Идентификация факторов риска смертности после инфаркта миокарда с использованием методов машинного обучения

И. Л. Каширина[1], М. А. Фирюлина,
Ю. В. Бондаренко
Воронежский государственный университет
[1]kash.irina@mail.ru

Е. Н. Десятирикова[1], О. Е. Ефимова
Воронежский государственный технический университет
[1]science2000@ya.ru

Л. В. Черненькая
Санкт-Петербургский политехнический университет Петра Великого
mv@qmd.spbstu.ru

***Аннотация*. В статье представлены результаты исследования по прогнозированию риска смертности после инфаркта миокарда по деперсонифицированным данным, включающим описание 11457 случаев инфарктов в Воронежской области в 2015–2017 годы. Анализируется несколько методов прогнозирования риска: модель Каплана-Мейера, модель Кокса, логистическая регрессия, модель градиентного бустинга. Для определения значимых факторов, влияющих на выживаемость после инфаркта миокарда, рассматривается несколько методов определения важности признаков, проводится их сравнительный анализ. Результаты исследования демонстрируют определяющее влияние тяжести инфаркта по шкале KILLIP, возраста пациента, проведенного чрескожного коронарного вмешательства и наличия в анамнезе пациента артериальной гипертензии на краткосрочный 20-дневный прогноз выживаемости после инфаркта миокарда.**

Ключевые слова: анализ выживаемости; градиентный бустинг; логистическая регрессия; модель Кокса; метод Каплана-Мейера; значимость признаков

I. Введение

Как известно, сердечно-сосудистые заболевания являются ведущей причиной смертности населения в России [1]. Данное исследование посвящено наиболее тяжелому осложнению ишемической болезни сердца – инфаркту миокарда (ИМ). Отличительная черта ИМ в том, что риск смертности остается высоким не только в первые дни после наступления ИМ, но и на более поздних сроках. Актуальность исследования заключается в том, что адекватная оценка риска смертности способствует своевременному принятию терапевтических методов для улучшения состояния пациента. Оценке прогноза жизни больных, перенесших ИМ, и влияющих на него факторов посвящено уже большое количество исследований [2, 3]. Однако итоговые результаты могут отличаться, по причине различий социально-экономических, географических показателей регионов. [3]

Цель предлагаемого исследования заключается в построении эффективной модели для прогнозирования риска смертности после ИМ и ее наглядной интерпретации, за счет выделения наиболее значимых факторов, влияющих на смертность, и оценке характера этого влияния. Анализ проведен на основе данных о зафиксированных случаях инфаркта миокарда по Воронежской области. 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 Google Colab. В исследовании использовалась статистическая модель прогнозирования риска Каплана-Мейера и три модели машинного обучения: регрессия Кокса, логистическая регрессия, градиентный бустинг Catboost. Высокая точность результатов была достигнута за счет балансировки данных, в ходе исследования проведен детальный анализ по выявлению наиболее значимых клинических факторов, влияющих на оценку риска смертности после инфаркта миокарда.

II. Материалы и методы

A. Описание исходных данных

Для анализа использовалась выборка пациентов, поступивших за 2015–2017 года в больницы Воронежской области с диагнозом ИМ. [4]. Всего в исследовании было рассмотрено 11457 случая инфаркта миокарда, из них 2025 (17.7 %) случаев с летальным исходом 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

Работа выполнена при финансовой поддержке РФФИ, проект 20-37-90029 Аспиранты (грант 20-37-90029).

болезнь легких (ХОБЛ), хроническая сердечно-сосудистая недостаточность (ХСН), локализация, тяжесть по KILLIP и проводилась ли пациенту тромболитическая терапия (ТЛТ) и чрескожные коронарные вмешательства (ЧКВ).

Процентное соотношение значений категориальных переменных представлены в табл. 1. Единственная числовая переменная в исходной выборке – возраст. Средний возраст пациентов 66 ± 12 лет. Для упрощения интерпретации полученных результатов возраст пациентов был разбит на 8 категорий.

