

УДК 004.852

doi:

АНАЛИЗ МЕТОДОВ ОЦЕНКИ ВАЖНОСТИ ПРЕДИКТОРОВ НЕБЛАГОПРИЯТНЫХ СОБЫТИЙ В КАРДИОХИРУРГИИ¹

Б.В. Потапенко (bvpotapenko@gmail.com)^A
К.И. Шахгельдян (carinashakh@gmail.com)^{A,B}
Б.И. Гельцер (boris.geltser@vvsu.ru)^{A,B}

^A Владивостокский государственный университет, Владивосток

^B Дальневосточный федеральный университет,
Владивосток

В работе исследованы методы оценки важности предикторов моделей машинного обучения. Рассмотрены как подходы зависящие от архитектуры модели, так и независящие от неё. Модели обучались предсказывать вероятность наступления летальности в послеоперационный период для пациентов с инфарктом миокарда с подъемом сегмента ST, которым выполнено чрескожное коронарное вмешательство. Результаты демонстрируют заметное расхождение в ранжировании признаков по их важности в зависимости от применяемых методов. Особенно это заметно для признаков, которые влияют на предсказание нелинейно, и связаны с другими признаками. В работе поднимаются вопросы проблемы интерпретации важности в клинической медицине. Результаты указывают в пользу применения комбинированных методов оценки важности для повышения доверия к системам поддержки принятия врачебных решений.

Ключевые слова: Объяснимый искусственный интеллект, Важность предикторов, Прогностические модели в клинической медицине, Интеллектуальный анализ данных, Машинное обучение.

Введение

¹ Работа выполнена при финансовой поддержке проекта FZNS-2023-0010 Госзадания Дальневосточного федерального университета (ДВФУ)

Искусственный интеллект (ИИ) нашёл широкое применение в системах поддержки принятия врачебных решений (СППВР) в последнее десятилетие [Chen et al., 2023]. В здравоохранении применяются классические алгоритмы машинного обучения (МО): линейной и логистической регрессии, SVM, системы, основанные на правилах, деревья решений случайный лес [Papadopoulos et al., 2022], стохастический градиентный бустинг, методы глубокого обучения нейросетей [Shamshirband et al., 2021].

Для системы здравоохранения и клинической медицины особенно важны объяснения генерируемых моделями МО заключений [Pierce et al., 2022]. Сложность объяснения результатов работы моделей ограничивает применимость моделей машинного обучения в здравоохранении [Wubineh et al., 2024] [Khan et al., 2024] [Albahri et al., 2023].

Одной из концепций объяснения заключений, генерируемых моделями МО, является оценка важности признаков [Saarela et al., 2021], а именно какие признаки являются для модели наиболее важными в целом (глобальная важность), и что побудило модель сделать конкретное предсказание в частном случае (локальная важность). Важность, представленная численно, может быть отображена на диаграмме, сравнена с важностью другого признака; может быть измерено её изменение при обновлении параметров модели, или в процессе обучения, делая работу эксперта более эффективной [Wang et al., 2021].

Самыми популярными подходами к оценке важности являются оценки весовых коэффициентов логистической регрессии, методы, встроенные в ансамбли деревьев решений, семейство методов основанных на методе аддитивного объяснения Шепли (SHAP) [Lundberg et al., 2017] [Lundberg et al., 2020], оценка важности методом перестановок значений [Breiman et al., 2001].

Целью данного исследования является сравнительный анализ методов оценки важности предикторов, популярных в индустрии, их сильные и слабые стороны на примере задачи бинарной классификации на данных клинической медицины.

