Инженерный вестник Дона, №10 (2024)
ivdon.ru/ru/magazine/archive/n10y2024/9524

Применение методов машинного обучения к распознаванию сердечно-сосудистых заболеваний

А.А. Алымова, К.М. Зубарев, Т.Л. Иванова

Московский государственный технический университет имени Н.Э. Баумана

Аннотация: Настоящая работа посвящена исследованию возможности определения сердечных заболеваний на основе 13 категориальных и численных признаков. Мы представляем подробный анализ набора данных, включающий разделение данных на обучающую и тестовую выборки, разбиение признаков на численные и категориальные, применение 4 различных алгоритмов классификации, проверку качества модели двумя техниками – отложенной выборки и кросс-валидацией. Для оценки качества модели обращаем внимание на значение метрики recall и на матрицу ошибок, построенные на тестовом наборе данных из отложенной выборки или на каждом тестовом фолде при использовании кросс-валидации. Результаты исследования имеют значение как для глубинного понимания связи определённых медицинских показателей с заболеваниями сердечно-сосудистой системы, так и для разработки более точных методов их прогнозирования на основе определённых симптомов.
Ключевые слова: сердечные заболевания, задача классификации, метрики качества, кросс-валидация, recall, машинное обучение, случайный лес.

Введение.

В современном мире проблема здоровья населения становится всё более актуальной [1,2]. Для диагностики заболеваний активно внедряются методы искусственного интеллекта [3], а также большой объём данных, содержащийся в электронных медицинских картах, упрощает мониторинг здоровья пациента [4]. Болезни сердечно-сосудистой системы являются одними из основных причин смертности пациентов разных возрастов. В связи с этим возникает необходимость разработки и применения новых методов диагностики сердечных заболеваний [5,6].

В данной работе рассматривается решение задача классификации [7] для определения сердечных заболеваний с использованием таких показателей, как возраст, артериальное давление, максимальная частота сердечных сокращений, уровень сахара в крови, холестерин и наличие сопутствующих заболеваний. Благодаря классификации можно разделить

пациентов на группы с повышенным и пониженным риском развития сердечно-сосудистых заболеваний, что способствует раннему выявлению и лечению этих заболеваний.

Для проведения классификации используются следующие методы: логистическая регрессия, метод ближайших соседей, случайные леса и наивный Байесовский классификатор [8].

Методы классификации и используемые метрики качества модели.

Метрика Recall (полнота) в задачах классификации используется для измерения того, насколько хорошо модель обнаруживает все объекты положительного класса [9,10]. Модель с высоким значением recall хорошо обнаруживает положительные объекты, что может быть полезно в ситуациях, где пропуск положительного объекта более критичен, чем ложноположительное предсказание. Recall определяется формулой (1):

$$recall = \frac{TP}{TP+FN}, \quad (1)$$

где TP – количество объектов положительного класса, классифицированных верно; FN – количество объектов положительного класса, классифицированных ошибочно.

Логистическая регрессия. Принцип работы метода основан на оценке параметров логистической модели, которая использует логистическую функцию для преобразования логарифмических шансов в вероятность.

Модель логистической регрессии для задачи двух классовой классификации $Y \in \{0,1\}$, $x \in \mathbb{R}^n$ описывается формулой (2):

$$P(y_i = 1|x_i, \theta) = \frac{1}{1+e^{-x_i\theta}} = \sigma(x_i\theta), \quad (2)$$

где $y_i = y(x_i)$ – значение целевой функции на объекте обучающей выборки; x_i – объект обучающей выборки; $\theta = (\theta_0, \theta_1, \dots, \theta_n) \in \mathbb{R}^n$ – вектор параметров; Y – множество ответов.

*Метод k ближайших соседей***.** Метод k-ближайших - это метрический алгоритм для автоматической классификации объектов или регрессии путём учёта принадлежности объекта к классам его соседей. Объект относится к классу, наиболее распространённому среди его k соседей с уже известными классами [10].

Случайный лес. Данный метод основан на подходе «разделяй и властвуй». Каждое дерево из ансамбля деревьев создаётся путём случайной выборки данных и отбора признаков на каждом узле дерева. Эти деревья-классификаторы образуют лес. Для решения задачи классификации используется голосование большинства, что позволяет получить более точный прогноз. [10].

Наивный Байесовский классификатор. Наивный байесовский классификатор — это вероятностный классификатор, базирующийся на формуле Байеса и предполагающий независимость признаков друг от друга в рамках заданного класса. В методе производится оценка одномерных вероятностных плотностей вместо многомерной. Пространство разбивается на непересекающиеся части так, чтобы минимизировать среднюю ошибку неправильной классификации.