ТАБЛИЦА I РАСПРЕДЕЛЕНИЕ КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ

Признак	**Выжившие (%)**	**Умершие (%)**
Пол		
Мужской	6015 (63.8 %)	1049 (51.8 %)
Женский	3417 (36.2 %)	976 (48.2 %)
ИМ(повторный)		
Нет	8204 (87.0 %)	1606 (79.3 %)
Да	1228 (13.0 %)	419 (20.7 %)
ОНМК		
Нет	8932 (94.7 %)	1852 (91.5 %)
Да	500 (5.3 %)	173 (8.5 %)
ХСН		
Нет	4222 (44.8 %)	818 (40.4 %)
Н I	1172 (12.4 %)	91 (4.5 %)
Н IIA	3783 (40.1 %)	900 (44.4 %)
Н III	255 (2.7 %)	216 (10.7 %)
Локализация		
Не определена	397 (4.2 %)	73 (3.6 %)
Передней стенки	4254 (45.1 %)	963 (47.6 %)
Заднебазальный	399 (4.2 %)	104 (5.1 %)
Боковой стенки	479 (5.1 %)	77 (3.8 %)
Нижней стенки	2773 (29.4 %)	548 (27.1 %)
Переднебоковой	1130 (12.0 %)	260 (12.8 %)
ТЛТ		
Не проводилось	8186 (86.8 %)	1775 (87.7 %)
Актилизе	538 (5.7 %)	97 (4.8 %)
Пуролаза	428 (4.5 %)	82 (4.0 %)
Метализе	280 (3.0 %)	71 (3.5 %)
Артериальная гипертензия		
Нет	1893 (20.1 %)	365 (18.0 %)
Да	7539 (79.9 %)	1660 (82.0 %)
Фибриляция предсердий		
Нет	8585 (91.0 %)	1683 (83.1 %)
Да	847 (9.0 %)	342 (16.9 %)
ХОБЛ		
Нет	8751 (92.8 %)	1780 (87.9 %)
Да	681 (7.2 %)	245 (12.1 %)
ЧКВ		
Нет	8643 (91.6 %)	1977 (97.6 %)
БАП	223 (2.4 %)	20 (1.0 %)
ЧКВ голаметал. стент	177 (1.9 %)	4 (0.2 %)
ЧКВ стент с покр-ем	389 (4.1 %)	24 (1.2 %)
Киллип		
Нет	391 (4.1 %)	66 (3.3 %)
I	5025 (53.3 %)	435 (21.5 %)
II	2907 (30.8 %)	496 (24.5 %)
III	936 (9.9 %)	415 (20.5 %)
VI	173 (1.8 %)	613 (30.3 %)
Сахарный диабет		
Нет	8214 (87.1 %)	1631 (80.5 %)
I-типа	162 (1.7 %)	68 (3.4 %)
II-типа	1056 (11.2 %)	326 (16.1 %)

B. Методы оценки значимости признаков

Одна из основных проблем при решении задач медицинского характера – интерпретируемость. В сфере медицины недостаточно предоставить модель как «черный ящик», необходимо объяснить полученные результаты. Для этого используются графическое представление результатов, построение деревьев принятия решений, а также выделяются признаки, которые имеют наибольшее влияние на итоговый результат. Проведя анализ значимости признаков, можно повысить точность итоговой модели, убрав из модели признаки, которые не влияют на результат. Кроме того, если определить, какие факторы влияют на риск смертности каждого пациента, можно напрямую учитывать эти факторы риска и корректировать терапевтические назначения [5]. В ходе данного исследования анализ значимости признаков проводился для всех построенных моделей. Для метода Каплана-Мейера, модели Кокса и логистической регрессии оценка значимости признаков осуществлялась с использованием p-уровня значимости. Если $p<0.05$, то верна гипотеза о влиянии данного признака на смертность, если $p>0.05$, то верна альтернативная гипотеза – нет существенного влияния. В целом, чем меньше этот показатель, тем более значим признак. В модели градиентного бустинга для оценки значимости признаков предусмотрено несколько альтернативных методов, представленных далее.

1. Метод анализа PredictionValuesChange показывает, насколько в среднем изменяется выходная переменная при изменении значения данного входного признака. Чем больше значение важности, тем сильнее эта переменная будет влиять на прогноз. Метод отличается высокой производительностью и не требователен к ресурсам, однако он может присвоить высокую важность признакам, которые мало влияют на величину итоговой метрики качества модели (функции потерь).

2. Принцип работы метода LossFunctionChange заключается в том, что для получения значимости признака сравнивается значение функции потерь модели с использованием данного признака и без него. Чем больше разница, тем важнее характеристика. Данный метод хорошо работает на различных типах наборов данных, но требует больше ресурсов.

3. Метод Interaction показывает насколько влияет на качество модели взаимодействие различных пар признаков, этот метод самый ресурсозатратный.

4. SHAP – значения важности показывают, насколько выбранный признак изменил результат прогнозирования (по сравнению с тем, как отработала бы модель при некотором базовом значении этого признака) [6]. Данный метод можно изобразить схематически (рис. 1). На основе исходных данных строится модель, затем составляется прогноз для одного объекта и с помощью функции SHAP можно интерпретировать полученный результат прогнозирования.

