УДК 004.4:616.127

ОЦЕНКА ВЫЖИВАЕМОСТИ ПАЦИЕНТОВ С ИНФАРКТОМ МИОКАРДА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Пекось О.А.

Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта, Ташкент, Узбекистан

Аннотация. Высокая летальность в остром и постгоспитальном периодах инфаркта миокарда обусловливает актуальность проблемы точной риск-стратификации пациентов и оценки их выживаемости. В работе сопоставлены два комплементарных подхода: регрессия Кокса, в том числе с временно-зависимыми ковариатами, и ансамблевый метод случайного леса для бинарной классификации и анализа выживаемости. Показано, что учёт динамики клинико-лабораторных показателей (АД, ЧСС, маркеров некроза и др.) повышает дискриминацию и калибровку прогнозов. Приведены варианты уравнения мгновенного риска, частичного правдоподобия, а также формулы для оценки выживаемости и кумулятивной интенсивности. Иллюстративные примеры на смешанном наборе реальных и синтетических данных демонстрируют преимущество комбинированного подхода при наличии цензурирования и неоднородных профилей риска. Результаты подтверждают целесообразность интеграции моделей в клинический рабочий процесс для персонализации тактики лечения и реабилитации после инфаркта миокарда.

Ключевые слова: машинное обучение, анализ выживаемости, прогнозирование летальности, модель Кокса, случайный лес выживания.

1 ВВЕДЕНИЕ

Инфаркт миокарда (ИМ) как наиболее тяжелое осложнение ишемической болезни сердца, занимает лидирующие позиции в структуре смертности кардиологических больных. Несмотря на значительные успехи, достигнутые в области кардиологической помощи за последние годы, острый период ИМ характеризуется высоким риском летального исхода. Однако, и после его купирования, в среднесрочной и отдаленной постгоспитальной перспективе, сохраняется значительная угроза развития фатальных осложнений, включая повторный инфаркт, разрыв миокарда и внезапную сердечную смерть.

В этой связи, максимально точная стратификация индивидуального риска для каждого пациента приобретает критическое значение. От качества прогноза напрямую зависит выбор адекватной тактики ведения больного, интенсивность диспансерного наблюдения, своевременность коррекции медикаментозной терапии и, в конечном счете, успех всего реабилитационного процесса [1, 2].

В современной кардиологической практике для оценки прогноза широко используются унифицированные прогностические шкалы, такие как EuroSCORE II, STS Score, GRACE и TIMI. Эти шкалы, основанные на результатах масштабных клинических исследований, интегрируют множество клинических, анамнестических и инструментальных предикторов и демонстрируют приемлемую валидность на различных когортах пациентов, перенесших ИМ и чрескожные коронарные вмешательства (ЧКВ) [3].

Тем не менее, ни одна из существующих шкал не является «идеальной». Их прогностическая точность может варьироваться в зависимости от специфики популяции, а включение большого числа переменных не всегда гарантирует высокую дискриминационную способность для конкретного пациента. Это стимулирует непрерывный научный поиск, направленный как на совершенствование классических моделей, так и на разработку принципиально новых подходов к стратификации риска.

В последние годы одним из наиболее динамично развивающихся направлений в этой области стало применение методов искусственного интеллекта и машинного обучения. В отличие от традиционных статистических моделей, алгоритмы машинного обучения способны выявлять сложные, нелинейные взаимосвязи и скрытые закономерности в больших массивах гетерогенных медицинских данных. Весьма распространенными и доказавшими свою эффективность подходами являются модели, основанные на: логистической регрессии, регрессии Кокса, алгоритмах дерева решений и случайного леса, а также на различных архитектурах нейронных сетей [4]. Эти методы не только

позволяют строить более точные прогнозы выживаемости, но и способствуют выявлению новых, наиболее значимых клинических факторов риска, что углубляет патофизиологическое понимание течения постинфарктного периода.

Таким образом, разработка и валидация прогностических моделей на основе методов машинного обучения представляется крайне актуальной задачей, открывающей новые возможности для персонализации кардиологической помощи и улучшения исходов у пациентов, перенесших инфаркт миокарда.

2 МЕТОДЫ

2.1 Оценка выживаемости на основе модели рисков Кокса

К числу базовых и широко используемых подходов к оценке выживаемости пациентов после ИМ относится метод Каплана–Мейера. Его ключевое достоинство – построение эмпирической функции выживаемости непосредственно по индивидуальным наблюдениям с учётом цензуры, без предварительной группировки. Получающаяся ступенчатая кривая даёт наглядное представление о динамике выживаемости по времени, позволяет оценивать ей медиану и удобно сравнивать подгруппы пациентов (например, по полу, возрасту или факту ревскуляризации) с помощью лог-ранговых критериев. Именно благодаря простоте, устойчивости при малых объёмах данных и прозрачной интерпретации метод Каплана–Мейера часто выбирают в исследованиях, где выборка невелика и требуется первичная ориентировочная оценка выживаемости.

