МИНОБРНАУКИ РОССИИ

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

Кафедра теории функций и стохастического анализа

МЕТРИКИ РАСХОЖДЕНИЯ В КРЕДИТНОМ СКОРИНГЕ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

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

механико-математического факультета
Дедюриной Ксении Алексеевны

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

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

Саратов 2026

Конечно, вот измененный автореферат с увеличенным объемом текста, который более подробно отражает содержание обновленной дипломной работы. Основные изменения коснулись более детального описания теоретических разделов и практической реализации.

ВВЕДЕНИЕ

В условиях цифровизации финансового сектора возрастает роль автоматизированных систем оценки заемщиков. Увеличение объема клиентских данных, усложнение структуры кредитных продуктов и необходимость оперативного принятия решений обусловливают применение методов математической статистики, теории вероятностей и машинного обучения в задачах кредитного анализа.

Одной из центральных задач кредитного скоринга является построение количественной оценки риска клиента до момента предоставления кредитного продукта. При решении данной задачи требуется учитывать большое число характеристик клиента, извлекаемых из исторических табличных данных. Существенное значение имеет не только построение классификационной модели, но и анализ различий между распределениями риска, между группами клиентов и между сегментами клиентского портфеля, возникающими после применения кредитной политики. В этой связи особую роль приобретают метрики расхождения, которые выступают мостом между качеством прогноза и его бизнес-интерпретацией.

Метрики расхождения позволяют анализировать различия между распределениями score, между клиентскими сегментами и между группами итоговых кредитных решений. Они используются как при оценке качества модели, так и при прикладной интерпретации результатов кредитного скоринга. Исследование подобных характеристик требует опоры как на аппарат теории вероятностей и математической статистики, так и на методы машинного обучения, ориентированные на обработку табличных данных.

Объектом исследования являются процессы оценки кредитного риска клиентов.

Предметом исследования являются метрики расхождения, применяемые в кредитном скоринге, а также методы машинного обучения, используемые для построения вероятностной модели риска.

Целью работы является исследование метрик расхождения в кредитном скоринге и построение модели оценки кредитного риска клиентов на основе методов машинного обучения.

Для достижения поставленной цели решаются следующие задачи:

1. рассмотреть теоретические основы вероятностного и статистического анализа данных, необходимые для исследования скоринговых моделей;
2. исследовать структуру выборок, распределений и выборочных характеристик, применяемых при обработке данных;
3. рассмотреть основные идеи регрессионного анализа как части статистического аппарата анализа зависимостей;
4. исследовать понятие кредитного скоринга и особенности построения признакового пространства;
5. изучить основные метрики качества бинарной классификации и метрики расхождения;
6. исследовать применение методов машинного обучения в задачах кредитного скоринга;
7. разработать процедуру агрегации исходных данных по идентификатору клиента;
8. построить модель бинарной классификации на основе CatBoostClassifier;
9. разработать схему формирования кредитных решений на основе прогнозируемого риска;
10. выполнить анализ различий между сегментами клиентского портфеля.

1 Основное содержание работы

В первом разделе выпускной квалификационной работы рассматривается математический аппарат, необходимый для последующего анализа скоринговых моделей. Поскольку практическая часть работы связана с обработкой большого массива наблюдений по клиентам, исследование опирается на базовые понятия теории вероятностей и математической статистики. Вводятся ключевые конструкции, позволяющие формализовать случайный эксперимент: пространство элементарных событий Ω, алгебра событий $\mathfrak{A}$ и вероятность $\mathbf{P}$. Вероятностное пространство определяется как тройка $(\Omega, \mathfrak{A}, \mathrm{P})$.

Рассматриваются понятия генеральной совокупности и случайной выборки $\vec{X}_n = (X_1, \ldots, X_n)$. Подчеркивается, что элементы выборки являются независимыми и одинаково распределенными случайными величинами, что отражается в формуле для функции распределения выборки:

$$F_{\vec{X}}(t_1, \ldots, t_n) = \prod_{i=1}^{n} F(t_i). \tag{1}$$

Детально изучаются основные распределения, используемые в статистическом анализе: нормальное распределение, χ^2-распределение, распределение Стьюдента и распределение Фишера. Для нормального распределения приводится плотность вероятности, зависящая от математического ожидания и дисперсии:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}. \tag{2}$$

Также вводятся выборочные числовые моменты. Выборочное среднее и выборочная дисперсия определяются как:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\sigma}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{3}$$

Показано, что выборочные моменты являются состоятельными оценками теоретических моментов, сходясь к ним по вероятности при $n \to \infty$.

В завершение раздела рассматриваются основы парной линейной регрессии. Уравнение регрессии имеет вид $Y' = a + bX$. Коэффициенты нахо-

дятся методом наименьших квадратов (МНК), который минимизирует сумму квадратов отклонений:

$$\min_{a,b} \sum_{i=1}^{n} (Y_i - Y_i')^2. \tag{4}$$

Отмечается, что аппарат линейной регрессии задает формальную основу для анализа связей в данных, однако в условиях множества тесно связанных признаков и нелинейных зависимостей его применение ограничено, что требует перехода к более современным алгоритмам машинного обучения.