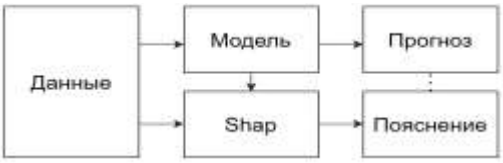

Рис. 1. Значимость признаков методом SHAP

III. Результаты и их обсуждение

При построении моделей машинного обучения исходная выборка разбиралась на тестовое и обучающее множества в соотношении 20:80. Оценка качества модели производилась по стандартным метрикам (точность, чувствительность и специфичность), на тестовой выборке [7]. В качестве выходной переменной использовалось бинарное значение – выживет ли пациент в течении 20 дней после инфаркта (1 – умрет, 0 – выживет), в качестве входных параметров использовались исходные клинические показатели. На первом этапе построения, несмотря на высокую точность, качество полученных моделей оказалось неудовлетворительным, так как, при высокой специфичности (>0.9) показатель чувствительности всех моделей был низким (<0.5). Для улучшения качества моделей была произведена балансировка данных [8], что позволило повысить чувствительность моделей на тестовой выборке. Весь последующий анализ проводился на улучшенных с помощью балансировки моделях. Метрики качества первоначальных и итоговых моделей представлены в табл. 2.

ТАБЛИЦА II Метрики качества моделей МО

Метрика	Гр. бустинг (исх.)	Лог. регрессия (исх.)	Гр. бустинг (итог)	Лог. регрес. (итог)	Модель Кокса
Чувств.	0.497549	0.357843	0.70098	0.713235	0.67523
Спец-ть	0.972399	0.964437	0.852442	0.757346	0.87231
AUC	0.808144	0.803867	0.798167	0.80776	0.783
Точность	0.887871	0.856457	0.82548	0.741274	0.83551

На первом этапе исследования была построена регрессионная модель пропорциональных рисков Кокса. Построение модели Кокса производилось с помощью модуля CoxPHFitter библиотеки lifelines.

На рис. 2 отношения рисков представлены в виде точек-прямоугольников, а полосы доверительного интервала – в виде усов. Центральная вертикальная линия указывает на базовый уровень риска. Можно сделать выводы, что отсутствие артериальной гипертензии (АГ) и проведенная процедура ЧКВ снижает риск смертности после ИМ, а высокий класс тяжести по шкале KILLIP и пожилой возраст существенно повышает риск. Такие показатели, как хроническая обструктивная болезнь легких и проводилась ли пациенту тромболитическая терапия, на результаты влияют не так существенно.

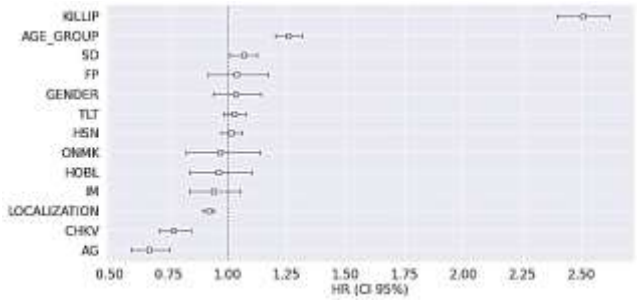

Рис. 2. Результаты модели Кокса

Как упоминалось ранее, для метода Каплана-Мейера и модели Кокса важность предикторов оценивалась, опираясь на показатель P-value. Для построения моделей логистической регрессии применялся модуль LogisticRegression. Значимость признака оценивалась по столбцу «P>|z|» в обобщающей таблице на выходе (функция Summary) – значимыми являются все признаки, для которых $p < 0.05$.

Для модели градиентного бустинга значимость признаков анализировалась несколькими способами, описанными в п. II. В. Для этого использовался модуль get_feature_importance, библиотеки CatBoostClassifier. Результаты методов PredictionValuesChange и LossFunctionChange для метода градиентного бустинга представлены на рис. 3. Они показывают, что изменение таких показателей как степень тяжести по шкале Киллип, возрастная группа, локализация возникновения заболевания, имеется ли ХСН, сильнее всего влияют на выходной результат.