В то же время, когда задача усложняется – требуется точное прогнозирование исхода ИМ, выявление периодов повышенного риска и количественная оценка вклада отдельных факторов – возможностей метода Каплана–Мейера становится недостаточно. Он описывает «среднюю» кривую выживания для группы и не регулирует влияние ковариат на уровне индивидуального пациента. Между тем клинические и биохимические показатели (возраст, пол, АД, ЧСС, тропонин, креатинин, липидный профиль, наличие СД2 и др.) существенно меняют риск и должны учитываться явно. Для стабильной оценки таких эффектов обычно нужна достаточно крупная выборка, а также инструменты, которые встроят эти признаки непосредственно в модель выживаемости [5, 6].

После того как функция выживаемости $S(t)$ и функция мгновенного риска $h(t)$ описаны в целом, закономерно встаёт вопрос: как зависят эти величины от конкретных факторов – пола, возраста, клинико-лабораторных показателей, способов лечения и, главное, как меняется эта зависимость во времени. На этом этапе в анализ выживаемости вводятся регрессионные модели, позволяющие связать индивидуальный профиль признаков пациента с его траекторией риска. Такие модели делают возможным не только сравнение групп, но и персонализированный прогноз с количественной оценкой вклада каждого предиктора.

Разрабатывая регрессионные модели для оценки выживаемости после ИМ, важно учитывать специфические свойства данных. Во-первых, многие связи между предикторами и риском нелинейны и содержат взаимодействия (например, возраст × СД2). Во-вторых, присутствуют цензурированные наблюдения (пациенты, у которых на момент окончания наблюдения событие не наступило). Эти особенности делают классические методы множественной регрессии (ориентированные на непрерывные исходы без цензуры) методологически неприменимыми или, по крайней мере, недостаточно точными [7]. Нужны модели, изначально «заточенные» под анализ времени до события.

В современной практике выделяют три комплементарных подхода: параметрические модели (например, экспоненциальная или Вейбулла), где полностью задаётся форма базового риска; полупараметрическая модель пропорциональных рисков Кокса, не требующая явной формы базовой интенсивности; расширения модели Кокса с временно-зависимыми ковариатами, позволяющие учитывать динамику клинических показателей во времени [8, 9]. Такой спектр методов охватывает как ситуации с предположительно простой формой риска, так и сценарии со сложной динамикой и богатым набором признаков.

В регрессии Кокса центральной категорией остаётся время, но вклад признаков вводится в виде мультипликативного эффекта к базовому риску. Добавление временно-зависимых ковариат (например, меняющихся значений АД, ЧСС, маркёров некроза в остром периоде) делает модель чувствительной к клинической динамике: риск обновляется в те моменты, когда обновляются и сами показатели. Практическое ограничение здесь не концептуальное, а измерительное: если данные о пациентах поступают редко или нерегулярно, информативность временной компоненты снижается. Тем не менее для прогнозирования летальности после ИМ модель Кокса остаётся мощным и интерпретируемым инструментом, так как позволяет оценить влияние каждого фактора до наступления события, то есть «на лету» формировать оценку риска.

Модель пропорциональных рисков Кокса строится как произведение двух множителей: один описывает базовую функцию мгновенного риска $h_0(t)$, общий для всей когорты, а второй – экспонента линейного прогностического индекса от ковариат пациента. Важное допущение – пропорциональность рисков: отношение интенсивностей для двух пациентов с разными наборами признаков не меняется во времени. Это означает постоянство относительного риска, но не предполагает, что сам по себе абсолютный риск остаётся неизменным после ИМ, то есть базовая интенсивность $h_0(t)$ может возрастать или убывать, отражая естественную клиническую динамику.

Типовой рабочий процесс построения модели Кокса включает несколько этапов. Сначала формируют и очищают набор данных о пациентах, перенёсших ИМ, собирая клинические, демографические и лабораторные параметры (возраст, пол, артериальное давление, частоту сердечных сокращений, липидный профиль, маркёры некроза, креатинин, индекс массы тела, факт курения, наличие СД2 и т. п.). Затем чётко задают зависимую переменную – «время до события» (например, до смерти) или время наблюдения, а также индикатор события: 1 – событие наступило, 0 – наблюдение цензурировано (событие не произошло до конца периода наблюдения). После этого строят собственно регрессионную модель, проверяя допущение пропорциональности рисков, при необходимости моделируя нелинейные эффекты (сплайны) и взаимодействия, а также рассматривая временно-зависимые ковариаты, если клиническая динамика важна.

