

УДК 004.852

## СТРАТИФИЦИРОВАННЫЕ СТАТИСТИЧЕСКИЕ МОДЕЛИ В АНАЛИЗЕ НАДЕЖНОСТИ ОБОРУДОВАНИЯ

© 2024 г. Ю. А. Васильев<sup>1,\*</sup>, И. О. Филимонова<sup>1,\*\*</sup>,  
М. И. Петровский<sup>1,\*\*\*</sup>, И. В. Машечкин<sup>1,\*\*\*\*</sup>

Поступило 27.09.2024 г.  
После доработки 02.10.2024 г.  
Принято к публикации 02.10.2024 г.

Анализ надежности становится первостепенной задачей для успешного функционирования систем. В данной работе рассматривается проблема поломки оборудования на примере жестких дисков (HDD) и твердотельных накопителей (SSD). Для прогнозирования деградации оборудования используются методы анализа выживаемости, которые позволяют оценивать вероятность наступления события во времени. Также модели выживаемости учитывают неполные данные об истинном времени события для цензурированных наблюдений. Однако популярные статистические методы не учитывают особенности реальных данных, таких как наличие пропусков и категориальных переменных. В данной работе мы предлагаем расширить классические статистические методы выживаемости с помощью введения интерпретируемого стратифицирующего дерева, каждому листу которого соответствует статистическая модель. Экспериментальное исследование основано на оценке зависимости качества моделей при увеличении глубины дерева. По результатам экспериментов предложенный метод превосходит по качеству классические статистические модели. Результаты исследования демонстрируют эффективность предложенного подхода и его потенциал в области обеспечения надежности сложных технических систем.

**Ключевые слова:** анализ надежности, оборудование, анализ выживаемости, параметрические модели, машинное обучение

**DOI:** 10.31857/S2686954324700437, **EDN:** HFXIVJ

### 1. ВВЕДЕНИЕ

В современном мире технологии становятся все более сложными и интегрированными, поэтому обеспечение надежной работы систем становится ключевым аспектом успешного функционирования. Анализ надежности играет решающую роль в этом процессе, позволяя предсказать вероятность сбоев и определить факторы, влияющие на безотказную работу оборудования. В данной работе мы сосредоточимся на анализе сбоев и поломок оборудования на примере жестких дисков (HDD) и твердотельных накопителей (SSD).

Одним из мощных инструментов моделирования процессов деградации оборудования является анализ выживаемости, применяемый для ана-

лиза времени до наступления события. В рамках рассматриваемой области событием является поломка или отказ оборудования. Уникальной особенностью методов выживаемости является работа с цензурированными наблюдениями, для которых не известно истинное время выхода из строя из-за преждевременного выхода из эксплуатации (например, по причине утилизации или продажи).

Модели анализа выживаемости используют признаки оборудования  $X$  в начале эксплуатации, а целевыми переменными являются время события  $T$  и флаг цензурирования  $\delta$ . Помимо оценки времени отказа и анализа влияющих факторов, модели позволяют прогнозировать вероятность возникновения события в любой момент времени. В частности, функция выживания  $S(t)$  описывает вероятность того, что событие не произойдет до определенного момента времени  $t$ , тогда как функция риска  $h(t)$  указывает на условный риск возникновения события в конкретный момент времени  $t$ .

Применение методов анализа выживаемости позволяет оценить вероятность отказа системы, исследовать время до отказа, выявить зависимо-

<sup>1</sup>Московский государственный университет им. М.В. Ломоносова; Факультет вычислительной математики и кибернетики, Москва, Россия

