МИНОБРНАУКИ РОССИИ

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

Кафедра теории функций и стохастического анализа

СОЗДАНИЕ СКОРИНГОВОЙ МОДЕЛИ НА ОСНОВЕ БИННИНГА ДАННЫХ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Власовой Елены Константиновны

Научный руководитель
доцент, к. ф.-м. н. _____ Н. Ю. Агафонова

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

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. Модели кредитного скоринга являются основой для таких финансовых учреждений, как банки розничного и потребительского кредитования. Цель этих моделей — оценить вероятность неплатежеспособности заявителей на получение кредита, чтобы принять решение о предоставлении им кредита.

Потребительский кредит — это деньги, выданные банком заёмщику на цели, не связанные с предпринимательством. В последние десятилетия спрос клиентов на личные займы возрос, рынок потребительского кредитования превратился в важный сектор финансовой сферы и сегодня представляет собой крупномасштабный бизнес. Такие изменения на рынке кредитования требуют автоматических, быстрых и последовательных решений, а также процессов для обработки огромного количества заявок. Использование моделей кредитного скоринга в настоящее время является ключевым компонентом в розничном банковском обслуживании. Таким образом, разработка оценочных карт представляет собой ключевую компетенцию управления рисками банка при оценке кредитоспособности физического лица.

Проблема заключается в том, что классические методы машинного обучения, такие как логистическая регрессия, предполагают линейную связь между признаками и целевой переменной. Однако в реальных кредитных данных эта предпосылка нарушается: зависимости являются нелинейными (например, риск дефолта имеет U-образную зависимость от возраста). Решением данной проблемы является применение WoE-бэннинга (Weight of Evidence Binning), который преобразует нелинейные зависимости в линейные.

Целью дипломной работы является разработка системы кредитного скоринга, использующей WoE-биннинг для повышения интерпретируемости и точности прогнозирования дефолта.

Для достижения поставленной цели необходимо решить следующие задачи:

— проанализировать существующие методы кредитного скоринга;
— реализовать алгоритм WoE-биннинга и расчёт информационной силы признаков;

— обучить модель на WoE-преобразованных данных;
— разработать Telegram-бот для оценки кредитного риска;
— создать пользовательское веб-приложение;
— реализовать экспорт отчётов в формате PDF.

Дипломная работа состоит из введения, 2 разделов, заключения и списка использованных источников.

В первом разделе рассматриваются теоретические аспекты кредитного скоринга и биннинга данных.

Во втором разделе описана практическая реализация системы: использование WoE-биннинга, создание Telegram-бота и разработка веб-интерфейса.

В заключении приведены основные результаты и выводы дипломной работы.

Список источников содержит 20 наименований, на которые в тексте работы присутствуют ссылки.

Основное содержание работы

В **первом** разделе раскрываются основные понятия скоринговых моделей и биннинга данных.

Скоринговой называют математическую модель, которая анализирует субъект проверки по разным параметрам, оценивает его качества и делает выводы о заданных характеристиках. Само слово «scoring» произошло от английского «score» — счёт, то есть скоринг — это подсчёт очков. Модель опирается на систему оценок, установленную для каждой характеристики: профессиональной, социальной, демографической, финансовой.

В работе основное внимание уделяется эксплуатационным характеристикам приёмника (ROC) и соответствующей площади под кривой (AUC) как наиболее широко применяемому показателю эффективности в практике кредитного скоринга. AUC измеряется от 0,5 до 1. Обычно считают, что значение площади от 0,9 до 1 соответствует отличному качеству модели, от 0,8–0,9 — очень хорошему, 0,7–0,8 — хорошему, 0,6–0,7 — среднему, 0,5–0,6 — неудовлетворительному.

Биннинг — это процесс дискретизации путём извлечения небольших групп (бинов) из непрерывной переменной. Применение биннинга в кредит-

ном скоринге позволяет:

— повысить точность моделей (WoE-преобразования делают признаки линейными по отношению к логарифмическим шансам);

— устранить влияние выбросов и шума;

— облегчить интерпретацию результатов.

В работе используется квантильный биннинг (quantile binning), при котором данные делятся на группы с равным количеством наблюдений. Этот метод обеспечивает устойчивость к выбросам, адаптивность и равномерное заполнение бинов.

Weight of Evidence (WoE) — это метод преобразования данных, который показывает предсказательную силу независимой переменной (фактора) по отношению к зависимой переменной.

Рассчитывается по формуле:

$$WoE = ln(\frac{d_1}{d_2}), \tag{1}$$

где d_1 – доля дефолтов при возврате кредита, d_2 - доля возвратов без обременений.

Исходные данные разбиваются на группы (бины). Для каждой группы рассчитывается значение WoE.