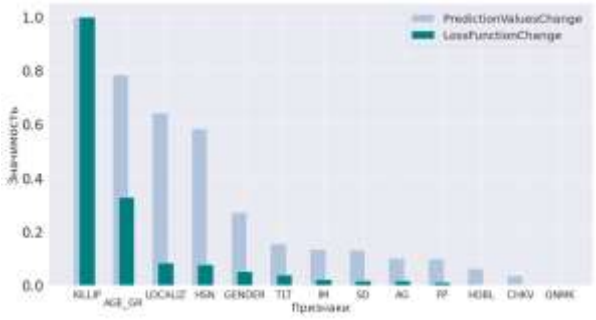

Рис. 3. Значимость признаков методом PredictionValueChange

Далее методом Interaction были сформированы пары признаков, которые имеют сильное воздействие на результат. В рамках текущей модели самой влиятельной взаимодействующей парой признаков оказались возраст пациента и тяжесть по шкале Киллип.

SHAP расшифровывается как SHapley Additive explanation. В основе данного метода лежит принцип теории игр для определения того, насколько каждый игрок при совместной игре способствует ее успешному исходу [9]. Прогноз представлен в виде графика, изображенного на рис. 4. На нем отмечено базовое значение риска смертности, которое составляет 0.2632 и является средним выходным значением модели по обучающему датасету. И отображено значение риска смертности для определенного

пациента, для примера был выбран пациент под номером 6, его риск смертности после ИМ равен 0.46. Показатели значимых признаков отображены ниже по оси X. Признаки, продвигающие риск к единице (то есть смертельному исходу), показаны красным цветом, а те, что понижают его значение – синим. Можно сделать выводы, что для данного пациента к повышающим риск факторам относится то, что он принадлежит к возрастной группе 8 (от 80 лет), что у него есть в анамнезе ХСН типа Н IIА, локализация его инфаркта – передней стенки и ему не проводилось ЧКВ. В то время как благоприятными для данного пациента факторами является незначительная тяжесть по Киллип и женский пол.

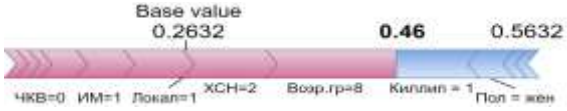

Рис. 4. График SHAP-значений

Чтобы оценить влияние признаков на модель в целом, можно рассмотреть агрегированные SHAP – значения, которые удобнее анализировать в форме сводного графика (рис. 5). График показывает, какие признаки являются наиболее важными, а также их диапазон влияния – насколько они могут снизить или повысить риск.

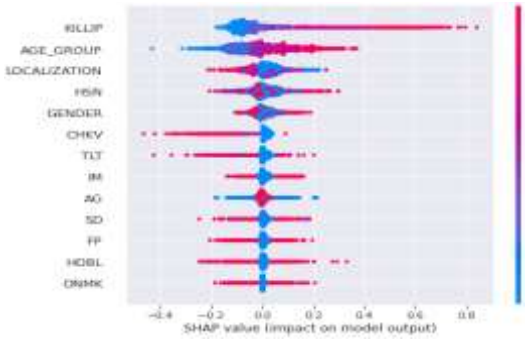

Рис. 5. Сводный график SHAP-значений

По оси Y отображены признаки, которые участвуют при построении модели, чем выше расположен признак, тем больше его влияние. Синими точками отмечены пациенты с низким значением соответствующего признака, а красным – с высоким. По оси X указано влияние входного признака на выходное значение модели. Толщина линий указывает на количество объектов для конкретного значения признака. Из графика очевидны выводы: чем выше степень тяжести состояния пациента по шкале Киллип, тем больше вероятность наступления смерти после ИМ; пациентов возраста чуть выше среднего (около 60 лет) больше всего в выборке и чем ниже возраст пациента, тем меньше вероятность смерти; пациенты у которых локализация возникновения ИМ соответствует передней или заднебазальной стенке левого желудочка, имеют выше вероятность смерти, а также больше шансов на выживание у пациентов с отсутствием сердечной недостаточности и если при поступлении проводилось чрескожное коронарное вмешательство.