1. Материалы и методы

1.1. Датасет

Для анализа методов оценки важности предикторов мы использовали данные о больных инфарктом миокарда с подъемом сегмента ST (ИМпST), которым была выполнена операция чрескожного коронарного вмешательства (ЧКВ) в Краевой клинической больнице №1 г. Владивостока в период с 2016 до 2022 гг. Перед обработкой все данные

были обезличены. В предыдущих исследованиях авторами были определены и валидированы предикторы внутригоспитальной летальности больных ИМпСТ после ЧКВ: возраст (Age), класс острой сердечно-сосудистой недостаточности (ОСН) по по T.Killip выше 2 (Killip_gt_2, бинарный признак), частота сердечных сокращений (HBR), систолическое артериальное давление (Systolic AP), уровень креатинина в крови (Creatinine), фракция выброса левого желудочка (EFLV), количество лейкоцитов в крови (WBC), относительное значение количества нейтрофилов (Neutrophils), относительное значение количества эозинофилов (Eosinophils), тромбоцит (Thrombocrit) [Shakhgeldyan et al., 2024].

В качестве зависимой переменной рассматривается внутригоспитальная летальность (ВГЛ) - летальность в больнице или в 30-дневный период после проведения операции. В итоговую выборку вошли: 4668 записей о пациентах с ИМпСТ после ЧКВ, из которых 4355 (93.3%) относились к группе выживших, а 313 (6,7%) - к группе ВГЛ.

1.2. Методы

Мы применили оценки статистической значимости: t-тест; U-тест; Хи2-тест. Модель-зависимые методы МО: однофакторная (LR) и многофакторная логистические регрессии (MLR); случайный лес (RF), CatBoost (CB), стохастический градиентный бустинг (XGB) и искусственная нейронная сеть - многослойный персептрон (NN). Для оценки важности предикторов в первых двух методах использовали весовые коэффициенты моделей, в оставшихся - характеристику importance, которая вычисляется в процессе обучения моделей. Модель-независимые методы: важность на основе метода перестановок и аддитивного объяснения Шепли (SHAP) [Lundberg et al., 2020].

1.3. Дизайн исследования

На первом этапе исследования статистическими методами оценивалось влияние признаков на зависимую переменную. Для непрерывных признаков были использовались: тест Стьюдента, тест Манна-Уитни. Для категориальных признаков - тест хи-квадрат. Для сравнения важности предикторов использовали p-value, минимальное значение которого соответствует максимальной важности предиктора. Кроме того, вычислены весовые коэффициенты LR на основе нормированных методом MinMax предикторов.

На втором этапе были применены методы МО. Данные были разделены на обучающие и валидационные с учётом стратификации. Для обучения методами MLR и NN данные были нормированы методом MinMax. В процессе обучения моделей выполнялся подбор

гипер-параметров с помощью стратифицированной k-fold кросс-валидации. Выбор лучшей модели осуществлялся с на основе максимизации метрики площади под ROC-кривой (AUC). Для обучения моделей использовали MLR, RF, CB, XGB, NN. Сравнили нормированные коэффициенты регрессии и важность признаков по собственной оценке моделей на основе деревьев решений.

На третьем этапе оценили важность признаков для этих моделей методами SHAP и методом перестановки. Метрикой для оценки важности методом перестановок была выбрана AUC на валидационном наборе данных

2. Результаты

2.1. Статистические метрики важности

Первый подход к оценке важности основывается на сравнении статистических параметров: t-статистика, g-значение и p-value, полученные методами межгрупповых сравнений - тестом Стьюдента, Манна-Уитни. Нормализованные значения обоих тестов и оценки p-value представлены в таблице 1. Результаты t-теста показали, что наиболее значимыми признаками являются: Creatinine, WBC, EF LV и Neutrophils . Наименее значимый — Thrombocrit. С точки зрения теста Манна-Уитни наиболее значимые предикторы - Neutrophils, Eosinophils, Systolic AP и Age, а наименее - Thrombocrit. Признаки, в важности которого уверены оба теста - это Neutrophils. Оба теста верифицировали Thrombocrit, как наименее важный предиктор. Самым противоречивым оказался признак Eosinophils.

Табл. 1.

