

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»**

Кафедра дифференциальных уравнений и математической экономики

Оценка кредитного риска с помощью алгоритмов

машинного обучения

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы

направление 09.03.03 — Прикладная информатика

механико-математического факультета

Журлова Александра Дмитриевича

Научный руководитель

профессор, д.э.н., профессор

В. А. Балаш

Зав. кафедрой

зав. каф., д.ф.-м.н., профессор

С. И. Дудов

Саратов 2025

Введение. Современный финансовый сектор стремительно развивается в условиях цифровизации, глобализации и высокой экономической нестабильности. При этом одной из ключевых задач для банков и других финансово-кредитных организаций остаётся эффективное управление рисками, в частности — кредитными. Кредитный риск, как вероятность невозврата заемщиком полученных денежных средств, оказывает существенное влияние на финансовую устойчивость и прибыльность учреждений, предоставляющих займы.

Традиционные методы оценки кредитного риска, основанные на статистических моделях и экспертных системах, часто не справляются с динамикой современных рынков, не учитывают множество взаимосвязанных факторов и могут давать неточные прогнозы. В связи с этим возрастает интерес к использованию более гибких и адаптивных инструментов, в том числе методов машинного обучения (ML), которые позволяют автоматически выявлять скрытые зависимости в данных, строить более точные прогнозы и учитывать изменчивость экономической среды.

Кроме того, для обучения моделей необходимы большие объемы разнообразных и качественных данных. Однако в условиях конфиденциальности и ограниченного доступа к финансовым и клиентским данным всё большую актуальность приобретает синтез искусственных данных, способных моделировать реальные ситуации без нарушения приватности. Такие данные позволяют не только восполнить дефицит информации, но и протестировать модели в различных гипотетических экономических сценариях.

В целом, применение машинного обучения в оценке кредитных рисков требует комплексного подхода, включающего как теоретическое обоснование используемых методов, так и практическую реализацию моделей с учётом особенностей доступных данных. Немаловажным аспектом является и анализ устойчивости построенных моделей при изменении внешних макроэкономических условий.

Цель работы — разработка и оценка моделей машинного обучения для прогнозирования кредитного риска на основе синтетических и реальных данных, с последующим анализом их устойчивости в различных экономических сценариях.

Для достижения цели необходимо решить следующие задачи:

- изучить теоретические основы кредитного риска, его виды и способы оценки;
- рассмотреть ключевые экономические и поведенческие факторы, влияющие на вероятность дефолта;
- проанализировать современные методы машинного обучения, применяемые в кредитном скоринге;
- исследовать подходы к синтезу искусственных данных и их применимость для обучения моделей;
- реализовать и протестировать ряд алгоритмов (логистическая регрессия, KNN, градиентный бустинг) на сгенерированных и реальных выборках;
- провести сравнительный анализ качества моделей и устойчивости результатов в условиях изменения экономических параметров.

Объект исследования — процессы оценки кредитного риска в банковской сфере.

Предмет исследования — применение методов машинного обучения и синтетических данных для прогнозирования вероятности дефолта заемщика.

Первый раздел работы посвящён рассмотрению теоретических основ оценки кредитных рисков. В нём раскрывается сущность понятия кредитного риска, его роль в деятельности финансовых организаций, а также анализируются существующие подходы к его оценке.

Кредитный риск возникает в случае, если заёмщик не выполняет свои обязательства по договору, что может привести к финансовым потерям для кредитора. Управление данным видом риска является одной из приоритетных задач в банковской системе, поскольку уровень кредитного риска напрямую влияет на устойчивость и надёжность финансовой организации.

В рамках данного раздела рассматриваются основные методы оценки кредитного риска, включая количественные и качественные подходы. Отдельное внимание уделяется показателю вероятности дефолта (PD), как одному из ключевых элементов в системах кредитного скоринга. Приводятся общие принципы его расчёта и значимость при прогнозировании рисков.

Кроме того, рассматривается влияние макроэкономических факторов на кредитный риск. Подчёркивается, что такие внешние условия, как инфля-

ция, уровень занятости и динамика процентных ставок, могут существенно изменить поведение заёмщиков и увеличить риск невозврата.

Завершающая часть раздела посвящена анализу роли машинного обучения в оценке кредитных рисков. Современные алгоритмы позволяют обрабатывать большие объёмы разнородных данных, выявлять сложные взаимосвязи и повышать точность прогнозов. Это делает машинное обучение перспективным инструментом в области риск-менеджмента и кредитного скоринга.

