

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

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

Кафедра дифференциальных уравнений и

математической экономики

**ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ ПРИ АНАЛИЗЕ
КРЕДИТНОГО СКОРИНГА КЛИЕНТА**

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 248 группы

направления 09.04.03 — Прикладная информатика

механико-математического факультета

Нарватова Вадима Валерьевича

Научный руководитель

профессор, д. э. н., профессор _____

В. А. Балаш

Заведующий кафедрой

к. ф. -м. н., доцент _____

В. С. Рыхлов

Саратов 2026

ВВЕДЕНИЕ

В условиях цифровой трансформации банковского сектора и роста объёмов обрабатываемых данных кредитный скоринг превращается из изолированной статистической модели в комплексный программно-аналитический конвейер. Современные системы оценки кредитного риска должны объединять загрузку и очистку данных, генерацию синтетических выборок (в условиях ограничений доступа к реальным портфелям), обучение нескольких моделей машинного обучения, сценарное стресс-тестирование и визуальный анализ результатов.

Актуальность работы обусловлена необходимостью создания воспроизводимого, безопасного и прозрачного инструментария для кредитного скоринга, который может использоваться в учебных и исследовательских целях без раскрытия конфиденциальной информации заёмщиков.

Цель работы – разработка и исследование программного конвейера кредитного скоринга, обеспечивающего загрузку и предобработку табличных данных, генерацию синтетических выборок, обучение моделей машинного обучения, сценарное стресс-тестирование и интерактивный разведочный анализ.

Задачи исследования:

1. анализ предметной области кредитного скоринга, регуляторных требований Банка России и ограничений на использование реальных банковских данных;
2. формализация задачи бинарной классификации заёмщиков и определение состава признаков;
3. исследование методов генерации синтетических табличных данных (Faker+Noise, Gaussian Copula, CTGAN, TVAE);
4. проектирование модульной архитектуры программного комплекса;
5. реализация приложения на Python с веб-интерфейсом Gradio;
6. экспериментальное сравнение моделей (логистическая регрессия, случайный лес, XGBoost, TabNet);
7. реализация механизма сценарного стресс-тестирования с тремя макроэкономическими сценариями;

8. визуальный и статистический анализ качества синтетических данных и устойчивости моделей.

Объект исследования – процесс построения систем кредитного скоринга на основе табличных финансовых данных.

Предмет исследования – методы машинного обучения, методы генерации синтетических данных и программные средства организации скорингового ML-конвейера.

Научная новизна заключается в создании программного конвейера, объединяющего полный цикл работы с кредитными данными в единой исследовательской среде, включая четыре метода синтеза, четыре модели классификации, сценарное стресс-тестирование и интерактивную визуализацию, что отличает его от существующих разрозненных учебных реализаций.

Практическая значимость: разработанный локальный комплекс может применяться в учебном процессе, при подготовке специалистов по анализу данных, для демонстрации методов кредитного скоринга, исследования устойчивости моделей к макроэкономическим шокам, а также для проведения лабораторных и научно-исследовательских работ.

Методы исследования: статистическая обработка данных, машинное обучение (логистическая регрессия, случайный лес, XGBoost, TabNet), генеративное моделирование (CTGAN, TVAE, Gaussian Copula), разведочный анализ данных (EDA), программная инженерия (Python, Gradio, JSON-конфигурации). Для оценки качества синтеза использованы статистика Колмогорова–Смирнова и расстояние полной вариации (TV-distance).

В первой главе проведён системный анализ места кредитного риска в иерархии банковских рисков, рассмотрены теоретические основы обработки данных, требования к качеству информации и регуляторные ограничения, а также обоснована роль синтетических данных в разработке скоринговых систем.

Кредитный риск трактуется как вероятность неисполнения заёмщиком обязательств по кредиту. Кредитный скоринг представляет собой формализованную оценку вероятности дефолта (Probability of Default, PD) на основе совокупности финансовых, демографических и поведенческих признаков. В работе показано, что в современной банковской практике скоринг не сводит-

ся к выдаче одного числового значения, а представляет собой непрерывный процесс — программно-аналитический конвейер, включающий сбор признаков, проверку полноты информации, преобразование переменных, расчёт PD, принятие решения, установление лимита, формирование цены риска и контроль качества модели.

В исследовательском стенде целевая переменная дефолта формируется по пороговому правилу (формула 1):

$$y_i = \begin{cases} 1, & \text{если } score_i < 600, \\ 0, & \text{если } score_i \geq 600. \end{cases} \quad (1)$$

Таким образом, задаётся воспроизводимая бинарная метка для проверки работы конвейера в управляемых условиях.