Статистики межгрупповых сравнений тестами Стьюдента и Манна-Уитни				
Предикторы	t-test rank	t-test p-value	U-test rank	U-test p-value
Age	0.57	2.676e-37	0.72	6.271E-35
HBR	0.77	7.162e-55	0.68	5.567E-32
Systolic AP	0.84	1.516e-61	0.75	2.342E-37
Creatinine	1.00	9.353e-79	0.72	1.532E-34
EF LV	0.85	4.705e-63	0.74	5.185E-35
WBC	0.89	1.759e-66	0.70	4.184E-33
Neutrophils	0.84	4.162e-61	1.00	2.622E-55
Eosinophils	0.30	6.130e-19	0.88	1.082E-44
Thrombocrit	0.00	5.603e-06	0.00	5.707E-03

Для категориального признака (Killip_gt_2) вычислили значения χ^2 , (p-value =5.699e-80). Согласно p-value Killip_gt_2 является наиболее значимым среди всех предикторов ВГЛ.

2.2. Коэффициенты однофакторной логистической регрессии

Важность предикторов может ассоциироваться с модулем весового коэффициента LR. На первое место по значимости выходят Systolic AP (7.02), Neutrophils (6.97), на третьем месте по важности Creatinine (6.76). Далее WBC (5.94), HBR (b) (5.81), EF LV (6.02), Eosinophils (5.35), Age (4.26), Thrombocrit (2.11), Killip_gt_2 (2.01). Таким образом, наиболее важными с точки зрения LR являются Systolic AP, Neutrophils и Creatinine, наименее Thrombocrit и Killip_gt_2.

2.3. Многофакторные модели машинного обучения

Для оценки важности мы рассматриваем не только изолированное влияние предикторов на конечную точку, но и то, как это влияние проявляется при работе в многофакторных моделях. Для MLR - мы рассматриваем весовые коэффициенты, для ансамблевых методов - внутренний механизм расчета важности (importance). Для сравнения все оценки были нормированы на максимальные значения (Табл.2).

Среди коэффициентов MLR с заметным отрывом лидирует Creatinine (относительная важность =1), за ним следуют EF LV (относительная важность = 0.379) и Neutrophils (относительная важность = 0.374). Наименьший коэффициент был получен при Killip_gt_2 (относительная важность = 0). По мнению трех ансамблевых моделей на основе деревьев решений Neutrophils - самый важный признак, за ним следует Eosinophils. Признак Age занимает третье место в оценке CatBoost, при этом XGBoost относит его к наименее важным, при этом на третье место последний ставит Killip_gt_2. Показатель Creatinine занимает третье место согласно RF. Наименее важными признаками являются согласно ансамблевым методам - Systolic AP и Thrombocrit.

Табл. 2.

Оценка важности предикторов многофакторных моделей				
Предикторы	MLR	RF	CB	XGB
Age	0.278	0.197	0.634	0.045
Killip_gt_2	0.000	0.111	0.136	0.287
HBR	0.304	0.140	0.248	0.003
Systolic AP	0.183	0.050	0.000	0.000
Creatinine	1.000	0.518	0.441	0.088
EF LV	0.379	0.470	0.490	0.267

WBC	0.239	0.143	0.065	0.019
Neutrophils	0.374	1.000	1.000	1.000
Eosinophils	0.178	0.782	0.873	0.377
Thrombocrit	0.227	0.000	0.009	0.008

2.4. Глобальные оценки SHAP

Метод SHAP позволяет оценить важность признаков в многофакторной модели, независимо от модели и после ее обучения. Метод оценивает степень влияния признака по величине shap-value, которая при положительном значении описывает влияние на риск развития неблагоприятного события. По оценке SHAP для положительного класса в MLR самыми важными являются признаки Neutrophils, EF LV, HBR (Табл. 3). Данный результат расходится с оценкой на основании весовых коэффициентов регрессии, где лидером был Creatinine. В то же время 2 других признака - Neutrophils, EF LV повторяются. Согласно SHAP в MLR Creatinine занимает 5 место в ранге важности.

Табл. 3.