Подводя итог, первый раздел формирует теоретическую базу, необходимую для дальнейшего анализа и практической реализации моделей оценки кредитных рисков на основе данных.

Второй раздел работы посвящён описанию процесса формирования обучающего набора данных, используемого для построения моделей оценки кредитного риска. В условиях ограниченного доступа к реальным клиентским и региональным данным было принято решение о генерации синтетических данных на основе математического моделирования. Такой подход позволяет создать контролируемую среду для тестирования алгоритмов и варьировать ключевые параметры, влияющие на вероятность дефолта.

Основой для синтеза данных стала логистическая модель, с помощью которой рассчитывается вероятность дефолта клиента:

$$P(\text{default}) = \frac{1}{1 + e^{-z}},$$

где z представляет собой линейную комбинацию переменных, отражающих индивидуальные характеристики клиента и макроэкономические условия ре-

гиона его проживания:

$$\begin{aligned} z = & -1.2 + 0.000001 \cdot (\text{regional_gdp} - 1\,000\,000) \\ & - 1.4 \cdot \log\left(\frac{\text{income}}{40\,000} + 1\right) + 0.9 \cdot \left(\frac{\text{regional_unemployment}}{10}\right) \\ & - 0.5 \cdot \left(\frac{\text{credit_history_length}}{20}\right) + 0.4 \cdot \text{existing_loans} \\ & + 0.25 \cdot \text{dependents} - 0.1 \cdot \left(\frac{\text{age} - 40}{40}\right) \\ & + 1.0 \cdot \log\left(1 + \frac{\text{loan_amount}}{\text{income} + 1}\right) + 0.3 \cdot \left(\frac{12}{\text{loan_term}}\right) \end{aligned}$$

Каждый коэффициент в данной формуле отражает вклад соответствующего признака в уровень кредитного риска. Например, увеличение дохода снижает вероятность дефолта, а высокая региональная безработица, напротив, повышает её. Использование логарифмической и нормализующей трансформации обеспечивает стабильность модели при генерации данных.

Для повышения реалистичности была смоделирована структура данных по 25 регионам России, сгруппированных по уровню экономического развития. Каждый регион характеризовался значениями валового регионального продукта (ВРП), уровня безработицы, численности населения и средней заработной платы. Эти макроэкономические параметры использовались совместно с индивидуальными переменными.

Профиль каждого клиента формировался с учётом принадлежности к региону. В таблице ниже приведены все переменные, включённые в итоговый датасет, с разделением на ключевые (значимые для модели) и дополнительные (шумовые):

Таблица 1 — Описание переменных в синтетическом датасете

Ключевые переменные	Описание
age	Возраст клиента
income	Уровень дохода клиента
credit_history_length	Продолжительность кредитной истории
existing_loans	Количество активных кредитов
dependents	Количество иждивенцев
loan_amount	Сумма запрашиваемого кредита
loan_term	Срок кредитного договора
regional_gdp	Валовой региональный продукт
regional_unemployment	Уровень безработицы в регионе
gdp_per_capita	ВРП на душу населения
unemployment	Общий уровень безработицы
avg_salary	Средняя заработная плата в регионе
population	Численность населения региона
default_prob	Расчётная вероятность дефолта
default	Факт дефолта (бинарная переменная)
credit_score	Комплексная оценка платёжеспособности клиента
type	Тип клиента (например, физическое или юридическое лицо)
Дополнительные (шумовые) переменные	Описание
full_name	ФИО клиента (синтетическое)
passport	Паспортные данные (фиктивные)
education	Уровень образования
favorite_color	Любимый цвет клиента
car_number	Номер автомобиля

Дополнительные признаки не использовались при обучении моделей, однако они были добавлены в датасет с целью имитации структуры реальных банковских массивов данных. Их наличие повышает визуальное правдоподобие выборки и делает её более близкой к тем условиям, в которых работают современные алгоритмы машинного обучения в промышленной практике.

Полученная итоговая выборка представляет собой сбалансированный и логически выстроенный набор данных, сочетающий как индивидуальные характеристики клиентов, так и региональные макроэкономические параметры. Это позволило моделировать как персональные кредитные риски, так и влияние внешних факторов.

В **третьем разделе работы** рассматривается этап загрузки сгенерированных данных в реляционную базу данных PostgreSQL, а также процесс последующей предобработки, необходимой для подготовки данных к машинно-

му обучению. Использование полноценной СУБД позволяет структурировать данные, обеспечить их сохранность, реализовать связи между сущностями и ускорить выполнение аналитических запросов.