Проанализированы типовые проблемы качества табличных данных: пропуски, выбросы, разнородная природа признаков (числовые, категориальные, временные, макроэкономические), а также дисбаланс классов (дефолт — редкое событие). Для их решения в разработанном конвейере применяются замена пропусков медианой (для числовых) и модой (для категориальных), а также фильтрация выбросов методом межквартильного размаха (IQR). Межквартильный размах определяется как разность между третьим и первым квартилями (формула 2):

$$IQR = Q_3 - Q_1, \quad (2)$$

а границы допустимых значений задаются интервалом (формула 3):

$$x \in [Q_1 - 1,5 \cdot IQR; Q_3 + 1,5 \cdot IQR]. \quad (3)$$

Выбор метода IQR обоснован его робастностью и отсутствием требований к нормальности распределения данных.

На рисунке 1 представлена иерархия банковских рисков.

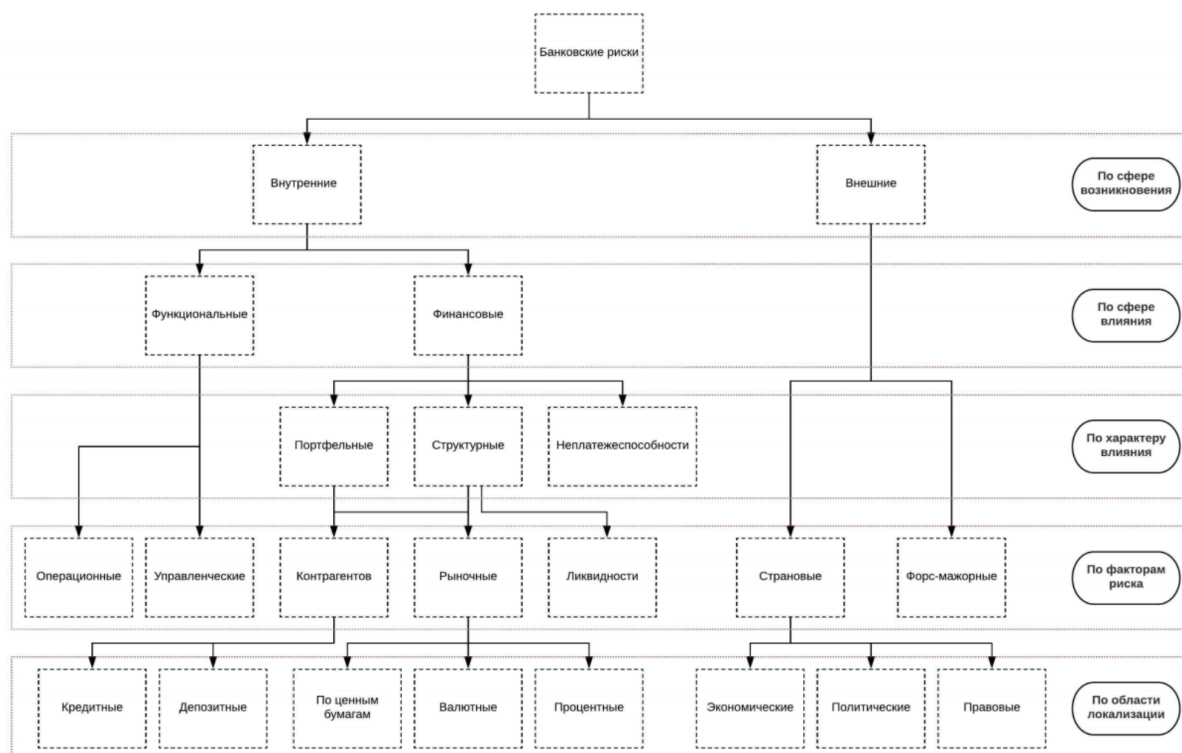


Рисунок 1 – Иерархия банковских рисков

Из приведённой схемы видно, что кредитный риск занимает центральное место в системе банковских рисков. Это объясняется тем, что кредитные операции, по данным Банка России, составляют порядка 80% всех операций коммерческих банков. Таким образом, снижение кредитного риска является одной из приоритетных задач банковской системы.

Особое внимание в первой главе уделено регуляторным требованиям Банка России, которые непосредственно влияют на проектирование скоринговых систем. Рассмотрены ключевые положения нормативных актов: порядок формирования резервов на возможные потери, подход на основе внутренних рейтингов, требования к управлению операционным риском и системе управления рисками и капиталом. Отдельный пласт требований формирует Федеральный закон «О персональных данных», ограничивающий использование реальных клиентских записей в исследовательских целях.

На основе анализа этих ограничений обоснована ключевая роль синтетических табличных данных как инструмента безопасной разработки аналитических систем. Синтетические данные позволяют воспроизводимо тестировать архитектуру конвейера без раскрытия конфиденциальной информации,

моделировать редкие или экстремальные экономические ситуации (рецессию, высокую инфляцию) и проводить эксперименты в условиях ограниченного доступа к реальным банковским портфелям. Таким образом, первая глава закладывает теоретическую и методологическую основу для последующей разработки программного конвейера кредитного скоринга.

