ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ

УДК 519.23, 004.85

МЕТОДЫ АНАЛИЗА ВЫЖИВАЕМОСТИ В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ ВЫХОДА ИЗ СТРОЯ ОБОРУДОВАНИЯ ПРОМЫШЛЕННЫХ ПРЕДПРИЯТИЙ

Андронов А.Н., Бадокина Т.Е.
Мордовский государственный университет им. Н.П. Огарева, г. Саранск

Поступила в редакцию 16.03.2025, после переработки 25.05.2025.

В статье исследуется применение методов анализа выживаемости для прогнозирования времени до отказа промышленного оборудования. Рассмотрены классические подходы, такие как метод Каплана-Мейера и модель Кокса, а также их модификации и методы машинного обучения, включая случайный лес выживаемости (RSF). На реальных данных мясоперерабатывающего предприятия показано, что оригинальные детали имеют меньший риск отказа по сравнению с неоригинальными. В работе также исследовано влияние различных факторов на вероятность выхода из строя промышленного оборудования методами анализа выживаемости. Модели Каплана-Мейера и Кокса продемонстрировали схожую точность, а взвешенные методы оказались более адаптивными к цензурированным данным. Для оценки качества использовались метрики Concordance Index, Brier Score и Time-Dependent AUC.

Ключевые слова: анализ выживаемости, отказ оборудования, функция выживаемости, функция риска, модель Каплана-Мейера, модель Кокса, модель Random Survival Forest.

Вестник ТвГУ. Серия: Прикладная математика. 2025. № 2. С. 65–83.
https://doi.org/10.26456/vtpmk737

Введение

Анализ выживаемости, традиционно применяемый в медицинских и биологических исследованиях для оценки времени наступления событий [1], таких как ремиссия заболеваний или выживаемость пациентов, в последние годы находит всё более широкое применение в промышленной аналитике. В контексте прогнозирования времени поломок оборудования [2, 3] этот метод позволяет учитывать специфику данных, связанных с техническими системами, включая цензурированность [4, 5] (например, оборудование, которое не вышло из строя к моменту наблюдения) и влияние внешних факторов (нагрузка, температура, возраст оборудования). Используя модели регрессии Кокса [6] или параметрических распределений,

можно не только оценивать вероятность отказа в определённый период, но и выявлять ключевые предикторы, ускоряющие или замедляющие износ. В контексте машинного обучения анализ выживаемости позволяет не только предсказывать время до отказа оборудования, но и выявлять ключевые факторы, влияющие на этот процесс, что делает его мощным инструментом для повышения надёжности, оптимизации обслуживания промышленных систем и снижения затрат на обслуживание.

1. Основы анализа выживаемости в контексте технических систем

Анализ выживаемости – это совокупность статистических методов, предназначенных для изучения времени до наступления определенного события, такого как отказ оборудования или рецидив заболевания [7]. Эти методы позволяют выявить факторы, влияющие на продолжительность времени до события, и прогнозировать вероятность его наступления.

Время до отказа оборудования является случайной величиной, которая может быть описана через функцию выживаемости $S(t)$ и функцию риска $h(t)$. Функция $S(t) = P(T > t)$ отражает вероятность того, что объект проработает дольше времени t, а $h(t)$ характеризует мгновенную вероятность наступления события (например, отказа) в момент времени t, при условии, что событие не произошло до этого момента. Функция риска может увеличиваться или уменьшаться со временем в зависимости от долгосрочных или краткосрочных рисков, либо оставаться постоянной. Ключевая особенность данных в промышленных задачах — цензурирование, когда информация о времени отказа доступна не для всех объектов (например, часть оборудования ещё функционирует на момент завершения исследования).

1.1. Цензурирование в анализе выживаемости

Одной из ключевых особенностей данных о времени до отказа оборудования является их неполнота. Например, при анализе времени безотказной работы промышленных станков невозможно отследить точный момент поломки для всех устройств. Некоторые станки могут оставаться работоспособными на момент завершения исследования, другие — ломаться до начала наблюдений или между плановыми проверками. Такие случаи требуют специальных методов обработки, известных как цензурирование.

Выделяют три типа цензурирования:

1. Правое цензурирование

   Правое цензурирование возникает, когда событие (поломка) не наступило до завершения наблюдения. Предположим, компания проводит исследование надежности конвейерных лент. Исследование длится 2 года. Из 100 лент за это время сломались 60, а 40 остались работоспособными. Для этих 40 лент известно только, что их время безотказной работы превышает 2 года. Это классический случай правого цензурирования. Важно учесть, что цензурированные данные не игнорируются, но учитываются в анализе как «время работы без поломки до момента окончания наблюдения». Предполагается,

что оборудование, не вышедшее из строя к концу исследования, имеет ту же функцию риска, что и оборудование, за которым продолжают наблюдать.

2. Левое цензурирование

   Левое цензурирование имеет место, когда событие (поломка) произошло до начала наблюдения, но точное время неизвестно. Такой тип цензурирования часто встречается при анализе оборудования, которое эксплуатировалось до передачи в текущий парк.

3. Интервальное цензурирование

   В случае интервального цензурирования событие (поломка) произошло в известном интервале, но точное время не зафиксировано. Такая ситуация типична для оборудования, которое проверяется дискретно, например, раз в неделю или месяц. Для анализа используются специальные модели, к которым относятся обобщённый метод моментов или интервальная регрессия Кокса.

В промышленности правое цензурирование встречается чаще, но левое и интервальное требуют особого внимания, так как они могут искажать оценки надежности, если это не учитывается в моделях. В проводимом исследовании использовались новые комплектующие, поэтому использовалось правое цензурирование.