\*E-mail: iuliivasilev@gmail.com

\*\*E-mail: filimonovaio@my.msu.ru

\*\*\*E-mail: michael@cs.msu.su

\*\*\*\*E-mail: mash@cs.msu.su

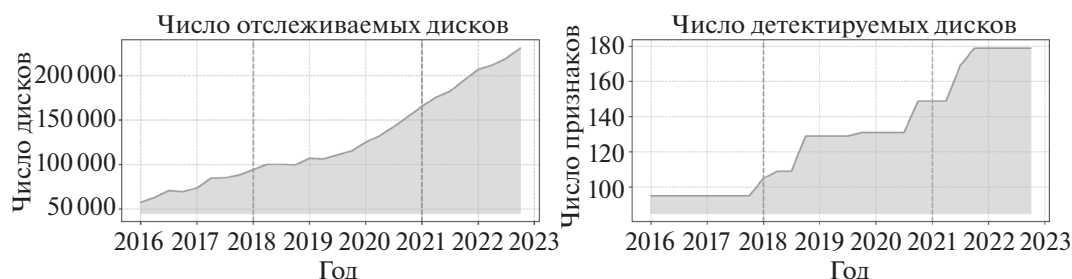


Рис. 1. Демонстрация роста количества наблюдений и детектируемых признаков в наборе Backblaze.

сти между различными факторами, влияющими на надежность, и оптимизировать процессы проектирования, производства и обслуживания оборудования. Однако данные анализа надежности имеют ряд особенностей, таких как: непрерывный рост объема данных, наличие категориальных признаков, пропуски и несбалансированность классов.

Классические статистические методы анализа выживаемости не учитывают данные особенности. В данной работе мы предлагаем метод расширения функционала статистических моделей путем построения стратифицирующего дерева выживания для обработки категориальных и пропущенных значений. В листах дерева выживаемости строятся классические статистические модели.

В разделе 2 мы представляем собранный набор данных мониторинга состояния жестких дисков. В разделе 3 приводится краткий обзор существующих статистических методов анализа выживаемости и их недостатков. В разделе 4 приводится обзор четырех метрик качества для оценки прогнозируемых величин анализа выживаемости. В разделе 5 предлагается собственный метод построения стратифицированных статистических моделей на основе дерева выживаемости. В разделе 6 мы проводим экспериментальное исследование зависимости качества моделей при увеличении числа страт. Наконец, в 7 формулируются основные результаты работы.

## 2. НАБОРЫ ДАННЫХ

Для обеспечения бесперебойной работы системы и сохранности данных существует необходимость в системе мониторинга и предупреждения о возможных отказах. Технология самоконтроля, анализа и отчетности (S.M.A.R.T.) — один из таких инструментов. Это система мониторинга, интегрированная в компьютерные жесткие диски и твердотельные накопители, которая предоставляет информацию о состоянии диска с целью предупреждения неизбежных отказов оборудования.

Одним из крупнейших источников данных S.M.A.R.T. для анализа надежности жестких дисков является набор открытых квартальных отчетов компании Backblaze. Эта компания специализируется на облачном хранении и резервном копировании данных и предоставляет данные о работе тысяч жестких дисков, что позволяет моделировать надежность устройств и разрабатывать эффективные стратегии обслуживания и обеспечения надежности систем хранения данных.

Отчеты хранилища Backblaze включают ежедневные снимки работающих жестких дисков. Все снимки дисков за заданный день собираются в файл, состоящий из строк для каждого активного жесткого диска. Каждый отчет содержит основную информацию о диске: дату файла в формате ГГГГ-ММ-ДД; уникальный серийный номер диска; номер модели; емкость диска в байтах; индикатор сбоя (содержит 1, если отчет получен в последний день работы диска); положение диска в хранилище — номер центра, кластера, модуля и слота хранения диска.

Также отчет содержит данные S.M.A.R.T. статистики. Каждый столбец представлен в нормализованном и сыром виде (например, *smart\_9\_normalized* и *smart\_9\_raw*). Методы перевода сырых значений в нормализованные скрыты от пользователя.

Непрерывный рост объема данных подтверждается увеличением числа отслеживаемых дисков и параметров S.M.A.R.T., детектируемых Backblaze (рис. 1). Исходя из изменений структуры признакового пространства во времени, исходные данные были представлены в виде трех наборов данных за 2016–2018, 2018–2021, 2021–2023 гг.

Сбор данных проводился согласно следующему алгоритму. Для каждого диска выделялось первое и последнее наблюдение при обслуживании. Целевая переменная времени вычислялась как разность между датами первого и последнего вхождения, а флаг цензурирования повторял значение столбца “failure” для последнего наблюдения. Отметим, что из наборов данных удалялись

Таблица 1. Характеристики наборов Backblaze

| Временной период                | 2016–2018 | 2018–2021 | 2021–2023 |
|---------------------------------|-----------|-----------|-----------|
| Количество наблюдений           | 33 392    | 66 434    | 33 288    |
| Количество признаков            | 64        | 72        | 78        |
| Количество терминальных событий | 4233      | 7098      | 9429      |

жесткие диски, для которых не наблюдались случаи отказа или цензурирования за все время исследования. Также были исключены признаки, содержащие более 99% пропущенных значений. Характеристики подготовленных наборов данных представлены в табл. 1.