Во второй главе задача кредитного скоринга формализована как бинарная классификация. Пусть \mathcal{X} — пространство признаков заёмщика и заявки, $\mathcal{Y} = \{0, 1\}$, где $y = 1$ соответствует дефолту. Требуется построить отображение $f : \mathcal{X} \rightarrow [0, 1]$ и решающее правило $\hat{y}(x) = \mathbb{1}\{f(x) > \tau\}$ с порогом τ , определяемым экономикой кредитного продукта.

В таблице 1 детально описана структура демонстрационного набора данных. Набор объединяет пять групп признаков, каждая из которых играет свою роль в оценке кредитного риска.

Таблица 1 – Структура исходного набора данных

Поле	Тип	Содержательное значение
age	вещественный	возраст заёмщика, лет
income	вещественный	годовой доход, у.е.
loan_amount	вещественный	запрашиваемая сумма кредита, у.е.
credit_score	вещественный	кредитный рейтинг (300–850)
years_employed	вещественный	стаж работы на текущем месте, лет
gender	категориальный	пол заёмщика (Male / Female / Other)
marital_status	категориальный	семейное положение
children_count	целочисленный	количество детей в семье
timestamp	дата-время	момент подачи заявки
gdp_growth	вещественный	темп роста ВВП, %
core_inflation	вещественный	базовая инфляция, %
unemployment	вещественный	уровень безработицы, %
conversion_rate	вещественный	показатель конверсии заявок
performance_score	вещественный	показатель эффективности

Признаки разделены на пять групп: демографические (age, gender, marital_status, children_count), финансовые (income, loan_amount, years_employed), кредитный индикатор (credit_score), макроэкономические (gdp_growth, core_inflation, unemployment) и поведенческие (conversion_rate, performance_score).

Для воспроизводимости эксперимента целевая переменная сформирована синтетически: $Default_i = \mathbb{1}\{credit_score_i < 600\}$.

Для экспериментального сравнения в конвейер включены четыре модели машинного обучения.

Логистическая регрессия используется как интерпретируемый бенчмарк. Модель оценивает вероятность дефолта через логистическую функцию:

$$P(y = 1|x) = \frac{1}{1 + \exp(-(\beta_0 + \beta^T x))}. \quad (4)$$

Преимущество — прозрачность коэффициентов β_j , важная для регуляторного аудита.

В конвейер включены четыре модели: логистическая регрессия, случайный лес, XGBoost и TabNet.

Случайный лес — ансамбль деревьев, устойчивый к выбросам. XGBoost — градиентный бустинг с регуляризацией, обеспечивающий высокую точность. TabNet — сеть с механизмом внимания для табличных данных.

Для генерации синтетических данных применены методы Faker+Noise (ресемплинг с шумом), Gaussian Copula, CTGAN, TVAE. Качество синтеза оценивается статистикой Колмогорова–Смирнова и расстоянием полной вариации.

Стресс-тестирование включает три макроэкономических сценария (рост, рецессия, высокая инфляция), при которых макроэкономические признаки умножаются на заданные коэффициенты, после чего пересчитывается средняя вероятность дефолта по портфелю.

В заключительной части главы описана схема стресс-тестирования. Пусть X — исходная матрица признаков, S — сценарий с коэффициентами $k_j^{(S)} \geq 0$ для макроэкономических признаков. Шокированная матрица:

$$X_{ij}^{(S)} = X_{ij} \cdot k_j^{(S)}. \quad (5)$$

Определены три сценария: экономический рост ($k_{unemployment} = 0,5, k_{gdp_growth} = 1,2, k_{core_inflation} = 1,0$), рецессия (2,0, 0,8, 1,1) и высокая инфляция (1,2, 0,9, 1,5). После применения шоков пересчитывается вероятность дефолта $\hat{p}(y_i = 1|x_i^{(S)})$, а среднее по портфелю —

$$\overline{PD}^{(S)} = \frac{1}{n} \sum_{i=1}^n \hat{p}(y_i = 1|x_i^{(S)}). \quad (6)$$

Таким образом, во второй главе формализована задача кредитного скоринга, описана структура демонстрационных данных, представлены четыре модели (от интерпретируемой логистической регрессии до TabNet), рассмотрены методы генерации синтетических данных и сценарного стресс-тестирования, что составляет научно-методическую основу разработанного программного конвейера.

В третьей главе представлена архитектура разработанного программного комплекса. Система реализована как локальное приложение на языке Python с модульной структурой, объединяющей этапы загрузки данных, предобработки, генерации синтетических выборок, обучения моделей, стресс-тестирования и визуального анализа. Такой подход обеспечивает гибкость, воспроизводимость экспериментов и возможность независимого тестирования каждого компонента.

