

56. РЕГРЕССИОННЫЕ МОДЕЛИ В МАШИННОМ ОБУЧЕНИИ

Бабиц Д.Д., Михалюк А.В.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Шинкевич Е.А. — канд. физ.-мат. наук

Аннотация. Регрессионный анализ и логистическая регрессия играют важную роль в принятии решений в финансовом секторе, особенно в сфере кредитного скоринга. Кредитный скоринг представляет собой процесс количественной оценки риска предоставления кредита на основе анализа финансового состояния и поведения заемщика. В данной работе рассматривается применение логистической регрессии как метода машинного обучения для предсказания вероятности невыплаты кредита заемщиком.

Машинное обучение — это подраздел теории искусственного интеллекта, в котором изучаются и исследуются алгоритмы, способные обучаться без прямого программирования [1]. Одним из ключевых разделов машинного обучения является регрессионный анализ.

Регрессионный анализ — это статистический метод, позволяющий исследовать связь переменных. Основная цель регрессионного анализа — предсказать значение одной переменной на основе других переменных [2].

Регрессия — это один из ключевых методов анализа данных, который позволяет выявлять и количественно оценивать зависимость между переменными. Регрессия позволяет моделировать зависимость между входными данными и целевой переменной. Этот метод, как правило, используется для прогнозирования и выяснения причинно-следственных связей между переменными [3].

Существуют разные модели машинного обучения, каждая из которых применяется в зависимости от цели исследования:

- регрессионные модели предсказывают числовую характеристику;
- модели классификации предсказывают категорию объекта по заранее заданным параметрам;
- регрессионно-классификационные модели умеют решать оба типа задач [4].

В линейной регрессии (регрессионная модель) каждая независимая переменная напрямую связана с зависимой и не связана с другими независимыми переменными. Эта взаимосвязь известна как линейная зависимость. Зависимая переменная обычно представляет собой значение из диапазона непрерывных значений.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (1)$$

где, y — прогнозируемая зависимая переменная; β_0 — пересечение по оси y , когда все независимые входные переменные равны нулю; $\beta_1 \dots \beta_n$ — коэффициенты регрессии при независимых переменных; ε — ошибка модели.

В логистической регрессии (модель классификации) зависимая переменная представляет собой одну из ограниченного набора категорий, называемых категориальными переменными. Этот метод применяется для задач бинарной классификации, где необходимо определить, к какому из двух возможных классов относится наблюдение [5].

$$y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon} / (1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon}) \quad (2)$$

где, y дает вероятность успеха категориальной переменной y .

Используются также следующие регрессионно-классификационные модели:

1 **Дерево решений.** В этой модели принятие решений зависит от «узлов» и «листьев». На вершине дерева решений — начальный узел, в него попадает вся выборка. Затем она проверяется на выполнение условия или наличие признака.

2 **Модель случайного леса.** Более усовершенствованная версия модели «Дерево решений» — ей реже нужна перенастройка. Состоит из нескольких деревьев решений, связанных между собой.

3 **Наивный байесовский классификатор.** Основана на теореме Байеса. Помогает понять, каким будет событие X , если наступит связанное с ним событие Y . В классификации Байеса X — это класс объекта, Y — его признак. Также существуют и другие модели. [4].

Регрессионный анализ и логистическая регрессия широко применяется в банковской сфере для различных задач, включая прогнозирование, оценку рисков и принятие управленческих решений. В частности, использоваться в кредитном скоринге для оценки вероятности дефолта клиента.

Скоринг представляет собой статистическую модель, с помощью которой на основе кредитной истории прошлых клиентов банк пытается определить, насколько велика вероятность, что конкретный потенциальный заемщик вернет кредит в срок.

С помощью алгоритмов машинного обучения можно автоматизировать процесс кредитного скоринга. Данное решение позволит:

- сократить время, требуемое для кредитного скоринга;
- уменьшить вероятность ошибки принятого решения;
- работать с большим количеством данных заемщиков [7].

В работе реализована модель логистической регрессии для бинарной классификации, целью которой является предсказание вероятности дефолта у клиентов на основе их социально-демографических и финансовых характеристик.

Исследуемые данные включают количественные и категориальные характеристики заемщиков, такие как возраст, пол, стаж работы, уровень дохода, уровень образования, семейное положение, тип занятости, цель кредита. Целевая переменная принимает значения 1 (одобрение кредита) и 0 (отказ кредита). Для повышения точности модели выполняется предобработка данных, включающая нормализацию числовых признаков, кодирование категориальных данных, а также масштабирование данных перед обучением модели.

После предобработки данных, выполняется разделение на обучающую (80%) и тестовую (20%) выборки. Оценка качества включает следующие метрики: F1-мера (0,76), ROC-AUC (0,95), точность (0,89), полнота (0,74), точность прогноза (0,78), а также визуализация матрицы ошибок и ROC-кривой (рисунок 1).

