МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ПРИМЕНЕНИЕ МОДЕЛИ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ В КРЕДИТНОМ СКОРИНГЕ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Осиповой Татьяны Дмитриевны

Научный руководитель
доцент, к. ф.-м. н. _____ О. А. Мыльцина

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2022

ВВЕДЕНИЕ

Актуальность темы. Разработка моделей кредитного скоринга стала одним из основных направлений деятельности финансовых учреждений. Для решения этой задачи исследовались различные алгоритмы классификации, однако в литературе слабо освещен вопрос, посвященный использованию для оценки клиентской кредитоспособности таких больших данных, как карточные транзакции. За последние десятилетия банки собрали множество информации, описывающей поведение своих клиентов. Поскольку истории карточных транзакций накапливаются по каждому клиенту, то их использование в оценке кредитного риска могло бы дать существенный прирост информации и, как следствие, повысить прогнозную точность моделей. Главной задачей данного исследования является установление целесообразности использования карточных транзакций для оценки кредитного скоринга. С этой целью написана программа, основанная на модели логистической регрессии.

Данная работа представляет интерес поскольку предложенныая модель кредитного скоринга имеют научную новизну. Data Mining **Целью бакалаврской работы** является разработка и анализ модели кредитного скоринга.

Объект исследования — кредитный скоринг.

Предмет исследования — особенности построения модели логистической регрессии.

Для достижения указанной цели были поставлены следующие задачи:

- рассмотреть понятие кредитного риска и его классификацию;
- ознакомиться с системой риск-менеджмента;
- описать количественную оценку кредитного риска — кредитный скоринг;
- ознакомиться с моделью логистической регрессии;
- реализовать алгоритм по данной теме на языке программирования Python.

Практическая значимость проводимого исследования состоит в том, что на основании построенной модели в области оценки кредитоспособности заемщиков возможно снизить кредитный риск и в итоге улучшить рейтинг организации. По результатам вычислений имеется возможность сделать вы-

воды о будущих потерях и прибыли.

Структура и содержание бакалаврской работы. Работа состоит из введения, двух разделов, пяти параграфов, заключения, списка использованных источников и приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы работы, формулируется цель работы и решаемые задачи, отмечается практическая значимость полученных результатов.

В **первом** разделе работы приводятся основные понятия из финансовой сферы, связанные с кредитным риском, и количественная оценка кредитного риска — кредитный скоринг.

В зависимости от доступности информации методики оценки кредитного риска подразделяются на четыре категории:

1. Полностью субъективная оценка порождается дефицитом данных. Основывается на мнении, без каких-либо моделей или правил.
2. Экспертные системы оценки — в этой методике специалисты полагаются на имеющийся опыт, разработанные правила и модели, которые используются либо для принятия решения, либо для «подсказки» специалистам при принятии решения.
3. Гибридные методики — доступность данных варьируется. Как правило, объединяются экспертные системы оценки и статистические модели.
4. Статистические модели применяются при достаточном объеме структурированных данных. Такая оценка кредитного риска на сегодняшний день является наиболее объективной и точной.

Наиболее эффективной оценкой заемщика признана скоринговая модель, представляющая собой отработанный числовой алгоритм для оценки покупательной способности физического лица. С ее помощью банк, выступающий в роли кредитора, оценивает возможности клиента (заемщика), выступающего в роли покупателя, и по итогам оценки дает согласие на сделку либо отказывает выдачу средств. Кредитоспособность заемщика оценивается по основным параметрам: сумма среднемесячного дохода, продолжительность

трудового стажа и время занятости клиента на последнем месте работы, возраст получателя кредита и его семейное положение и т. д. Цель кредитного скоринга состоит в том, чтобы разделить претендентов на две группы: первая группа — это те, кто наиболее вероятно сможет погасить свои финансовые обязательства в будущем, вторая — те, кому стоит отказать в кредите, так как велика вероятность того, что будут не выполнены финансовые обязательства.