### 3. СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ВЫЖИВАЕМОСТИ

Наиболее популярным непараметрическим методом оценки функции выживаемости является **метод Каплана–Мейера** (Kaplan–Meier, КМ) [1]. В рамках метода анализируются выборки наблюдений с различным временем возникновения события. Для каждого момента времени  $T_i$  определяются количество оставшихся наблюдений  $N_i$  и количество наступивших событий  $O_i$ .

Функцию выживаемости в конкретный момент времени  $t$  можно определить через кумулятивное произведение долей выживших до предыдущих моментов времени:

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

Полупараметрическая **модель пропорциональных рисков Кокса** (Cox proportional hazards, СохРН) [2] позволяет строить индивидуальный прогноз функции риска, предполагая, что отношение функций риска остается постоянным во времени. Прогноз модели основан на вертикальном масштабировании базисной функции риска  $h_0(t)$  на основе коэффициента пропорциональности  $\alpha_i = X_i \cdot \beta$ :

$$h(t | X_i) = h_0(t) \exp(X_i \cdot \beta). \quad (2)$$

Функция  $R(X) = \exp(X^T \beta)$  также называется функцией относительного риска. Базовые функции риска  $h_0(t)$  и кумулятивного риска  $H_0(t)$  строятся по оценке Бреслоу [3]. Прогноз функции выживания основан на оценке базовой функции  $S_0(t)$  по методу Каплана–Мейера:

$$S(t | X_i) = S_0(t)^{\exp(X_i \cdot \beta)}. \quad (3)$$

Вектор коэффициентов  $\beta$  оценивается путем минимизации отрицательного логарифмического частичного правдоподобия:

$$LL(\beta) = -\sum_{j=1}^N \delta_j \left[ X_j \beta - \lg \left[ \sum_{T_j \geq T_i} \exp(X_i \cdot \beta) \right] \right]. \quad (4)$$

Параметрический **метод ускоренного времени отказа** (Accelerated Failure Time, AFT) [4] предполагает, что связь между логарифмом времени выживания  $T$  и ковариатами является линейной:

$$\lg(T_i) = X_i \cdot \beta + \sigma \epsilon, \quad (5)$$

где  $X$  – матрица ковариат,  $\beta$  – вектор коэффициентов,  $\sigma$  ( $\sigma > 0$ ) – параметр масштабирования,  $\epsilon$  – переменная ошибки. Также модель предполагает, что  $\epsilon$  и  $\lg(T)$  одинаково распределены и параметрическое семейство распределений известно. Наиболее популярно использование экспоненциального, логнормального и лог-логистического распределения, а также распределения Вейбулла [5], на основе которых рассчитываются базовые функции выживаемости  $S_0(t)$  и риска  $h_0(t)$ . Прогноз функций вычисляется по сдвигу базовой функции с учетом коэффициента  $X\beta$ :

$$S(t | X) = S_0(e^{-X\beta} t), \quad h(t | X) = e^{-X\beta} h_0(e^{-X\beta} t). \quad (6)$$

Параметры  $\beta$  и  $\sigma$  определяются с помощью метода максимального правдоподобия на основе параметрического частичного правдоподобия [6].

### 4. МЕТРИКИ КАЧЕСТВА

Для оценки качества моделей используются метрики анализа выживаемости, которые позволяют оценивать одну из прогнозируемых величин: ожидаемое время события  $T$ , функцию выживания  $S(t)$ , функцию риска  $h(t)$ . Далее рассмотрим наиболее популярных представителей.

