DOI: 10.58168/MIST2026_332-339
УДК 004.02

ЛОГИСТИЧЕСКАЯ МОДЕЛЬ ПРОГНОЗИРОВАНИЯ САХАРНОГО ДИАБЕТА

В.М. Леонов[1], С.Л. Сахаров[1]

[1]ФГБОУ ВО «Воронежский государственный лесотехнический университет имени Г.Ф. Морозова»

Аннотация. Логистическая регрессия является одним из наиболее распространённых методов анализа бинарных исходов в медицине и биологии. Метод обеспечивает вероятностную интерпретацию результата и позволяет оценивать вклад факторов риска через отношение шансов. В статье предложена воспроизводимая методика построения и валидации логистической модели для медицинских данных, включающая этапы предобработки, обучения и оценки качества по ROC-анализу и калибровке вероятностей. Практическая часть продемонстрирована на задаче прогнозирования сахарного диабета. Полученные результаты подтверждают пригодность логистической регрессии как интерпретируемого инструмента поддержки принятия медицинских решений.

Ключевые слова: логистическая регрессия, биостатистика, ROC-анализ, AUC, медицинские данные, факторы риска.

LOGISTIC REGRESSION MODEL FOR DIABETES MELLITUS PREDICTION

V.M. Leonov[1], S.L. Sakharov[1]

[1]Voronezh State University of Forestry and Technologies named after G.F. Morozov

Abstract. Logistic regression is one of the most widely used methods for analyzing binary outcomes in medicine and biology. The method provides a probabilistic interpretation of the result and allows the contribution of risk factors to be assessed through the odds ratio. The paper proposes a reproducible methodology for constructing and validating a logistic model for medical data, including data preprocessing, model training, and quality assessment using ROC analysis and probability calibration. The practical application is demonstrated on the problem of diabetes mellitus prediction. The obtained results confirm the suitability of logistic regression as an interpretable tool for supporting clinical decision-making.

Keywords: logistic regression, biostatistics, ROC analysis, AUC, medical data, risk factors.

© Леонов В. М., Сахаров С. Л., 2026

Введение

Медицинские исследования часто связаны с анализом бинарных исходов: наличия заболевания, развития осложнения, токсического эффекта или выживаемости пациента. В клинической практике важно не только отнести пациента к категории «здоров» или «болен», но и оценить вероятность наступления события, поскольку именно вероятность используется при принятии диагностических и профилактических решений.

Среди методов машинного обучения логистическая регрессия занимает особое место благодаря интерпретируемости параметров. Коэффициенты модели напрямую связаны с отношением шансов, что позволяет трактовать их как факторы риска. По этой причине метод широко применяется в эпидемиологии и доказательной медицине [3]. Одним из наиболее распространённых хронических заболеваний является сахарный диабет. Раннее выявление пациентов группы риска позволяет снизить частоту осложнений и повысить эффективность профилактических мероприятий [2].

Цель работы – построение и исследование логистической модели прогнозирования сахарного диабета.

Разработанная система относится к узконаправленным системам искусственного интеллекта, формирующим прогноз на основе входных данных, в соответствии с терминологией ГОСТ Р 71476-2024. Модель машинного обучения обучалась методом логистической регрессии на размеченных медицинских данных [7], после чего использовалась для вычисления вероятности наличия сахарного диабета у пациента по значениям признаков [8]. Таким образом, логистическая регрессия выступает алгоритмом обучения, а итоговая обученная модель выполняет функцию прогноза заболевания.

Материалы и методы

В работе использован открытый медицинский набор данных Pima Indians Diabetes Database. Выборка содержит 768 пациентов, из них 268 имеют диабет (34,9%), 500 – здоровы (65,1%). Используемые в модели клинические признаки представлены в таблице 1. Они отражают основные физиологические и демографические характеристики пациента, потенциально влияющие на риск развития сахарного диабета. Целевая переменная – наличие диабета.

Построение модели осуществлялось с позиций системного подхода, предполагающего формализацию структуры диагностической задачи и выделение значимых факторов риска [7].