Разработка скоринговых моделей происходит по следующему плану:

1. Сбор и подготовка данных: большую часть этой работы берет на себя банк.
2. Анализ данных и статистических показателей.
3. Построение модели.
4. Оценка модели.

Подготовка и анализ данных. Все переменные проверяются на наличие взаимосвязей. Корреляционная связь может существовать между двумя переменными, а может и между несколькими. Последнее явление называется мультиколлинеарностью, и для его измерения используется фактор инфляции дисперсии (VIF):

$$VIF_j = \frac{1}{1 - R_j^2},\ R_j^2 = 1 - \frac{\sum\limits_{i=1}^{m} (x_{ij} - \hat{x}_{ij})^2}{\sum\limits_{i=1}^{m} (x_{ij} - \overline{x}_j)^2}.$$

где R_j^2 — коэффициент детерминации j-го признака относительно остальных признаков, x_{ij} — значение j-ой независимой переменной для i-го наблюдения, $\overline{x}_j$ — среднее значение j-го признака, $\hat{x}_{ij}$ — оценка j-ой независимой переменной для i-го наблюдения.

Тем не менее избежать отсутствия связей между признаками практически невозможно, поэтому коррелирующие показатели можно включить в модель, если выполняются следущие соотношения:

$$\begin{cases} r_{yx_i} > r_{x_i x_j}, \text{при } i \neq j; \\ \\ r_{yx_j} > r_{x_i x_j}, \text{при } i \neq j; \end{cases}$$

$$VIF_j \leq 3,$$

r_{yx_i} — коэффициент корреляции между зависимой (y) и независимой (x_i) переменными, $r_{x_ix_j}$ — теснота связи между признаками x_i и x_j. Наиболее популярными показателями оценки статистической значимости признаков являются Weight of Evidence (WoE) и Information Value (IV).

WoE показывает насколько экзогенная переменная способна спрогнозировать значение эндогенной переменной. Рассчитывается как логарифм от отношения частот «хороших» кредитов к частотам «плохих» кредитов по каждому из признаков:

$$WoE_k = \ln\left(\frac{p_k}{q_k}\right) = \ln\left(\frac{Event\%}{NonEvent\%}\right),$$

k — номер группы, p_k — доля платежеспособных клиентов среди всех плетежеспособных (частота «хороших» кредитов), q_k — доля неплатежеспособных клиентов среди всех неплетежеспособных (частота «плохих» кредитов). Далее произоводится подсчет IV, величины, определяющей значимость переменной в модели бинарной классификации:

$$IV = \sum_{k=1}^{l} (p_k - q_k) * WoE_k.$$

Затем происходит отбор признаков IV по Таблице 1:

Таблица 1 – Значения IV

Интервал IV	Влияние
<0,02	Бесполезно для предсказания
0,02 – 0,1	Слабое
0,1 – 0,3	Среднее
0,3 – 0,5	Хорошее
>0,5	Отличное

Признаки со значением IV от 0,3 до 1 принято брать для прогнозирования.

Далее происходит разбиение данных на группы внутри признаков по следующим критериям:

1. размер группы должен быть не менее 5% от общего объема выборки;
2. объединение в группы осуществляется по близким показателям WoE. Это необходимо для максимизации расстояний между группами.

Модель логистической регрессии. Логистическая регрессия — модель регрессии, общее назначение которой состоит в анализе связи между независимыми переменными и зависимой переменной. В данной работе используется бинарная логистическая регрессия, т. к. зависимая переменная может принимать только два значения: 0 или 1.

Математическая модель логистической регрессии имеет вид:

$$P_i = \frac{1}{1 + exp(-(w_0 + w_1 x_{i1} + w_2 x_{i2} + ... + w_n x_{in} + \epsilon_i))},$$

где i — порядковый номер заемщика, P_i — вероятность наступления дефолта по кредиту для i-го заемщика, n — количество признаков, w_0 — независимая константа модели, w_j — параметры модели или веса модели, x_{ij} — значение j-ой независимой переменной для i-го наблюдения.