В простейшей форме уравнение регрессии Кокса записывается так:

$$h(t \mid \mathbf{x}) = h_0(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{x}\right), \tag{1}$$

где $h(t \mid \mathbf{x})$ – условный риск в момент времени t при наборе ковариат $\mathbf{x} \in \mathbb{R}^p$; $h_0(t)$ – базовая функция риска; $\boldsymbol{\beta} \in \mathbb{R}^p$ – вектор регрессионных коэффициентов; $\mathbf{x} = (x_1, \ldots, x_p)^\top$ – вектор ковариат.

Оценка $\boldsymbol{\beta}$ может выполняться методом максимального (частичного) правдоподобия. Компоненты $\boldsymbol{\beta}$ интерпретируются как логарифмы отношений рисков: $\exp\{\beta_k\}$ показывает, во сколько раз меняется мгновенный риск при увеличении соответствующей ковариаты на одну единицу (при прочих равных).

Качество модели оценивают индексами дискриминации (например, C -индексом) и проверкой калибровки. В итоге полученные коэффициенты позволяют количественно оценить вклад факторов риска в летальность, а отношения рисков делают интерпретацию клинически прозрачной.

Предположим, у нас есть следующие данные для модели: возраст (age); пол (gen); наличие СД2 (dia); уровень холестерина (chol), артериальное давление (АД), частота сердечных сокращений (ЧСС) и т.д. Тогда, уравнение (1) примет иметь вид:

$$h(t \mid \mathbf{x}) = h_0(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{x}\right) =$$

$$= h_0(t)\exp\left(\beta_{\text{age}} \cdot \text{age} + \beta_{\text{gen}} \cdot \text{gen} + \beta_{\text{dia}} \cdot \text{dia} + \beta_{\text{chol}} \cdot \text{chol}+\ldots\right).$$

Условно, если коэффициент β_{age} (возраст) равен 0,05, это означает, что увеличение возраста на один год связано с увеличением риска смерти на 5%. А положительный коэффициент β_{gen} (пол) может показать, что мужчины имеют повышенный риск по сравнению с женщинами.

Использование регрессии Кокса при прогнозировании летальности пациентов после инфаркта миокарда позволяет учитывать множество совместно действующих факторов и формировать более точные персонализированные прогнозы. Это помогает в планировании лечения и мониторинга, улучшая результаты реабилитации и снижая риск повторных событий.

В классической постановке ковариаты $\mathbf{x}$ предполагаются фиксированными. Однако многие клинико-лабораторные показатели (АД, ЧСС, маркёры некроза и др.) изменяются в процессе лечения, особенно в первые дни после госпитализации. В этом случае рассматривают временно-зависимые ковариаты $\mathbf{x}_i(t)$, принимая их кусочно-постоянными на интервалах наблюдения:

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i(t)\right), \tag{2}$$

где индекс i – пациент, а $\mathbf{x}_i(t) \in \mathbb{R}^p$ обновляется при переходе к следующему временному окну (например, каждые 48 часов при плановых измерениях).

Пусть моменты событий упорядочены: $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$. Тогда для j-го события частичное правдоподобие имеет вид

$$P(\boldsymbol{\beta}) = \prod_{j=1}^{D} \frac{\exp\left(\boldsymbol{\beta}^{\top} \mathbf{x}_{i_j}\left(t_{(j)}\right)\right)}{\sum_{i \in R_j} \exp\left(\boldsymbol{\beta}^{\top} \mathbf{x}_i\left(t_{(j)}\right)\right)}, \tag{3}$$

где $\mathbf{x}_{i_j}\left(t_{(j)}\right)$ – вектор признаков пациента, у которого наступило событие в момент $t_{(j)}$; $R_j = \left\{i : T_i \geq t_{(j)}\right\}$ – риск-сет (пациенты, ещё «под риском» непосредственно перед $t_{(j)}$.

Далее $P(\boldsymbol{\beta})$ максимизируют, используя итерационные методы, например алгоритм Ньютона–Рафсона с пересчётом градиента и гессиана на каждом шаге. После получения оценки $\hat{\boldsymbol{\beta}}$ вычисляют индивидуальный риск нового пациента: $\hat{h}_i(t) = h_0(t)\exp\left(\hat{\boldsymbol{\beta}}^{\top}\mathbf{x}_i(t)\right)$. Если $\mathbf{x}_i(t)$ улучшается или ухудшается со временем (например, снижается маркёр некроза), это немедленно отражается в пересчёте $\hat{h}_i(t)$.

Итоговая кривая выживания пациента имеет вид

$$\hat{S}_i(t) = \exp\left(-\int_0^t \hat{h}_i(u)\,du\right). \tag{4}$$

Описанный подход позволяет учесть динамику состояния пациента на госпитальном этапе, когда показатели могут меняться ежедневно, что даёт более реалистичные прогнозы, поскольку факторы риска в острой фазе и восстановительной фазе могут заметно отличаться.

