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 МОДЕЛЕЙ БИНАРНОГО ВЫБОРА К ЗАДАЧЕ КРЕДИТНОГО СКОРИНГА

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Осиповой Татьяны Дмитриевны

Научный руководитель
доцент, к. ф.-м. н. _____ Д. В. Мельничук

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

Саратов 2024

# 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 написаны программы, основанные на моделях логистической регрессии и случайного леса.

Данная работа представляет интерес поскольку предложенныая модель кредитного скоринга имеют научную новизну. Работа прошла апробацию на различных конференциях, в частности, на ежегодной студенческой конференции «Актуальные проблемы математики и механики», которую проводил механико-математический факультет СГУ в апреле 2024 года, в секции «Анализ данных», в XII Международной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками», ноябрь 2023 года.

**Целью бакалаврской работы** является разработка и анализ моделей бинарного выбора для решения задачи кредитного скоринга, направленной на повышение точности предсказания кредитоспособности заемщиков и минимизацию кредитных рисков.

**Объект исследования** — кредитный скоринг.

**Предмет исследования** — модели бинарного выбора такие, как логистическая регрессия и случайный лес.

Для достижения указанной цели были поставлены следующие задачи:

— рассмотреть математический аппарат моделей бинарного и множествен-

ного выбора;

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

**Практическая значимость** проводимого исследования состоит в том, что на основании построенных моделей в области оценки кредитоспособности заемщиков возможно снизить кредитный риск и в итоге улучшить рейтинг организации. По результатам вычислений имеется возможность сделать выводы о будущих потерях и прибыли.

**Структура и содержание бакалаврской работы.** Работа состоит из введения, трех разделов, восьми подразделов, 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 представлены математические основы оценивания моделей бинарного и множественного выбора.

Во **втором** разделе работы приводятся основые 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.

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**. Разработка скоринговой модели строится на анализе данных с полученного кредитного опыта. Качество исходных показателей определяет точность прогнозирования скоринговой системы.

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 такой показатель, как фактор инфляции дисперсии ($VIF$):

$$VIF_j = \frac{1}{1 - R_j^2}, \tag{1}$$

где

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

Фактор инфляции дисперсии $VIF$ (Variance Inflation Factor) показывает, во сколько раз увеличивается дисперсия коэффициента регрессии за счёт коррелированности регрессоров $x_1, ..., x_n$ по сравнению с дисперсией этого коэффициента, если бы регрессоры были некоррелированы.

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.

**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

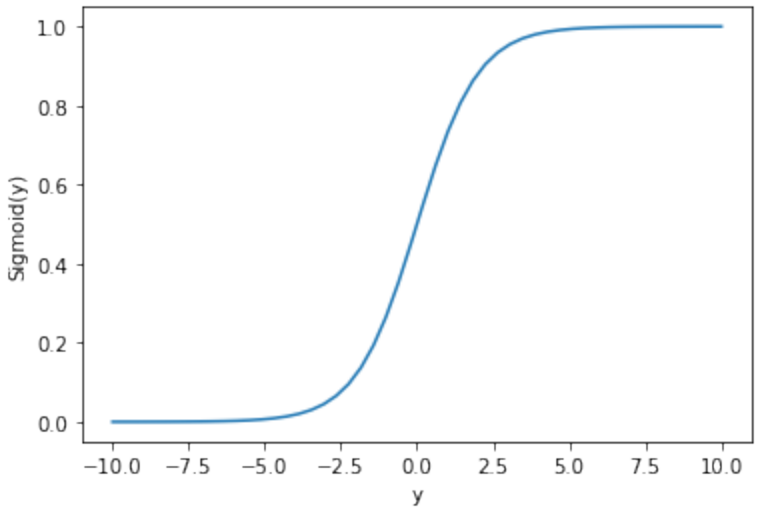


Рисунок 1 – Сигмоид-функция

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

**Модель случайного леса.** Это метод машинного обучения, который строит дерево решений на основе обучающих данных.

Теоретическую составляющую этого алгоритма можно описать следующим образом. Пусть у нас есть случайный лес с $B$ деревьями. Предположим, что для объекта $x$ каждое дерево $T_b$ предсказывает класс $y_b$. Тогда окончательное предсказание $\hat{y}_x$ определяется как класс, который получает наибольшее количество голосов среди всех предсказаний деревьев.

Формула классификатора в алгоритме случайного леса может быть

представлена следующим образом:

$$\hat{y}_x = argmax_y \sum_{b=1}^{B} I\left(y_b = y\right), \tag{3}$$

где $\hat{y}_x$ — предсказанный класс для объекта $x$, $y_b$ — предсказанный класс деревом $T_b$, $I\left(\cdot\right)$ — индикаторная функция, которая возвращает 1, если условие истинно, и 0 в противном случае.

Каждое дерево $T_b$ строится рекурсивным разделением на основе некоторого критерия информативности (например, индекс Джини или прирост информации). Рекурсивное разделение происходит до достижения критериев остановки.

