Прогнозирование риска смертности после инфаркта миокарда с использованием методов машинного обучения

М. А. Фирюлина, email: mashafiryulina@mail.ru
И. Л. Каширина, email: kash.irina@mail.ru

Воронежский государственный университет

Аннотация. *В данной статье представлены результаты прогнозирования риска смертности после инфаркта миокарда. В ходе исследования использовалось несколько методов машинного обучения: регрессия Кокса, логистическая регрессия, градиентный бустинг Catboost, а также статистический метод Каплана-Мейера для определения наиболее значимых факторов влияния на выживаемость после инфаркта миокарда. По итогам сравнения метрик качества наиболее эффективной является модель градиентного бустинга, однако регрессия Кокса показала лучшую интерпретируемость.*

Ключевые слова: *анализ выживаемости, градиентный бустинг, логистическая регрессия, модель Кокса, метод Каплана-Мейера.*

Введение

Одним из наиболее тяжелых осложнений ишемической болезни сердца является инфаркт миокарда (ИМ). Случаи, связанные с сердечно-сосудистыми заболеваниями, отличаются высоким уровнем смертности. Риск летального исхода остается высоким не только в первые дни после наступления ИМ, но и в более поздние сроки [1]. Адекватная оценка риска смертности способствует своевременному принятию терапевтических методов для улучшения состояния пациента.

Большое количество исследований посвящено оценке прогноза жизни больных, перенесших ИМ, и влияющих на него факторов. Однако итоговые результаты могут отличатся, по причине различий социально-экономических, географических показателей региона [2,3]. Некоторые исследования основаны на использовании шкалы GRACE – Global registry of acute coronary events. В шкале GRACE оценивают клинические показатели пациентов, согласно сумме баллов определяется низкий, средний или высокий риск летального исхода пациентов с ИМ [3,4].

Цель исследования, описанного в данной статье, заключается в построении эффективной модели машинного обучения для прогнозирования риска смертности после ИМ. Для этого было

построено три модели: регрессия Кокса, логистическая регрессия, градиентный бустинг Catboost, и оценена точность каждой модели, для выявления наиболее эффективной. Также проведен анализ по выявлению наиболее значимых клинических факторов, влияющих на оценку риска смертности после инфаркта миокарда.

Анализ проведен на основе данных о зафиксированных случаях инфаркта миокарда по Воронежской области за 2015-2017 года. Выводы сделаны на основе результатов на основе общей статистики за три года. Предобработка данных проводилась с помощью СУБД Oracle 19c в среде разработки SQL Developer. Построение моделей машинного обучения, статистический анализ данных, построение графических материалов проводилось с помощью различных библиотек языка Python на платформе Goole Colab.

1. Материалы и методы

Для анализа использовалась выборка пациентов, поступивших за 2015-2017 года в больницы Воронежской области с диагнозом ИМ. Вся информация с неперсонифицированными данными пациентов была предоставлена областным кардиологическим диспансером ОКБ№1. Также для получения наиболее достоверных результатов исходная выборка была дополнена информаций о зарегистрированных смертельных случаях после выписки пациентов на основе данных по всем смертельным случаям в Воронежской области. Случаи летального исхода через несколько дней после выписки важны для анализа [5]. Всего в исследовании было рассмотрено 11457 случая инфаркта миокарда, из них 2025 (17.7%) случаев с летальным исходом (в течение года) и 9432 (82.3%) – выжившие пациенты.

Анализ проводился по следующим факторам: пол, возрастная группа, артериальная гипертензия (АГ), является ли инфаркт миокарда повторным (ИМ), сахарный диабет (СД), фибрилляция предсердий (ФП), острое нарушение мозгового кровообращения (ОНМК), хроническая обструктивная болезнь легких (ХОБЛ), хроническая сердечно-сосудистая недостаточность (ХСН), локализация, тяжесть по KILLIP и проводилась ли пациенту тромболитическая терапия (ТЛТ) и чрескожные коронарные вмешательства (ЧКВ).

Ранее авторами проводилось подобное исследование, в ходе чего были выявлены наиболее значимые факторы для оценки риска смертности с помощью метода Каплана-Мейера [6]. В предыдущем исследовании горизонт прогнозирования смертности после ИМ составлял год, но было обнаружено, что наблюдение динамики свыше 20 дней приводит к растущей неточности результатов, поэтому было решено ограничиться 20 днями.