1.2. Непараметрические методы: метод Каплана-Мейера

Метод Каплана-Мейера [8] — это непараметрический подход для оценки функции выживаемости $S(t)$, который учитывает цензурированные данные. Он широко используется в анализе времени до события, например, для прогнозирования безотказной работы оборудования или оценки долговечности технических систем - позволяет оценить вероятность того, что оборудование не выйдет из строя раньше определенного времени. Метод не требует предположений о форме распределения времени до события (в отличие от экспоненциального метода или метода Вейбулла). Это делает его универсальным для анализа данных с неизвестной структурой.

Оценка Каплана-Мейера строится как произведение условных вероятностей выживания на каждом интервале времени:

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right),$$

где t_j — моменты времени, когда произошли события (например, поломки оборудования), d_j — число событий в момент t_j, n_j — число объектов, которые ещё работали (находились под наблюдением) непосредственно перед моментом времени t_j.

Результаты метода визуализируются в виде кривых выживаемости, где по оси x откладывается время, а по оси y — оценка $\hat{S}(t)$, таким образом каждое событие (поломка) отмечается «ступенькой» на графике.

1.2.1. Взвешенная оценка Каплана-Мейера

Одним из недостатков классической оценки Каплана-Мейера является предположение о том, что все наблюдения имеют одинаковую весовую значимость, что на практике труднодостижимо. В работе [9] показано, что в случае, когда значительная доля наблюдений цензурирована, оценка Каплана-Мейера ненадежна и неэффективна. Вводится взвешенная оценка Каплана-Мейера - модификация классической оценки Каплана-Мейера, в которой учитываются веса наблюдений.

Взвешенная оценка Каплана-Мейера задается следующей формулой:

$$\hat{S}(t) = \prod_{j:\, t_j \le t} w_j \left(1 - \frac{d_j}{n_j}\right),$$

где $w_j = \dfrac{n_j - c_j}{n_j}$ — доля нецензурированных наблюдений (вес для j-го наблюдения), d_j — множество событий, произошедших в момент времени t_j, n_j — множество объектов, находящихся под наблюдением непосредственно перед t_j.

Функцию, задающую веса наблюдений, можно определять и альтернативными способами. Так, можно учитывать разную важность наблюдений (в промышленности это может означать, что старые модели оборудования более критичны для анализа надежности, чем новые), неравномерное распределение в данных (дисбаланс классов, в этом случае меньший вес присваивается более многочисленным группам), сложные дизайны исследований (взвешенная оценка может быть использована для анализа данных, собранных через стратифицированную выборку и другие методы).

1.2.2. Модифицированная взвешенная оценка Каплана-Мейера

Недостаток взвешенной оценки в формулировке из пункта 1.2.1 заключается в том, что для последнего цензурированного наблюдения вес $w_j = 0$.

Модификация весовой функции, предложенная в статье [10], позволяет устранить этот недостаток. Весовая функция более точно учитывает сложные временные зависимости в данных и имеет вид

$$w_j = 1 - \sin\left(\frac{c_j * P_j}{n_j}\right).$$

1.3. Полупараметрические модели: модель Кокса (пропорциональных рисков)

Модель Кокса [11] является полупараметрическим методом анализа выживаемости, который позволяет оценивать влияние ковариат (например, температуры, возраста оборудования, нагрузки) на риск наступления события (поломки или отказа), не делая строгих предположений о форме базовой функции риска. Она сочетает гибкость непараметрических методов с возможностью моделирования факторов, как в регрессии.

Модель Кокса имеет вид:

$$h(t \mid \mathbf{X}) = h_0(t) \cdot e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p},$$

где: $h(t \mid \mathbf{X})$ — индивидуальный риск (интенсивность события) для объекта с ковариатами $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, $h_0(t)$ — базовая функция риска (не зависит от ковариат, её форма не задаётся), $\beta_1, \beta_2, \ldots, \beta_p$ — коэффициенты, показывающие вклад каждой ковариаты в риск.

Ключевая идея модели заключается в том, что отношение рисков для двух объектов постоянно во времени (пропорциональность рисков). Например, если оборудование A имеет риск в 2 раза выше, чем B, это соотношение сохраняется на всём временном интервале.

В отличие от параметрических моделей, таких как модель Вейбулла, в модели Кокса базовая функция риска $h_0(t)$ не параметризуется. Вместо этого используется метод частичного правдоподобия, который фокусируется на порядке событий, а не на их точном времени.

Формула частичного правдоподобия [12]:

$$L(\beta) = \prod_{i:\text{событие в } t_i} \frac{e^{\beta X_i}}{\sum_{j \in R(t_i)} e^{\beta X_j}},$$

где $R(t_i)$ — риск-множество в момент времени t_i: все объекты, которые ещё не имели события и не были цензурированы к моменту t_i.

Для каждого события t_i вычисляется вероятность того, что именно объект i «победил» в конкурентной борьбе за наступление события среди всех, кто был в риске в этот момент.

Каждый коэффициент β_k показывает, как изменение ковариаты X_k на 1 единицу влияет на риск наступления поломки. Экспонента коэффициента $HR_k = e^{\beta_k}$ показывает отношение рисков, таким образом, при $HR_k > 1$ увеличение ковариаты X_k повышает риск наступления поломки, а при $HR_k < 1$ увеличение X_k снижает риск.

