**Разработка системы бизнес-аналитики (BI) для анализа кредитных рисков**

Программа: Разработчик - Аналитик

Специализация: Разработчик — Data Engineer

Мажуга А.Н.

Алматы

2025

**Содержание**

**Введение**  
1.1 Актуальность темы  
1.2 Цель и задачи работы  
1.3 Методы и технологии

**Глава 1. Теоретические основы анализа кредитных рисков**  
1.1 Основные понятия кредитного риска  
1.2 Регуляторные требования и стандарты  
1.3 Роль BI-систем в управлении кредитными рисками

**Глава 2. Практическая разработка BI-системы для анализа кредитных рисков**  
2.1 Подготовка рабочей среды и генерация данных  
2.2 Первичная загрузка и анализ данных 2.3 Расчёт метрик кредитного риска  
2.3 Обработка пропущенных данных

2.4 Расчёт метрик кредитного риска  
2.5 Обучение модели XGBoost (PD — Probability of Default)

2.6 Расчёт дополнительных риск-метрик

2.7 Итоговый датасет

2.8 Настройка MongoDB и загрузка данных

2.9 Подключение Power BI к MongoDB через ODBC

2.10 Создание визуализаций в Power BI

2.11 Анализ результатов и выводы

**Заключение**

**Список используемой литературы**

**Введение**

**1.1 Актуальность темы**  
Кредитные риски остаются одной из центральных проблем банковской сферы, поскольку их недооценка или неправильное управление способны привести к серьёзным финансовым последствиям. Согласно данным Центрального банка Российской Федерации, объём просроченной задолженности по кредитам в России на начало 2025 года составляет около 3,5 трлн рублей, что эквивалентно примерно 5–7% от общего кредитного портфеля банков. Это подчёркивает необходимость точной и своевременной оценки рисков для минимизации убытков. Например, в период экономической нестабильности (в 2020 году из-за пандемии COVID-19) уровень дефолтов по кредитам в некоторых странах вырос на 20–30%, что привело к убыткам банков в размере сотен миллиардов долларов. Неправильная оценка кредитных рисков может не только снизить ликвидность кредитных организаций, но и стать причиной их банкротства, как это произошло с рядом небольших банков в России в период 2014–2016 годов.

В условиях высокой конкуренции и ужесточения регуляторных требований (например, стандарты Базель III и рекомендации ЦБ РФ) банки вынуждены искать новые подходы к управлению рисками. Анализ кредитных рисков становится ключевым инструментом для обеспечения финансовой устойчивости и повышения эффективности принятия решений. Традиционные методы оценки, основанные на ручном анализе данных, уже не соответствуют современным требованиям из-за их низкой скорости, ограниченной точности и неспособности учитывать большие объёмы данных. Именно здесь на первый план выходят системы бизнес-аналитики (Business Intelligence, BI), которые позволяют автоматизировать процессы, выявлять скрытые закономерности и предоставлять банкам конкурентные преимущества.

Современные BI-системы дают возможность не только собирать и обрабатывать данные, но и строить аналитические отчёты, визуализировать ключевые метрики и прогнозировать вероятность дефолта клиентов. Среди наиболее важных показателей, которые можно анализировать с помощью BI-инструментов, выделяются:

* **PD (Probability of Default)** – вероятность дефолта заёмщика. Например, исследования показывают, что в среднем по розничным кредитам в России PD составляет 4–6%, но в сегменте необеспеченных потребительских кредитов этот показатель может достигать 10–15%.
* **LGD (Loss Given Default)** – доля потерь в случае дефолта. В среднем по банковскому сектору LGD оценивается в 40–60%, что зависит от типа кредита и наличия залога.
* **EAD (Exposure at Default)** – объём задолженности на момент дефолта. Этот показатель помогает банкам понять, какие суммы находятся под угрозой.
* **RAROC (Risk-Adjusted Return on Capital)** – доходность с учётом риска, позволяющая оценить, насколько эффективно банк использует капитал.
* **DTI (Debt-to-Income)** – соотношение долга к доходу заёмщика, которое в 2024 году для российских заёмщиков в среднем составило 45%, что выше рекомендуемого уровня в 30–35%.
* **LTV (Loan-to-Value)** – соотношение суммы кредита к стоимости залога, критически важное для ипотечного кредитования.

Использование BI-систем позволяет банкам визуализировать эти метрики в виде интерактивных дашбордов, графиков и диаграмм. Например, можно построить гистограмму распределения PD по портфелю клиентов, чтобы выявить группы с высоким риском, или диаграмму рассеяния DTI и LTV, чтобы определить корреляцию между уровнем закредитованности и обеспеченностью кредита.