**Concordance index (CI)** [7] оценивает качество ранжирования наблюдений относительно прогнозируемого времени события. Метрика равна отношению числа верно упорядоченных пар по истинному времени события  $T_i$  относительно прогнозируемого времени  $\eta_i$  к количеству всех пар:

$$CI = \frac{\sum_{i,j} I(T_j < T_i) \cdot I(\eta_j < \eta_i)}{\sum_{i,j} I(T_j < T_i)}. \quad (7)$$

**Integrated AUC (IAUC)** [8] распространяет вычисления ROC-кривой [9] 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  $t$ . Временная метрика  $AUC(t)$  измеряет взвешенную долю пар наблюдений из каждого множества, имеющих со-

гласованный порядок рисков. Пусть  $\hat{h}(t | X_i)$  — оценка кумулятивного риска для наблюдения  $X_i$ ,  $w_i = G(t_i)$  — обратная вероятность цензурирования, тогда Integrated AUC равен:

$$IAUC(t_{\min}, t_{\max}) = \int_{t_{\min}}^{t_{\max}} \frac{\sum_{i=1}^n \sum_{j=1}^n I(T_j > t) I((T_i \leq t) \cdot \delta_i) w_i I(\hat{h}(t | X_j) \leq \hat{h}(t | X_i))}{\left(\sum_{j=1}^n I(T_j > t)\right) \left(\sum_{i=1}^n I((T_i \leq t) \cdot \delta_i) w_i\right)} dt. \quad (8)$$

**Integrated Brier score (IBS)** [10] оценивает квадрат отклонения между прогнозом функции выживания  $S(t | X_i)$  и эталоном. Эталоном является ступенчатая функция, равная единице до наступления события и нулю после. Далее мы используем модификацию метрики  $IBS_{RM}$  [11], определяющую равный вклад наблюдений независимо от времени события ( $N(t)$  — число наблюдаемых событий в момент  $t$ ):

$$IBS = \int_0^{t_{\max}} \frac{1}{t_{\max} \cdot N(t)} \times \sum_i \begin{cases} (0 - S(t | X_i))^2, & \text{if } T_i < t | \delta_i = 1, \\ (1 - S(t | X_i))^2, & \text{if } T_i > t, \\ 0, & \text{if } T_i = t | \delta_i = 0. \end{cases} dt. \quad (9)$$

**Survival AUPRC** [12] оценивает качество порогости функции выживания  $S(t)$  относительно истинного времени события  $T_i$ , сравнивая значения в точках  $T_i/\phi$  и  $T_i \cdot \phi$ , для всех  $\phi$  в интервале  $[0, 1]$ :

$$AUPRC = \frac{1}{N} \sum_{i=0}^N \int \hat{S}(T_i \cdot \phi) - \delta_i \cdot S(T_i/\phi) d\phi. \quad (10)$$

## 5. ПРЕДЛОЖЕННЫЙ ПОДХОД

Традиционные статистические методы не позволяют работать с пропущенными значениями и категориальными признаками. Для обработки номинальных признаков используются стратифицированные параметрические методы [13], основанные на построении множества независимых параметрических моделей на подвыборках с фиксированной категорией признака. Однако повышение количества категорий приводит к излишнему расслоению данных и понижению репрезентативности.

В данной работе предлагается использовать дерево выживаемости в качестве стратификатора. Модель дерева выживаемости [14] является расширением классического дерева решения, разделяя рекурсивно выборку на группы с разной выживаемостью. Наиболее популярным критерием разбиения является семейство логранк критериев [15], основанных на максимизации расстояния между функциями риска дочерних узлов. При разбиении узла оцениваются все потенциальные промежуточные значения для каждого признака в наборе  $X$ . Для каждого значения создаются две ветви разбиения и вычисляется значение статистики логранк. Оптимальное разбиение для выборки выбирается по наибольшему значению статистики.

Метод построения деревьев выживаемости позволяет обрабатывать категориальные и пропущенные значения. При поиске разбиения используется метод *weight of evidence* [16], отображающий категории на числовую шкалу относительно значений индикатора события. Для обработки пропущенных значений перебираются варианты добавления данных значений в каждую ветвь разбиения с целью минимизации статистического значения логранк критерия. Схема предложенного метода представлена на рис. 2. Первым этапом является построение дерева выживаемости на исходных гетерогенных данных, включающих категориальные и пропущенные значения. Далее каждому листовому узлу дерева сопоставляется собственная статистическая модель, построенная на заполненных числовых признаках листовой выборки. Для прогнозирования ожидаемого времени события, функции выживаемости или функции риска, входному наблюдению  $x_i$  сопоставляется лист  $L(x_i)$  исходя из структуры  $L$  дерева выживаемости. Прогнозом модели является прогноз статистической модели в выбранном листе.