Проиллюстрируем на таком примере. Каждые 2 дня у пациента i регистрировали креатинин $\mathrm{Cr}_i(t)$. В классической модели Кокса $\mathbf{x}_i$ включал «креатинин на момент поступления» и «креатинин на момент выписки», не учитывая плавную динамику. Теперь же можно подставлять $\mathrm{Cr}_i(t)$ на каждом временном отрезке. Если $\beta_{\mathrm{Cr}} > 0$, рост креатинина в динамике свидетельствует об увеличении риска смерти, что напрямую вовлекается в расчёт.

Таким образом, получаем модель Кокса, учитывающую временно-зависимые факторы риска, что особенно ценно в острой фазе после ИМ.

2.2 Стратификация риска летальности методом случайного леса

Помимо регрессии Кокса, рассмотренной выше, для построения прогностических моделей всё активнее применяют алгоритмы машинного обучения [10–12]: деревья решений, случайные леса, машины опорных векторов, наивные байесовские классификаторы, искусственные нейронные сети и др. Ниже подробно остановимся на алгоритме случайного леса, который зарекомендовал себя как практичный и устойчивый инструмент в задачах клинического прогнозирования.

Случайный лес – это ансамблевый метод, объединяющий множество деревьев решений. Идея проста: отдельное дерево подвержено высокой дисперсии и легко переобучается, но усреднение большого числа «разнообразных» деревьев резко снижает разброс и повышает устойчивость прогноза. Благодаря этому случайный лес одинаково хорошо применим к широкому спектру задач (классификация, регрессия), способен улавливать нелинейности и взаимодействия признаков, устойчив к шуму и выбросам, работает с переменными разного типа и, как правило, требует минимум настройки.

Суть алгоритма заключается в следующем.

Пусть дано обучающее множество $D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, где x_i – вектор признаков i-го образца, а y_i – целевая переменная (метка класса или значение для регрессии).

Для каждого из B решающих деревьев создается новая выборка D_b путем бутстраппинга – случайного выбора N примеров из D с возвращением, то есть некоторые наблюдения могут повторяться в каждой выборке.

Далее выполняется построение решающего дерева для каждой бутстрап выборки D_b. На каждом узле дерева случайным образом выбирается подмножество признаков. Пусть m – количество выбранных признаков на каждом узле ($m \ll d$, где d – общее количество признаков). В отличие от

обычного построения деревьев, при каждом разбиении выбирается случайное подмножество признаков, и лучшее разбиение находится только среди этих признаков. Это увеличивает разнообразие в ансамбле.

Построение начинается с корневого узла, содержащего все примеры из D_b. Далее, для каждого узла выполняются следующие действия: выбирается случайное подмножество признаков $M \subset \{1,2,\ldots,d\}$ размера m. Для каждого признака $j \in M$ и порога t вычисляется прирост информации или снижение индекса Джини [13] в зависимости от задачи. Пусть L и R – множества индексов примеров, попадающих в левый и правый дочерние узлы после разделения: $L = \{i | \ x_{ij} \le t\}$, $R = \{i | \ x_{ij} > t\}$. Выбирается разделение (j^*, t^*), которое максимизирует прирост информации:

$$\Delta I = I(N) - \left(\frac{|L|}{|N|} I(L) + \frac{|R|}{|N|} I(R) \right), \tag{5}$$

где I – мера неопределенности (энтропия или индекс Джини).

Узел v становится листом дерева (терминальным узлом), если выполнено одно из следующих условий:

- глубина дерева достигла заранее заданного максимума $d_v \ge d_{\max}$, где d_v – текущая глубина узла v, а $d_{\max}$ – максимальная глубина дерева;

- количество примеров в узле меньше минимально допустимого значения $N_v < N_{\min}$, где N_v – количество примеров в узле v, а $N_{\min}$ – минимально допустимое количество примеров в узле;

- прирост информации меньше порогового значения $\Delta I < \Delta I_{\min}$, где ΔI – прирост информации при разделении узла v, а $\Delta I_{\min}$ – пороговое значение прироста информации.

После построения всех деревьев $\{T_b\}_{b=1}^{B}$ в случайном лесе, они комбинируются для предсказания.

При этом, для задачи классификации используется метод голосования большинства – каждое дерево «голосует» за класс, и класс с наибольшим количеством голосов становится предсказанным классом.

$$\hat{y} = \text{mode}\left(\{T_b(\mathbf{x})\}_{b=1}^{B} \right), \tag{6}$$

где $T_b(\mathbf{x})$ – предсказание b-го дерева для примера $\mathbf{x}$.

Для задачи регрессии используется усреднение предсказаний:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}). \tag{7}$$

После того как модель обучена, ее можно использовать для прогнозирования на новых данных, применяя описанные выше методы голосования большинства или усреднения.

Преимущества метода случайного леса заключаются в снижении риска переобучения за счет усреднения предсказаний множества деревьев; устойчивости к шуму за счет бутстрап-выборок и случайного выбора признаков; оценке важности признаков, что помогает понимать, какие признаки вносят наибольший вклад в предсказания модели.