Для начала был проведен первичный статистический анализ данных с помощью метода Каплана-Мейера. Метод множительных оценок Каплана-Мейера производит оценку функции выживаемости, основываясь на времени выживания для полных и цензурированных данных. В нашем исследовании данные по тем пациентам, которые выписались ранее 20 дней, считаются цензурированными, а которые умерли – полные. Для полных данных – время жизни пациента – это число дней от наступления ИМ до смерти, для цензурированных – это число дней от ИМ до выписки. Для оценки значимости признаков использовался показатель p – уровень значимости предиктора. Чтобы определить имеется ли различие в рассматриваемых группах необходимо воспользоваться критерием Гехана-Вилкоксона. Сравнение выживаемости в группах проводится при уровне значимости критерия p = 0.05. Если p < 0.05, то верна гипотеза о различии выживаемости в группах, если p > 0.05, то верна альтернативная гипотеза – нет существенной разницы в выживаемости больных в группах.

С помощью метода Каплана-Мейера можно отобрать значимые признаки для построения моделей прогнозирования смертности. В данном исследовании сравниваются три модели: регрессия Кокса (традиционный подход), логистическая регрессия и модель градиентного бустинга.

Регрессия Кокса – это метод, который предполагает прогнозирование риска наступления события для рассматриваемого объекта и оценивает влияние независимых переменных на этот риск. При этом риск наступления события является функцией, зависимой от времени, и выявляет вероятность наступления события для объектов, которые находятся в группе риска. Данный метод отличается также возможностью работы с цензурированными данными. Основное преимущество регрессии Кокса для оценки рисков состоит в том, что для моделирования ситуации по данному методу требуется меньшее количество наблюдений, чем для многих других методов. При этом данная модель обладает высокой точностью оценки пропорциональных рисков [8].

Логистическая регрессия – это один из наиболее популярных методов линейной бинарной классификации. Результатом предсказания данного метода является вероятность того, что входной объект принадлежит к определенному классу. Это свойство важно в медицинских приложениях, где наряду с классификацией объекта требуется оценить связанный с этим риск ошибочной классификации.

Градиентный бустинг – это техника машинного обучения, основная идея которой заключается в итеративном процессе последовательного

построения частных моделей, которые используют информацию об ошибках на предыдущем этапе обучения [9]. Данный алгоритм выделяется высокой точностью, в большинстве случаев превосходящей точность остальных методов. В данном исследовании рассматривалась модель градиентного бустинга CatBoost. Это довольно новый метод, который отличается высокой производительностью и точностью, а также способностью обрабатывать категориальные переменные.

Для сравнения различных моделей прогнозирования смертности использовалась метрика AUC_ROC. Величина AUC ROC – площадь под ROC - кривой является компромиссной метрикой, широко применяемой в медицинских исследованиях.

2. Результаты и их обсуждение

На первом этапе был построен график оценочной функции выживания. При анализе выживаемости, как и при других методах статистического анализа, вся информация о выборке содержится в соответствующей ей функции распределения вероятности времен жизни, но используется она в виде функции выживания (survival function). Построение функции Каплана-Мейера производилось с помощью модуля KaplanMeierFitter библиотеки lifelines языка Python. На рис. 1 представлен график функции выживаемости. По рис. 1 видно, что функция резко убывает до 5 дней от момента наступления инфаркта миокарда, затем убывание замедляется. Можно сделать вывод, что данный период является наиболее опасным.

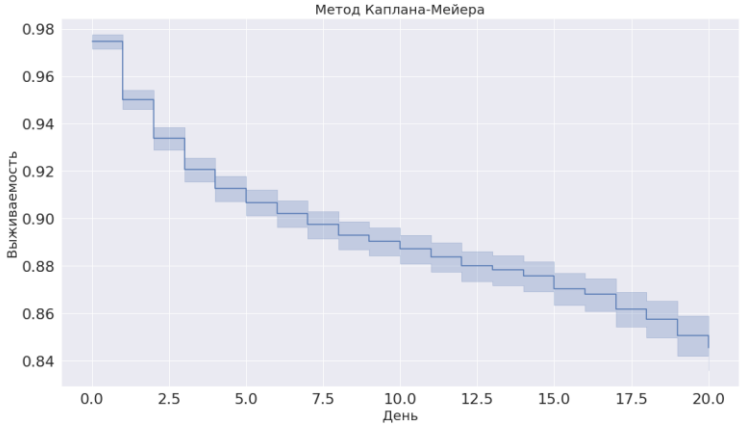

Рис. 1. График функции выживания

С помощью расчетной функции выживаемости Survival_function_ библиотеки KaplanMeierFitter было определено процентное соотношение выживших пациентов в определенный день.