Построение множества статистических моделей на выборках с близкими наблюдениями позволяет учесть неоднородность и локальные свойства данных. Предложенный подход позволяет обрабатывать пропущенные и номинальные значения без модификации статистических листо-

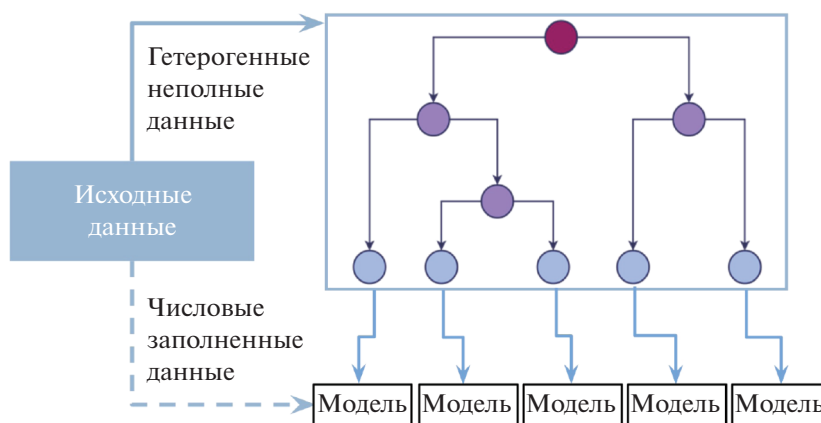


Рис. 2. Предложенный метод стратификации статистических моделей.

вых моделей. Также использование дерева выживаемости обеспечивает гибкость в управлении количеством и размером страт, что позволяет избежать излишнего расслоения данных. Наконец, дерево имеет строгую интерпретацию и позволяет оценить влияние признаков на выживаемость наблюдений.

Стоит отметить, что параметрические методы очень чувствительны к данным, а построение моделей на схожих наблюдениях может повлиять на сходимость модели. Также использование дерева выживаемости приводит к увеличению вычислительных ресурсов и необходимости дополнительного подбора гиперпараметров.

## 6. ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ

Целью экспериментального исследования является оценка зависимости качества прогнозирования моделей при увеличении глубины стратифицирующего дерева. Также необходимо сравнить качество предложенного подхода с классическими статистическими методами. Экспериментальное исследование проводится на кросс-валидации [17] по 5 блокам. При кросс-валидации исходные данные делятся на  $n$  блоков, модель  $n$ -кратно обучается на  $n - 1$  части и тестируется на оставшейся. Среднее качество вычисляется по всем итерациям кросс-валидации.

В данной работе в качестве листовых моделей рассматриваются оценка Каплана–Мейера, модель пропорциональных рисков Кокса (CoxPH) и модель ускоренного времени отказа (AFT). Реализации CoxPH и AFT были взяты из open-source библиотеки Lifelines [18], а модель Каплана–Мейера (Kaplan–Meier) и дерево выживаемости (CRAID) — из библиотеки Survivors [19]. Для оценки качества рассматриваются 4 метрики анализа выживаемости: CI (формула (7)), IBS\_RM

(формула (9)), IAUC (формула (8)), AUPRC (формула (10)). Реализации метрик были взяты из библиотеки Survivors.

В ходе экспериментального исследования проводился перебор гиперпараметров по сетке. Для модели CoxPH проводился перебор коэффициента регуляризации, для модели AFT рассматривались 3 семейства распределений (Вейбулла, логнормального и лог-логистического) и коэффициент регуляризации. Для модели стратификации CRAID рассматривались гиперпараметры глубины дерева, минимального размера листа и типа листовой модели.

На рис. 3–5 представлены результаты экспериментального исследования качества моделей на наборах Backblaze 2016–2018, Backblaze 2018–2021, Backblaze 2021–2023. По горизонтальной оси расположены метки глубины дерева решения, по вертикальной оси — одной из четырех метрик качества. Цвет линии отражает тип листовой модели в дереве выживаемости. Нулевая глубина соответствует случаю, когда дерево определяется только корневым узлом и дополнительная стратификация не проводится.