В целом, алгоритм позволяет с достаточной степенью вероятности правильного прогноза идентифицировать больных с неблагоприятным течением ИБС, что в свою очередь дает возможность своевременно рекомендовать больному более агрессивное лечение и возможно инвазивные вмешательства либо ограничиться медикаментозным лечением при невозможности и нецелесообразности реализации инвазивных методов лечения.

Рассмотренный алгоритм классического случайного леса применялся в контексте классификации («умрёт в течение года / не умрёт»). Однако для задач анализа выживаемости (учитывающих цензуру, то есть ситуации, когда пациент не умер, но наблюдение закончилось) более подходящим выбором является алгоритм случайного леса выживания (англ. Random Survival Forest).

Допустим имеются N пациентов и для каждого известно: T_i – время наблюдения (до смерти или до окончания исследования); $\delta_i \in \{0,1\}$ – индикатор события (1, если наступила смерть; 0, если цензурирован); x_i – вектор признаков.

Наша цель – оценить функцию выживания $S_i(t)$ или вероятность «дожить до времени t, а также, при желании – кумулятивный риск смерти за определённый период.

Механизм построения случайного леса выживания, как и в классическом алгоритме, начинается с формирования B бутстрап-выборок D_b на каждой из которых строим отдельное дерево выживания.

В данном случае, при разделении узлов индекс Джинни или энтропия не используются, вместо этого берут, к примеру, лог-ранговый критерий, оценивающий различия во времени наступления события между левой и правой ветвями.

В листе ℓ собирается подмножество пациентов. На основе них можно оценить «локальную кривую выживания», например, по формуле Каплана–Мейера:

$$\hat{S}_\ell(t) = \prod_{\tau \le t} \left(1 - \frac{d_\ell(\tau)}{r_\ell(\tau)}\right), \tag{8}$$

где $d_\ell(\tau)$ – число событий (смертей) в момент τ среди пациентов листа ℓ; $r_\ell(\tau)$ – число ещё «выживающих» к моменту τ.

Для нового пациента x_i каждое дерево выживания даёт $\hat{S}_i^{(b)}(t)$ – оценку функции выживания, взятую из листа, куда попал x_i:

$$\hat{S}_i(t) = \frac{1}{B}\sum_{b=1}^{B} \hat{S}_i^{(b)}(t).$$

Отсюда легко получить вероятность смерти к моменту t как $1 - \hat{S}_i(t)$.

Таким образом, преимуществом данной модификации алгоритма случайного леса является учет цензурированных наблюдений (пациенты, которые не умерли, но время наблюдения закончилось); предоставление полной кривой выживания $\hat{S}_i(t)$, что полезно для оценки риска на разных горизонтах; нелинейное комбинирование ковариат, что позволяя учесть сложные взаимосвязи.

3 РЕЗУЛЬТАТЫ

3.1 Оценка риска по модели Кокса

Проиллюстрируем модель оценки выживаемости пациентов в течении ближайших трёх лет после эпизода ИМ на следующем ограниченном смешанном наборе реальных и синтетических данных (таб. 1), всего 1000 пациентов.

Таблица 1. Сокращённый образец набора признаков для обучения модели

Ковариаты	Пациенты						
Возраст	65	59	66	75	76	68	…
Пол	0	1	0	1	0	0	…
СД2	0	0	0	0	1	0	…
Гипертония	1	1	0	1	1	0	…
Холестерин	209	167	221	201	290	190	…
АД	99	92	144	92	134	139	…
ЧСС	51	70	86	59	77	64	…
ИМ в анамнезе	1	1	1	0	0	0	…
Курение	1	0	1	1	0	0	…
Прием препарата А	1	0	0	0	1	0	…
Прием препарата Б	1	0	1	1	1	1	…
Время наблюдения (мес.)	9,4	0,0	2,6	11,9	3,8	2,1	…

В результате расчетов получена нижеследующая таблица значений коэффициентов β с их статистической значимостью p и доверительными интервалами (ДИ).

Таблица 2. Значения коэффициентов регрессии, их значимость и ДИ

Ковариаты	β	p	e^{β}	95% ДИ ниж. гран.	95% ДИ верх. гран.
Возраст	-7,2E-05	0,987104	0,999928	0,991253	1,008679
Пол	-0,15147	0,091545	0,859442	0,72078	1,024779
СД2	0,106992	0,228534	1,112926	0,935048	1,324642
Гипертония	-0,04655	0,603261	0,954521	0,800851	1,137677
Холестерин	-0,00084	0,347676	0,999158	0,997402	1,000917
АД	0,000267	0,899797	1,000267	0,996116	1,004435
ЧСС	-0,00187	0,679982	0,998128	0,989282	1,007054
ИМ в анамнезе	0,263174	0,003409	1,301054	1,090919	1,551665
Курение	0,024331	0,784395	1,024629	0,860737	1,219727
Прием препарата А	0,004525	0,959462	1,004535	0,843708	1,196019
Прием препарата Б	0,157868	0,07856	1,171012	0,982136	1,396211

В таб. 2 значения коэффициентов β указывают на направление и силу эффекта ковариат на риск летальности: положительное значение повышает риск, отрицательное – снижает; p – значение указывает на статистическую значимость ковариаты – менее 0,05 считается значимым эффектом; e^{β_i} указывает на отношение риска летального исхода $h(t)$ при увеличении ковариаты на единицу, относительно базового риска $h_0(t)$.