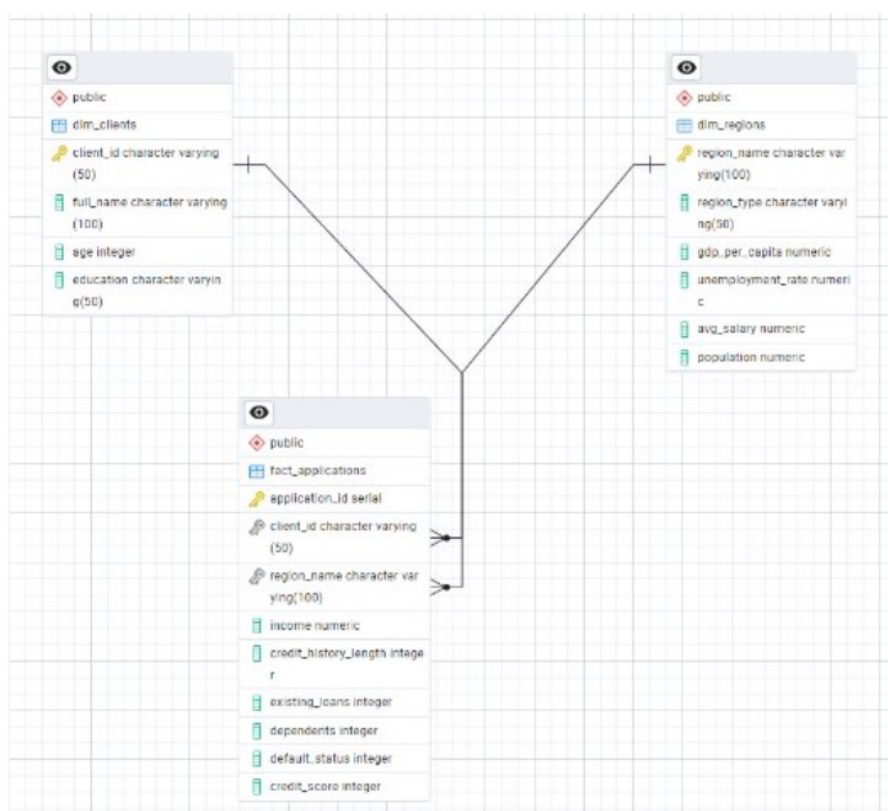


Рисунок 1 — Схема базы данных для хранения синтетических данных

таблиц заполняется соответствующим подмножеством данных: `dim_regions` — уникальными регионами, `dim_clients` — уникальными записями клиентов, а `fact_applications` — основной фактологической информацией, содержащей внешние ключи на регионы и клиентов.

После загрузки создаётся материализованная витрина данных `credit_mart`, которая объединяет ключевые признаки из всех таблиц. Это упрощает дальнейший доступ к данным при аналитических запросах и ускоряет работу моделей. Витрина включает персональные, региональные и финансовые параметры в едином представлении.

Следующий этап — предобработка данных, выполняемая с помощью библиотек `sklearn.pipeline` и `sklearn.compose`. Эти инструменты позволяют выстроить последовательность преобразований и реализовать масштабируемый и воспроизводимый пайплайн подготовки данных.

Сначала данные извлекаются из материализованной витрины `credit_mart` с использованием библиотеки `SQLAlchemy`, которая обеспечивает надёжное соединение с базой данных и выполнение запросов. Загруженные данные преобразуются в формат `DataFrame` библиотеки `Pandas`.

Числовые признаки проходят обработку, включающую заполнение пропущенных значений медианой, логарифмическую трансформацию для сглаживания распределений и стандартизацию признаков до единого масштаба. Категориальные признаки очищаются от пропусков путём замены на наиболее частое значение и преобразуются методом `OneHot-Encoding` в числовое представление.

Для борьбы с выбросами применяется метод винзоризации — ограничение значений в допустимом диапазоне, что минимизирует искажения при обучении модели. Дополнительно создаются новые признаки, отражающие комплексные зависимости: отношение дохода к ВРП (`income_to_gdp`), нагрузка на клиента с учётом иждивенцев (`loan_per_dependent`) и агрегированный риск-индекс (`risk_score`), объединяющий кредитный рейтинг и уровень безработицы.

Объединение числовой и категориальной обработки реализуется с помощью объекта `ColumnTransformer`, позволяющего параллельно обрабатывать разные типы данных. После завершения пайплайна целевая переменная и

признаки сохраняются в локальную директорию `complite_data` для последующего использования на этапе моделирования.

