Angela, Katzmann Jerry A, and Melton III L Joseph // New England Journal of Medicine. — 2006. — Vol. 354, no. 13. — P. 1362–1369.

[121] The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective / Mahmood Syed S, Levy Daniel, Vasan Ramachandran S, and Wang Thomas J // The lancet. — 2014. — Vol. 383, no. 9921. — P. 999–1008.

[122] Raschka Sebastian. Model evaluation, model selection, and algorithm selection in machine learning // arXiv preprint arXiv:1811.12808. — 2018.

[123] Andronov Mikhail, Kolesnikov Sergey. CVTT: Cross-validation through time // arXiv preprint arXiv:2205.05393. — 2022.