Модель Кокса требует, чтобы отношение рисков между группами было постоянным во времени. Если это нарушается, выводы могут быть некорректными. В статье для проверки используется графический анализ остатков Шенфельда от времени.

Модель Кокса, являясь гибким инструментом для анализа рисков в условиях цензурирования, хорошо подходит для задач, где важно ранжировать факторы по их влиянию на надёжность, но точная форма функции риска неизвестна.

1.4. Параметрические модели в анализе выживаемости

Параметрические методы [2] анализа выживаемости основываются на предположении, что время до события следует определённому статистическому распределению. В отличие от непараметрических подходов (как Каплана-Мейера), которые не требуют знания формы распределения, или полупараметрических моделей Кокса, которые фокусируются на соотношении рисков, параметрические методы напрямую моделируют форму функции выживаемости, используя такие распределения, как экспоненциальное, Вейбулла или лог-логистическое. Распределение

Вейбулла подходит для описания оборудования, изнашивающегося со временем (монотонный рост риска), экспоненциальное – для моделирования случайных поломок, а лог-логистическое – для случаев с «пиками» отказов в начале эксплуатации или в определённых условиях.

Этот подход позволяет не только оценить влияние факторов (например, температуры или нагрузки) на риск, но и сделать точные прогнозы остаточного срока службы или вероятности события на конкретный момент времени. Однако ключевым условием является корректный выбор распределения: ошибочный выбор может искажать результаты, поэтому часто применяют графические методы или тесты для проверки соответствия данных выбранной модели. Таким образом, при правильном использовании параметрические модели дают более информативные выводы, особенно при работе с небольшими выборками или когда требуется прогнозирование индивидуальных траекторий.

1.5. Деревья выживаемости

Деревья выживаемости (Survival Trees) [13] — это деревья решений, специально разработанные для анализа данных времени до события. В контексте анализа выживаемости цель построения дерева выживаемости заключается в выявлении подгрупп экземпляров, которые различаются по риску наступления события на основе их базовых характеристик.

Основная идея дерева выживаемости заключается в рекурсивном разделении [14] выборки на всё более однородные подгруппы относительно интересующего события. Разделение достигается путем повторного разбиения данных на основе значений одной или нескольких прогностических переменных, так чтобы внутриузловая гетерогенность по исходам выживаемости была минимальной. На каждом этапе разбиения алгоритм выбирает переменную и точку разбиения, которые максимально дифференцируют исходы выживаемости подгрупп, определяемых разбиением. Этот процесс продолжается до тех пор, пока не будет достигнуто улучшение внутриузловой однородности или выполнен критерий остановки. Результатом является древовидная структура, где каждый конечный узел представляет собой отдельную подгруппу с уникальным профилем риска для интересующего события.

После построения дерева его необходимо обрезать, чтобы предотвратить переобучение и улучшить его обобщающую способность на новых данных. Наиболее часто используемым методом обрезки является обрезка с учетом стоимости сложности, которая включает добавление штрафного слагаемого к мере неоднородности для отдачи предпочтения более простым деревьям, избегающим переобучения. Для обрезки также могут использоваться другие методы, такие как кросс-валидация или методы повторной выборки.

Деревья выживаемости имеют несколько преимуществ перед традиционными регрессионными моделями. Они обладают способностью обрабатывать нелинейные и взаимодействующие эффекты, выявлять различные подгруппы с разными профилями риска и представлять легко интерпретируемые результаты в виде дерева решений. Однако они склонны к созданию переобученных деревьев, обладают чувствительностью к выбору критериев разбиения и правила остановки, а также характеризуются неспособностью учитывать изменяющиеся во времени ковариаты или конкурирующие риски [15].

1.6. Случайный лес выживаемости

Случайный лес выживаемости (RSF, Random Survival Forest) [15, 16] является методом машинного обучения для прогнозирования исходов выживаемости, расширяющим классический алгоритм случайного леса для обработки цензурированных данных выживаемости. RSF строит ансамбль деревьев решений, где каждое дерево выращивается с использованием случайной подвыборки данных и случайной подвыборки признаков. Для обработки цензурированных данных RSF вводит новый критерий разбиения, который учитывает распределение времени выживаемости в каждом узле дерева. В частности, критерий разбиения основан на статистике лог-ранга, которая измеряет различие во времени выживаемости между двумя группами наблюдений. Статистика лог-ранга рассчитывается как:

$$LR = \frac{(O - E)^2}{V},$$

где O – наблюдаемое число событий в группе, E – ожидаемое число событий на основе оценки Каплана-Мейера, а V – дисперсия числа событий по формуле Гринуорда.

На каждом узле дерева алгоритм рассматривает все возможные разбиения по всем возможным признакам и выбирает то, которое максимизирует статистику лог-ранга. В частности, критерий разбиения определяется как:

$$S = \frac{(LR_{left} - LR_{right})}{SE},$$

где LR_{left} и LR_{right} – статистики лог-ранга для левого и правого дочерних узлов, а SE – стандартная ошибка статистики лог-ранга.

После построения ансамбля деревьев RSF на его основе может делать прогнозы для новых наблюдений. Вероятность выживаемости для наблюдения рассчитывается как средняя вероятность выживаемости, предсказанная всеми деревьями в ансамбле. Вероятность выживаемости для заданного времени t оценивается как:

$$P(t \mid X) = e^{-H(t|X)},$$

где $H(t \mid X)$ — функция риска, предсказанная ансамблем деревьев для наблюдения с ковариатами X.