Четвёртый раздел посвящён обучению моделей машинного обучения на предобработанных данных и анализу полученных результатов. Целью данного этапа является сравнительная оценка эффективности различных алгоритмов классификации при решении задачи предсказания вероятности дефолта по кредиту.

В рамках эксперимента были обучены и протестированы четыре модели: логистическая регрессия, метод опорных векторов (Linear SVC), алгоритм ближайших соседей (K-Nearest Neighbors) и градиентный бустинг. Для обучения использовалась библиотека `scikit-learn`. Особое внимание при оценке качества моделей уделялось миноритарному классу (класс 1), который соответствует клиентам, склонным к дефолту. Использовались следующие метрики: *Precision*, *Recall* и *F1-score*.

Результаты моделей представлены на рисунке 2:

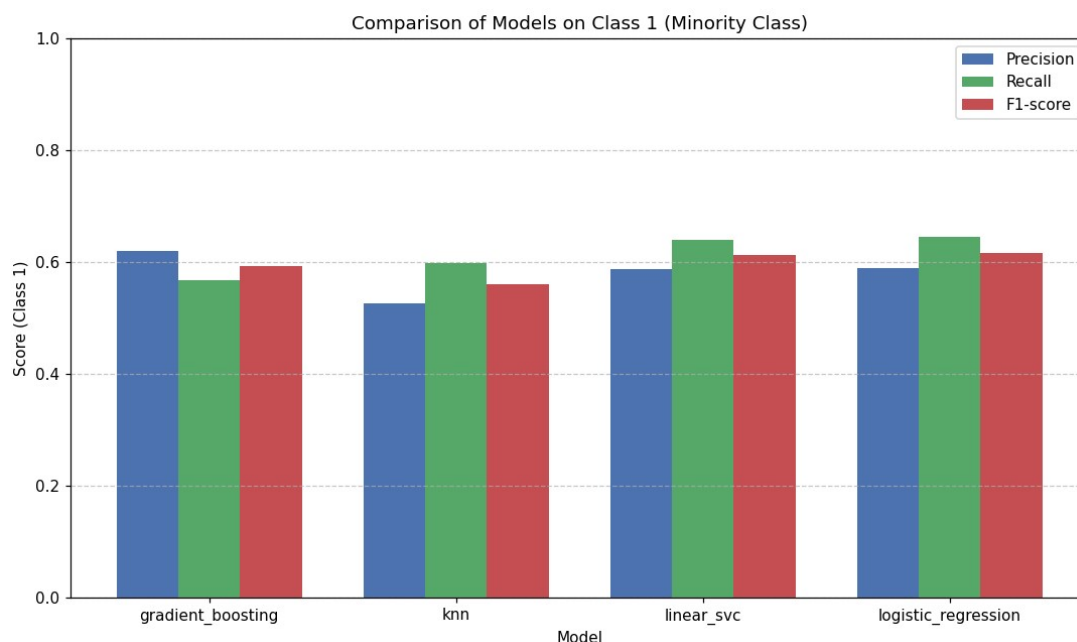


Рисунок 2 — Сравнение моделей по метрикам Precision, Recall и F1-score для класса 1

Модель градиентного бустинга показала наивысшее значение точности (Precision = 0,62), что свидетельствует о её способности минимизировать количество ложных положительных классификаций. Полнота модели состави-

ла 0,57, что несколько ниже, чем у остальных моделей, но при этом достигается хороший компромисс между метриками. Итоговый F1-score составил 0,59, что делает модель сбалансированной и надёжной. Бустинг рекомендуется для сценариев, где особенно важно избегать ложных срабатываний и обеспечивать устойчивость к шуму в данных.

Метод ближайших соседей (KNN) достиг максимальной полноты (Recall = 0,60), при этом показав самую низкую точность (Precision = 0,53). Это означает, что модель склонна к избыточным срабатываниям, но способна выявлять большее число клиентов с потенциальным дефолтом. Итоговый F1-score составил 0,56.

Linear SVC показал сбалансированные метрики: Precision = 0,59, Recall = 0,64 и наивысший F1-score = 0,61. Это делает модель особенно подходящей для задач, в которых важно как выявить как можно больше дефолтных клиентов, так и не создавать чрезмерное количество ложных срабатываний.

Логистическая регрессия продемонстрировала аналогичное поведение: точность — 0,59, полнота — 0,65, F1-score — 0,61. Таким образом, этот алгоритм также обеспечивает устойчивый и интерпретируемый результат при хорошем балансе между чувствительностью и точностью.

