|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_\_РАДИОТЕХНИЧЕСКИЙ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_\_\_\_\_\_\_\_\_\_\_Разработка и оценка моделей\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_машинного обучения\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_РТ5-61Б\_\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_ И.С. Викулин\_\_\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** **\_\_\_\_\_****Ю.Е. Гапанюк\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2024 г.***Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_ИУ5\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_В.И. Терехов\_\_

(И.О.Фамилия)

«\_07\_» \_\_\_\_февраля\_\_\_\_\_\_\_ 2024 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_Разработка и оценка моделей машинного обучения\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_РТ5-61Б\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Викулин Иван Сергеевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ИССЛЕДОВАТЕЛЬСКАЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_КАФЕДРА\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к \_\_\_\_ нед., 50% к \_\_\_\_ нед., 75% к \_\_\_ нед., 100% к \_\_\_\_ нед.

***Техническое задание*** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на \_\_\_\_\_ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «\_07\_» \_\_\_\_февраля\_\_\_\_\_\_\_ 2024 г.

**Руководитель НИР** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_Ю.Е. Гапанюк\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_\_\_\_\_\_\_И. С. Викулин**\_\_**

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**СОДЕРЖАНИЕ**

[Введение 3](file:///C:\Users\Ivan\Downloads\Отчет%20по%20НИР%20(1).docx#_Toc166636917)

[Основная часть 5](file:///C:\Users\Ivan\Downloads\Отчет%20по%20НИР%20(1).docx#_Toc166636918)

[Заключение 14](file:///C:\Users\Ivan\Downloads\Отчет%20по%20НИР%20(1).docx#_Toc166636919)

**Введение**

В современном банковском секторе удержание клиентов является одной из ключевых задач. Привлечение новых клиентов обходится значительно дороже, чем удержание существующих. Поэтому предсказание оттока клиентов и своевременное принятие мер для его предотвращения может существенно снизить затраты и увеличить доходность банка. Более того, непредсказуемый отток клиентов – одна из главных проблем банков. Разработка модели для прогнозирования оттока клиентов требует анализа большого объема данных и выявления ключевых факторов, влияющих на решение клиента прекратить использование банковских услуг.

Целью данного проекта является создание модели машинного обучения, способной предсказывать вероятность оттока клиентов банка на основе их транзакционной активности и демографических характеристик. Предполагается, что с помощью такой модели банк сможет идентифицировать клиентов, находящихся в зоне риска, и применять соответствующие меры для их удержания.

Поставленные задачи:

1. Провести сбор и предварительную обработку данных.
2. Исследовать взаимосвязи между различными признаками и оттоком клиентов.
3. Построить и оценить базовые модели классификации.
4. Провести подбор гиперпараметров для улучшения качества моделей.
5. Сравнить качество моделей и выбрать наилучшую для дальнейшего использования.

**Основная часть**

Выбранный датасет содержит информацию о 10 000 клиентах банка, пользующихся кредитными картами. В нем указаны различные характеристики клиентов, такие как возраст, доход, семейное положение, лимит по кредитной карте и другие. Цель анализа — предсказать, какие клиенты собираются уйти (т.е. перестать пользоваться услугами банка), чтобы банк мог проактивно принять меры и удержать этих клиентов.

- CLIENTNUM — Номер клиента (уникальный идентификатор).

- Attrition\_Flag — Флаг оттока (показывает, ушел клиент или нет).

- Customer\_Age — Возраст клиента.

- Gender — Пол (Male — мужчина, Female — женщина).

- Dependent\_count — Количество иждивенцев.

- Education\_Level — Уровень образования.

  - Unknown — Неизвестно

  - Uneducated — Без образования

  -  High School — Средняя школа

  - College — Колледж

  - Graduate — Выпускник

  - Post-Graduate — Аспирантура

  - Doctorate — Докторантура

- Marital\_Status — Семейное положение.

  - Married — Женат/Замужем

  - Single — Не женат/Не замужем

  - Divorced — Разведен(а)

  - Unknown — Неизвестно

- Income\_Category — Категория дохода.