График функции P_i имеет вид

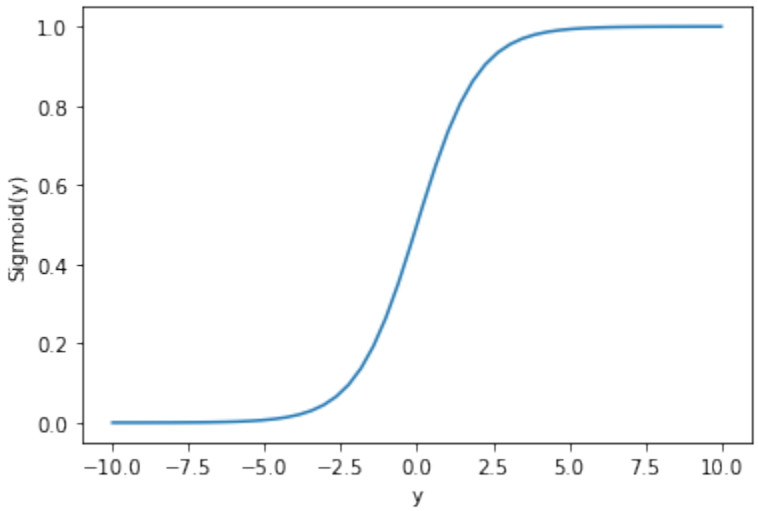

Рисунок 1 – Сигмоид-функция

Требуется найти $\boldsymbol{w} = (w_0, ..., w_n)$ — $(n+1) \times 1$ вектор коэффициентов модели.

Оценка работы модели. Одной из полезных метрик для оценки прогностических моделей является кривая ROC (Receiver Operating Characteristic).

AUC (Area Under Curve) — это площадь под кривой ROC.

ROC-кривая представляет из себя линию от (0,0) до (1,1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR):

$$TPR = \frac{TP}{TP + FN} * 100\%, \qquad FPR = \frac{FP}{TN + FP} * 100\%.$$

Где метрики означают следущее:

1. True Positive (TP) — сколько раз модель правильно классифицировала Positive как Positive.
2. False Negative (FN) — сколько раз модель неправильно классифицировала Positive как Negative.
3. False Positive (FP) — сколько раз модель неправильно классифицировала Negative как Positive.
4. True Negative (TN) — сколько раз модель правильно классифицировала Negative как Negative.

На их основе можно рассчитать другие метрики, которые предоставляют дополнительную информацию о поведении модели:

1. Accuracy (доля правильных ответов) — это показатель, который описывает общую точность предсказания модели по всем классам. Это особенно полезно, когда каждый класс одинаково важен. Он рассчитывается как отношение количества правильных прогнозов к их общему количеству.
$$Accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}.$$
2. Precision представляет собой отношение числа семплов, верно классифицированных как Positive, к общему числу результатов с меткой Positive (распознанных правильно и неправильно). Precision измеряет точность модели при определении класса Positive.
$$Precision = \frac{TP}{TP + FP}.$$
3. Recall рассчитывается как отношение числа Positive результатов, корректно классифицированных как Positive, к общему количеству Positive семплов. Показатель измеряет способность модели обнаруживать ре-

зультаты, относящиеся к классу Positive. Чем выше Recall, тем больше Positive семплов было найдено.

$$Recall = \frac{TP}{TP + FN}.$$

Площадь под кривой (AUC) в данном случае показывает качество алгоритма, кроме этого, важной является крутизна самой кривой — требуется максимизировать TPR, минимизируя FPR, а значит, кривая в идеале должна стремиться к точке (0,1). В Таблице 2 приведены интервалы и оценка по ним качества модели.