Вывод. Выбор модели зависит от контекста применения:

- если ключевым приоритетом является высокая точность при минимуме ложных тревог — предпочтительна модель градиентного бустинга;
- если критически важно выявить максимум дефолтных клиентов — подойдут Linear SVC и логистическая регрессия.

В рамках данной работы для интеграции в пользовательский интерфейс была выбрана модель градиентного бустинга. Это обусловлено её высокой точностью, способностью работать с разнородными признаками и устойчивостью в условиях ограниченного количества информации в онлайн-режиме.

В **пятом разделе** приведён этап реализации пользовательского интерфейса, обеспечивающего удобство пользователя при взаимодействии с моделью оценки кредитного риска. Интерфейс разработан с использованием фреймворка Streamlit, что позволило быстро создать наглядное и интуитивно понятное веб-приложение.

Основной целью интерфейса является предоставление возможности пользователю ввести параметры нового клиента или выбрать существующего из базы данных, после чего получить прогноз модели — вероятность дефолта и итоговую рекомендацию по выдаче кредита.

Интерфейс включает в себя:

- загрузку информации о клиентах, регионах и историях заявок из базы данных PostgreSQL;
- автоматическое подставление региональных экономических показателей;
- предобработку данных с использованием сохранённого пайплайна, идентичного использовавшемуся при обучении модели;
- визуализацию промежуточных данных и отображение вероятности дефолта;
- хранение и отображение истории предыдущих кредитных заявок клиента.

Прогнозирование осуществляется с помощью модели градиентного бустинга, предварительно обученной на сбалансированном датасете. В зависимости от рассчитанной вероятности дефолта интерфейс выдает соответствующее текстовое сообщение — о возможной выдаче кредита или о повышенном риске.

На рисунке 3 представлен внешний вид интерфейса в момент ввода параметров клиента и отображения результата прогноза.

Интерфейс скоринговой модели

Выбрать клиента из базы или оставить поле пустым для нового

Ковалева Алла Афанасьевна (301081112)

Данные клиента: Ковалева Алла Афанасьевна

- Возраст: 43
- Образование: высшее
- Доход: 63,299.0 руб.
- Регион: Нижегородская область

История заявок

	application_id	income	credit_history_length	existing_loans	dependents	credit_score	def_pred
11	12	63299	1	2	0	545	0

Регион клиента

Москва

Сумма кредита (руб.)

1000000

Срок кредита (мес.)

6

Оценить заявку

Промежуточные данные:

	full_name	region_name	income	credit_history_length	existing_loans	dependents
0	Ковалева Алла Афанасьевна	Москва	63299	1	2	

Размер X: (1, 19)

Вероятность дефолта: 0.601

Рисунок 3 — Пример работы пользовательского интерфейса

Разработанный интерфейс позволяет применять скоринговую модель в реальном времени, обеспечивая доступность результатов как для специалистов, так и для сотрудников, не обладающих техническими навыками.

Заключение. В ходе выполнения работы была реализована полнофункциональная система для оценки вероятности дефолта заемщика, включающая генерацию данных, предобработку, обучение моделей и разработку пользовательского интерфейса. Такой комплексный подход позволил не только достичь поставленной цели, но и на практике реализовать все намеченные задачи. Была реализована архитектура, охватывающая ключевые этапы жизненного цикла скоринговой модели: от генерации синтетических данных до построения модели и создания интерфейса для пользователей.

Анализ существующих подходов к оценке кредитного риска позволил выделить важнейшие признаки, влияющие на вероятность дефолта. На основе этих признаков была сгенерирована репрезентативная выборка данных, в которой учтены как индивидуальные характеристики клиента, так и макроэкономические параметры региона проживания.

Загруженные данные были структурированы в PostgreSQL с использованием принципов нормализации и витринирования, что обеспечило удобство для аналитики и предобработки. С применением современных инструментов машинного обучения были обучены и протестированы четыре модели.

Особое внимание в работе было уделено практической реализации решения. Разработанный пользовательский интерфейс на базе Streamlit обеспечивает возможность интерактивного ввода данных клиента, их предобработки и оценки кредитного риска в режиме реального времени. Интерфейс поддерживает как работу с новыми пользователями, так и анализ повторных обращений.

Результаты данной работы демонстрируют, что построенная система способна эффективно решать прикладные задачи кредитного скоринга. Предложенное решение может быть адаптировано под различные бизнес-сценарии и масштабировано в условиях реальных финансовых организаций.