2. Прогнозирование поломки оборудования

В статье представлены результаты применения методов анализа выживаемости для прогнозирования времени выхода из строя промышленного оборудования. Исследование проводилось на реальных данных, полученных с производственных линий предприятия мясоперерабатывающей промышленности. Источниками данных являются датчики производственных линий, а также журналы обслуживания. Агрегированный датасет содержит информацию о следующих признаках:

1. Линия – производственная линия, принимает значения из множества ['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H'].

2. Дата – дата в формате YYYY-MM-DD.
3. Время – время работы в днях.
4. Оригинальность – индикатор оригинальности детали (оригинал — 1, аналог — 0).
5. Выработка (ресурс детали, кг) – количество килограммов продукции, произведенной во время работы детали.
6. Моточасы – количество часов наработки.
7. Поломка – целевая переменная = 1, если деталь вышла из строя.

Характеристики данных приведены в Таблице 1.

Таблица 1: *Характеристики данных*

Описание	Среднее значение ± стандартное отклонение	Медиана	Диапазон [мин., макс.]
Время (дни)	1.2862 ± 1.7157	1.00	[0, 30.0]
Моточасы (часы)	2200.9517 ± 4434.8854	659.5	[9.0, 26280.0]
Выработка (кг.)	14821.4369 ± 19534.2096	12432.9897	[0.0, 335294.793]

Датасет содержит цензурированные данные — это записи, в которых значение признака «Поломка» равно 0.

Основной целью анализа, выполненного в рамках данной статьи, является оценка влияния различных факторов на вероятность отказа оборудования и построение моделей, позволяющих прогнозировать время до наступления поломки. В качестве временных параметров рассматривались два показателя: количество дней с момента замены детали и количество моточасов наработки. Для анализа использовались как непараметрические методы (оценка Каплана-Мейера), так и полупараметрические модели (модель Кокса). Результаты визуализированы в виде кривых выживаемости, что позволяет наглядно оценить динамику риска отказа в зависимости от времени и ключевых факторов.

2.1. Модель Каплана-Мейера

Датасет содержит два признака, которые могут быть использованы в качестве временной шкалы: время в днях и моточасы. В Таблице 2 представлены значения функции выживания, где t — время в днях, построенные методом Каплана-Мейера (из библиотеки lifelines — KM lifelines, самостоятельная реализация — KM), взвешенным методом Каплана-Мейера (Weighted KM) и модифицированным взвешенным методом Каплана-Мейера (Modified Weighted KM).

Кривая выживаемости, показывающая изменение вероятности выживания на временном горизонте, является невозрастающей функцией (Рис. 1, 2). Как видно из Таблицы 2, вычисленные значения $\hat{S}(t)$, где t — время в днях, для модифицированного взвешенного и простого метода, отличаются незначительно, но в общем случае в модифицированном взвешенном методе $\hat{S}(t) \neq 0$ для данных, в которых последнее по времени наблюдение является цензурированным.

Таблица 2: *Значения функции выживаемости* $\hat{S}(t)$

Время (в днях)	KM (lifelines)	KM	Weighted KM	Modified Weighted KM
0.0	0.909091	0.781034	0.610015	0.772621
1.0	0.818182	0.279066	0.077878	0.276060
2.0	0.727273	0.089512	0.008012	0.088548
3.0	0.636364	0.056164	0.003154	0.055559
4.0	0.545455	0.024572	0.000604	0.024307
5.0	0.454545	0.019306	0.000373	0.019098
6.0	0.363636	0.010531	0.000111	0.010417
7.0	0.272727	0.005265	0.000028	0.005209
8.0	0.181818	0.003510	0.000012	0.003472
11.0	0.090909	0.001755	0.000003	0.001736
30.0	0.000000	0.000000	0.000000	0.000000

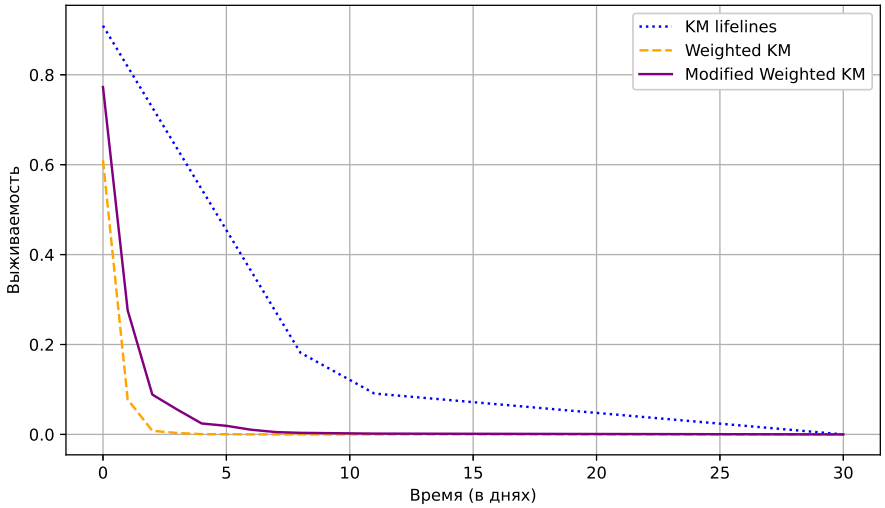


Рис. 1: *Кривые выживаемости, построенные по признаку «Дни»*

Медиана времени выживаемости равна 5.0 и 95%-ный доверительный интервал имеет вид $[1.0, 8.0]$ для случая t в днях, и 663.0 и $[579.0, 731.0]$ для случая t в моточасах.

В связи с большим количеством уникальных значение (469 различных значений в диапазоне [9, 26280], таблица значений $\hat{S}(t)$ для t в моточасах не приводится.

