**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:**

к.э.н., к.ф.-м.н., к.ю.н., 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. ТЕСТИРОВАНИЕ АЛГОРИТМОВ МАШИННОГО И ГЛУБОКОГО ОБУЧЕНИЯ НА ДАННЫХ ПО КРЕДИТНОМУ СКОРИНГУ

В предыдущей главе было выяснено, что алгоритмы градиентного бустинга и нейронных сетей не имеют эффектов переобучения и недообучения. В случае градиентного бустинга используются гиперпараметры по умолчанию, результаты метрик которых не ниже метрик других алгоритмов. Это означает, что не требуется поиск гиперпараметров для данного алгоритма и можно остановиться на гиперпараметрах по умолчанию. Идентичная ситуация возникает у нейронной сети с отличием в том, что у нее метрики точности немного хуже, чем у алгоритма градиентного бустинга. Поэтому в этой главе будут подобраны гиперпараметры для алгоритмов деревьев решений и случайного леса, поскольку они обладают эффектом переобучения. Основным инструментом для поиска оптимальных гиперпараметров используются инструменты *GridSearchCV*[1] и *RandomizedSearchCV*[2]. Также в конце главы будут проанализированы прибыль и убыток банка на основании предсказаний четырех рассмотренных алгоритмов.

## 1.1 Подбор гиперпараметров для алгоритмов машинного обучения при оценке возврата кредита

Гиперпараметры – это такие параметры модели, которые задаются до начала работы модели. Правильно подобранные параметры позволяют улучшить качество предсказаний моделей. Каждому алгоритму МО присуще собственные наборы гиперпараметров, которые будут рассматриваться ниже. Существуют два метода подбора гиперпараметров:

1. *GridSearchCV* – метод подбора оптимальных гиперпараметров для модели с помощью перебора всех возможных комбинаций из заданного набора, что является ресурсозатратным способом.

2. *RandomizedSearchCV* – метод, позволяющий выбирать количество случайных комбинаций из заданного набора гиперпараметров. Является наиболее эффективным методом по времени при работе с большими данными.

[1] GridSearchCV [Электронный ресурс] / Python-библиотека для машинного обучения. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (дата обращения: 25.11.2025).

[2] RandomizedSearchCV [Электронный ресурс] / Python-библиотека для машинного обучения. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html (дата обращения: 25.11.2025).

### 1.1.1 Деревья решений

Для алгоритма классификации решающих деревьев (*Decision Tree Classifier*) основными гиперпараметрами являются:

1. *criterion* – критерий качества разбиения, определяющая показатель, по которому алгоритм лучше классифицирует клиентов.
   a) *gini* – индекс Джини[3], измеряющий неоднородность двух классов (0 и 1) внутри узла. Если индекс равен 0, значит в узле содержатся объекты одного класса. Чем индекс ближе к 0, тем лучше алгоритм определяет класс клиента. Индекс Джини в алгоритме *Decision Tree* используется по умолчанию, поскольку быстрее вычисляется;
   b) *entropy* – энтропия Шеннона[4], измеряющий степень смешанности классов. Высокая энтропия свидетельствует о высокой неоднородности классов, чем ниже энтропия, тем узел более однородный. Алгоритм выбирает разбиения с наиболее низкой энтропией, т.е. содержатся большинство объектов одного класса;
   c) *log_loss* – логарифмическая функция потерь. В данном случае алгоритм не просто выбирает класс, а оценивает их вероятности. За каждый неверный прогноз вводит штраф.
2. *splitter* – параметр, задающий способ разбиения каждого узла дерева решений.
   a) *best* – алгоритм последовательно перебирает все возможные признаки и пороги, которые лучше всего разделяют клиентов по целевой переменной. При данном параметре требуется больше времени на обучение.
   b) *random* – в каждом узле дерева рассматривается случайная часть разбиения, среди которых выбирается лучшее значение. За счет этого модель быстрее обучается, но качество разбиений может быть немного хуже, чем при *best*.
3. *max_depth* – максимальная глубина дерева решений. Чем больше глубина, тем длиннее цепочка вопросов, чтобы выдать прогноз по клиенту. При небольших значениях алгоритм делает мало разбиений и недообучается, т.е. плохо объясняет данные и дает невысокую точность на обучающей и тестовой выборках. При высоких значениях алгоритм, наоборот, переобучается, запоминает все особенности обучающей выборки, но плохо классифицирует клиентов на тестовых данных.
4. *min_samples_split* – параметр задает минимальное число наблюдений в узле, при котором возможно следующее разбиение на два дочерних. Если в узле наблюде-