Расчётная часть.

В данной практической части нашего исследования мы обращаем внимание на конкретные данные — медицинские показатели людей, страдающих и не страдающих сердечно-сосудистыми заболеваниями. Данные были взяты с сайта Kaggle [11].

Будем оценивать качество модели с помощью матрицы ошибок и метрики recall, так как в медицинской среде ложноотрицательные показатели наиболее опасны [6]. Если модель отметит пациента с наличием заболевания (метка 1) как человека с его отсутствием (метка 0), то это может быть серьёзной угрозой для его жизни и здоровья [5,7].

Строить метрики будем двумя способами – на тестовом наборе данных из отложенной выборки и на каждом тестовом фолде при использовании кросс-валидации [10], а затем вычислять среднее значение метрики.

В таблице 1 отражены значения метрики recall для различных методов решения задачи классификации.

Таблица №1

Значение метрики recall в зависимости от метода и проверки

	Лог. регрессия	К-ближайших соседей	Случ. лес	Наивный Байесовский классификатор
Отложенная выборка	0,9097	0,9742	0,9484	0,8903
Кросс-валидация	0,9002	0,9703	0,9946	0,8787

Отметим, что оценка качества модели с помощью кросс-валидации наиболее репрезентативна, так как модель тестируется n раз на различных частях набора данных. В данной задаче n = 10.

Посмотрим теперь на матрицы ошибок, построенные на отложенной выборке. В общем случае матрица ошибок имеет вид:

$$Confusion\ matrix = \begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix}, \quad (3)$$

где True Positive (TP) – количество правильно диагностированных заболеваний; False Positive (FP) – число ошибочно поставленных диагнозов; False Negative (FN) – число пропущенных заболеваний; True Negative (TN) – число верно не поставленных диагнозов [10].

При оценке качества работы моделей будем обращать особое внимание на показатель FN: нам нужно, чтобы количество ложноотрицательных ошибок было минимальным.

На Рисунке 1 видно, что логистическая регрессия определяет значения целевой функции недостаточно полно и количество ложноотрицательных ошибок достаточно велико.

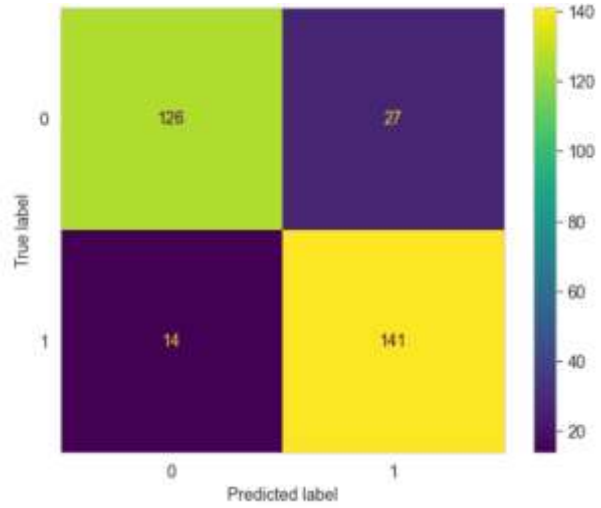

Рис. 1. – Матрица ошибок в модели логистической регрессии.

На рисунке 2 представлена матрица ошибок при использовании метода k-ближайших соседей.

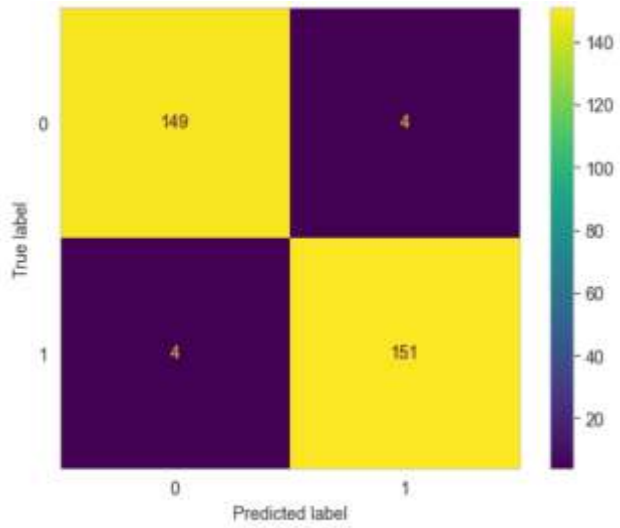

Рис. 2. – Матрица ошибок в модели k-ближайших соседей.