Например в день наступления ИМ доля выживших пациентов в выборке составляет 97.5 %, а на 20-ый день всего – 85 %.

Так как часто анализ выживаемости рассматривается относительно гендерного распределения, на рис. 2 представлен график выживаемости для мужчин и женщин отдельно. На рис. 2 видно, что уровень выживаемости женщин ниже, чем у мужчин.

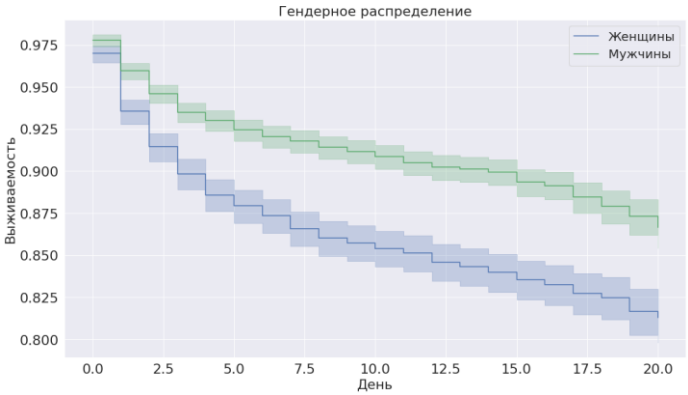

Рис. 2. График функции выживания для мужчин и женщин

Наиболее значимым признаком влияющим на риск смертности от инфаркта миокарда является оценка тяжести пациента по шкале KILLIP.

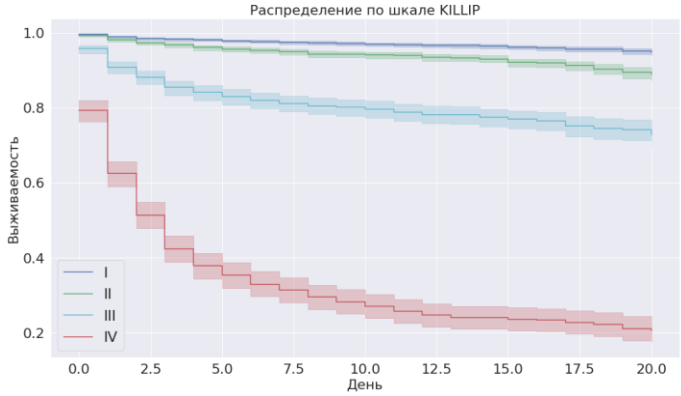

Рис. 3. График функции выживания по показателю KILLIP

Во время госпитальной терапии состояние пациента оценивается по 4-балльной шкале в зависимости от результатов физикального исследования, IV – наиболее тяжелое состояние. На рис. 3 изображен график выживаемости относительно каждого класса тяжести по KILLIP. Ожидаемо, что наибольший риск летального исхода у пациентов с высокой степенью тяжести.

Следующим этапом была построена регрессионная модель Кокса. Построение регрессионной модели Кокса производилось с помощью модуля CoxPHFitter библиотеки lifelines.

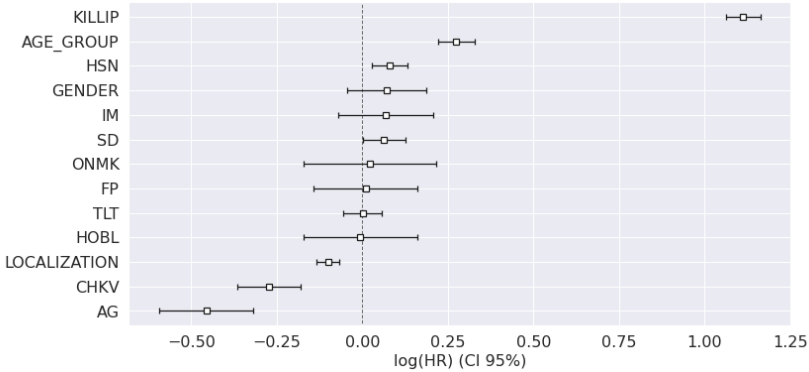

Рис. 4. Результаты модели Кокса

На рис. 4 показаны результаты модели Кокса. Отношения рисков представлены в виде черных прямоугольников, а полосы доверительного интервала - в виде усов. Центральная вертикальная линия указывает на базовый уровень риска 1. По рис 4. можно сделать выводы что отсутствие артериальной гипертензии снижает риск смертности после ИМ почти, а высокий класс тяжести по шкале KILLIP существенно повышает риск. Такие показатели, как хроническая обструктивная болезнь легких и проводилась ли пациенту тромболитическая терапия на результаты на влияют.

