Московский государственный университет имени М.В. Ломоносова

Направление 38.03.01 Экономика

Программа «Национальная экономика»

Прогнозирование рисков банковского кредитования с использованием технологий искусственного интеллекта

Выпускная квалификационная работа

Студент

Касенова Асем Ардаковна

Научный руководитель:

Кандидат экономических наук, доцент

Сидоренко Владимир Николаевич

Астана

2026

ОГЛАВЛЕНИЕ

Введение

Актуальность темы исследования. В настоящее время, в условиях технологического развития, искусственный интеллект (далее ИИ) стал необходимым инструментом, позволяющим с высокой скоростью обрабатывать большие массивы данных. Для банковского сектора управление кредитными рисками является одной из ключевых задач, в частности оценка кредитоспособности заемщиков. В современном финансовом секторе Казахстана использование кредитного скоринга поднимает важный вопрос о справедливом доступе к потребительскому кредитованию. Кредитный скоринг – это статистический процесс, позволяющий прогнозировать вероятность дефолта заемщика банка. Значимость кредитного скоринга для Казахстана определяется его ролью в снижении долговой нагрузки населения, ограничении массового заимствования и обеспечении устойчивого и эффективного развития финансового сектора.

С развитием услуг микрокредитования и предоставления кредитов в рассрочку банки стали получать значительно больше заявок на кредит, что на первоначальном этапе было выгодно финансовым учреждениям. Однако по мере роста количества заявок анализ разнообразных кредитных историй клиентов стал сложной задачей для сотрудников банка, что приводило к снижению прибыли. В связи с этим возникает необходимость в автоматизации и оптимизации процесса оценки и выдачи кредитов с помощью технологий ИИ, что позволяет минимизировать кредитные риски для обеих сторон. Практическая значимость работы заключается в том, что результаты исследования могут быть использованы финансовыми организациями для повышения эффективности принятия кредитных решений. Применение алгоритмов машинного и глубокого обучения позволит банкам принимать более точные и обоснованные решения о кредитоспособности клиентов, снижая вероятность возникновения финансовых рисков.

Степень разработанности темы исследования. Применение искусственного интеллекта в банковском кредитовании рассматривались в работах H. Sadok, F. Sakka и M.E. El Maknouzi, в магистерской диссертации Ш. Сяоюй, а также в исследованиях Г.З. Зиятбековой, А.А. Давыдовой и О.Л. Ксенофонтовой, посвященных использованию методов машинного обучения и интеллектуального анализа данных в банковской сфере. Данные работы свидетельствуют об устойчивом научном интересе внедрения ИИ в деятельность финансовых учреждений. Тем не менее для Казахстана данное направление является новым в контексте цифровизации финансового сектора и, следовательно, требует дальнейшего исследования.

Цель и задачи исследования – разработать метод прогнозирования дефолта клиента банка при выдаче кредита на основе методов машинного и глубокого обучения. Для достижения поставленной цели были определены следующие задачи:

1. Провести анализ и предобработку данных о заемщиках.
2. Обучить и протестировать модели классификации для прогнозирования дефолта.

3. Подобрать гиперпараметры моделей для повышения качества прогнозирования дефолтных клиентов.

4. Оценить качество моделей и проанализировать полученные результаты.

Объектом исследования является выданный кредит банком и связанное с ним наступление либо отсутствие дефолта по выданным кредитам.

Предметом исследования является кредитный скоринг для оценки риска невозврата кредита.

Для достижения целей и задач исследования использовались следующие методы:

- машинного и глубокого обучения;
- статистического анализа;
- оценивания моделей;
- визуализации исторических данных.

Научная новизна исследования исследования заключается в следующем:

- после завершения работы добавим

Информационной базой исследования являлись платформы: kaggle, github, stackoverflow, нормативно-правовая база, научные статьи и монографии.

Структура и объем работы. Выпускная квалификационная работа состоит из 3 глав, заключения, списка литературы, X таблиц, Y рисунков и Z приложений.

Объем исследовательской работы: 50 страниц.