Можно отметить, что данный метод определяет значения целевой функции достаточно точно – только 8 объектов были классифицированы неверно, при этом FN = 4, что является хорошим показателем (рис.2).

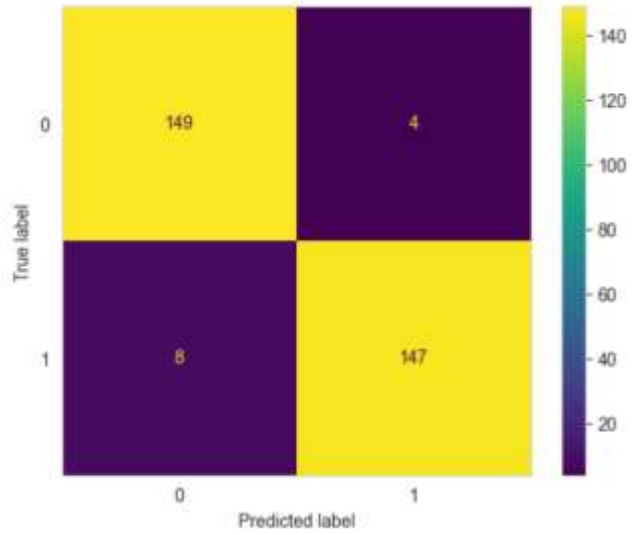

Рис. 3. – Матрица ошибок в модели случайного леса.

Метод случайного леса по данной отложенной тестовой выборке определяет значения целевой немного хуже, чем метод k-ближайших соседей: FN = 8 (рис.3).

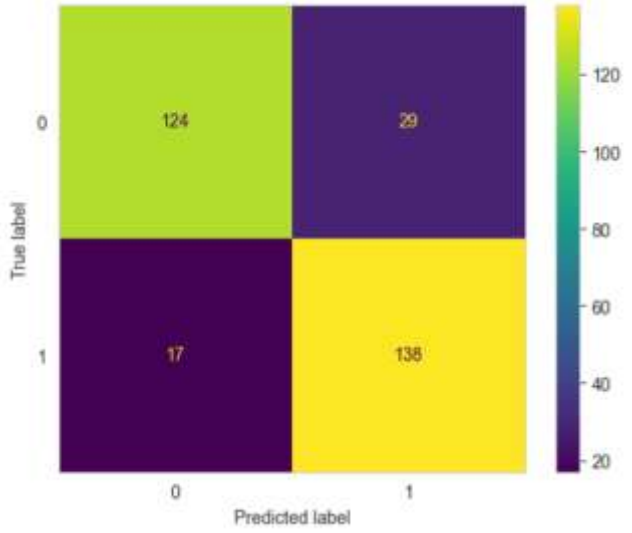

Рис. 4. – Матрица ошибок в модели наивного Байесовского классификатора.

Наивный Байесовский классификатор хуже остальных изученных моделей подходит для решения данной задачи (рис.4).

Таким образом, при тестировании методом отложенной выборки наибольшее значение метрики recall достигается в методе К-ближайших

соседей - 0.9742, затем в методе Случайного леса - 0.9484. При тестировании с использованием кросс-валидации наилучшее значение recall в методе Случайного леса - 0.9946, затем в методе К-ближайших соседей - 0.9757.

Оценим, какие признаки наиболее значимы для определения значения целевой переменной на примере Случайного леса. Это особенно важно, если собрать всю информацию о пациенте не представляется возможным. На рисунке 5 можно видеть, какие из факторов являются наиболее репрезентативными при выявлении сердечно-сосудистых заболеваний.