Для набора Backblaze 2016–2018 (рис. 3) лучшее качество метрики CI для модели CoxPH достигается при глубине 7, KM — при глубине 10, AFT — при глубине 5. Лучшее качество IBS для всех моделей достигается при глубине 2. Лучшее качество IAUC для моделей CoxPH и KM достигается при глубине 5, а для AFT — при глубине 2. Лучшее качество по AUPRC для всех моделей достигается при глубине 1 (дерево содержит два листовых узла, а качество сравнимо с базовой моделью). Отметим, что по всем метрикам лучшее качество достигается стратифицированной моделью Каплана–Мейера.

Для набора Backblaze 2018–2021 4 лучшее качество метрики CI для модели CoxPH достигается

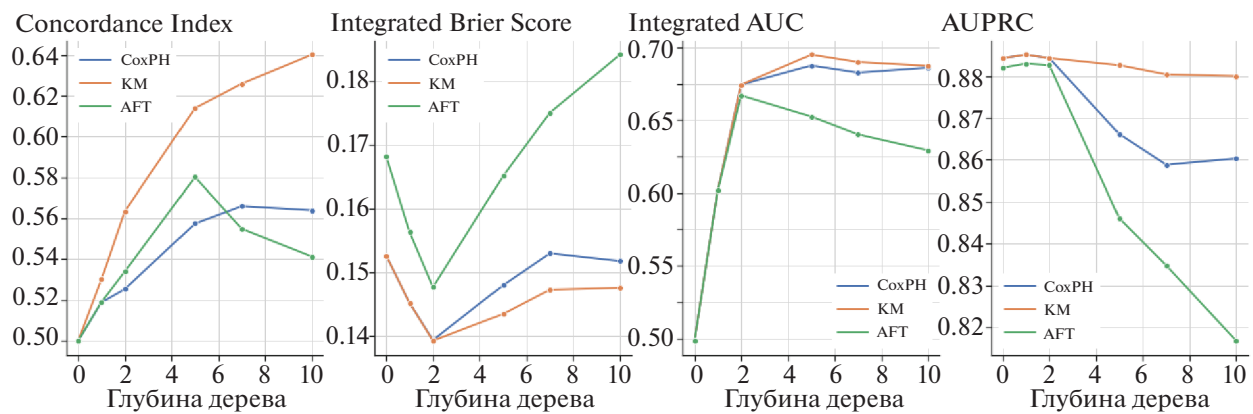


Рис. 3. Результаты экспериментов на наборе Backblaze 2016–2018.

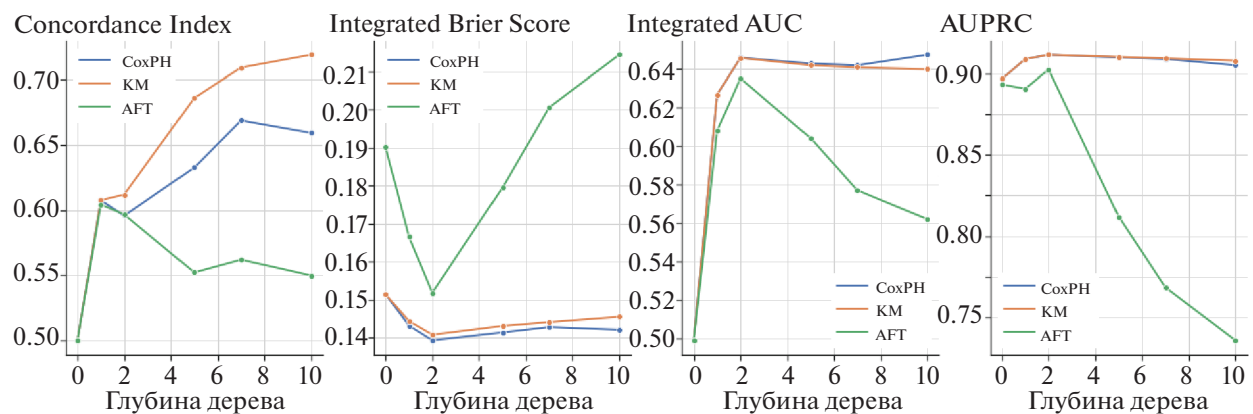


Рис. 4. Результаты экспериментов на наборе Backblaze 2018–2021.