Таблица 1 – Используемые признаки

Признак	**Описание**
Pregnancies	число беременностей
Glucose	уровень глюкозы
BloodPressure	артериальное давление
SkinThickness	толщина кожной складки
Insulin	уровень инсулина
BMI	индекс массы тела
DiabetesPedigreeFunction	наследственный фактор
Age	возраст

Данные и предобработка

При построении прогностической модели важно, чтобы этапы обработки данных не использовали информацию из тестовой выборки. Все операции, которые «настраиваются» по данным, выполняются только на обучающей части выборки. В противном случае возникает утечка информации, и качество модели оказывается завышенным [6]. Предобработка включает следующие этапы:

– определение единицы наблюдения (пациент, визит или анализ), чтобы не использовать показатели, измеренные после постановки диагноза;

– проверку допустимых диапазонов медицинских показателей;

– кодирование категориальных признаков при необходимости;

– масштабирование числовых признаков (особенно при использовании регуляризации).

В медицинских данных пропуски встречаются часто и обычно не являются случайными. Кроме явных отсутствующих значений возможны скрытые пропуски – физиологически невозможные показатели, например нулевой индекс массы тела или нулевая концентрация глюкозы. Такие значения рассматриваются как отсутствующие. Предобработка выполняется следующим образом:

– анализируется доля пропусков по каждому признаку;

– при небольшом числе пропусков допускается удаление наблюдений;

– при большем объёме используется заполнение пропусков, например медианой [7];

– при необходимости применяется более сложное множественное заполнение.

Поскольку число пациентов с заболеванием существенно меньше числа здоровых, метрика общей точности классификации не является информативной: модель может демонстрировать высокую точность при неудовлетворительном

выявлении больных. В связи с этим оценка проводилась по клинически значимым показателям. Использовалось взвешивание классов и подбор порога классификации с приоритетом чувствительности, а также анализ прогностической ценности положительного и отрицательного результатов.

При наличии сильной взаимосвязи признаков оценки коэффициентов логистической регрессии становятся нестабильными и затрудняют интерпретацию влияния факторов риска. Для выявления мультиколлинеарности анализировалась корреляция признаков и рассчитывался коэффициент инфляции дисперсии. При необходимости зависимые признаки исключались либо применялась регуляризация, стабилизирующая параметры модели [4].

Оценка качества модели

Основной характеристикой способности модели различать больных и здоровых пациентов является ROC-кривая и площадь под ней (AUC – *Area Under the Curve*) [5]. Значение AUC можно интерпретировать как вероятность того, что случайный пациент с заболеванием получит более высокий прогнозируемый риск, чем здоровый. Для более детальной оценки дополнительно рассчитываются чувствительность и специфичность, позволяющие определить, насколько хорошо модель выявляет заболевших и исключает здоровых [3]. Однако одной только способности различать классы недостаточно. В медицинских задачах важно, чтобы предсказанные вероятности соответствовали реальной частоте заболевания. Поэтому проводится калибровка модели: строится калибровочная кривая и вычисляется показатель Brier score (метрика качества вероятностных прогнозов). Даже при высоком значении AUC плохо откалиброванные вероятности могут привести к неверным клиническим решениям.

Надёжная оценка качества модели также требует корректной процедуры проверки. Простого разделения данных на обучающую и тестовую выборки недостаточно, поскольку результат может зависеть от конкретного разбиения. Поэтому используется кросс-валидация или bootstrap-валидация, позволяющие снизить риск переобучения и получить более устойчивые характеристики модели. Для оценки способности модели различать пациентов с заболеванием и без него была построена ROC-кривая (рисунок 1).