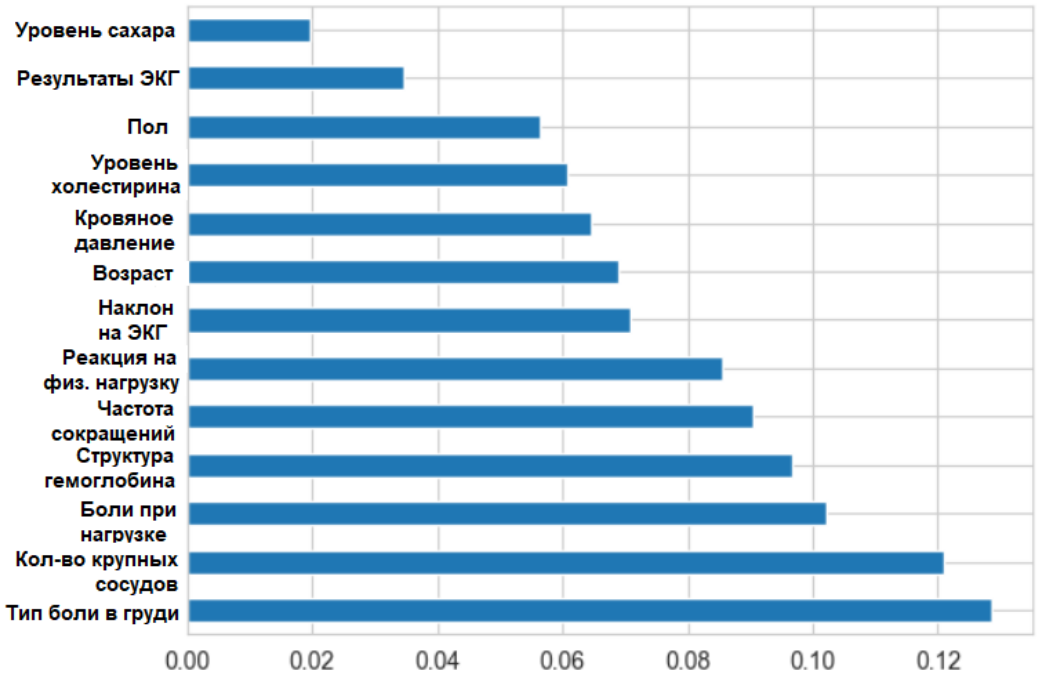

Рис. 5. – График влияния признаков на значение целевой переменной.

Очевидно, что для диагностирования сердечных заболеваний наиболее значим признак, отражающий тип боли в груди пациента. Также важно количество крупных сосудов сердца, хорошо пропускающих кровь. Здесь также важно отметить, что результаты ЭКГ сами по себе не являются решающим фактором, гораздо важнее реакция на физическую нагрузку. Также авторы отмечают структуру гемоглобина, как важный фактор,

известно, что у 48% пациентов, у которых диагностированы сердечные заболевания, отмечается дефицит гемоглобина. Меньше всего на результат влияет уровень сахара в крови пациента. Рассмотрим наиболее информативный признак подробнее. На рисунке 6 представлена связь типа боли в груди с наличием заболевания:

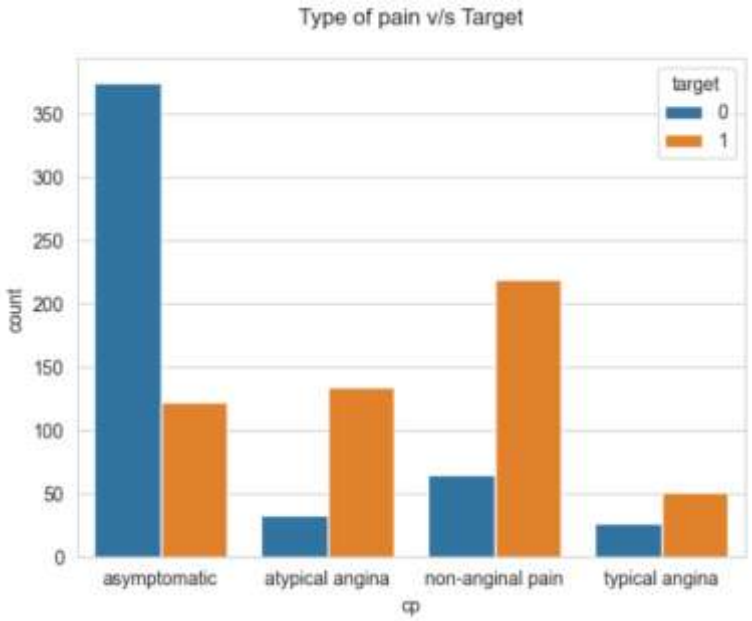

Рис. 6. – Связь типа боли в груди с наличием сердечного заболевания (0 – заболевания нет, 1 – заболевание есть).

Из рисунка 6 видно, что у здоровых пациентов зачастую отсутствует боль в груди (asymptomatic) и редко встречается атипичная ангинозная боль (atypical angina). У пациентов с заболеваниями сердечно-сосудистой системы наиболее часто наблюдается боль не ангинозного характера (non-anginal pain) и реже всего встречается типичная ангинозная боль (typical angina). Таким образом, данные показатели имеют большое значение для диагностики заболеваний сердца.

Заключение.

Наиболее информативными признаками оказались тип боли в груди и количество крупных сердечных сосудов, хорошо пропускающих кровь. У

здоровых пациентов зачастую боли в груди отсутствуют, а у пациентов с болезнями сердца часто наблюдается боль не ангинозного характера.