Information Value (IV) или информационное значение — это статистический показатель в кредитном скоринге, используемый для оценки предсказательной силы (значимости) независимой переменной (фактора) при прогнозировании вероятности дефолта заемщика.

Рассчитывается по формуле:

$$IV = \sum_{i=1}^{k}(d_1 - d_2) * WoE_i, \tag{2}$$

где:

— k - количество групп (бинов) переменной,

— d_1 – доля дефолтов при возврате кредита в бине,

— d_2 - доля возвратов без обременений в бине,

— WoE_i – вес доказательства бина.

IV рассчитывается на этапе подготовки данных, до обучения модели. Он помогает отобрать признаки, которые действительно влияют на дефолт.

Во **втором** разделе приведен процесс реализации программы кредитного скоринга.

Для обучения и тестирования модели использовался набор данных, содержащий 5 890 записей о кредитных заёмщиках (sample_data.csv). Каждая запись включает 4 параметра:

— age - возраст заёмщика на момент подачи заявки (от 22 до 70 лет);
— income - годовой доход заёмщика (от 80000 до 1440000, в рублях);
— debt_ratio - долговая нагрузка (отношение ежемесячных платежей к доходу, от 0.00 до 0.80);
— default - целевая переменная (1 — дефолт, 0 — нет дефолта).

Доля дефолтов составляет 15,1%, что соответствует реальной банковской практике.

В ходе работы было установлено, что не для всех признаков удаётся получить 10 бинов. Это не является недостатком метода, а скорее демонстрирует его адаптивность — количество бинов определяется структурой данных, а не задаётся жёстко.

На основе обучающей выборки были рассчитаны значения IV для каждого признака. Все три признака были оставлены для обучения модели.

Обучение модели в разработанной системе включает следующие последовательные этапы:

— загрузка и предобработка обучающей выборки;
— WoE-биннинг каждого признака с расчётом IV;
— преобразование всех данных в WoE-значения;
— обучение логистической регрессии на WoE-признаках;
— оценка качества модели;
— сохранение модели и WoE-объектов для использования в телеграм-боте и веб-интерфейсе.

Обучение логистической регрессии — это процесс подбора таких коэффициентов $\beta_0, \beta_1, \beta_2, \beta_3$, которые минимизируют ошибку модели. Иными словами, модель «учится» на исторических данных, подбирая наилучшие веса для каждого признака.

Коэффициенты β — это веса, которые определяют силу влияния каждого признака на итоговую вероятность дефолта.

Процесс подбора коэффициентов можно представить в следующем виде:

— берем начальные β (случайные значения);
— модель делает предсказания для всех клиентов;
— производим сравнение предсказания с реальными ответами;
— корректируем значения β (если ошибка Log Loss большая);
— повторяем шаги несколько тысяч раз;
— сохраняем финальные значения β.

После процесса обучения (обычно после нескольких тысяч или миллионов итераций) получаются оптимальные коэффициенты β, которые лучше всего разделяют дефолтных и успешных заёмщиков.

Обучение логистической регрессии основано на автоматическом процессе подбора коэффициентов β, который происходит внутри метода fit(). На вход подаются WoE-значения признаков и реальные ответы, на выходе — оптимальные веса, которые минимизируют ошибку модели. Эти коэффициенты сохраняются в файл credit_model.pkl и затем используются ботом и веб-приложением для предсказания вероятности дефолта новых клиентов.

Далее, на основе обученной модели создан Telegram-бот и веб-интерфейс.

Telegram-бот, представляющий собой пользовательский интерфейс для системы кредитного скоринга. Основные функции бота:

— сбор данных о заёмщике (возраст, доход, долговая нагрузка);
— преобразование введённых данных в WoE-значения;
— расчёт вероятности дефолта с использованием обученной модели;
— отображение кредитного балла и решения (одобрено / на рассмотрении / отклонено);
— формирование и отправка PDF-отчёта с детализацией и графиками.

Архитектура бота построена по принципу клиент-серверного взаимодействия с использованием Telegram Bot API в качестве транспортного уровня

Для реализации бота использовалась библиотека python-telegram-bot версии 20.7. Выбор обусловлен следующими факторами:

— асинхронность (поддержка async/await для высоконагруженных сценариев);
— стабильность (регулярные обновления и большая пользовательская база);
— функциональность (встроенная поддержка клавиатур, callback-запросов, отправки документов);
— документация (подробная официальная документация с примерами).

Для сериализации модели и WoE-объектов используется библиотека joblib, которая обеспечивает:

— быструю загрузку и выгрузку больших объектов;
— совместимость с моделями scikit-learn;
— сохранение структуры словарей и пользовательских объектов.