ГЛАВА 1. АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ В КРЕДИТНОМ СКОРИНГЕ

1.1 Предобработка данных

В данной работе используются обезличенные данные АО «Альфа-Банка» [1]. Данные состоят из 12 файлов (`train_data_0.pq - train_data_11.pq`), содержащих информацию о платежах клиентов банка. В каждом из 12 файлов содержится информация о 250 000 клиентах. При этом один клиент может иметь несколько кредитов, и каждому такому клиенту соответствует персональный `id` (идентификационный номер). Отдельно имеется файл `train_target.csv`, который состоит из 3 млн строк, и каждая строка соответствует клиенту с меткой (флагом) равной 0 (отсутствие дефолта) или 1 (наличие дефолта). Задача предобработки данных состоит в структурировании исходной информации, т.е. формирование единого датасета, выделение важных признаков (колонок), выявление аномальных клиентов (определение аномальности будет приведено ниже), визуализация данных и тестирование моделей МО и ГО на этих данных. Программный код формирования единого датасета реализован в листинге 1 (см. приложение А).

Комментарий 1.1. *Пояснение к листингу: 1*

1. Строки 1 – 3. Импортируются необходимые библиотеки: `pandas` – для работы с табличными данными, `os` – для работы с файловой системой и `pyarrow.parquet` – для чтения файлов формата `.parquet`.

2. Строка 5. Задается путь `path = "train_data"` к папке, в которой находятся исходные файлы формата `.pq`.

3. Строки 6 – 13. Запускается цикл `for`, который перебирает все файлы в папке `train_data`. Формируется имя поочередного файла `train_data_i.pq`, создается объект `ParquetDataset` для текущего файла, из которого данные считываются в `DataFrame (df)`. Затем выполняется агрегация данных по признаку *id*, вычисляются средние значения признаков, после чего данные сохраняются в соответствующий `csv` – файл в папку `train_data_csv_all`.

4. Строка 14. Выводится список файлов каталога `train_data` для проверки корректности формирования файлов.

5. Строки 16 – 21. Задается путь к папке с полученными `csv`-файлами `train_data_csv_all`. Создается пустой список `frames` для последующего хранения отдельных `DataFrame`.

[1] Соревнование на данных кредитных историй [Электронный ресурс] / Open Data Science. URL: https://ods.ai/competitions/dl-fintech-bki (дата обращения: 25.11.2025) Загл. с экрана.

Затем запускается цикл по всем `csv`-файлам в указанной папке и формируются файлы `train_data_i.csv`. Эти файлы считываются в `DataFrame(df)` и добавляются в список `frames`.

6. Строки 23 – 26. Все элементы списка `frames` объединяются в единый `DataFrame result` с помощью функции `pd.concat`. Полученный датасет сохраняется в файл `1_data_csv_all.csv`, после чего заново считывается в переменную `df_all` для последующего анализа.

Комментарий 1.2. *Пояснение к агрегированию клиентов в листинге: 1*

Агрегация клиентов по идентификатору `id` в строке 11 необходима для определения итоговой метки (флага) каждого клиента, поскольку одному заемщику может соответствовать несколько кредитных договоров. Задача состоит в присвоении каждому кредиту одного заемщика единую итоговую. метку. В данном исследовании в качестве общего значения для всех кредитов используется среднее исходных значений. Для понимания идеи приводится следующий пример в виде таблицы 1 :

Таблица 1 – Дисциплина оплаты кредитов клиента (id = 1)

id	N	M1	M2	M3	M4	M5	M6	flag
1	1	1	0	0	1	0	1	
1	2	0	1	0	1	1	2	0
1	3	1	2	0	0	3	2	
		Среднее значение						
1		0.67	1	0	1	1.33	1.67	0

Источник: составлено автором на основе: Соревнование на данных кредитных историй [Электронный ресурс] / Open Data Science. – URL: https://ods.ai/competitions/dl-fintech-bki (дата обращения: 25.11.2025).