Для обоих рассматриваемых временных параметров характерно, что кривая выживаемости, построенная на основе взвешенной оценки Каплана-Мейера, дает самый пессимистичный вариант прогноза.

Метод Каплана-Мейера часто используется для сравнения выживаемости между группами. Например, необходимо сравнить время безотказной работы двух мо-

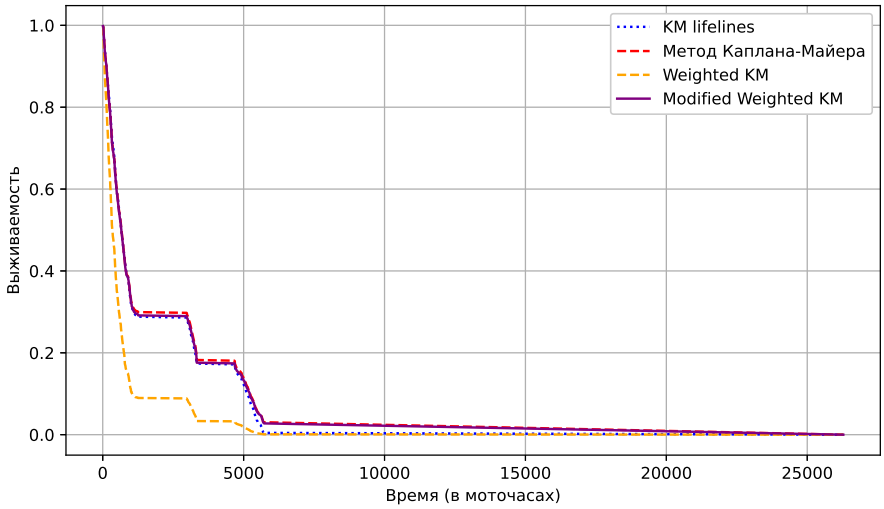


Рис. 2: *Кривые выживаемости, построенные по признаку «Моточасы»*

делей оборудования, проанализировать влияние условий эксплуатации, таких как температура и нагрузка на надёжность. Для проверки статистической значимости различий применяется тест лог-ранга, который оценивает, отличаются ли кривые выживаемости для разных групп.

2.2. Модель пропорциональных рисков Кокса

Модель пропорциональных рисков Кокса широко применяется для анализа выживаемости, поскольку позволяет учитывать влияние объясняющих переменных (ковариат) на вероятность отказа оборудования. Модель является полупараметрической и представляет собой суммы линейного риска, зависящего от времени, и частичного риска, включающего ковариаты.

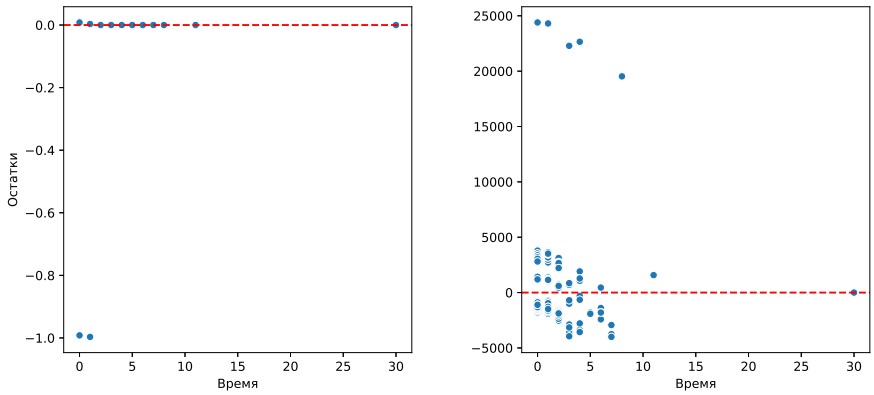
В качестве ковариат были выбраны признаки «Моточасы» и «Оригинальность», таким образом функция риска имеет следующий вид

$$h(t|X) = h_0(t) \cdot e^{\beta_1 X_1 + \beta_2 X_2},$$

где $h_0(t)$ – базовая функция риска, t – количество дней, прошедших с момента замены детали, X_1 – количество часов, прошедших с момента замены, X_2 – логическая переменная, определяемая следующим образом

$$X_2 = \begin{cases} 1, \text{если деталь оригинальная,} \\ 0, \text{в противном случае.} \end{cases}$$

Построенные диаграммы остатков Шенфельда (Рис. 3) для признаков «Оригинальность» и «Моточасы» не демонстрируют различимую закономерность, тем самым подтверждая предположение модели Кокса о независимости от времени.



(А) (В)

Рис. 3: *Остатки Шенфельда для ковариат: (А) «Оригинальность», (В) «Моточасы»*

Вычисленные значения коэффициентов ковариат $\beta_1 = -0.000025$ и $\beta_2 = -1.402785$ показывают, что ковариата количества отработанных часов X_1 слабо влияет на количество отказов, а ковариата оригинальности детали X_2 снижается вероятность отказа, то есть это является защитным фактором. Экспоненциальная форма коэффициента β_2 показывает, что для оригинальных деталей риск отказа в $e^{-1.402785} = 0,24$ раза меньше, чем для неоригинальных. Полученный результат подтверждает график (Рис. 4), на котором наглядно продемонстрировано, что неоригинальные детали требуют более частой замены по сравнению с оригинальными.