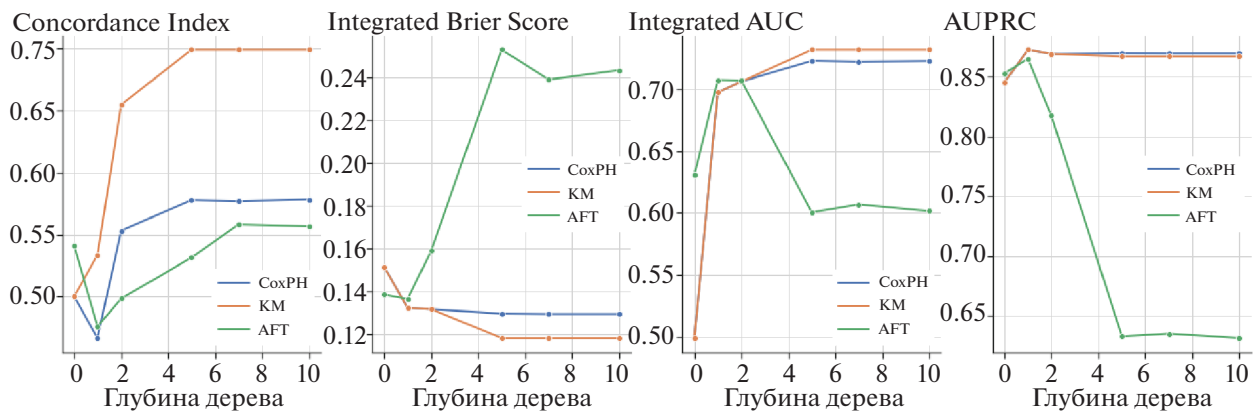


Рис. 5. Результаты экспериментов на наборе Backblaze 2021–2023.

при глубине 7, *KM* – при глубине 10, *AFT* – при глубине 1. Лучшее качество *IBS* для всех моделей достигается при глубине 2. Лучшее качество *IAUC* для модели *CoxPH* достигается при глубине 10, а для *KM* и *AFT* – при глубине 2. Лучшее ка-

чество по *AUPRC* для всех моделей достигается при глубине 2.

Для набора Backblaze 2021–2023 5 лучшее качество метрики *CI* для моделей *CoxPH* и *KM* достигается при глубине 10, *AFT* – при глубине 7.

Лучшее качество *IBS* для моделей *CoxPH* и *KM* достигается при глубине 10, *AFT* — при глубине 1. Лучшее качество *IAUC* для модели *CoxPH* и *KM* достигается при глубине 5, а для *AFT* — при глубине 2. Лучшее качество по *AUPRC* для всех моделей достигается при глубине 1.

Таким образом, для всех наборов данных, методам и метрикам наблюдается рост качества при использовании стратификации по сравнению с классической моделью. Наибольший прирост качества наблюдается на метриках ранжирования *CI* и *IAUC*. Отметим, что при повышении числа страт, при построении модели *AFT* могут наблюдаться проблемы со сходимостью, поскольку модель сильно зависит от объема данных и выполнения предположения о форме распределения времени.

## 7. ЗАКЛЮЧЕНИЕ

Предложенный метод построения стратифицированных статистических моделей с помощью дерева выживаемости позволяет строить оценки ожидаемого времени события, функции выживания и риска. По сравнению с классическими статистическими моделями, метод обрабатывает категориальные и пропущенные значения на этапе разделения данных и строит статистические модели на листовых выборках с однородными и схожими наблюдениями. Также наличие строгой интерпретации дерева выживаемости не нарушает свойств интерпретируемости статистических моделей.

По результатам экспериментов, для всех наборов данных и рассмотренных непараметрических, полупараметрических и параметрических моделей, предложенный метод демонстрирует прирост качества по всем метрикам. Таким образом, предложенный метод может стать важным инструментом для исследователей и специалистов в области анализа отказоустойчивости оборудования.

## ИСТОЧНИК ФИНАНСИРОВАНИЯ

Исследование выполнено при финансовой поддержке Некоммерческого Фонда развития науки и образования “Интеллект”.