**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 результаты, 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.

В **третьем** 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).

В Таблице 3 представлены результаты оценки качества модели логистической регрессии:

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.

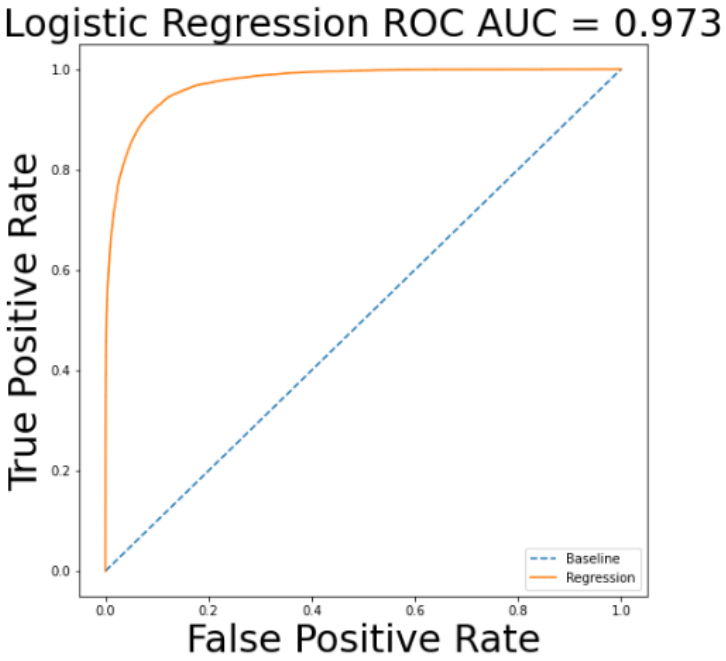


Рисунок 2 – Кривая ROC

В таблице 4 представлены результаты оценки качества модели случайного леса:

Таблица 4 – Метрики для оценки модели

| Название метрики | Значение |
|---|---|
| accuracy | 0.9996 |
| precision | 1 |
| recall | 1 |
| auc | 0.9997 |

Из таблицы 4 видно, что общая точность предсказания модели составляет 99,9%, точность модели при определении класса платежеспособных клиентов — 100%, способность модели обнаруживать заемщиков, относящиеся к классу платежеспособных клиентов — 100%, качество модели составляет 99,9%.

График кривой ROC представлен на рисунке 3.

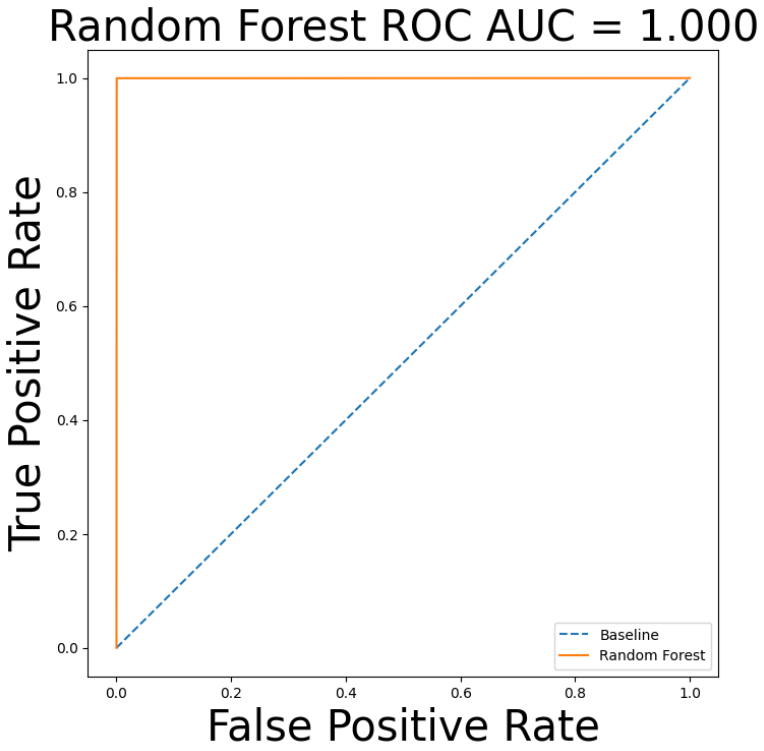


Рисунок 3 – Кривая ROC

В **заключении** приведены результаты магистерской работы.

### Основные результаты

1. Рассмотрен математический аппаратом моделей бинарного и множественного выбора.
2. Рассмотрены основные понятия, связанные с кредитным риском и его классификация.
3. Рассмотрена система риск-менеджмента банка по кредитам.
4. Описана количественная оценка кредитного риска — кредитный скоринг.
5. Изучена модель логистической регрессии.
6. Изучена модель случайного леса.
7. Реализованы алгоритмы по данной теме на языке программирования Python. Разработаны программы, моделирующие работу логистической регрессии и случайного леса. Программный код приводится в приложениях А, Б, В, Г, Д.