На рисунке 2 показана логическая схема ML-конвейера, демонстрирующая последовательное прохождение данных от загрузки исходного датасета до визуализации результатов. Из схемы видно, что данные могут поступать как из пользовательского файла, так и из встроенного генератора демонстрационного набора. Синтетические выборки создаются на основе очищенной таблицы, после чего признаки подготавливаются, модели обучаются, выполняются стресс-тесты и формируются отчёты.

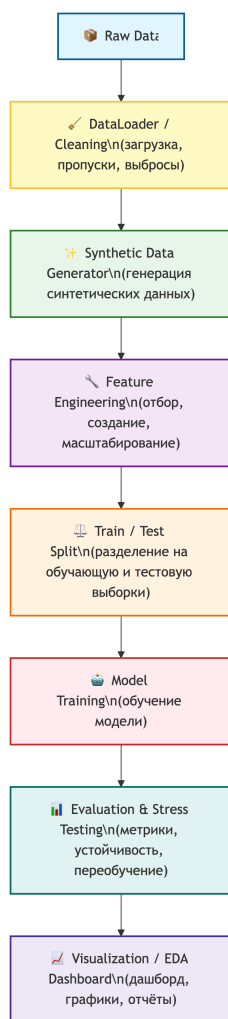


Рисунок 2 – Логическая схема ML-конвейера программного комплекса

Данная архитектура положена в основу пользовательского интерфейса, который реализован с помощью библиотеки Gradio. Интерфейс состоит из пяти последовательных вкладок, каждая из которых соответствует отдельному этапу конвейера. Такой дизайн позволяет аналитику пошагово выполнять эксперимент, контролируя каждый этап и при необходимости возвращаясь к предыдущим шагам.

Перейдём к рассмотрению первой вкладки. На рисунке 3 представлена вкладка загрузки данных. Здесь пользователь может выбрать один из двух режимов: загрузить собственный CSV или XLSX-файл либо сгенерировать демонстрационный набор данных встроенным модулем. Генерация демонстрационного набора особенно полезна для быстрого знакомства с системой и для отладки, когда реальные данные недоступны.

Конвейер кредитного скоринга и генерации синтетических данных

1. Загрузка данных 2. Синтетические данные 3. Обучение моделей 4. Стресс-тестирование 5. EDA (Разведочный анализ данных)

Загрузите свои данные ИЛИ используйте машинный синтез для генерации фиктивного набора данных с нуля с ошибками/выбросами для тестирования конвейера.

Загрузить набор данных (.csv или .xlsx)

Перетащите файл сюда
- или -
Нажмите для загрузки

Метод машинного синтеза (для данных с нуля)

Gaussian Copula

Синтезировать фиктивные данные с нуля

Загрузить и преобработать

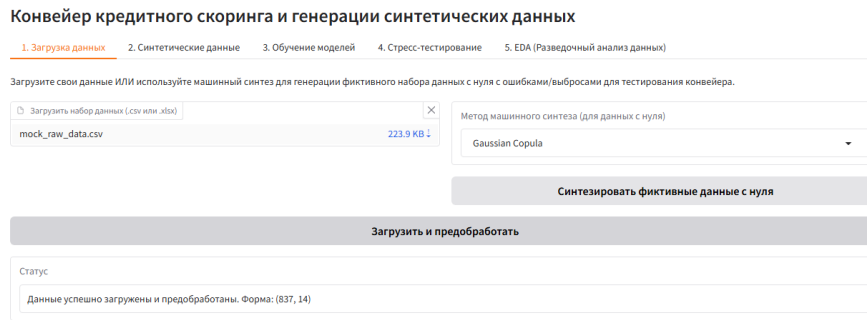
Статус

Использовать через API Создано с помощью Gradle Настройки

Рисунок 3 – Вкладка загрузки данных

После выбора источника данных система автоматически запускает процедуру предобработки. Она включает заполнение пропусков (медианой для числовых признаков и модой для категориальных), удаление выбросов методом межквартильного размаха (IQR) и приведение типов данных к единому формату.