Модели машинного обучения, градиентный бустинг и логистическая регрессия, были построены с помощью модулей CatBoostClassifier библиотеки catboost и LogisticRegression библиотеки scikit-learn соответственно. Для оценки полученных моделей логистической регрессии и градиентного бустинга использовались стандартные метрики оценки качества: Accuracy, чувствительность, специфичность и AUC_ROC. При построении моделей исходная выборка разбиралась на тестовое и обучающее множества в

соотношении 20:80. Метрики качества рассчитаны на тестовой выборке. На основе элементов матрицы ошибок классификации для каждого класса были рассчитаны показатели: (TP) - True Positive (число верно предсказанных примеров класса 1), (FN) - False Negative (число ложноотрицательных примеров, неверно предсказанный класс 0), (TN) - True Negative (верно предсказанный класс 0), (FP) - False Positive (число ложноположительных примеров, неверно предсказанный класс 1). Основная метрика задач классификации – Accuracy (доля правильных ответов). Так как исходная выборка не сбалансирована (доля выживших пациентов существенно выше доли умерших), эта метрика недостаточно показательно, поэтому помимо нее были рассмотрены метрики: чувствительность (Sensitivity) – доля истиноположительных примеров от общего числа положительно предсказанных примеров, и специфичность (Specificity) – доля правильно классифицированных объектов негативного класса. В табл. 1 представлены полученные результаты. Прогнозируемая переменная– выживет ли пациент в течение 20 дней (1- умрет, 0 – выживет), в качестве входных параметров использовались исходные клинические показатели.

Таблица 1

Метрики качества моделей машинного обучения

Метрика	**Градиентный бустинг (Catboost)**	**Логистическая регрессия**
Accuracy	0.89	0.86
Специфичность	0.97	0.96
Чувствительность	0.53	0.4
AUC_ROC	0.842	0.804

Для оценки построенных моделей на рис. 5 изображены графики ROC-кривых для моделей градиентный бустинг, логистическая регрессия и регрессия Кокса. По графикам видно, что наилучшую точность показывает модель градиентного бустинга, а хуже всего прогноз дает модель Кокса.

В рамках задач медицинского характера важна не только точность построенной модели, но и интерпретируемость. Врачам недостаточно использовать систему как «черный ящик», важно понять почему модель предсказала определенный результат. Модель Кокса позволяет более детально анализировать влияние нескольких факторов на результат исследования. Поэтому одна из задач исследования состоит в том, чтобы построить модель, которая достаточно сложная, чтобы иметь высокую точность прогнозирования, но в то же время достаточно простая, чтобы

обеспечить медицинскую интерпретацию. Логистическая регрессия и регрессионная модель Кокса, несмотря на чуть более низкую точность, больше подходят для достижения поставленной цели.