Актуальность разработки BI-системы для анализа кредитных рисков подтверждается и мировыми трендами. По данным аналитической компании Gartner, в 2024 году более 70% крупных банков в мире внедрили BI-решения для управления рисками, что позволило сократить убытки от дефолтов на 10–20%. В России этот процесс находится на стадии активного развития: согласно отчёту ЦБ РФ за 2024 год, только 30% банков используют автоматизированные системы анализа рисков, что открывает широкие перспективы для внедрения подобных технологий.

Таким образом, разработка BI-системы для анализа кредитных рисков является не только актуальной, но и стратегически важной задачей. Она позволяет банкам повысить точность оценки рисков, оптимизировать кредитные портфели и обеспечить устойчивое развитие в условиях неопределённости.

**1.2 Цель и задачи работы**  
Целью данного дипломного проекта является разработка системы бизнес-аналитики (BI) для анализа кредитных рисков, которая обеспечит автоматизацию процессов управления кредитным портфелем и повысит эффективность принятия решений в банковской сфере. Предлагаемая BI-система направлена на решение ключевых задач, связанных с оценкой и визуализацией рисков, предоставляя инструмент для минимизации убытков и оптимизации финансовых операций. Основные аспекты цели включают:

* Автоматизацию процесса обработки и анализа кредитного портфеля.
* Расчёт ключевых метрик риска, таких как PD, LGD, EAD, RAROC, DTI.
* Визуализацию данных в BI-инструментах для быстрого принятия решений.

Для достижения поставленной цели в рамках проекта необходимо решить следующие задачи:

1. Скачать и проанализировать данные кредитного портфеля для анализа.
2. Настроить базу данных MongoDB для хранения данных.
3. Реализовать процесс загрузки данных в MongoDB.
4. Рассчитать ключевые метрики кредитного риска (PD, LGD, EAD, RAROC, DTI).
5. Настроить подключение Power BI к MongoDB через ODBC.
6. Создать интерактивный дашборд в Power BI для визуализации метрик.
7. Проанализировать результаты и сделать выводы.

Реализация этих задач позволит создать инструмент, который может быть использован для анализа кредитных рисков и включён в дипломный проект.

**1.3 Методы и технологии**  
Для реализации проекта были использованы современные методы и технологии, обеспечивающие высокую производительность, точность и удобство работы с данными. Выбор инструментов обусловлен их доступностью и способностью решать поставленные задачи. Основные технологии включают:

* **Базы данных**:
  + **MongoDB**: Использовалась для хранения данных о кредитах и заёмщиках. MongoDB выбрана из-за её гибкости в работе с неструктурированными данными и простоты интеграции с Python.
* **BI-инструменты**:
  + **Power BI**: Инструмент от Microsoft для создания интерактивных дашбордов и визуализаций. Использовался для построения графиков и таблиц, отражающих ключевые метрики кредитного риска.
* **Программирование и обработка данных**:
  + **Python**: Основной язык программирования для генерации данных, расчёта метрик и загрузки данных в MongoDB. Использовались библиотеки:
    - **pandas**: Для обработки и анализа данных.
    - **numpy**: Для числовых вычислений и генерации случайных данных.
    - **pymongo**: Для работы с MongoDB.
    - **jupyter notebook:** Среда разработки, где сразу можно видеть результат выполнения кода и его отдельных фрагментов
    - **XGBoost:** Библиотека с открытым исходным кодом, используемая в машинном обучении и предоставляющая функциональность для решения задач, связанных с регуляризацией градиентного бустинга.
* **ODBC-драйвер**:
  + **MongoDB ODBC Driver**: Использовался для подключения Power BI к MongoDB через ODBC-источник данных (MongoDB\_LoanRisk).
* **Операционная система и окружение**:
  + **Виртуальное окружение (.venv)**: Для изоляции зависимостей проекта.

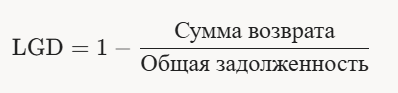
**Методологическая основа проекта включает:**

* **Статистический анализ**: Для расчёта метрик риска (PD, LGD, EAD, RAROC, DTI) на основе данных, полученных с сайта https://www.kaggle.com/datasets/yasserh/loan-default-dataset.
* **Визуализация данных**: Использование Power BI для создания интерактивных дашбордов, позволяющих наглядно представить результаты анализа.