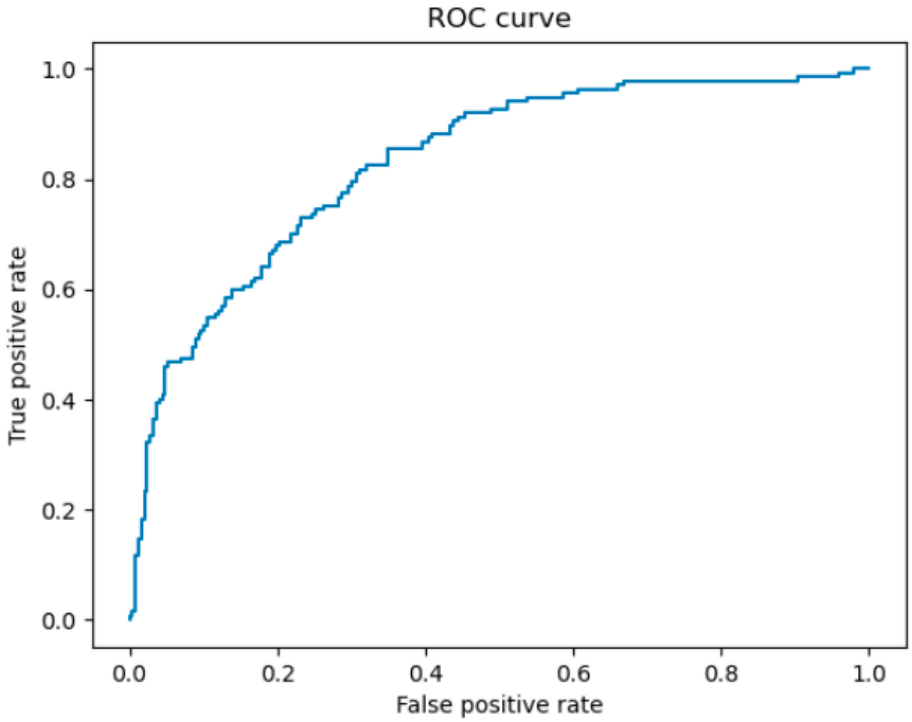

Рисунок 1 – ROC-кривая модели прогнозирования сахарного диабета

Из рисунка 1 и таблицы 2 следует, что модель обеспечивает приемлемое разделение пациентов по наличию заболевания, что подтверждается значением площади под ROC-кривой. Верификация модели заключалась в подтверждении корректности реализации алгоритма обучения и вычислений: проверялась корректность подготовки выборок, устойчивость обучения при повторных запусках и воспроизводимость получаемых параметров модели. Валидация проводилась с точки зрения соответствия модели поставленной медицинской задаче прогнозирования заболевания. Для этого применялась перекрестная проверка и анализ способности модели корректно разделять пациентов на группы риска.

Оценка качества выполнялась на основе метрик дискриминационной способности и калибровки вероятностей. Использовались ROC-AUC, чувствительность, специфичность, доля правильных классификаций и показатель Brier score. Такой подход позволяет оценить, как способность модели различать классы, так и точность вероятностных прогнозов. Количественные показатели качества классификации представлены в таблице 2.

Интерпретация влияния признаков на вероятность заболевания приведена в таблице 3.

Таким образом видно, что наибольший вклад в прогноз вносит уровень глюкозы, что соответствует медицинским представлениям о факторах риска сахарного диабета.

Таблица 2 – Метрики качества

Метрика	Значение
AUC	0.79
Чувствительность	0.82
Специфичность	0.74
Доля правильных классификаций	0.77
Brier score	0.16

Таблица 3 – Интерпретация коэффициентов

Признак	OR	Интерпретация
Glucose	1.08	повышение глюкозы увеличивает риск диабета
BMI	1.04	увеличение массы тела повышает риск
Age	1.03	риск возрастает с возрастом

Корректность применения логистической регрессии в медицинских исследованиях определяется не только математической формой модели, но и соблюдением процедуры анализа данных. Использование обучающей выборки для оценки качества приводит к систематическому завышению результатов, поэтому необходима раздельная, а при возможности и внешняя валидация на независимых данных. Площадь под ROC-кривой характеризует дискриминационную способность модели, однако не отражает точность вероятностных оценок. В связи с этим наряду с ROC-анализом обязательна проверка калибровки предсказанных вероятностей. При ограниченном объёме выборки или большом числе признаков возрастает риск переобучения, что требует применения регуляризации, обеспечивающей устойчивость коэффициентов и переносимость модели. Дополнительные трудности возникают при редких событиях и почти полном разделении классов; в этих случаях используются взвешивание классов и регуляризованные версии логистической регрессии.

Заключение

В работе показано, что логистическая регрессия является эффективным и интерпретируемым методом анализа медицинских данных, позволяющим оценивать вероятность заболевания и количественно характеризовать вклад факторов риска. Надёжность получаемых результатов обеспечивается корректной предобработкой данных, применением процедур валидации и оценкой калибровки

вероятностей. Соблюдение указанных требований делает логистическую регрессию пригодным инструментом для задач скрининга и поддержки принятия клинических решений.

Список литературы

1. Лапач С. Н. Статистические методы в медико-биологических исследованиях с использованием Excel Эксперим. исслед. Клин. испытания. Анализ фармац. рынка / С.Н. Лапач, А.В. Чубенко, П.Н. Бабич. – 2-е изд., доп., с магнит. носителем. – Киев : МОРИОН, 2001. – 407 с. ил.; 26. – (Медицина и статистика); ISBN 966-7632-33-4.