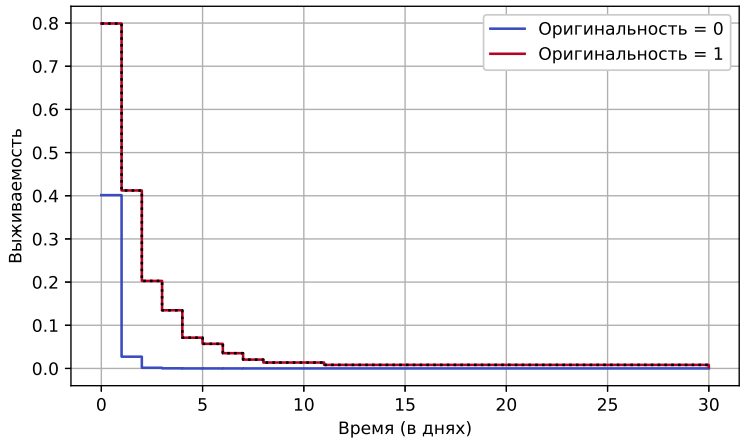


Рис. 4: *Кривые выживаемости для оригинальных и неоригинальных деталей (модель Кокса)*

Кривые выживаемости для построенных моделей Каплана-Майера и модели Кокса по ковариатам «Оригинальность» и «Моточасы» представлены на Рис. 5. Модели показывают схожие результаты, что указывает на их сопоставимую точность в данных условиях. Однако взвешенные методы (Weighted KM и Modified Weighted KM) демонстрируют отклонения, что свидетельствует об их адаптивности к специфическим характеристикам данных, таким как цензурирование.

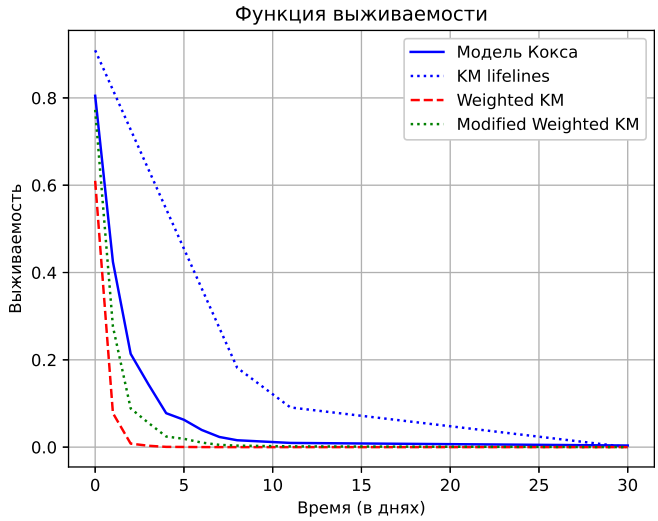


Рис. 5: *Кривые выживаемости для построенных моделей*

Небольшой объём данных, содержащий ограниченное число признаков, ограничивает возможность построения более точных моделей предсказания поломок, поэтому представляется перспективным преодоление данного ограничения в будущих исследованиях, в том числе за счет интеграции данных с IoT-устройств.

2.3. Случайный лес выживаемости

В исследовании использовался метод Random Survival Forest из библиотеки scikit-survival со следующими гиперпараметрами, подобранными через RandomizedSearchCV: количество деревьев (n_estimators = 156), минимальное количество образцов для разделения узла (min_samples_split = 8), минимальное количество образцов в листовом узле (min_samples_leaf = 1), максимальная глубина дерева (max_depth = 9), количество признаков при каждом разделении (max_features='log2') и фиксированное зерно для генератора случайных чисел (random_state = 42) для обеспечения воспроизводимости результатов. Данный подход обеспечивает устойчивость к переобучению и высокую производительность при прогнозировании времени до наступления событий в условиях цензурированных данных.

В данной работе для оценки качества моделей RSF использовались следующие метрики: Concordance Index (C-index), Brier Score, Integrated Brier Score (IBS) и Time-Dependent AUC.

1. C-Index [17] измеряет способность модели корректно ранжировать порядок событий, принимая значения из интервала $[0, 1]$, где значение $C\_Index = 1$ указывает на идеальное предсказание. Он измеряет, насколько хорошо модель предсказывает порядок событий, но если в данных много цензурированных наблюдений, C-index может быть менее информативным.

2. Brier Score оценивает точность предсказанных вероятностей выживания в конкретный момент времени, вычисляя среднее квадратичное отклонение между предсказанными и наблюдаемыми значениями, таким образом чем ближе значение к 0, тем выше точность модели.

3. Integrated Brier Score (IBS) расширяет Brier Score, интегрируя ошибку по всем временным точкам, что позволяет оценить общую точность модели на протяжении всего временного интервала.

4. Time-Dependent AUC позволяет оценить дискриминационную способность модели в зависимости от времени, что особенно полезно для анализа динамических изменений в данных.

Эти метрики обеспечивают комплексную оценку качества модели, учитывая как точность предсказаний, так и их согласованность с наблюдаемыми данными.

Показатели качества полученных моделей по различным наборам ковариат приведены в Таблице 3, где в столбце AUC указано среднее значение AUC по всем временным точкам дает, которое дает общую оценку качества модели.