Во втором разделе работа переходит от общего статистического аппарата к теоретическим основам кредитного скоринга. Под кредитным скорингом понимается система количественной оценки заемщика, предназначенная для прогнозирования вероятности неблагоприятного кредитного события. В математической постановке скоринг описывается как отображение из пространства признаков в вероятность:

$$f : \mathbb{R}^d \to [0, 1], \quad f(x) \approx P(y = 1 \mid x). \tag{5}$$

Акцентируется внимание на том, что в реальных задачах данные по одному клиенту часто распределены по нескольким таблицам и содержат множество записей. Поэтому для построения признакового пространства используются агрегированные характеристики: сумма, минимум, максимум и среднее:

$$x^{sum} = \sum_{j=1}^{k} x_j, \quad x^{mean} = \frac{x^{sum}}{k}. \tag{6}$$

Такой подход позволяет перейти от событийного описания клиента к фиксированному набору числовых признаков.

Проведен обзор методов машинного обучения, включая логистическую регрессию, деревья решений и нейронные сети. Подчеркивается, что для табличных данных с большим числом агрегированных признаков особенно эффективными оказываются методы градиентного бустинга. Ансамблевая мо-

дель записывается как:

$$F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x), \tag{7}$$

с логистическим преобразованием для получения вероятностной интерпретации:

$$p(x) = \frac{1}{1 + e^{-F_M(x)}}. \tag{8}$$

Исследованы метрики качества бинарной классификации и метрики расхождения. Помимо стандартных Accuracy, TPR и FPR, особое внимание уделено AUC как мере ранжирующей способности модели. Под метриками расхождения понимаются числовые характеристики, отражающие различие между распределениями score, надежными и проблемными клиентами, группами кредитных решений и сегментами клиентского портфеля.

В третьем разделе описывается разработка скоринговой модели, логика отбора переменных, формирование кредитной политики и анализ полученных результатов. В прикладной части исследования решалась задача построения модели кредитного скоринга на основе данных соревнования «Кредитный скоринг на основе кредитных историй».

Описана процедура агрегации исходных parquet-файлов по идентификатору клиента. Для каждого числового признака вычислялись сумма, минимум, максимум и среднее значение, что привело к формированию расширенного признакового пространства. Обсуждена проблема взаимосвязи признаков. В отличие от классической линейной регрессии, где мультиколлинеарность является серьезным ограничением, было показано, что использование алгоритма CatBoostClassifier позволяет работать с широким и частично зависимым признаковым пространством, полагаясь на внутренние механизмы регуляризации и отбора признаков.

Обучение модели CatBoostClassifier выполнялось с параметрами: 500 итераций, глубина деревьев 5, шаг обучения 0.05, функция потерь Logloss и метрика контроля качества AUC. На основе полученной вероятностной оценки score была разработана система прикладных кредитных решений, включающая полное одобрение, одобрение с уменьшенным лимитом, ручную про-

верку и отказ. Правило принятия решения задавалось процентильными порогами:

$$decision(score) = \begin{cases} \text{approve\_full}, & score \le t_{approve}, \\ \text{approve\_reduced}, & t_{approve} < score \le t_{reduce}, \\ \text{manual\_review}, & t_{reduce} < score \le t_{manual}, \\ \text{decline}, & score > t_{manual}. \end{cases} \tag{9}$$

Дополнительно были введены рекомендуемый кредитный лимит и процентная надбавка.

Проведенный графический и табличный анализ результатов подтвердил, что сформированные клиентские сегменты существенно различаются не только по типу решения, но и по количественным характеристикам: распределению score, средней условной платежеспособности и рекомендуемому лимиту. Было показано, что по мере увеличения риска клиенты последовательно переходят из одной зоны решений в другую, а наиболее надежные и наиболее рискованные клиенты образуют качественно различные группы.

ЗАКЛЮЧЕНИЕ

В выпускной квалификационной работе исследованы метрики расхождения в кредитном скоринге и рассмотрены подходы к построению модели оценки кредитного риска клиентов на основе методов машинного обучения.

В теоретической части работы были рассмотрены вероятностные и статистические основы анализа данных. Было показано, что аппарат теории вероятностей и математической статистики является необходимой базой для описания структуры наблюдаемых данных, построения признакового пространства и интерпретации количественных характеристик клиентского портфеля. Также были исследованы теоретические основы кредитного скоринга как задачи вероятностной бинарной классификации, где особое внимание уделено метрикам расхождения, позволяющим анализировать различия между распределениями риска и клиентскими сегментами.

В практической части была разработана процедура агрегации исходных данных по идентификатору клиента. Было показано, что использование расширенного набора агрегированных переменных является оправданным в рамках бустинговой модели. Выбор CatBoostClassifier был обоснован его способностью работать с табличными данными, большим числом признаков и сложными нелинейными зависимостями, что делает его более адекватным инструментом по сравнению с классической регрессией для данной задачи.

Была реализована процедура перехода от вероятностного прогноза риска к прикладной системе кредитных решений. Проведенный анализ показал, что клиентские сегменты существенно различаются по распределению score, условной платежеспособности и группам кредитных решений. Было установлено, что вероятностный прогноз риска может быть непосредственно встроен в контур принятия кредитных решений, обеспечивая эффективную сегментацию клиентского портфеля.

Таким образом, по результатам работы можно сделать следующие выводы:

1. вероятностный и статистический аппарат является необходимой теоретической базой для исследования скоринговых моделей;
2. агрегирование событийных данных по клиенту позволяет сформиро-

вать информативное признаковое описание заемщика;

3. использование расширенного набора переменных является оправданным в рамках бустинговой модели;
4. CatBoostClassifier является пригодным инструментом для построения модели кредитного скоринга на табличных данных подобного типа;
5. метрики расхождения проявляются в различиях между распределениями score, клиентскими сегментами и группами кредитных решений;
6. вероятностный прогноз риска может быть непосредственно встроен в контур принятия кредитных решений.