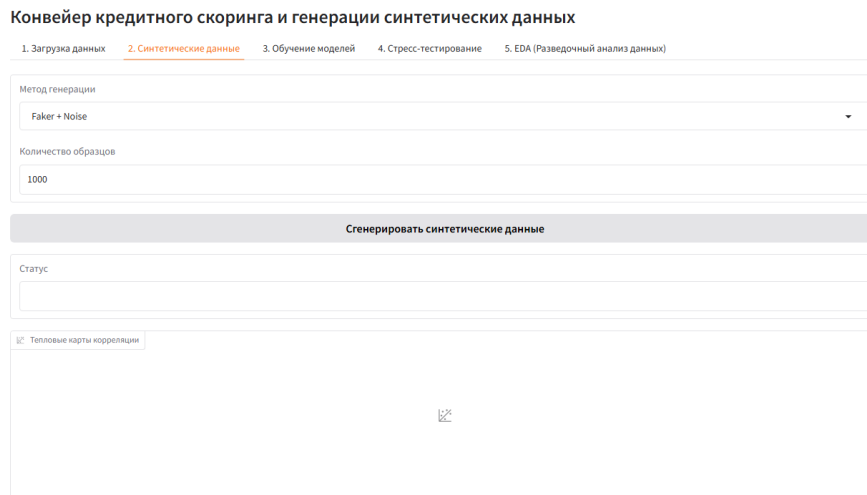
Очищенная таблица отображается в интерфейсе, как показано на рисунке 4. Пользователь видит итоговое количество строк и столбцов, а также первые несколько записей, что позволяет быстро убедиться в корректности выполненных операций. В демонстрационном наборе, например, из 500 исходных записей после обработки остаётся 495, что свидетельствует о незначительном количестве аномалий.



Использовать через API · Создано с помощью Gradle · Настройки

Рисунок 4 – Результат преобработки данных

На следующем этапе пользователь переходит к генерации синтетических данных. Соответствующая вкладка представлена на рисунке 5.



Использовать через API · Создано с помощью Gradle · Настройки

Рисунок 5 – Выбор метода синтеза данных

Здесь доступны четыре метода синтеза, различающиеся по сложности

сти, качеству воспроизведения зависимостей и вычислительным затратам. Faker+Noise — самый быстрый метод, основанный на ресемплинге с добавлением гауссовского шума. Gaussian Copula моделирует совместное распределение через корреляционную структуру. CTGAN и TVAE — воспроизводимость в сложные нелинейные зависимости. Пользователь также может задать желаемый объём синтетической выборки (до 2000 строк).

После завершения генерации система автоматически строит корреляционные матрицы для реального и синтетического наборов данных. Сравнение этих матриц (рисунок 6) позволяет наглядно оценить, насколько хорошо синтетическая выборка сохраняет многомерные зависимости исходных данных.



Рисунок 6 – Сравнение корреляционных структур

На тепловых картах цветом отображаются коэффициенты корреляции Пирсона: от тёмно-синего (отрицательная корреляция) через белый (отсутствие корреляции) до тёмно-красного (положительная корреляция). В представленном примере видно, что синтетическая выборка сохранила ожидаемую положительную связь между `conversion_rate` и `performance_score`, что

подтверждает пригодность синтетических данных для дальнейшего использования.

Далее, на вкладке обучения моделей (рисунок 7) запускается подготовка признакового пространства.