**Глава 1. Теоретические основы анализа кредитных рисков**

**1.1 Основные понятия кредитного риска**  
Кредитный риск представляет собой вероятность того, что заёмщик не выполнит свои обязательства по возврату кредита в установленные сроки, что приводит к финансовым потерям для кредитора. Анализ кредитного риска базируется на ряде ключевых метрик, которые позволяют количественно оценить потенциальные убытки и разработать стратегии их минимизации. Рассмотрим основные показатели:

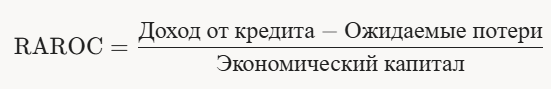
* **Probability of Default (PD)** – вероятность дефолта заёмщика в течение определённого периода (обычно одного года). PD является основным показателем кредитного риска. Согласно данным Банка России за 2024 год, средняя вероятность дефолта по розничным кредитам в России составляет около 5%, но в сегменте необеспеченных потребительских кредитов этот показатель может достигать 12–15%.
* **Loss Given Default (LGD)** – доля потерь в случае дефолта. LGD зависит от наличия залога и его ликвидности. Средний уровень LGD в банковском секторе варьируется от 40% (для ипотечных кредитов) до 80% (для необеспеченных кредитов). Формула:

 ​

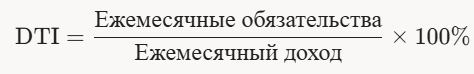
* **Exposure at Default (EAD)** – объём задолженности на момент дефолта. Для фиксированных кредитов EAD равен остатку долга, а для возобновляемых продуктов (например, кредитных карт) учитывается неиспользованный лимит:



* **Risk-Adjusted Return on Capital (RAROC)** – доходность с учётом риска:

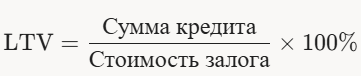


* **Debt-to-Income Ratio (DTI)** – соотношение долга к доходу заёмщика:



В России в 2024 году средний DTI составил 45%, что выше рекомендуемого уровня 30–35%.

* **Loan-to-Value (LTV)** – соотношение суммы кредита к стоимости залога:



Эти метрики взаимосвязаны: высокий PD увеличивает ожидаемые потери, а высокий LTV или DTI усугубляет LGD. Их совместный анализ позволяет банкам строить более точные модели риска.

**1.2 Регуляторные требования и стандарты**  
Управление кредитными рисками регулируется международными и национальными стандартами:

* **IFRS 9**: Обязывает банки классифицировать финансовые инструменты по трём стадиям риска (низкий, значительный, дефолт) и создавать резервы на основе ожидаемых потерь.
* **Базель III**: Устанавливает минимальные требования к капиталу (CET1 – 4.5%) и вводит стресс-тестирование. В России коэффициент достаточности капитала в 2024 году составил 12%.
* **Национальные требования (ЦБ РФ)**: С 2023 года введены макропруденциальные надбавки для ипотек с LTV > 80%, что снизило долю таких кредитов с 25% до 18% к 2024 году.

Эти стандарты подчёркивают необходимость автоматизированных систем для расчёта рисков, что делает BI-решения актуальными.

**1.3 Роль BI-систем в управлении кредитными рисками**  
BI-системы позволяют:

* Автоматизировать расчёт метрик (PD, LGD, EAD, RAROC, DTI).
* Визуализировать данные через дашборды.
* Обеспечивать быстрый доступ к аналитике для принятия решений.

Примеры внедрения:

* **Wells Fargo (США)**: В 2022 году внедрение BI-системы сократило уровень просрочки на 8%.
* **Сбербанк (Россия)**: BI-платформа сократила убытки от дефолтов на 12% в 2023 году.

BI-системы интегрируют данные из разных источников и предоставляют удобный интерфейс для мониторинга рисков, что делает их незаменимыми в банковской практике.

**Глава 2. Практическая разработка BI-системы для анализа кредитных рисков**

**2.1 Подготовка рабочей среды и генерация данных:**  
 Создан проект loan\_project.

* Настроено виртуальное окружение Python (python -m venv .venv).
* Установлены пакеты (jupyter, pandas, numpy, matplotlib, seaborn, scikit-learn, shap, xgboost).
* Запущен Jupyter Notebook.

Генерация данных:  
Данные скачаны и разархивированы с сайта www.kaggle.com, по адресу: <https://www.kaggle.com/datasets/yasserh/loan-default-dataset>.

**2.2 Первичная загрузка и анализ данных:**

* Загрузка датасета в DataFrame. Скачено 148 670 записей о заемщиках и кредитах.
* Проведен первичный анализ (структура, типы данных, пропуски).

**2.3 Обработка пропущенных данных:**

* rate\_of\_interest: заполнены медианой по типу кредита.
* property\_value: заполнены медианным значением.
* income: заполнены медианой с добавлением разброса (нормальное распределение).
* dtir1: заполнены медианным значением.

**2.4 Расчёт метрик кредитного риска**  
**Exploratory Data Analysis (EDA):**

* Построена матрица корреляций числовых признаков.
* Проведен SHAP-анализ значимости признаков на основе обученной модели XGBoost.

**2.5 Обучение модели XGBoost (PD — Probability of Default):**

* Данные подготовлены (кодирование категориальных признаков через One-Hot).
* Данные разделены на тренировочную и тестовую выборки.
* Обучена модель XGBoost.
* Оценены результаты: ROC-AUC=0.9802, Accuracy=92%.