## СПИСОК ЛИТЕРАТУРЫ

1. Kaplan E.L., Meier P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*. 1958. Т. 282. № 53. P. 457–81.
2. ElHafeez S. et al. Methods to analyze time-to-event data: the Cox regression analysis // *Oxidative Medicine and Cellular Longevity*. 2021. Т. 2021. P. 1–6.
3. Lin D.Y. On the Breslow estimator. *Lifetime data analysis*. 2007. Т. 13. P. 471–80.
4. Wei Lee-Jen. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis // *Statistics in medicine*. 1992. Т. 11. № 14–15. P. 1871–1879.
5. Clement N.L., Lasky R.C. Weibull distribution and analysis: 2019. In *2020 Pan Pacific Microelectronics Symposium (Pan Pacific)*. 2020. P. 1–5.
6. Ma D., Ma J., Graham P.L. On semiparametric accelerated failure time models with time-varying covariates: A maximum penalised likelihood estimation. *Statistics in medicine*. 2023. Т. 42. № 30. P. 5577–95.
7. Hartman N. et al. Pitfalls of the concordance index for survival outcomes // *Statistics in Medicine*. 2023. Т. 42. № 13. P. 2179–2190.
8. Heagerty P.J., Zheng Y. Survival model predictive accuracy and ROC curves // *Biometrics*. 2005. Т. 61. № 1. P. 92–105.
9. Carrington A.M. et al. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022. Т. 45. № 1. P. 329–341.
10. Haider H. et al. Effective Ways to Build and Evaluate Individual Survival Distributions // *J. Mach. Learn. Res.* 2020. Т. 21. P. 85.
11. Vasilev I., Petrovskiy M., Mashechkin I. Sensitivity of survival analysis metrics // *Mathematics*. 2023. Т. 11. № 20. P. 4246.
12. Avati A. et al. Countdown regression: sharp and calibrated survival predictions // *Uncertainty in Artificial Intelligence / PMLR*. 2020. P. 145–155.
13. Harrell, Jr F.E., Harrell F.E. Cox proportional hazards regression model. *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. 2015. P. 475–519.
14. LeBlanc M., Crowley J. Survival trees by goodness of split // *Journal of the American Statistical Association*. 1993. Т. 88. № 422. P. 457–467.
15. Lee S. Weighted Log-Rank Statistics for Accelerated Failure Time Model // *Stats*. 2021. Т. 4. № 2. P. 348–358.
16. Weed D.L. Weight of evidence: a review of concept and methods // *Risk Analysis: An International Journal*. 2005. Т. 25. № 6. P. 1545–1557.
17. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*. 2018.
18. Davidson-Pilon C. lifelines: survival analysis in Python. *Journal of Open Source Software*. 2019. Т. 4. № 40. P. 1317.
19. Васильев Ю.А. Разработка библиотеки древовидных моделей анализа выживаемости // *Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика*. 2024. № 3. С. 60–72.

## STRATIFIED STATISTICAL MODELS IN HARDWARE RELIABILITY ANALYSIS

**I. A. Vasilev<sup>a</sup>, I. O. Filimonova<sup>a</sup>, M. I. Petrovskiy<sup>a</sup>, and I. V. Mashechkin<sup>a</sup>**

*<sup>a</sup>Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics,  
Moscow, Russian Federation*

Reliability analysis is becoming paramount to the successful operation of systems. This paper considers the problem of hardware failure using hard disc drives (HDD) and solid state drives (SSD) as examples. Survival analysis methods are used to predict hardware degradation by estimating the probability of an event occurring over time. Also, survival models account for incomplete data about the true time of an event for censored observations. However, popular statistical methods do not account for features of real data such as the presence of outliers and categorical variables. In this paper, we propose to extend classical survival statistical methods by introducing an interpretable stratifying tree, each leaf of which corresponds to a statistical model. The experimental study is based on the evaluation of the dependence of the quality of the models as the depth of the tree increases. According to the experimental results, the proposed method outperforms classical statistical models. The results of the study demonstrate the effectiveness of the proposed approach and its potential in the field of reliability of complex technical systems.

*Keywords:* reliability analysis, hardware, survival analysis, parametric models, machine learning