2. Сахарный диабет : диагностика, лечение, профилактика : руководство / И. И. Дедов [и др.] ; под ред. И. И. Дедова, М. В. Шестаковой. – Москва : Мед. информ. агентство, 2011. – 801 с. : цв. ил., портр. – ISBN 978-5-9986-0061-6.

3. Clinical prediction models: A practical approach to development, validation, and updating Ewout W. Steyerberg (2019). Second Edition, Springer Series Statistics for Biology and Health. Cham: Springer. 558 pages. ISBN: 978-3-030-16398-3. https://doi.org/10.1007/978-3-030-16399-0.

4. Kim J. H. Multicollinearity and misleading statistical results // Korean Journal of Anesthesiology. – 2019. – Vol. 72, No. 6. – P. 558–569.

5. Lachiche N., Flach P. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves // Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). – Washington : AAAI Press, 2003. – Vol. 20. – P. 416–424.

6. Kuhn M., Johnson K. Applied Predictive Modeling. – New York : Springer, 2013. – 600 p.

7. Коробова Л. А., Эйнуллаев Т. А. О., Тэн Р. В. Системный подход при решении прикладных медицинских диагностических задач // Моделирование систем и процессов. – 2024. – Т. 17, № 3. – С. 52–61. – DOI: 10.12737/2219-0767-2024-50-59.

8. Арапов Д. В., Курицын В. А., Скоробогатов С. А. Автоматизированная мобильная система принятия решений для использования в специальной медицине // Моделирование систем и процессов. – 2024. – Т. 17, № 3. – С. 7–16. – DOI: 10.12737/2219-0767-2024-5-13.

9. Юрчишина М. В., Бушмелева К. И. Алгоритмическая модель СППР «Оптимальный учебный план» // Моделирование систем и процессов. – 2024. – Т. 17, № 4. – С. 84–95. – DOI: 10.12737/2219-0767-2024-17-4-84-95.

References

1. Lapach S. N. Statistical methods in biomedical research using Excel Experiments. research. The wedge. trials. Pharmaceutical market analysis / S.N. Lapach, A.V. Chubenko, P.N. Babich. – 2nd ed., supplement, with magnet. the carrier. Kiev : MORION, 2001. 407 p. ill.; 26. –Medicine and statistics); ISBN 966-7632-33-4.

2. Diabetes mellitus : diagnosis, treatment, prevention : a guide / I. I. Dedov [et al.]; edited by I. I. Dedov, M. V. Shestakova. – Moscow : Med. inform. Agency, 2011. – 801 p. : tsv. il., portr. – ISBN 978-5-9986-0061-6.

3. Clinical prediction models: A practical approach to development, validation, and updating Ewout W. Steyerberg (2019). Second Edition, Springer Series Statistics for Biology and Health. Cham: Springer. 558 pages. ISBN: 978-3-030-16398-3. https://doi.org/10.1007/978-3-030-16399-0.

4. Kim J. H. Multicollinearity and misleading statistical results // Korean Journal of Anesthesiology. – 2019. – Vol. 72, No. 6. – P. 558–569.

5. Lachiche N., Flach P. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves // Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). – Washington : AAAI Press, 2003. – Vol. 20. – P. 416–424.

6. Kuhn M., Johnson K. Applied Predictive Modeling. – New York : Springer, 2013. – 600 p.

7. Korobova L. A., Eynullaev T. A. O., Ten R. V. System approach in applied medical diagnostic tasks // Modeling of Systems and Processes. – 2024. – Vol. 17, No. 3. – P. 52–61. – DOI: 10.12737/2219-0767-2024-50-59.

8. Arapov D. V., Kuritsyn V. A., Skorobogatov S. A. Automated mobile decision-making system for special medicine applications // Modeling of Systems and Processes. – 2024. – Vol. 17, No. 3. – P. 7–16. – DOI: 10.12737/2219-0767-2024-5-13.

9. Yurchishina M. V., Bushmeleva K. I. Algorithmic DSS model “Optimal Curriculum” // Modeling of Systems and Processes. – 2024. – Vol. 17, No. 4. – P. 84–95. – DOI: 10.12737/2219-0767-2024-17-4-84-95.: 10.12737/2219-0767-2024-17-4-84-95.