[3]Джини К. (1884-1965 гг.) – итальянский статистик, экономист, социолог и демограф.
[4]Шеннон К.Э. (1916-2001 гг.) – американский инженер, криптоаналитик и математик, заложил основы теории информации.

ний меньше заданного порога, то дальнейшее разбиение не выполняется, и узел становится конечным.

5. $min\_samples\_leaf$ – параметр с минимальным заданным количеством наблюдений после разбиения в каждом узле, что позволяет дереву не создавать слишком маленькие группы клиентов и служит инструментов против переобучения алгоритма.

6. $max\_features$ – параметр задает случайное количество признаков, среди которых алгоритм ищет оптимальное разделение.

   a) $sqrt$ – в каждой вершине выбирается количество признаков, равное квадратному корню из общего числа признаков.

   b) $log2$ – число признаков, равное двоичному логарифму от их общего количества.

7. $random\_state = 42$ – произвольное число[5], задаваемое программе, чтобы при каждом запуске программы конечный результат оставался неизменным.

В таблице указаны найденные гиперпараметры алгоритма деревья решений двумя методами $GridSearchCV$ и $RandomizedSearchCV$.

Таблица 1 – Сравнение гиперпараметров по двум методам

| Гиперпараметр \ Метод | GridSearchCV | RandomizedSearchCV |
|---|---|---|
| criterion | entropy | log_loss |
| max_depth | 15 | 20 |
| max_features | `sqrt` | `sqrt` |
| min_samples_leaf | 25 | 20 |
| min_samples_split | 10 | 25 |
| random_state | 42 | 42 |

### 1.1.2 Случайный лес

### 1.1.3 Градиентный бустинг

### 1.1.4 Нейронная сеть

## 1.2 Ограничения и возможности моделей

[5]Почему именно «42» - широко известное и запоминающееся значение по умолчанию среди исследователей. Это ответ на главный вопрос жизни, Вселенной и всего сущего в романе Дугласа Адамса «Автостопом по Галактике».

# СПИСОК ЛИТЕРАТУРЫ


1. Соревнование на данных кредитных историй [Электронный ресурс] / Open Data Science. — URL: https://ods.ai/competitions/dl-fintech-bki. — Загл. с экрана.
2. PCA [Электронный ресурс] / Python-библиотека для машинного обучения. — URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html. — Загл. с экрана.
3. Смирнов *С.В.* Методы машинного обучения в макроэкономическом прогнозировании: предварительные итоги // Вопросы экономики. — 2025. — С. 131—154. — URL: https://doi.org/10.32609/0042-8736-2025-10-131-154. — Загл. с экрана.
4. Можно ли «отбелить» кредитную историю и есть ли черный список должников в Казахстане [Электронный ресурс] / Информационный интернет-портал. — URL: https://tengrinews.kz/kazakhstan_news/li-otbelit-kreditnuyu-istoriyu-est-chernyiy-spisok-doljnikov-560869/. — Загл. с экрана.
5. Wang *H.* Application of Decision Tree Model in Personal Credit Scoring and Its Fairness Optimization // Advances in Economics, Management and Political Sciences. — 2025. — С. 109—118. — URL: https://www.researchgate.net/publication/390979182_Application_of_Decision_Tree_Model_in_Personal_Credit_Scoring_and_Its_Fairness_Optimization. — Загл. с экрана.
6. GridSearchCV [Электронный ресурс] / Python-библиотека для машинного обучения. — URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. — Загл. с экрана.
7. RandomizedSearchCV [Электронный ресурс] / Python-библиотека для машинного обучения. — URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html. — Загл. с экрана.