Таким образом, данные трех строк, соответствующих трем кредитам клиента, в таблице 1 были усреднены и преобразованы в одну строку, которая содержит агрегированную информацию по клиенту с `id = 1`. В целом агрегировать клиентов можно не только по среднему значению, но и по моде или медиане. Однако в данной работе выбрано среднее значение, поскольку оно обладает важными статистическими свойствами, такими как несмещенность и состоятельность. После группирования данных был сформирован новый датасет, состоящий из 3 млн клиентов, каждому из которых соответствует одна итоговая метка. Получившийся датасет содержит 61 признак, из которых далее необходимо отобрать наиболее информативные, т.е. такие признаки, существенно влияющие на точность алгоритмов МО и ГО.

1.1.1 Метод главных компонент

1.1.2 Определение аномальности клиента

1.2 Базовые алгоритмы

1.2.1 Деревья решений

1.2.2 Случайный лес

1.2.3 Градиентный бустинг

1.2.4 Нейронные сети

ГЛАВА 2. ТЕСТИРОВАНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ НА ВСЕХ ДАННЫХ

2.1 Подбор гиперпараметров для алгоритмов машинного и глубокого обучения (нейронная сеть)

2.2 Деревья решений. Подбор гиперпараметров

2.3 Случайный лес. Подбор гиперпараметров

2.4 Градиентный бустинг. Подбор гиперпараметров

2.5 Нейронная сеть. Подбор гиперпараметров

ГЛАВА 3. РАЗРАБОТКА ВЕБ-ПРИЛОЖЕНИЯ

3.1 Библиотека *Streamlit* языка Python для разработки приложения

3.2 Дизайн и реализация приложения

Заключение

СПИСОК ЛИТЕРАТУРЫ

Книги:

1. Сяоюй, Ш. Разработка моделей кредитного скоринга заёмщиков коммерческих банков с использованием методов машинного обучения: магистерская диссертация // БГУ, Факультет прикладной математики и информатики, Кафедра математического моделирования и анализа данных. URL: https://elib.bsu.by/handle/123456789/331793 (дата обращения: 14.11.2025).

Интернет-ресурсы:

2. Метрические методы [Электронный ресурс] / Яндекс Практикум. URL: https://education.yandex.ru/handbook/ml/article/metricheskiye-metody. — Загл. с экрана.

3. DecisionTreeClassifier [Электронный ресурс] / Python-библиотека для машинного обучения. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. — Загл. с экрана.

4. Default of Credit Card Clients Dataset [Электронный ресурс] / Бизнес и конкурсная платформа по исследованию данных «Kaggle». – URL: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/data?select=UCI_Credit_Card.csv. — Загл. с экрана.

5. GridSearchCV [Электронный ресурс] / Python-библиотека для машинного обучения. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. — Загл. с экрана.

6. KNeighborsClassifier [Электронный ресурс] / Python-библиотека для машинного обучения. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. — Загл. с экрана.

7. PCA [Электронный ресурс] / Python-библиотека для машинного обучения. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html. — Загл. с экрана.

8. What is the k-nearest neighbors (KNN) algorithm? [Электронный ресурс] / International Business Machines. – URL: https://www.ibm.com/think/topics/knnl. — Загл. с экрана.

Приложение А. Предобработка данных

Листинг 1 – Формирование единого датасета

```
import pandas as pd
import os
import pyarrow.parquet as pq

path = "train_data"
for i,file in enumerate(os.listdir(path)):
    print('train_data_' + str(i) + '.pq')
    file = 'train_data_' + str(i) + '.pq'
    dataset = pq.ParquetDataset(os.path.join(path,file))
    df = dataset.read(use_threads=True).to_pandas()
    df_gr = df.groupby('id').agg('mean')
    file_csv = file.replace('.pq', '.csv')
    df_gr.to_csv(os.path.join('train_data_csv_all',file_csv))
os.listdir(path)

path = 'train_data_csv_all'
frames = []
for i,file_csv in enumerate(os.listdir(path)):
    file_csv = 'train_data_' + str(i) + '.csv'
    df = pd.read_csv(os.path.join('train_data_csv_all',file_csv) )
    frames.append(df)

result = pd.concat(frames)
file_csv_all = '1_data_csv_all.csv'
result.to_csv(file_csv_all)
df_all = pd.read_csv(file_csv_all)
```

Приложение Б