Глобальная оценка SHAP для положительного класса					
Предикторы	MLR	RF	CB	XGB	NN
Age	0.52	0.04	0.35	0.32	4.04
Killip gt 2	0.32	0.04	0.35	0.27	3.98
HBR	0.61	0.04	0.39	0.32	4.5
Systolic AP	0.24	0.02	0.14	0.13	1.56
Creatinine	0.38	0.07	0.46	0.42	1.48
EF LV	0.64	0.06	0.48	0.39	2.57
WBC	0.29	0.03	0.18	0.16	2.15
Neutrophils	0.72	0.11	0.54	0.75	9.26
Eosinophils	0.13	0.09	0.41	0.41	10.83
Thrombocrit	0.24	0.01	0.13	0.14	1.82

Для моделей RF, CatBoost и XGBoost важность по SHAP предиктора Neutrophils подтверждается, но заметно отличается от MLR переходом признака Eosinophils с последнего места на второе, четвёртое и третье соответственно. Среди значений SHAP для признаков NN выделяется вышедший на первое место предиктор Eosinophils, Neutrophils на вторую позицию и заметное изменение рангов для остальных признаков.

2.5. Важность на основе метода перестановок

Метод перестановок подразумевает искажение по очереди каждого признака и оценку снижения при этом точности прогностической модели. В качестве базовой метрики качества моделей используется площадь под ROC-кривой (AUC). Важными признаками для всех моделей по этому методу оценки являются: Neutrophils, Eosinophils, HBR, Age, Creatinine (табл.4). Остальные признаки не попали в тройку наиболее важных ни для одной из моделей.

Табл. 4.

Важность признаков по методу перестановок					
Предикторы	LR	RF	CB	XGB	NN
Age	0.030	0.014	0.020	0.023	0.035
Killip_gt 2	0.001	0.004	0.001	0.001	0.002
HBR	0.022	0.016	0.020	0.017	0.035
Systolic AP	0.005	0.003	0.004	0.002	0.000
Creatinine	0.016	0.022	0.020	0.014	0.019
EF LV	0.011	0.009	0.011	0.014	0.021
WBC	0.010	0.006	0.003	0.002	0.022
Neutrophils	0.026	0.037	0.024	0.057	0.100
Eosinophils	0.004	0.023	0.015	0.023	0.079
Thrombocrit	0.001	0.000	-0.001	0.000	-0.003

Метод перестановок подтвердил значимость Neutrophils, поставив его на первое место для всех моделей кроме MLR, где он поставил его на второе место. Показатель Eosinophils занимал вторую позицию за исключением MLR и CatBoost. Все модели демонстрировали низкий уровень значимости для показателей Killip_gt_2 и Thrombocrit.

Можно заметить различия между глобальной важностью по SHAP и важностью на основе перестановок. Так для MLR Age и HBR оказались важнее, чем EF LV (второй по важности по оценке SHAP); у модели CatBoost признаки EF LV и Eosinophils в оценке перестановками уступили HBR; у модели XGBoost признак Age по оценке перестановками оказался важнее, чем EF LV и Creatinine; а для NN при оценке перестановками признак Neutrophils оказался важнее, чем Eosinophils.

Обсуждение

В данной работе мы рассмотрели несколько подходов к оценке важности предикторов на примере задачи прогнозирования ВГЛ у пациентов с ИМпСТ после ЧКВ. Обобщая важность предикторов, полученных разными методами, можно представить их ранг (табл. 5).

Табл. 5.

Обобщенный ранг важности предикторов						
Предикторы	Статистика	LR	ML R	Ансамбли	SHA P	Перест ановки
Age	8	8	5	5	6	4
Killip gt 2	1	10	10	6	7	7
HBR	7	5	4	7	5	3
Systolic AP	5	1	8	10	10	8
Creatinine	2	3	1	3	2	5
EF LV	4	6	2	4	4	6
WBC	3	4	6	8	8	9
Neutrophils	6	2	3	1	1	1
Eosinophils	9	7	9	2	3	2
Thrombocrit	10	9	7	9	9	10