Понимание общего тренда выживаемости, применительно к наблюдаемой группе пациентов (таб. 1) без привязки к конкретному пациенту можно получить путем построения обобщённой кривой выживаемости, которая отражает поведение функции выживания для «усреднённого» профиля пациента (рис. 1).

Что касается индивидуальных прогнозов, то обученная модель позволяет получать результат при вводе конкретного идентификатора пациента. Например, в данном случае для пациента № 1 получен следующий прогноз:

«*Индивидуальный прогноз риска для пациента № 1:* ***1,285207543812622***».

Подчеркнем, что значение индивидуального прогноза риска (за время наблюдения) интерпретируется относительно других пациентов в выборке, так как модель Кокса предсказывает относительный риск.

Сам риск делится на три категории относительно среднего: при < 0,75 – низкий риск; от 0,75 до 1,25 – средний; > 1,25 – высокий.

Следовательно значение 1,285 для пациента № 1 попадает в категорию высокого риска летального исхода в течении ближайших трех лет.

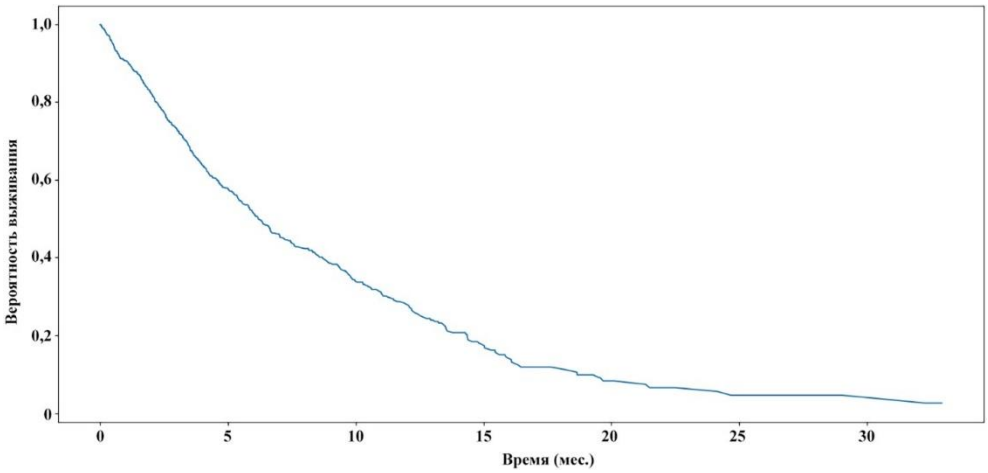

Рис. 1. Кривая выживаемости в среднем для наблюдаемой группы из 1000 пациентов

3.2 Оценка риска летальности методом случайного леса

Проиллюстрируем реализацию модели прогнозирования риска летального исхода больного в течении года после перенесенного ИМ на следующем ограниченном наборе реальных и синтетических данных:

Таблица 3. Сокращённый образец набора признаков для обучения модели

Возраст	55	67	48	72	60	…
Пол	м	ж	м	ж	м	…
СД2	нет	да	нет	да	да	…
Гипертония	да	да	нет	да	нет	…
Холестерин	норма	высок.	норма	высок.	повышен.	…
АД	120/80	140/90	118/75	150/95	130/85	…
ЧСС	75	85	80	90	88	…
ИМ в анамнезе	нет	де	нет	нет	да	…
Курение	нет	да	да	нет	да	…
Прием препарата А	да	да	нет	да	да	…
Прием препарата Б	нет	да	да	нет	да	…

На рис. 3 приведена гистограмма, демонстрирующая важность каждого из признаков, согласно рассмотренной модели случайного леса.