Таблица 2 – Значения AUC

Интервал AUC	Качество модели
0,9 - 1	Отличное
0,8 - 0,9	Очень хорошее
0,7 - 0,8	Хорошее
0,6 - 0,7	Среднее
0,5 - 0,6	Неудовлетворительное

Результат 0,5 говорит о том, что модель бесполезна для предсказания. Если показатели оказались ниже границы 0,5, то модель работает в точности до наоборот. Чем ближе значение AUC к 1, тем лучше модель предсказывает вероятность дефорлта.

Во **втором** разделе работы представлен вычислительный эксперимент. Целью эксперимента является моделирование кредитного скоринга. Для этого была написана программа на языке Python с использованием датасета Train dataset Kaggle Credit Scoring (1).

Входными данными программы является $m \times n$ матрица объясняющих переменных

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

Столбцами матрицы $\boldsymbol{X}$ являются $m \times 1$ векторы регрессоры $\boldsymbol{x_j} = = (x_{1j}, ..., x_{mj})'$, $j = 1, ..., n$.

После проведения анализа данных и подсчета статистических показателей значимости признаков количество регрессоров уменьшилось до 18.

Для решения задачи и оценки основных характеристик были использованны формулы из раздела 1.

С помощью комбинации методов LogisticRegression и GridSearchCV были получены следущие веса модели:

```
array([[-1.18227343e-05,  5.10414312e-04, -3.25197473e-07,
         3.09213015e-03, -3.61361511e-05, -7.15695821e-07,
         1.30734856e-03, -5.81394876e-03,  2.24860732e-02,
         9.43709598e-04,  1.74930469e-02,  7.00525686e-04,
         5.87983648e-06,  1.11650004e-02, -1.88492909e-03,
        -1.58476171e-04,  1.28701762e-02, -6.07851614e-07]])
```

В Таблице 3 представлены результаты оценки качества модели:

Таблица 3 – Метрики для оценки модели

Название метрики	Значение
accuracy	0.9214
precision	0.8517
recall	0.8767
auc	0.9734

Из Таблицы 3 видно, что общая точность предсказания модели составляет 92%, точность модели при определении класса платежеспособных клиентов — 85%, способность модели обнаруживать заемщиков, относящиеся к классу платежеспособных клиентов — 88%, качество модели составляет 97,34%.

График кривой ROC представлен на Рисунке 2.

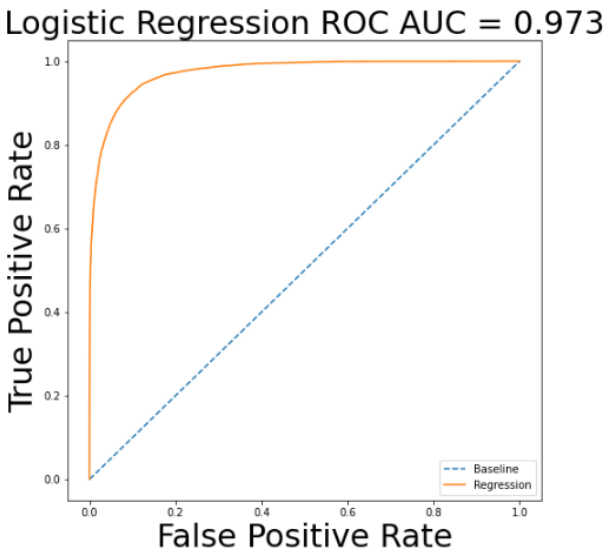

Рисунок 2 – Кривая ROC

В **заключении** приведены результаты бакалаврской работы.

Основные результаты

1. Рассмотрены основные понятия, связанные с кредитным риском и его классификация.
2. Рассмотрена система риск-менеджмента банка по кредитам.
3. Описана количественная оценка кредитного риска — кредитный скоринг;
4. Изучена модель логистической регрессии.
5. Реализован алгоритм по данной теме на языке программирования Python. Разработана программа, моделирующая работу логистической регрессии. Программный код приводится в **приложениях А, Б, В**. Разработанная программа также позволяет оценить качество модели. В результате работы программы были подсчитаны коэффициенты модели и метрики для оценки работы модели.