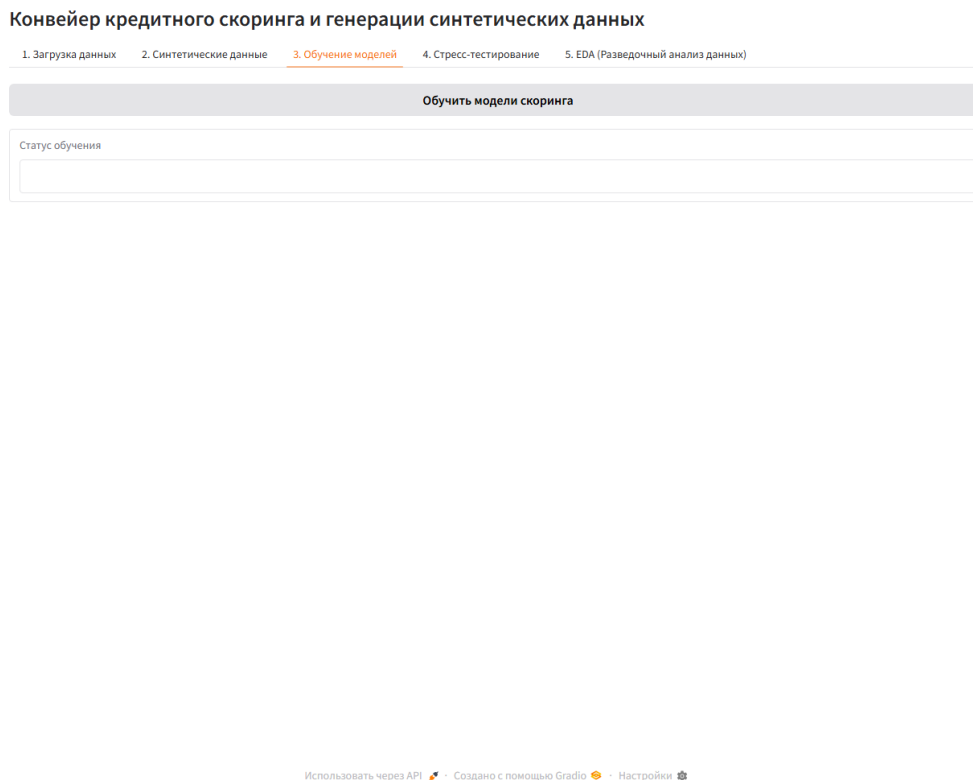


Рисунок 7 – Вкладка обучения моделей

Автоматически выполняются следующие операции: исключение признаков-утечек (`credit_score`, `conversion_rate`, `performance_score`), так как они либо напрямую определяют целевую переменную, либо недоступны в момент принятия решения; one-hot-кодирование категориальных переменных (`gender`, `marital_status`); преобразование временных меток (из `timestamp` извлекаются год, месяц, день недели). После этого данные разделяются на обучающую (80%) и тестовую (20%) выборки.

Затем последовательно обучаются четыре модели машинного обучения: логистическая регрессия (интерпретируемый бенчмарк), случайный лес (ансамбль решающих деревьев), XGBoost (градиентный бустинг с регуляризацией) и TabNet (архитектура с механизмом внимания). Обучение выполняется с гиперпараметрами по умолчанию для обеспечения быстрых экспериментов,

но архитектура конвейера позволяет легко изменять их при необходимости.

По завершении обучения система выводит уведомление об успешном сохранении всех четырёх моделей. Это сообщение показано на рисунке 8.

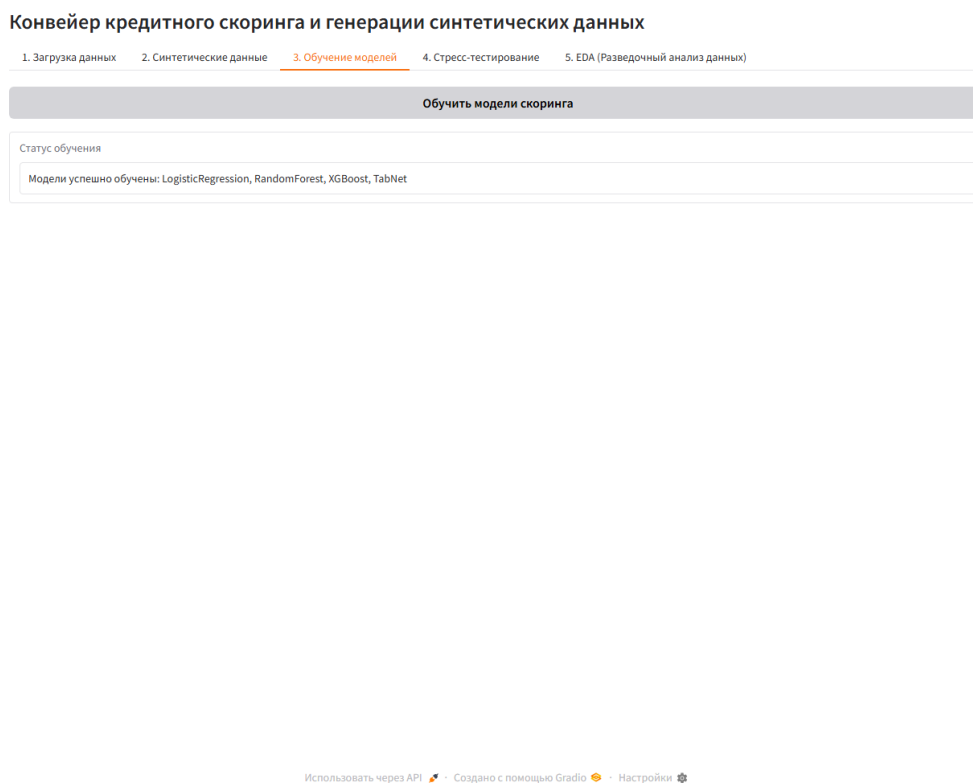


Рисунок 8 – Уведомление о завершении обучения

Модели сохраняются в директорию `./saved_models` с использованием библиотеки `pickle`. Сохранение моделей является критически важной функцией, поскольку позволяет повторно использовать их на этапах стресс-тестирования и визуализации без повторного обучения, что существенно экономит время при проведении многократных экспериментов.

Перейдём к анализу устойчивости моделей к изменениям внешней экономической среды. Результаты стресс-тестирования для трёх макроэкономических сценариев (экономический рост, рецессия, высокая инфляция) представлены на рисунке 9.