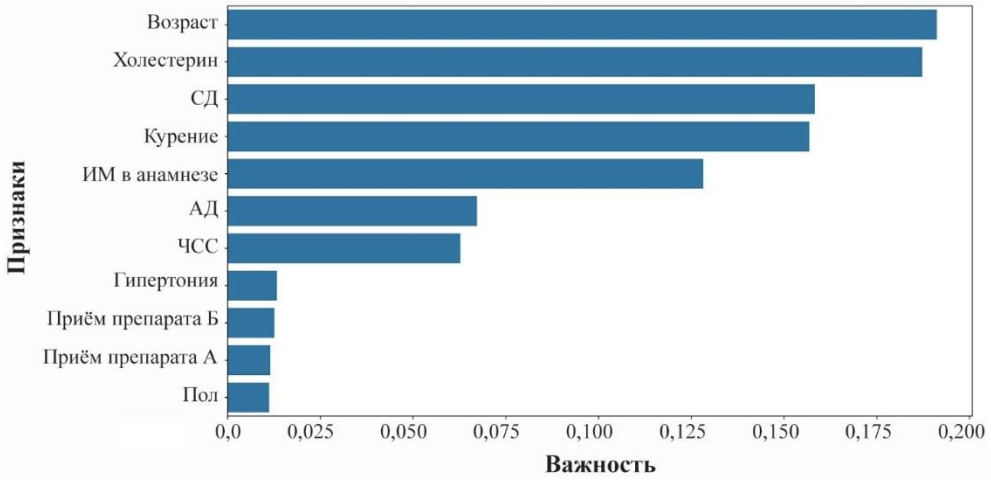

Рис. 2. Важность признаков по классификатору случайного леса

Сводка основных метрик классификации для оценки качества модели машинного обучения приведена в таб. 4, где «точность» – это доля правильных положительных прогнозов из всех прогнозов, классифицированных как положительные; «полнота» – это доля правильно предсказанных положительных примеров из всех фактических положительных примеров.

Таблица 4. Оценка качества модели

Точность: 0.99	Точность	Полнота	F1	Поддержка
0	0,90	0,80	0,83	158
1	0,90	0,90	0,83	142
Точность для всех классов			0,81	300
Макросреднее	0,85	0,85	0,81	300
Средневзвешенное	0,85	0,85	0,81	300

F1 или среднее гармоническое точности и полноты – помогает понять баланс между точностью и полнотой, что особенно полезно, если распределение классов неравномерное. «Поддержка» – количество фактических случаев в каждом классе. Показывает, сколько примеров каждого класса было в тестовом наборе данных, что важно для понимания представленности классов. «Точность для всех классов» – общая доля правильных предсказаний из всех прогнозов. Обычно указывается общая точность классификации по всем классам. «Макросреднее» – среднее значение метрик, рассчитанное отдельно по каждому классу без учета веса класса (поддержки). Наконец, «средневзвешенное» – среднее значение метрик, рассчитанное отдельно по каждому классу с учетом их поддержки (количество примеров в каждом классе).

Согласно требованиям методологии в клинических исследованиях, оценка эффективности разработанной модели требует апробации на отдельной независимой выборке больных (группе экзамена) для подтверждения результатов, полученных у пациентов (группы обучения). Группа экзамена была сформирована на основе базы данных отделения неотложной кардиологии РСНПМЦК из пациентов, находящихся под наблюдением научного отдела лаборатории ОИМ сопоставимая с группой обучения.

Для оценки валидности разработанного алгоритма у пациентов «группы экзамена» была сопоставлена структура прогнозных заключений, полученных с использованием разработанного алгоритма, со структурой оценочных заключений реальных исходов, полученных в результате наблюдения пациентов в течении 1 года. После оценки прогнозных заключений с использованием алгоритма распределение больных оказалось следующим: прогноз при благоприятном течении – у 124 (82,6%), прогноз при неблагоприятном течении (включая больных, нераспознанных алгоритмом) – у 26 (17,3%), число больных, не распознанных алгоритмом, составило 12 (8,0%).

4 ЗАКЛЮЧЕНИЕ

В заключение сформулируем основные результаты.

Подробно рассмотрены и сопоставлены два взаимодополняющих подхода к прогнозированию летальности пациентов после инфаркта миокарда: полупараметрическая регрессионная модель Кокса (включая постановку с временно-зависимыми ковариатами) и ансамблевый метод на основе случайного леса, адаптированный к задачам анализа выживаемости. Показано, что сочетание интерпретируемости первой и робастности второго расширяет спектр клинически значимых сценариев применения.

Предложена и реализована модель пропорциональных рисков Кокса для прогноза летальности у перенёсших ИМ. Использование этой модели позволяет учитывать целый комплекс факторов риска и получать более точные, персонализированные оценки, поддерживая выбор стратегии ведения и мониторинга пациента. Тем самым модель способствует улучшению результатов реабилитации и снижению вероятности повторных неблагоприятных событий.

Разработан алгоритм стратификации риска летального исхода на основе случайного леса в постановке выживаемости. Ансамбль демонстрирует достаточную точность при гетерогенности профилей риска и коморбидности (например, на фоне СД2), что даёт основания для своевременных клинических рекомендаций – от инвазивных вмешательств до оптимизации медикаментозной терапии – с учётом индивидуальной кривой риска во времени.