Анализ рейтинга важности предикторов в их влиянии на конечную точку позволяет сделать несколько замечаний. Разные подходы к оценке важности обеспечивают противоречивые результаты, вплоть до прямо противоположных. Так, например, предиктор Systolic AP имеет максимальную важность при сравнении весовых коэффициентов LR, и наименьшую важность при многофакторных ансамблевых моделях и при использовании SHAP. Класс OCCN по T.Killip имеет максимальную важность при статистической оценке и минимальную - при работе в MLR. Оценка важности зависит от методов МО, включая использования методов SHAP или перестановки. Наиболее существенные различия мы можем наблюдать в сравнении методов MLR и ансамблевых методов на основе деревьев решений. Это согласуется с выводами коллег [Saarela et al., 2021] [Khan et al., 2024]. Ансамблевые методы МО обеспечивают устойчивую оценку для наиболее важных предикторов, которая подтверждается при наложении методов SHAP и перестановки. Так, наиболее важными в этом случае были Neutrophils, Eosinophils и Creatinine. Наименее важные предикторы подтверждают свой рейтинг в большинстве методов оценки. Например, Thrombocrit имеют низкую важность согласно всем рассматриваемым подходам.

Наша работа подтверждает высокую значимость показателей EF LV и Creatinine, которая была оценена в работе по эпидемиологии сердечно-сосудистых заболеваний [Ziaieian et al., 2016]. При этом мы показали, что признак Neutrophils является одним из наиболее важных предикторов ВГЛ у пациентов с ИМпST после ЧКВ. Существенные различия в оценке методов, использующих MLR и ансамблевых методов на основе деревьев

решений можно объяснить учетом линейных и нелинейных отношений между предикторами и конечной точкой. При наличии линейных отношений рейтинговая оценка важности подтверждается разными подходами. Примером такой взаимосвязи служит Age, увеличение которого ведет к росту вероятности ВГЛ у пациентов с ИМПСТ после ЧКВ. Противоположным примером является признак Eosinophils, который демонстрирует низкий уровень значимости в MLR и статистике, но высокий - для ансамблевых методов. Аналогичную нелинейную зависимость можно предположить у Creatinine, который обладал наибольшим коэффициентом в MLR, но при этом в подходе на основе перестановки был лишь на 5-6 позиции.

Наше исследование демонстрирует, что рассмотренные оценки важности тем более стабильны, чем более сильна и более линейна связь зависимой переменной с конкретным признаком. И менее эффективны там, где предиктор нелинейно влияет на зависимую переменную, что типично для данных клинической медицины.

Заключение

В данной работе мы показали, что разные подходы к оценке важности предикторов прогностических моделей обеспечивают противоречивые результаты, вплоть до прямо противоположных. Оценка важности зависит от методов МО, выбранной архитектуры моделей. Ансамблевые методы МО обеспечивают устойчивую оценку для наиболее важных предикторов, которая подтверждается при наложении методов SHAP и перестановки. Методы оценки важности предикторов имеют демонстрируют разные результаты, когда связь между предикторами и предсказанием нелинейна. Таким образом, проблема выбора метода оценки важности актуальна, как и задача разработки новых более универсальных алгоритмов.

Список литературы

- [Albahri et al., 2023] A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion / A.S. Albahri [и др.] // Information Fusion. 2023. Vol. 96. pp. 156–191.
- [Breiman, 2001] Breiman L. Random Forests // Machine Learning. 2001. Vol. 45(1). pp. 5–32. doi: 10.1023/A:1010933404324.
- [Chen et al., 2023] Harnessing the power of clinical decision support systems: challenges and opportunities / Z. Chen [и др.] // Open Heart. 2023. doi: 10.1136/openhrt-2023-002432.
- [Khan et al., 2024] Guaranteeing Correctness in Black-Box Machine Learning: A Fusion of Explainable AI and Formal Methods for Healthcare Decision-Making / N. Khan [и др.] // IEEE Access. 2023. doi: 10.1109/ACCESS.2024.3420415.