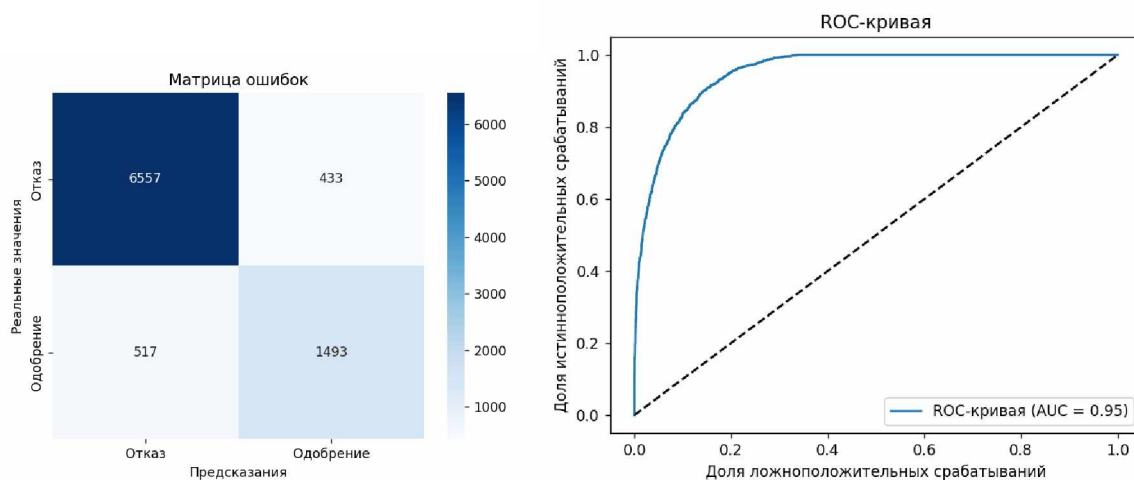


Рисунок 1 – Визуализация матрицы ошибок и ROC-кривой

Полученные результаты показывают, что модель способна качественно классифицировать заемщиков, однако имеются случаи ложноположительных и ложноотрицательных предсказаний. Матрица ошибок свидетельствует о некотором дисбалансе классов, однако на практике невозможно получить модель, которая обладает 100% чувствительностью и специфичностью [7]. При этом, показатель ROC-AUC составляет 0,95 при максимально возможном значении 1,00. В результате, логистическая регрессия продемонстрировала высокие показатели, что говорит о способности модели эффективно классифицировать заемщиков на тех, кому будет одобрен кредит, и тех, кому будет отказано.

Таким образом, проведенное исследование показало, что логистическая регрессия как модель машинного обучения является полезным инструментом для оценки кредитных рисков, помогая прогнозировать вероятность дефолта заемщиков. Это означает, что представленная модель может быть использована для задач кредитного скоринга.

Список использованных источников:

1. Базовые принципы машинного обучения на примере линейной регрессии. [Электронный ресурс]. – Режим доступа: <https://habr.com/ru/company/ods/blog/322076>. – Дата доступа: 25.02.2025.
2. Регрессионный анализ. [Электронный ресурс]. – Режим доступа: <https://practicum.yandex.ru/blog/chto-takoe-regressionnyj-analiz/>. – Дата доступа: 25.02.2025.
3. Ротова, О.М., Шибанова, А.Д. Основные принципы метода линейной регрессии / О.М. Ротова, А.Д. Шибанова // Теория и практика современной науки. – 2020. – №1 (55). – С. 483-487.
4. Модели машинного обучения. [Электронный ресурс]. – Режим доступа: <https://practicum.yandex.ru/blog/modeli-mashinnogo-obucheniya/#kakie-byvayut>. – Дата доступа: 11.03.2025.

*61-я Научная Конференция Аспирантов, Магистрантов и Студентов БГУИР,
Минск 2025*

5. Регрессионный анализ: как находить закономерности в данных и строить прогнозы. [Электронный ресурс]. – Режим доступа: <https://blog.skillfactory.ru/regressionnyy-analiz-kak-nahodit-zakonomernosti-v-dannyh-i-stroit-prognozy/>. – Дата доступа: 13.03.2025.

6. Воронин С.М., Кредитный скоринг, реализованный с помощью машинного обучения / С.М. Воронин // Столыпинский вестник. – 2022. – №10. – С. 5647-5648.

7. Логистическая регрессия и ROC-анализ — математический аппарат. [Электронный ресурс]. – Режим доступа: <https://loginom.ru/blog/logistic-regression-roc-auc>. – Дата доступа: 02.04.2025.