Для удобства пользователя в боте реализованы два типа клавиатур:

— reply-клавиатура - главное меню (появляется вместо стандартной клавиатуры телефона. Используется для выбора основных действий);
— inline-клавиатура - выбор долговой нагрузки (прикрепляется к конкретному сообщению, используется для выбора долговой нагрузки, чтобы пользователь не вводил числа вручную).

Бот состоит из следующих функциональных модулей:

— telegram_bot.py (включает в себя весь функционал: загрузку модели и бинов из файлов .pkl, маршрутизацию команд и кнопок, работу с данными от пользователя);
— core/exporters.py (необходим для создания PDF-отчёта);
— core/binning.py (передает функцию apply_woe()).

Для формирования отчёта реализован модуль exporters.py, который генерирует PDF-документ, содержащий:

— данные клиента;
— результат скоринга (балл, вероятность, решение);
— WoE-детализацию по каждому признаку;
— график вероятности (спидометр);
— график WoE-значений.

В коде бота вызываются функции из модуля exporters.py, которая создает отчет.

Для удобства кредитного эксперта все отчеты сохраняются в отдельную папку проекта «exports».

Наряду с Telegram-ботом, разработано веб-приложение для системы кредитного скоринга с использованием фреймворка Streamlit. Веб-приложение предоставляет следующие возможности:

— оценка кредитного риска для одного клиента;
— анализ информационной силы признаков (IV);
— визуализация WoE-бэннинга;
— детальная информация о модели.

Streamlit — это Python-фреймворк, который упрощает создание интерактивных веб-приложений. Разработчики могут создавать приложения, работающие на основе ИИ, не сталкиваясь с тонкостями фронтенд-разработки.

Он предназначен для бесшовной работы с моделями ИИ и машинного обучения. С помощью всего нескольких строк на Python разработчики могут создать интерфейс, в котором пользователи смогут загружать изображения, обрабатывать видео и взаимодействовать с моделями ИИ.

Одна из его ключевых особенностей - динамический рендеринг. Когда пользователи вносят изменения, приложение обновляется автоматически, не требуя ручной перезагрузки страницы.

Кроме того, благодаря своему легкому весу и простоте использования Streamlit эффективно работает как на локальных машинах, так и на облачных платформах. Это делает его отличным выбором для развертывания приложений ИИ, обмена моделями с другими людьми и обеспечения интуитивного, интерактивного пользовательского опыта.

Основные возможности Streamlit:

— быстрое развертывание: ML-модель или обычную программу можно быстро превратить в одностраничное веб-приложение и управлять им. Не нужно долго верстать и загружать модель, пользоваться традиционными инструментами для создания веб-интерфейсов. Несколько десятков строк кода — и веб-приложение готово, отрисовано и работает;
— использование скриптов: приложения можно делать интерактивными. Каждый раз, когда пользователь взаимодействует с получившимся веб-интерфейсом или разработчик меняет что-то в коде, Streamlit сам об-

новляет и перерисовывает нужные части страницы. Так интерфейс «откликается» на действия пользователя или на изменения модели в реальном времени. Поэтому с помощью фреймворка можно делать интерактивные визуализации, дашборды или простые пользовательские сервисы;

— виджеты и визуализация: в Streamlit есть встроенные стандартные виджеты для частых действий, например ползунки или поля для ввода текста. Можно взять готовые виджеты и собрать из них работающий интерфейс. Еще можно отрисовать график или картинку, вывести результат работы программы в виде схемы или таблицы. Есть и функция для отрисовки карты, на которой можно указать с помощью кода какие-то координаты, маршруты и линии.

Приложение реализовано в виде одностраничного сайта с боковой панелью для навигации по пяти разделам:

— кредитный скоринг — интерактивная форма для оценки нового клиента;
— анализ WoE/IV — загрузка CSV и расчёт IV;
— визуализация биннинга — построение WoE-графиков;
— теория WoE — справочная информация;
— информация о модели — коэффициенты и границы бинов.

Для оптимизации производительности используется кэширование, предотвращающее повторную загрузку модели при каждом действии пользователя.

Веб-интерфейс использует обученную модель данных как и Telegram-бот, но с возможностями используемого фреймворка удалось увеличить наглядность кредитного скоринга, что делает приложение наиболее удобным и понятным.

В **заключении** приведены результаты бакалаврской работы.

Основные результаты

1. Проанализированы существующие методы кредитного скоринга;
2. Реализован алгоритм WoE-биннинга и расчёт информационной силы признаков;
3. Обучена модель на WoE-преобразованных данных;

4. Разработан Telegram-бот для оценки кредитного риска;
5. Создано пользовательское веб-приложение;
6. Реализован экспорт отчётов в формате PDF.