Полученные результаты свидетельствуют о том, что учёт вышеуказанных медицинских показателей может значительно улучшить прогнозирование и диагностику заболеваний сердца. Это может способствовать созданию более результативных методов профилактики и лечения заболеваний сердечно-сосудистой системы, а также повышению качества жизни пациентов.

Литература

1. Сасов Д.А., Зубков А.В., Орлова Ю.А., Турицына А.В. Классификация рака молочной железы с помощью сверточных нейронных сетей // Инженерный вестник Дона. 2023. № 6. URL: ivdon.ru/ru/magazine/archive/n6y2023/8507.

2. Гусев А. В. Искусственный интеллект в медицине и здравоохранении // Информационное общество.2017.№ 4-5.С. 78-93.

3. Шараев, Д. А. Метод определения изогнутой линии черепного шва на основе сверточных нейронных сетей // Инженерный вестник Дона. 2021. № 6. URL: ivdon.ru/ru/magazine/archive/n6y2021/7023

4. Гусев А. В., Зингерман Б. В. Электронные медицинские карты как источник данных реальной клинической практики// Реальная клиническая практика: данные и доказательства. 2022. Т. 2, № 2. С. 8-20.

5. Sansone M., Fusco R., Pepino A. and Sansone C. Electrocardiogram Pattern Recognition and Analysis Based on Artificial Neural Networks and Support Vector Machines: A Review Journal of Healthcare Engineering. № 4. 2013. pp. 465-504.

6. Ayer T., Alagoz O., Chhatwal J., Shavlik J.W., Kahn C.E., Burnside E.S. Breast cancer risk estimation with artificial neural networks revisited. Cancer 2010. № 116(14). pp. 3310-3321

7. Aguiar F.S., Almeida L.L., Ruffino-Netto A., Kritski A.L., F. CQ Melloand L. Werneck. Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients. BMC Pulmonary Medicine, 2012. № 12. pp. 1-8.

8. Зубарев К.М., Безрученко Т.С. Анализ эффективности маркетинговой кампании методами машинного обучения // Дневник науки. 2024. № 6. URL: dnevniknauki.ru/index.php/number6-2024/fiziko-matematicheskie-nauki

9. Облакова, Т. В., Зубарев К.М. Анализ распределения высоты морских волн. Сравнение оценок и применение критерия согласия Пирсона // Дневник науки. 2023. № 12. URL: dnevniknauki.ru/index.php/number12-2023/fiziko-matematicheskie-nauki

10. Sylvain Arlot, Alain Celisse. A survey of cross-validation procedures for model selection. Statist. Surv. 2010. Vol. 4. pp. 40-79.

11. kaggle URL: kaggle.com/datasets/johnsmith88/heart-disease-dataset/data (дата обращения: 28.08.2024).

References

1. Sasov D.A., Zubkov A.V., Orlova Yu. A., Turicyna A.V. Inzhenernyj vestnik Dona. 2023. № 6. URL: ivdon.ru/ru/magazine/archive/n6y2023/8507.

2. Gusev A. V. Informacionnoe obshhestvo.2017.№ 4-5.pp. 78-93.

3. Sharaev, D. A. Inzhenernyj vestnik Dona. 2021. № 6. URL: ivdon.ru/ru/magazine/archive/n6y2021/7023

4. Gusev A. V., Zingerman B. V. Real'naya klinicheskaya praktika: dannye i dokazatel'stva. 2022. T. 2, № 2. pp. 8-20.

5. Sansone M., Fusco R., Pepino A. and Sansone C. A Review Journal of Healthcare Engineering. № 4. 2013 pp. 465-504.

6. Ayer T., Alagoz O., Chhatwal J., Shavlik J.W., Kahn C.E., Burnside E.S. Cancer 2010. № 116(14). pp. 3310-3321.

7. Aguiar F.S., Almeida L.L., Ruffino-Netto A., Kritski A.L., F. CQ Melloand L. Werneck. BMC Pulmonary Medicine, 2012. № 12. pp. 1-8.

8. Zubarev K.M., Bezruchenko T.S. Dnevniki nauki. 2024. № 6. URL: dnevniknauki.ru/index.php/number6-2024/fiziko-matematicheskie-nauki

9. Oblakova, T. V., Zubarev K.M. Dnevniki nauki. 2023. № 12. URL: dnevniknauki.ru/index.php/number12-2023/fiziko-matematicheskie-nauki

10. Arlot Sylvain, Celisse Alain. Statist. Surv. 2010. Vol. 4. pp. 40-79.

11. kaggle. URL: kaggle.com/datasets/johnsmith88/heart-disease-dataset/data (data obrashheniya: 28.08.2024).

Дата поступления: 28.08.2024

Дата публикации: 01.10.2024