Конвейер кредитного скоринга и генерации синтетических данных

1. Загрузка данных 2. Синтетические данные 3. Обучение моделей 4. Стресс-тестирование 5. EDA (Разведочный анализ данных)

Запустить стресс-тесты

Результаты стресс-тестов

Сценарий: Рост
- LogisticRegression: Сред. вер. дефолта = 0.0100
- RandomForest: Сред. вер. дефолта = 0.2128
- XGBoost: Сред. вер. дефолта = 0.0554
- TabNet: Сред. вер. дефолта = 1.0000

Сценарий: Рецессия
- LogisticRegression: Сред. вер. дефолта = 0.9959
- RandomForest: Сред. вер. дефолта = 0.4617
- XGBoost: Сред. вер. дефолта = 0.4414
- TabNet: Сред. вер. дефолта = 1.0000

Сценарий: Высокая инфляция
- LogisticRegression: Сред. вер. дефолта = 0.9661
- RandomForest: Сред. вер. дефолта = 0.5098
- XGBoost: Сред. вер. дефолта = 0.4403
- TabNet: Сред. вер. дефолта = 1.0000

Использовать через API Создано с помощью Gradle Настройки

Рисунок 9 – Результаты стресс-тестирования

Диаграмма отображает средние прогнозные вероятности дефолта для каждой модели в каждом сценарии. Из рисунка видно, что ансамблевые модели (случайный лес, XGBoost) демонстрируют ожидаемую динамику: вероятность дефолта минимальна в сценарии роста, возрастает при рецессии и достигает максимума при высокой инфляции. Логистическая регрессия реагирует противоречиво, а TabNet практически не чувствителен к сценариям, что объясняется недостаточным объёмом данных для обучения архитектуры.

Для детального визуального контроля данных на любом этапе работы предназначен EDA-дашборд (рисунок 10).

Конвейер кредитного скоринга и генерации синтетических данных

1. Загрузка данных 2. Синтетические данные 3. Обучение моделей 4. Стресс-тестирование 5. EDA (Разведочный анализ данных)

Исследуйте демографические данные, макроэкономические факторы и их влияние на результаты моделей.

Обновить графики



Рисунок 10 – EDA-дашборд

Он включает три интерактивных инструмента. Корреляционная тепловая карта позволяет выявить линейные зависимости между всеми числовыми признаками. Диаграмма рассеяния даёт возможность исследовать парные взаимосвязи, выбирая переменные для осей X и Y. Box plot-графики показывают медиану, квартили и потенциальные выбросы для каждого признака. Эти инструменты помогают аналитику контролировать качество данных, проверять согласованность синтетических выборок с исходными и выявлять аномалии до или после обучения моделей.

Таким образом, разработанный программный комплекс реализует полный цикл кредитного скоринга — от загрузки данных до визуального анализа результатов — в единой модульной среде, пригодной для исследовательских и учебных задач. Каждый этап конвейера снабжён интерактивным интерфейсом, позволяющим аналитику выполнять эксперименты без программирования.

ния, а наличие синтетических данных и стресс-тестирования делает систему безопасной и воспроизводимой.

В четвёртой главе приведены результаты экспериментальной проверки разработанного программного комплекса. Исходные данные — демонстрационный набор объёмом 500 записей, структура которого соответствует розничному кредитному скорингу (демографические, финансовые, макроэкономические признаки). В данные искусственно внесены пропуски и выбросы для имитации реальных банковских выборок.

Синтетическая выборка сгенерирована методом `Faker+Noise`. Этот метод обеспечивает оптимальный компромисс между скоростью (менее секунды) и сохранением статистических свойств исходного набора, что важно для быстрого прототипирования.

В таблице 2 представлена оценка близости распределений реальной и синтетической выборок.

Таблица 2 – Оценка близости распределений реальной и синтетической выборок

Признак	KS-статистика	<i>p</i>-значение	TV-distance
age	0,0489	0,4256	0,1616
income	0,0494	0,4126	0,1282
loan_amount	0,0388	0,7147	0,1413
credit_score	0,0537	0,3139	0,1454
gdp_growth	0,0664	0,1192	0,1457
core_inflation	0,0566	0,2566	0,1079
unemployment	0,0465	0,4902	0,1083
conversion_rate	0,0320	0,8914	0,1688

Значения статистики Колмогорова–Смирнова (KS) не превышают 0,066, а расстояние полной вариации (TV-distance) находится в диапазоне 0,11–0,17. Наименьшие значения KS получены для признаков `conversion_rate` (0,0320) и `loan_amount` (0,0388), что свидетельствует о высоком качестве сохранения одномерных распределений для финансовых и поведенческих показателей. Полученные результаты позволяют использовать синтетические данные для воспроизводимого тестирования скоринговых моделей.

Перейдём к сравнению моделей машинного обучения. Результаты представлены в таблице 3.