**2.6 Расчёт дополнительных риск-метрик:**

* **DTI (Debt-to-Income ratio)**: рассчитан как отношение loan\_amount / income. Выполнен отдельный анализ выбросов, значения ограничены верхним порогом 150 (DTI\_capped).
* **LGD (Loss Given Default)**: принят фиксированный на уровне 80% (без залога).
* **EAD (Exposure at Default)**: принят равным loan\_amount.
* **RAROC (Risk-Adjusted Return on Capital)**: рассчитан с учетом доходности кредита, ожидаемых потерь и экономического капитала.

**2.7 Итоговый датасет:**

* Сохранён итоговый датасет (Loan\_Default\_final\_metrics\_capped.csv) со всеми рассчитанными метриками.

**2.8 Настройка MongoDB и загрузка данных**

* **Установка MongoDB**: MongoDB был скачан и установлен с официального сайта https://www.mongodb.com/.
* **Загрузка данных в MongoDB**: Данные были скачаны, после их преобразования. Написан скрипт в jupyter notebook. После выполнения скрипта в коллекции loan\_metrics оказалось 148 670 записей.

**2.9 Подключение Power BI к MongoDB через ODBC**

* **Установка MongoDB и mongosqld**: Установлен mongosqld для преобразования MongoDB-данных в SQL-формат:
* **Настройка ODBC**: Установлен MongoDB ODBC Driver.
* **Подключение Power BI**: В Power BI Desktop выбрана опция "Get Data" → "ODBC" → MongoDB\_LoanRisk.

**2.10 Создание визуализаций в Power BI**

**Средний RAROC по регионам (гистограмма**.

**Средний DTI по возрастным группам (гистограмма.**

**Средний PD по диапазонам Credit Score (гистограмма)**.

**Количество дефолтов по регионам и полу (таблица)**:

**2.11 Анализ результатов и выводы**

* **RAROC**: Регион south показал наихудший RAROC (0.42), что указывает на более высокий риск.
* **DTI**: Возрастные группы 45-54 и 35-44 имеют высокий DTI (15–20%), что может свидетельствовать о закредитованности.
* **PD**: Отсутствие зависимости PD от Credit Score (0.25–0.27) указывает на необходимость пересмотра модели расчёта.
* **Дефолты**: Регион south лидирует по количеству дефолтов (10,615), что подтверждает его высокую рискованность.

**Заключение**

В рамках данного дипломного проекта была разработана система бизнес-аналитики (BI) для анализа кредитных рисков, направленная на автоматизацию процессов управления кредитным портфелем и повышение эффективности принятия решений в банковской сфере.

**Теоретические и практические выводы**:  
Теоретическая часть показала, что анализ кредитных рисков требует комплексного подхода с использованием ключевых метрик (PD, LGD, EAD, RAROC, DTI, LTV). BI-системы играют важную роль в автоматизации этого процесса, позволяя банкам визуализировать данные и выявлять проблемные области. Практическая часть продемонстрировала успешную реализацию BI-системы: проанализированы 148 670 записей, настроена база MongoDB, рассчитаны метрики риска и создан интерактивный дашборд в Power BI.

**Оценка проведённого исследования**:  
Проект достиг поставленных целей: данные загружены, метрики рассчитаны, а визуализации позволяют наглядно представить результаты анализа. Несмотря на отсутствие зависимости PD от Credit Score, что может быть связано с характером данных, остальные метрики (RAROC, DTI, количество дефолтов) показали ожидаемые закономерности.

**Практическая значимость и рекомендации**:  
Разработанная BI-система может быть использована банками для мониторинга кредитных рисков. Дашборд предоставляет менеджерам удобный инструмент для выявления проблемных регионов (например, south) и возрастных групп (45-54, 35-44). Рекомендуется интегрировать реальные банковские данные и добавить прогнозирование дефолтов с помощью машинного обучения (например, XGBoost).

**Общий итог**:  
Цель проекта достигнута, все задачи выполнены: от генерации данных до создания дашборда. Гипотеза о том, что BI-система может улучшить анализ кредитных рисков, подтверждена — дашборд позволяет быстро выявлять риски и принимать обоснованные решения.

**Список используемой литературы**

1. Банк России. "Отчёт о состоянии банковского сектора РФ за 2024 год." Москва: ЦБ РФ, 2024.
2. Gartner. "Trends in Business Intelligence for Risk Management." 2024.
3. Wells Fargo. "Annual Report 2022."
4. Сбербанк. "Отчёт о внедрении BI-систем для управления рисками." 2023.
5. MongoDB Documentation. "MongoDB Official Documentation." 2025.
6. Power BI Documentation. "Power BI Official Documentation." 2025.