Таблица 3: *Метрики качества моделей RSF*

Ковариаты	Метрики			
	C-Index	Brier Score	IBS	AUC
«Оригинальность», «Выработка»	0.8684	0.402	0.427	0.287
«Оригинальность», «Моточасы»	0.5888	0.386	0.400	0.505
«Оригинальность», «Моточасы», «Выработка»	0.6472	0.399	0.424	0.310
«Оригинальность», «Моточасы», «Выработка», «Линия»	0.3657	0.397	0.422	0.293

Для сравнения моделей BSF с моделями Кокса, построенными на аналогичных наборах ковариат, датасет был разбит на обучающую и тестовую часть в отношении 80:20. Анализ значений перечисленных выше метрик показал, что качество RSF незначительно отличается от соответствующих моделей Кокса. Модели с лучшим качеством построены по ковариатам «Оригинальность»и «Моточасы», однако и они не обладают высокой дискриминационной способностью. Поэтому для улучшения качества моделей необходимо увеличить датасет, используемый для построения моделей.

Заключение

В данной статье проведено исследование применения методов анализа выживаемости для прогнозирования времени до отказа промышленного оборудования и повышению надежности систем. В промышленности это напрямую переводится в экономию ресурсов за счёт оптимизации планово-предупредительных ремонтов и снижения простоев. На основе реальных данных о работе линий оборудовании предприятия мясоперерабатывающей промышленности были рассмотрены различные подходы, включая непараметрические методы (метод Каплана-Мейера), полупараметрические модели (модель Кокса) и методы машинного обучения (случайный лес выживаемости). Результаты показали, что оригинальные детали значительно снижают риск отказа по сравнению с неоригинальными, что подтверждается как классическими методами, так и их модификациями.

Модели, построенные на основе метода Каплана-Мейера и модели Кокса, продемонстрировали сопоставимую точность, однако взвешенные методы оказались более адаптивными к специфике данных, особенно в условиях цензурирования. Метод случайного леса выживаемости (RSF) показал близкие результаты к модели Кокса, но его применение требует более тщательной настройки и большего объёма данных для повышения точности прогнозов.

Основным ограничением исследования является небольшой объём данных и ограниченное количество признаков, что снижает дискриминационную способность моделей. В качестве перспективы предложено расширение набора данных за счёт интеграции информации с IoT-устройств, что позволит учитывать больше факторов и повысить точность прогнозирования. Дальнейшие исследования могут быть направлены на разработку более сложных моделей, учитывающих динамические изменения факторов и конкурирующие риски.

Список литературы

[1] Kartsonaki C. Survival analysis // Diagnostic Histopathology. 2016. Vol. 22, № 7. Pp. 263–270.

[2] Kalbfleisch J.D., Prentice R.L. The statistical analysis of failure time data. New Jersey: John Wiley and Sons, 2002. 462 p.

[3] Papathanasiou D., Demertzis K., Tziritas N. Machine Failure Prediction Using Survival Analysis // Future Internet. 2023. Vol. 15, № 5. ID 153.

[4] Klein J.P., Moeschberger M.L. Survival analysis: techniques for censored and truncated data. Springer Science and Business Media, 2006. 538 p.

[5] Nelson W. Theory and Applications of Hazard Plotting for Censored Failure Data // Technometrics. 1972. Vol. 14, № 4. Pp. 945–966.

[6] Cox D.R. Regression models and life-tables // Journal of the Royal Statistical Society: Series B (Methodological). 1972. Vol. 34, № 2. Pp. 187–202.

[7] Kleinbaum D.G., Klein M. Survival Analysis: A Self-Learning Text. Springer, 1996. 590 p.

[8] Kaplan E.L., Meier P. Nonparametric Estimation from Incomplete Observations // Journal of the American Statistical Association. 1958. Vol. 53, № 282. Pp. 457–481.

[9] Jan B., Shah S.W.A., Shah S., Qadir M.F. Weighted Kaplan Meier estimation of survival function in heavy censoring // Pakistan Journal of Statistics. 2005. Vol. 21, № 1. Pp. 55–63.

[10] Shafiq M., Shah S., Alamgir M. Modified Weighted Kaplan-Meier Estimator // Pakistan Journal of Statistics and Operation Research. 2007. Vol. 3, № 1. Pp. 39–44.

[11] Therneau T.M., Grambsch P.M. Modeling Survival Data: Extending the Cox Model. Springer Science and Business Media, 2012. 350 p.

[12] Cox D.R. Partial likelihood // Biometrika. 1975. Vol. 62, № 2. Pp. 269–276.

[13] Gordon L., Olshen R.A. Tree-structured survival analysis // Cancer Treatment Reports. 1985. Vol. 69, № 10. Pp. 1065–1069.

[14] LeBlanc M., Crowley J. Survival trees by goodness of split // Journal of the American Statistical Association. 1993. Vol. 88, № 422. Pp. 457–467.

[15] Bou-Hamad I., Larocque D., Ben-Ameur H. A review of survival trees // Statistics Surveys. 2011. Vol. 5. Pp. 44–71.

[16] Ishwaran H., Kogalur U.B., Blackstone E.H., Lauer M.S. Random survival forests // The Annals of Applied Statistics. 2008. Vol. 2, № 3. Pp. 841–860.

[17] Raykar V.C., Steck H., Krishnapuram B., Dehing-Oberije C., Lambin P. On ranking in survival analysis: bounds on the concordance index // Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS'07). New York: Curran Associates Inc., Red Hook, 2007. Pp. 1209–1216.

Образец цитирования

Андронов А.Н., Бадокина Т.Е. Методы анализа выживаемости в задаче прогнозирования выхода из строя оборудования промышленных предприятий // Вестник ТвГУ. Серия: Прикладная математика. 2025. № 2. С. 65–83. https://doi.org/10.26456/vtpmk737

Сведения об авторах

1. **Андронов Артем Николаевич**
доцент кафедры анализа данных и искусственного интеллекта факультета математики и информационных технологий МГУ им. Н.П. Огарёва.