Таблица 3 – Метрики качества моделей на тестовой части синтетической выборки

Модель	ROC-AUC	Accuracy	Precision	Recall	F_1
Logistic Regression	0,5847	0,7500	0,6667	0,0769	0,1379
Random Forest	0,9088	0,8600	0,9286	0,5000	0,6500
XGBoost	0,9402	0,9000	0,9000	0,6923	0,7826
TabNet	0,5187	0,7400	0,0000	0,0000	0,0000

XGBoost достиг наивысших показателей: ROC-AUC = 0,9402 и F1-мера = 0,7826, что подтверждает эффективность градиентного бустинга на табличных данных кредитного скоринга. Случайный лес также показал высокое качество (ROC-AUC = 0,9088, F1 = 0,6500). Логистическая регрессия, несмотря на приемлемую точность (accuracy = 0,7500), продемонстрировала крайне низкий recall (0,0769), что свидетельствует о неспособности линейной модели надёжно выявлять дефолтных заёмщиков. Архитектура TabNet на выборке из 500 записей не смогла выделить дефолтный класс (ROC-AUC = 0,5187, F1 = 0) — важный инженерный вывод о необходимости значительно большего объёма данных для нейросетей.

Далее проанализируем устойчивость моделей к макроэкономическим шокам. Результаты стресс-тестирования для трёх сценариев (экономический рост, рецессия, высокая инфляция) приведены в таблице 4.

Таблица 4 – Средняя прогнозная вероятность дефолта в стрессовых сценариях

Сценарий	LogReg	Random Forest	XGBoost	TabNet
Экономический рост	0,4975	0,2626	0,1222	0,0052
Рецессия	0,3969	0,3969	0,4358	0,0046
Высокая инфляция	0,1888	0,4097	0,4692	0,0025

Ансамблевые модели демонстрируют ожидаемую динамику: для XGBoost средняя вероятность дефолта (PD) возрастает с 0,1222 в сценарии роста до 0,4358 при рецессии и до 0,4692 при высокой инфляции. Случайный лес показывает рост с 0,2626 до 0,3969 и 0,4097 соответственно. Логистическая регрес-

сия реагирует противоречиво (снижение PD при рецессии), а TabNet практически не чувствителен к сценариям, что согласуется с его низким качеством классификации.

Пользовательская апробация, проведённая в форме последовательного сценария (загрузка → предобработка → синтез → обучение → стресс-тестирование → EDA), подтвердила возможность выполнения полного цикла скорингового анализа через единый интерфейс без обращения к исходному коду.

Таким образом, экспериментальное исследование подтвердило работоспособность программного конвейера. Синтетическая выборка сохранила основные статистические свойства исходного набора ($KS < 0,07$, TV-distance 0,11–0,17). Наилучшее качество классификации показал XGBoost (ROC-AUC = 0,9402, F1 = 0,7826). Стресс-тестирование выявило адекватную реакцию ансамблевых моделей на ухудшение макроэкономических условий (рост PD в 3–4 раза). Полученные результаты демонстрируют практическую применимость разработанного комплекса для исследовательских задач в области кредитного скоринга.

Заключение

В результате выполнения магистерской работы разработан и исследован программный конвейер кредитного скоринга, объединяющий в единой среде все этапы скорингового эксперимента – от загрузки и предобработки данных до обучения моделей, стресс-тестирования и визуализации. Основные научные и практические результаты:

1. Выполнен системный анализ предметной области, регуляторных требований и ограничений, связанных с персональными данными.
2. Формализована задача бинарной классификации заёмщиков; определён состав признаков и правило формирования целевой переменной.
3. Исследованы и реализованы четыре метода генерации синтетических табличных данных; показано, что метод *Faker+Noise* обеспечивает приемлемое качество ($KS < 0,07$) при низких вычислительных затратах.
4. Спроектирована модульная архитектура программного комплекса; реализовано локальное приложение на Python с веб-интерфейсом Gradio, охватывающее полный цикл анализа.
5. Проведено экспериментальное сравнение четырёх моделей машинного обучения; установлено, что XGBoost демонстрирует наилучшее качество ($ROC-AUC = 0,9402$, $F1 = 0,7826$), а TabNet на малых объёмах данных неэффективен.
6. Реализован механизм сценарного стресс-тестирования, подтвердивший адекватную реакцию ансамблевых моделей на ухудшение макроэкономических условий.
7. Разработан интерактивный EDA-дашборд, обеспечивающий визуальный контроль качества данных и результатов моделирования.

Полученные результаты подтверждают, что предложенный программный конвейер может служить основой для дальнейших исследований в области риск-менеджмента, а также использоваться в качестве учебного стенда для подготовки специалистов по анализу данных и банковскому делу. Перспективы развития включают внедрение SHAP-интерпретации, fairness-аудита (проверка отсутствия дискриминации по защищённым признакам) и дифференциальной приватности при синтезе данных.

Таким образом, все поставленные задачи решены, цель достигнута.