ТАБЛИЦА III ЗНАЧИМОСТЬ ПРИЗНАКОВ, ПОЛУЧЕННАЯ РАЗЛИЧНЫМИ МЕТОДАМИ

Признак	Метод К-М	Метод Кокса	Лог. регр.	Градиентный бустинг		
				SHAP	Prediction Values	LossFunction
Возраст	0	0	0	0.081	16.2311	0.044155
Пол	0,0015	0.5	0.618	0.024	6.996766	0.006824
АГ	0	0	0	0.018	4.095762	0.003036
ИМ (повт)	0,0008	0.29	0.011	0.018	5.235203	0.004577
СД	0.00005	0.02	0	0.016	4.854288	0.00296
ФП	0.0027	0.55	0.628	0.015	3.941026	0.003265
ОНМК	0,0951	0.7	0.237	0.005	2.38141	0.002095
ХОБЛ	0.0471	0.57	0.163	0.01	3.340096	0.000548
ХСН	0	0.5	0	0.039	12.30736	0.019888
Локал-я	0	0	0	0.04	13.68025	0.022737
KILLIP	0	0	0	0.2	19.19209	0.130625
ТЛТ	0.0006	0.25	0.328	0.021	5.003892	0.002874
ЧКВ	0	0	0	0.022	2.740771	0.008019

В сводной табл. 3 представлены все признаки со значимостью для каждой построенной модели. Цветом выделены наиболее значимые, чем темнее – тем более значим признак. Для методов Каплана–Мейера, модели Кокса и логистической регрессии указаны значения p-value, и чем меньше значение, тем значимее признак. Для всех методов градиентного бустинга указано значение важности, что означает, чем выше значение, тем более он значим. Так, можно выделить значимые признаки для большинства методов: степень тяжести по шкале Киллип, локализация возникновения ИМ, возраст пациента, наличие хронической сердечной недостаточности. Как минимум, три метода отмечают также высокую значимость наличия в анамнезе артериальной гипертензии, сахарного диабета и проведённого пациенту чрезкожного коронарного вмешательства.

IV. ЗАКЛЮЧЕНИЕ

В ходе исследования рассмотрено несколько моделей прогнозирования риска смертности после инфаркта миокарда: статистический метод Каплана-Мейера, и моделей машинного обучения – модель Кокса, модель градиентного бустинга и модель логистической регрессии. Наиболее критичный период составляет первые двадцать дней после ИМ. Точнее всего риск смертности прогнозирует модель градиентного бустинга Catboost, однако модели Кокса и логистической регрессии, позволяют выделять больше значимых признаков. Установлено, что наиболее значимое влияние на риск смертности после ИМ оказывают показатели степень тяжести по шкале KILLIP, возраст, локализация и наличие хронической сердечной недостаточности. Для логистической регрессии также значим показатель сахарного диабета, для модели Кокса и метода Каплана-Мейера наличие артериальной гипертензии и проводилось ли ЧКВ. Для градиентного бустинга также значим показатель гендерного распределения пациентов.

СПИСОК ЛИТЕРАТУРЫ

[1] Глазкова Т.Г. Оценка качества методов диагностики и прогноза в медицине. *Вестник ОНЦ АМН России*. 1994. № 2. С. 3–11.

[2] Pieszko K. Predicting Long-Term Mortality after Acute Coronary Syndrome Using Machine Learning Techniques and Hematological Markers. *Disease Markers*, 2019. no.1. pp. 1-10.

[3] Bhatt D.L., Eagle K.A., Ohman E.M. Comparative Determinants of 4-Year Cardiovascular Event Rates in Stable Outpatients at Risk of or With Atherothrombosis. *Elsevier - Journal of Vascular Surgery*, 2011. no.12. pp. 1350-1357.

[4] Каширина И.Л., Хохлов Р.А., Казакова А.О. Прогнозирование развития инфаркта миокарда на основании анализа метеорологических факторов и данных областного регистра. *Вестник Воронежского государственного университета*, 2016. вып. 3. С. 116-123.

[5] Фирюлина М.А., Каширина И.Л., Гафанович Е.Я. Применение методов машинного обучения при назначении терапии гипертонической болезни. *Моделирование, оптимизация и информационные технологии*, 2020. вып. 8(4). С. 1-17

[6] Swalin A. Deep Dive into Catboost Functionalities for Model Interpretation. Available at: https://towardsdatascience.com/deep-dive-into-catboost-functionalities-for-model-interpretation-7cdef669aeed (accessed 5 February 2021)

[7] Firyulina M.A., Kashirina I.L. Classification of cardiac arrhythmia using machine learning techniques. *Journal of physics: conference serie*, 2020. pp. 1167–1175.

[8] Kashirina I.L., Firyulina M.A. Building models for predicting mortality after myocardial infarction in conditions of unbalanced classes, including the influence of weather conditions. *CEUR Workshop Proceedings*, 2020 no.2790. pp. 188–197

[9] Pandey P. Interpretable Machine Learning, Extracting human understandable insights from any Machine Learning model. Available at: https://towardsdatascience.com/interpretable-machine-learning-1dec0f2f3e6b (accessed 3 February 2021)