Россия, 430005, г. Саранск, ул. Большевистская, д. 68, МГУ им. Н.П. Огарёва.
E-mail: arbox@inbox.ru

2. **Бадокина Татьяна Евгеньевна**
доцент кафедры анализа данных и искусственного интеллекта факультета математики и информационных технологий МГУ им. Н.П. Огарёва.

Россия, 430005, г. Саранск, ул. Большевистская, д. 68, МГУ им. Н.П. Огарёва.
E-mail: badokinate@yandex.ru

METHODS OF SURVIVAL ANALYSIS IN THE PROBLEM OF PREDICTING FAILURE OF EQUIPMENT IN INDUSTRIAL ENTERPRISES

Andronov A.N., Badokina T.E.
Ogarev Mordovia State University, Saransk

Received 16.03.2025, revised 25.05.2025.

The article explores the application of survival analysis methods for predicting the time until failure of industrial equipment. Classical approaches such as the Kaplan-Meier method and the Cox model, as well as their modifications and machine learning techniques, including Random Survival Forests (RSF), are examined. Using real-world data from a meat processing plant, it is demonstrated that original parts have a lower risk of failure compared to non-original ones. The study also investigates the impact of various factors on the likelihood of industrial equipment failure using survival analysis methods. The Kaplan-Meier and Cox models demonstrated comparable accuracy, while weighted methods proved to be more adaptable to censored data. For quality assessment, metrics such as the Concordance Index, Brier Score, and Time-Dependent AUC were utilized.

Keywords: survival analysis, equipment failure, survival function, hazard function, Kaplan-Meier estimator, Cox proportional hazards model, Random Survival Forest.

Citation

Andronov A.N., Badokina T.E., "Methods of survival analysis in the problem of predicting failure of equipment in industrial enterprises", *Vestnik TvGU. Seriya: Prikladnaya Matematika [Herald of Tver State University. Series: Applied Mathematics]*, 2025, № 2, 65–83 (in Russian). https://doi.org/10.26456/vtpmk737

References

[1] Kartsonaki C., "Survival analysis", *Diagnostic Histopathology*, **22**:7 (2016), 263–270.

[2] Kalbfleisch J.D., Prentice R.L., *The statistical analysis of failure time data*, John Wiley and Sons, New Jersey, 2002, 462 pp.

[3] Papathanasiou D., Demertzis K., Tziritas N., "Machine Failure Prediction Using Survival Analysis", *Future Internet*, **15**:5 (2023), 153.

[4] Klein J.P., Moeschberger M.L., *Survival analysis: techniques for censored and truncated data*, Springer Science and Business Media, 2006, 538 pp.

[5] Nelson W., "Theory and Applications of Hazard Plotting for Censored Failure Data", *Technometrics*, **14**:4 (1972), 945–966.

[6] Cox D.R., "Regression models and life-tables", *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**:2 (1972), 187–202.

[7] Kleinbaum D.G., Klein M., *Survival Analysis: A Self-Learning Text*, Springer, 1996, 590 pp.

[8] Kaplan E.L., Meier P., "Nonparametric Estimation from Incomplete Observations", *Journal of the American Statistical Association*, **53**:282 (1958), 457–481.

[9] Jan B., Shah S.W.A., Shah S., Qadir M.F., "Weighted Kaplan Meier estimation of survival function in heavy censoring", *Pakistan Journal of Statistics*, **21**:1 (2005), 55–63.

[10] Shafiq M., Shah S., Alamgir M., "Modified Weighted Kaplan-Meier Estimator", *Pakistan Journal of Statistics and Operation Research*, **3**:1 (2007), 39–44.

[11] Therneau T.M., Grambsch P.M., *Modeling Survival Data: Extending the Cox Model*, Springer Science and Business Media, 2012, 350 pp.

[12] Cox D.R., "Partial likelihood", *Biometrika*, **62**:2 (1975), 269–276.

[13] Gordon L., Olshen R.A., "Tree-structured survival analysis", *Cancer Treatment Reports*, **69**:10 (1985), 1065–1069.

[14] LeBlanc M., Crowley J., "Survival trees by goodness of split", *Journal of the American Statistical Association*, **88**:422 (1993), 457–467.

[15] Bou-Hamad I., Larocque D., Ben-Ameur H., "A review of survival trees", *Statistics Surveys*, **5** (2011), 44–71.

[16] Ishwaran H., Kogalur U.B., Blackstone E.H., Lauer M.S., "Random survival forests", *The Annals of Applied Statistics*, **2**:3 (2008), 841–860.

[17] Raykar V.C., Steck H., Krishnapuram B., Dehing-Oberije C., Lambin P., "On ranking in survival analysis: bounds on the concordance index", *Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS'07)*, Curran Associates Inc., Red Hook, New York, 2007, 1209–1216.

Author Info

1. **Andronov Artyom Nikolaevich**
Associate Professor at Data Analysis and Artificial Intelligence Department, Faculty of Mathematics and Information Technology, Ogarev Mordovia State University.

Russia, 430005, Saransk, Bolshevistskaya str., 68, Ogarev Mordovia State University. E-mail: arbox@inbox.ru

2. **Badokina Tatiana Evgen'evna**
Associate Professor at Data Analysis and Artificial Intelligence Department, Faculty of Mathematics and Information Technology, Ogarev Mordovia State University.

Russia, 430005, Saransk, Bolshevistskaya str., 68, Ogarev Mordovia State University. E-mail: badokinate@yandex.ru