Эксперименты показали, что расширенная регрессия Кокса с временными рядами (АД, маркёры некроза и др.) повышает точность и калибровку оценок по сравнению со статической постановкой. Алгоритм случайного леса для выживаемости, в свою очередь, снижает риск переобучения, устойчив к нелинейностям и взаимодействиям признаков и позволяет выявлять наиболее значимые предикторы. Практическая апробация на реальной клинической базе продемонстрировала высокие значения метрик и удовлетворительную калибровку; частично высокий уровень показателей объясняется упрощением выборок и добавлением синтетических данных, о чём отдельно оговорено.

Таким образом, комбинированное применение регрессии Кокса и случайного леса выживания обеспечивает более точную стратификацию пациентов с ИМ по степени риска и поддерживает своевременное принятие решений о тактике лечения и реабилитации, что в конечном счёте способствует снижению смертности и улучшению прогноза в реабилитационный период.

ЛИТЕРАТУРА

[1] *Фирюлина М.А.* Анализ показателей смертности Воронежской области в сравнении с развитыми странами // Вестник Воронежского института высоких технологий. – 2018. – № 2(25). – С. 150-153.

[2] *Фирюлина М.А.* Прогнозирование риска смертности после инфаркта миокарда с использованием методов машинного обучения // Информатика: проблемы, методы, технологии : материалы XXI междунар. науч.-метод. конф. – Воронеж: Вэлборн, 2021. – С. 1535-1544.

[3] *Ad N. et al.* Comparison of EuroSCORE II, Original EuroSCORE, and The Society of Thoracic Surgeons Risk Score in Cardiac Surgery Patients // The Annals of Thoracic Surgery. – 2016. – Vol. 102, Issue 2. – P. 573-579.

[4] *Alimova D. et al.* Prediction of diastolic dysfunction in patients with cardiovascular diseases and type 2 diabetes with respect to Covid-19 in anamnesis using artificial intelligence // ICMHI '23: Proc. of the 7th int. conf. on medical and health informatics. – Kyoto, 2023. – P. 61-65.

[5] *Сови А.* Старение населения и продление жизни // Методы демографических исследований. – М.: Статистика, 1989. – С. 48-58.
[6] *Медик В.* Математическая статистика в медицине. – М.: Финансы и статистика, 2007. – 800 с.
[7] *Гланц С.* Медико-биологическая статистика. – М.: Практика, 1999. – 459 с.
[8] *Дуброва Т.А*. Статистические методы прогнозирования. – М., 2003. – 206 с.
[9] Математика, статистика, экономика на компьютере / А.В. Каплан и др. – М.: ДМК Пресс, 2006. – 600 с.
[10] *Kilic A. et al.* Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery // The Annals of Thoracic Surgery. –2020. – Vol. 109, Issue 6. – P. 1811-1819.
[11] *Allyn J. et al.* A Comparison of a Machine Learning Model with EuroSCORE II in Predicting Mortality after Elective Cardiac Surgery: A Decision Curve Analysis // PLOS ONE. – 2017. – Vol. 12. – DOI: 10.1371/ journal.pone.0169772.
[12] *Гельцер Б.И. и др.* Методы машинного обучения в прогнозировании летальных исходов в стационаре у больных ишемической болезнью сердца после коронарного шунтирования // Кардиология. – 2020. – Т. 60, № 10. – С. 38-46.
[13] *Минашкин В.Г.* Джини коэффициент // Большая российская энциклопедия. Том 8. – Москва, 2007. – С. 661.

Поступила в редакцию 26.06.2025

Цитирование: *Пекось О.А.* (2025). Оценка выживаемости пациентов с инфарктом миокарда методами машинного обучения. *Международный журнал теоретических и прикладных вопросов цифровых технологий*, 8(4), –С. 48-57. https://doi.org/10.62132/ijdt.v8i4.302.

ASSESSMENT OF MYOCARDIAL INFARCTION PATIENT SURVIVAL USING MACHINE LEARNING METHODS

Pekos O.A.

Digital Technologies and Artificial Intelligence Development Research Institute, Tashkent, Uzbekistan

Abstract. High mortality in the acute and post-hospital periods of myocardial infarction underscores the relevance of accurately stratifying patient risk and assessing survival outcomes. This study compares two complementary approaches: the Cox regression model, including time-dependent covariates, and an ensemble random forest method for binary classification and survival analysis. It is shown that accounting for the dynamics of clinical and laboratory indicators (blood pressure, heart rate, necrosis markers, etc.) improves both the discrimination and calibration of prognostic models. Variants of the instantaneous hazard equation, partial likelihood formulation, and formulas for estimating survival and cumulative hazard are presented. Illustrative examples based on a mixed dataset of real and synthetic data demonstrate the advantages of the combined approach in the presence of censoring and heterogeneous risk profiles. The results confirm the feasibility of integrating such models into the clinical workflow to personalize treatment and rehabilitation strategies after myocardial infarction.

Keywords: machine learning, survival analysis, mortality prediction, Cox model, survival random forest.