- [Lundberg et al., 2017] Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. pp. 4768–4777. doi: 10.5555/3295222.3295230.
- [Lundberg et al., 2020] From local explanations to global understanding with explainable AI for trees / S.M. Lundberg [и др.] // Nature Machine Intelligence. 2020. Vol. 2. pp. 56–67. doi:10.1038/s42256-019-0138-9.
- [Papadopoulos et al., 2022] A systematic review of technologies and standards used in the development of rule-based clinical decision support systems / Papadopoulos, M [и др.] // Health Technology. 2022. Vol. 12. pp.713–727. doi:10.1007/s12553-022-00672-9.
- [Pierce et al., 2022] Explainability in medicine in an era of AI-based clinical decision support systems / R.L. Pierce [и др.] // Frontiers in Genetics. 2022. Vol. 13. 903600. doi:10.3389/FGENE.2022.903600/BIBTEX.
- [Saarela et al., 2021] Saarela M., Jauhiainen S. Comparison of feature importance measures as explanations for classification models // SN Applied Sciences. 2021. Vol. 3(2). 1-12. doi:10.1007/s42452-021-04148-9/TABLES/4.
- [Shakhgeldyan et al., 2024] Methods of prognostic analysis for the prediction of in-hospital mortality in patients with acute st-elevation myocardial infarction after percutaneous coronary interventions. / Shakhgeldyan K.J., Kuksin N.S., Domzhalov I.G., Geltser B.I. // Pattern Recognition and Image Analysis. 2024. T. 34. Vol. 3. 786-796. doi: 10.1134/S1054661824700676
- [Shamshirband et al., 2021] A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues / S. Shamshirband [и др.] // Journal of Biomedical Informatics. 2021. Vol. 113. 103627. doi:10.1016/J.JBI.2020.103627.
- [Wang et al., 2021] Wang X., Yin M. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making // International Conference on Intelligent User Interfaces, Proceedings IUI. 2021. pp. 318–328. doi: 10.1145/3397481.3450650/SUPL_FILE/P318-WANG.PDF.
- [Wubineh et al., 2024] Exploring the opportunities and challenges of implementing artificial intelligence in healthcare: A systematic literature review / B.Z. Wubineh, F.G. Deriba, M.M. Woldeyohannis // Urologic Oncology: Seminars and Original Investigations. 2024. Vol. 42(3). 48–56. doi: 10.1016/J.UROLONC.2023.11.019.
- [Ziaecian et al., 2016] Ziaecian B., Fonarow G. Epidemiology and aetiology of heart failure.// Nature Reviews Cardiology. 2016. Vol. 13(6). pp. 368–378. doi: 10.1038/nrcardio.2016.25.

ANALYSIS OF METHODS FOR ASSESSING THE IMPORTANCE OF PREDICTORS OF ADVERSE EVENTS IN CARDIAC SURGERY²

² This work was supported by the project FZNS-2023-0010 under the State Assignment of the Far Eastern Federal University (FEFU)

Bogdan V. Potapenko (*bvpotapenko@gmail.com*)^A
Karina J. Shakhgelyan (*carinashakh@gmail.com*)^{A,B}
Boris I. Geltser (*boris.geltser@vvsu.ru*)^{A,B}

^A Vladivostok State University, Vladivostok, Russia

^B Far Eastern Federal University, Vladivostok, Russia

This study investigates methods for assessing the importance of predictors in machine learning models. Both model-dependent and model-independent approaches were considered. The models were trained to predict the probability of mortality in the postoperative period for patients with ST-segment elevation myocardial infarction who underwent percutaneous coronary intervention. The results demonstrate a noticeable divergence in feature importance rankings depending on the applied methods. This is particularly evident for features that influence predictions non-linearly and are associated with other features. The study raises questions regarding the interpretation of feature importance in clinical medicine. The results indicate the advantage of using combined methods of importance assessment to increase trust in clinical decision support systems.

Keywords: Explainable Artificial Intelligence, Feature Importance, Clinical Prognostic Models, Machine Learning, Clinical Data Mining.