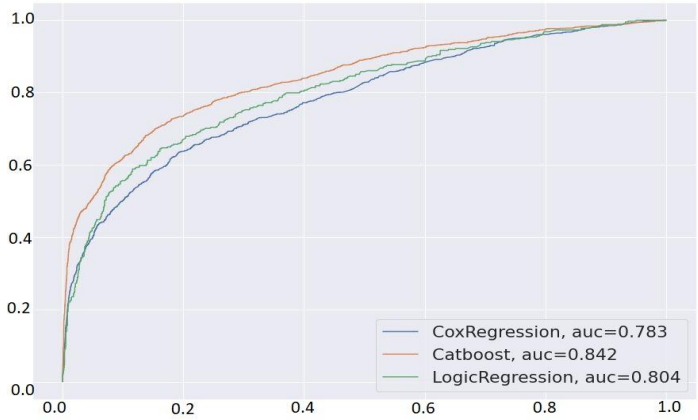

Рис. 5. ROC-кривые для моделей

Таблица 2

Значимость предикторов

метод Каплана-Мейера		**модель Кокса**	
предиктор	*p-value*	*предиктор*	*p-value*
KILLIP	0	KILLIP	0
Локализация	6.29e-26	Возраст	1.12e-23
АГ	7.17e-25	АГ	1.05e-10
ЧКВ	1.58e-29	ЧКВ	8.31e-09
ТЛТ	1.33e-03	ХСН	3.33e-03
градиентный бустинг		**логистическая регрессия**	
предиктор	*fea_imp*	*предиктор*	*p-value*
KILLIP	0.1825	KILLIP	0
Возраст	0.1718	Возраст	0
Локализация	0.1342	СД	0.001
ХСН	0.1185	ХСН	0.001
Пол	0.0643	Локализация	0.001

Для каждого построенного метода и модели были выделены наиболее значимые признаки. Для метода Каплана-Мейера, модели

Кокса и логистической регрессии отбор признаков производился на основе показателя p-value (чем меньше этот показатель, тем более значим признак). Для градиентного бустинга рассчитана важность предикторов feature importance (в данном случае, чем важнее признак, тем выше показатель). Результаты пяти наиболее значимых факторов приведены в табл. 2.

Заключение

В ходе данного исследования было построено несколько моделей машинного обучения для прогнозирования смертности после инфаркта миокарда. Наиболее критичный период составляет первые пять дней после ИМ. Точнее всего риск смертности прогнозирует модель градиентного бустинга Catboost, относительно модели Кокса и логистической регрессии, однако эти методы позволяют лучше объяснять получаемые прогнозы. Также были проанализированы исходные предикторы на оценку влияния прогнозирования смертности от ИМ. Наиболее значимы показатели KILLIP, возраст и локализация.

Благодарности

Выражаем благодарность РФФИ за поддержку проекта 20-37-90029 Аспиранты (грант 20-37-90029).

Список литературы

1. Фирюлина, М. А. Анализ показателей смертности Воронежской области в сравнении с развитыми странами / М. А. Фирюлина, И. Л. Каширина // Вестник Воронежского института высоких технологий. – 2018. – № 2 (25). – С. 150-153.
2. Pieszko, K. Predicting Long-Term Mortality after Acute Coronary Syndrome Using Machine Learning Techniques and Hematological Markers / K. Pieszko, J. Hiczkiewicz // Disease Markers. – 2019. – 1. – С. 1-10.
3. Bhatt, D. L. Comparative Determinants of 4-Year Cardiovascular Event Rates in Stable Outpatients at Risk of or With Atherothrombosis / D. L. Bhatt, K. A. Eagle, E. M. Ohman // Elsevier - Journal of Vascular Surgery. – 2011. – № 12. – С. 1350-1357.
4. Хоролец, Е. В. Прогноз пациентов с острым инфарктом миокарда на госпитальном этапе лечения / Е. В. Хоролец, С. В. Шлык // Современные проблемы науки и образования. – 2019. – № 1. – С. 1-11.
5. Каширина, И. Л. Прогнозирование развития инфаркта миокарда на основании анализа метеорологических факторов и данных областного регистра / И. Л. Каширина, Р. А. Хохлов, А. О. Казакова //

Вестник Воронежского государственного университета. – 2016. – № 3. – С. 116-123.

6. Фирюлина, М. А. Анализ значимости предикторов выживаемости после инфаркта миокарда с помощью метода Каплана-Мейера / М. А. Фирюлина, И. Л. Каширина, Е. Я. Гафанович // Моделирование, оптимизация и информационные технологии. – 2019. – № 7. – С. 7-20.

7. W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer, Detecting Hip Fractures with Radiologist-Level Performance Using Deep Neural Networks [Электронный ресурс] : многопредмет. науч. журн. – Режим доступа : https://arxiv.org/abs/1711.06504

8. Шарашова, Е. Е. Применение регрессии Кокса в здравоохранении с использованием пакета статистических программ SPSS / Е. Е. Шарашова, К. К. Холматова, М. А. // Наука и Здравоохранение. – 2017. – № 6. – С. 5-27

9. Shitikov V.K., Mastitsky S.E. Classification, regression, Data Mining algorithms using R. [Электронный ресурс] : многопредмет. науч. журн. – Режим доступа : https://github.com/ranalytics/data-mining/