  - Less than $40K — Менее $40K

  - $40K - $60K — От $40K до $60K

  - $60K - $80K — От $60K до $80K

  - $80K - $120K — От $80K до $120K

  - $120K + — Более $120K

  - Unknown — Неизвестно

- Card\_Category — Категория кредитной карты.

  -  Blue — Синяя

  - Silver — Серебряная

  - Gold — Золотая

  - Platinum — Платиновая

- Months\_on\_book — Количество месяцев, в течение которых клиент пользуется услугами банка.

- Total\_Relationship\_Count — Общее количество продуктов банка, которыми пользуется клиент.

- Months\_Inactive\_12\_mon — Количество месяцев бездействия за последние 12 месяцев.

- Contacts\_Count\_12\_mon — Количество контактов с банком за последние 12 месяцев.

- Credit\_Limit — Лимит по кредитной карте.

- Total\_Revolving\_Bal — Общий оборотный баланс.

- Avg\_Open\_To\_Buy — Средний доступный кредит.

- Total\_Amt\_Chng\_Q4\_Q1 — Изменение общей суммы транзакций с четвертого квартала до первого квартала.

- Total\_Trans\_Amt — Общая сумма транзакций.

- Total\_Trans\_Ct — Общее количество транзакций.

- Total\_Ct\_Chng\_Q4\_Q1 — Изменение общего количества транзакций с четвертого квартала до первого квартала.

- Avg\_Utilization\_Ratio — Средний коэффициент использования кредита.

- Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1 — Результат классификации наивного байесовского классификатора (используя категорию карты, количество контактов за 12 месяцев, количество иждивенцев, уровень образования и количество месяцев бездействия за последние 12 месяцев).

- Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2 — Результат классификации наивного байесовского классификатора (предсказание оттока).

**Последовательность действий**

1. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.

Размер датасета – 10127 строк, 23 колонки,

Для решения задачи классификации стоит удалить 2 последние колонки, которые содержит предсказания наивного байесовского классификатора по оттоку клиентов и, по сути, является результатом модели, а не исходным признаком клиента. Итого, 21 колонка

Список колонок с типами данных:

CLIENTNUM int64

Attrition\_Flag object

Customer\_Age int64

Gender object

Dependent\_count int64

Education\_Level object

Marital\_Status object

Income\_Category object

Card\_Category object

Months\_on\_book int64

Total\_Relationship\_Count int64

Months\_Inactive\_12\_mon int64

Contacts\_Count\_12\_mon int64

Credit\_Limit float64

Total\_Revolving\_Bal int64

Avg\_Open\_To\_Buy float64

Total\_Amt\_Chng\_Q4\_Q1 float64

Total\_Trans\_Amt int64

Total\_Trans\_Ct int64

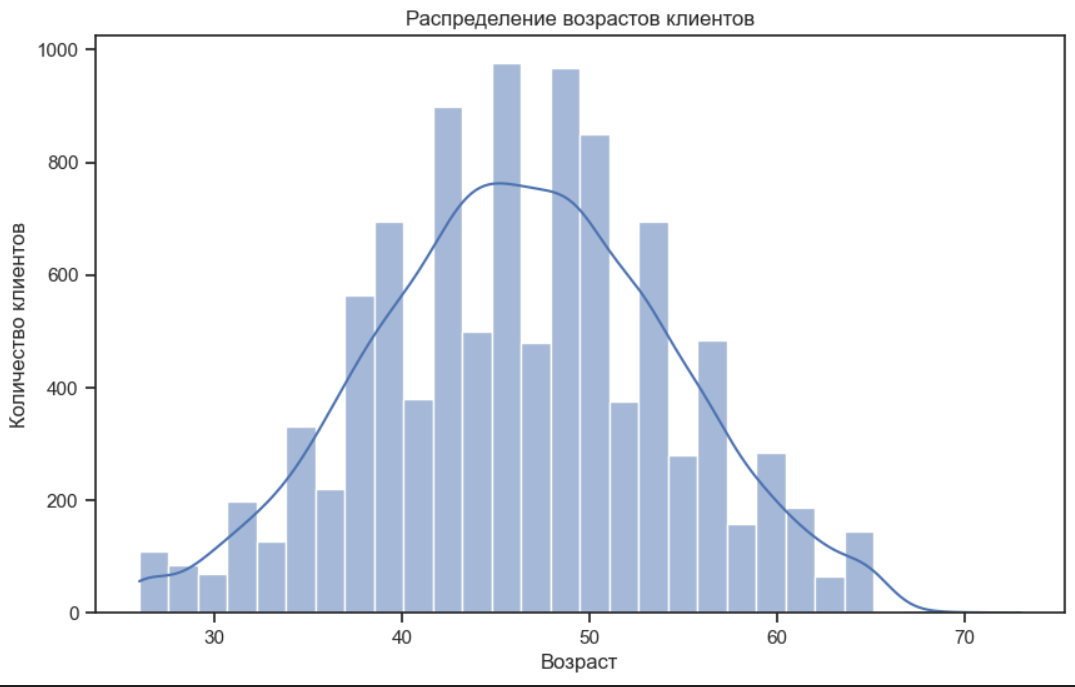
Total\_Ct\_Chng\_Q4\_Q1 float64

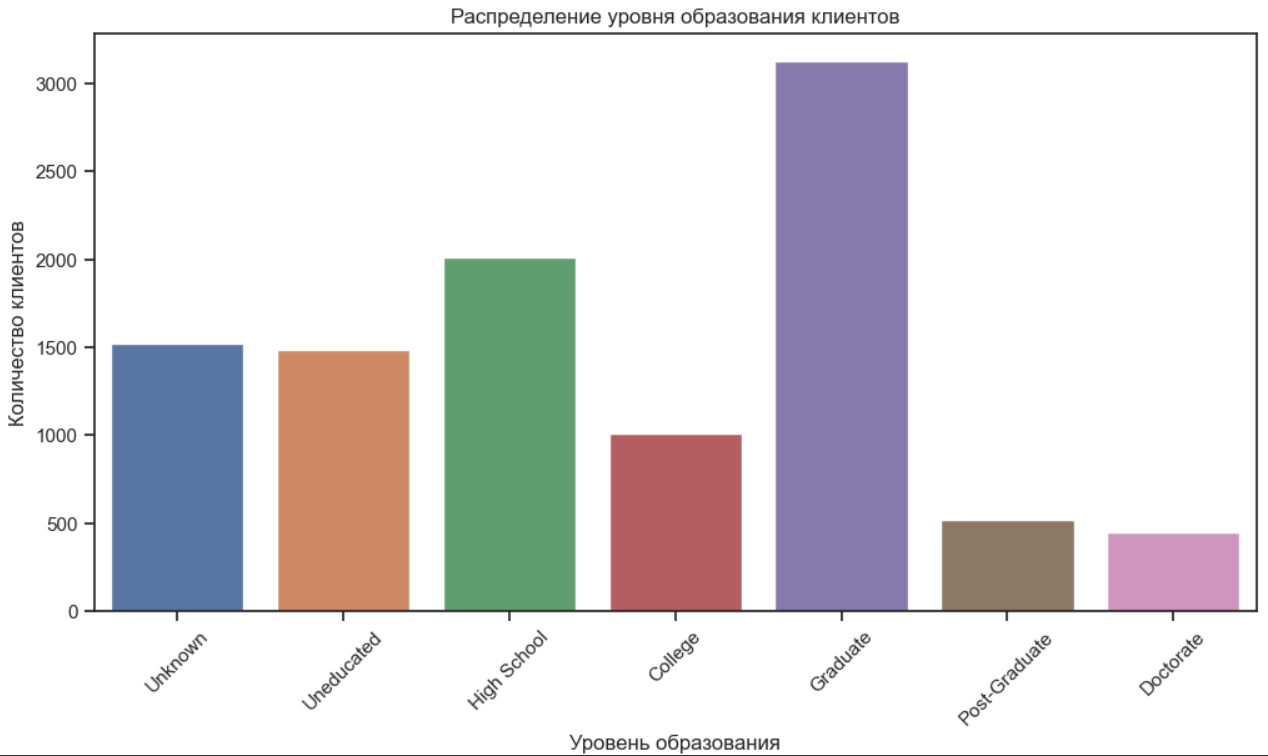
Avg\_Utilization\_Ratio float64

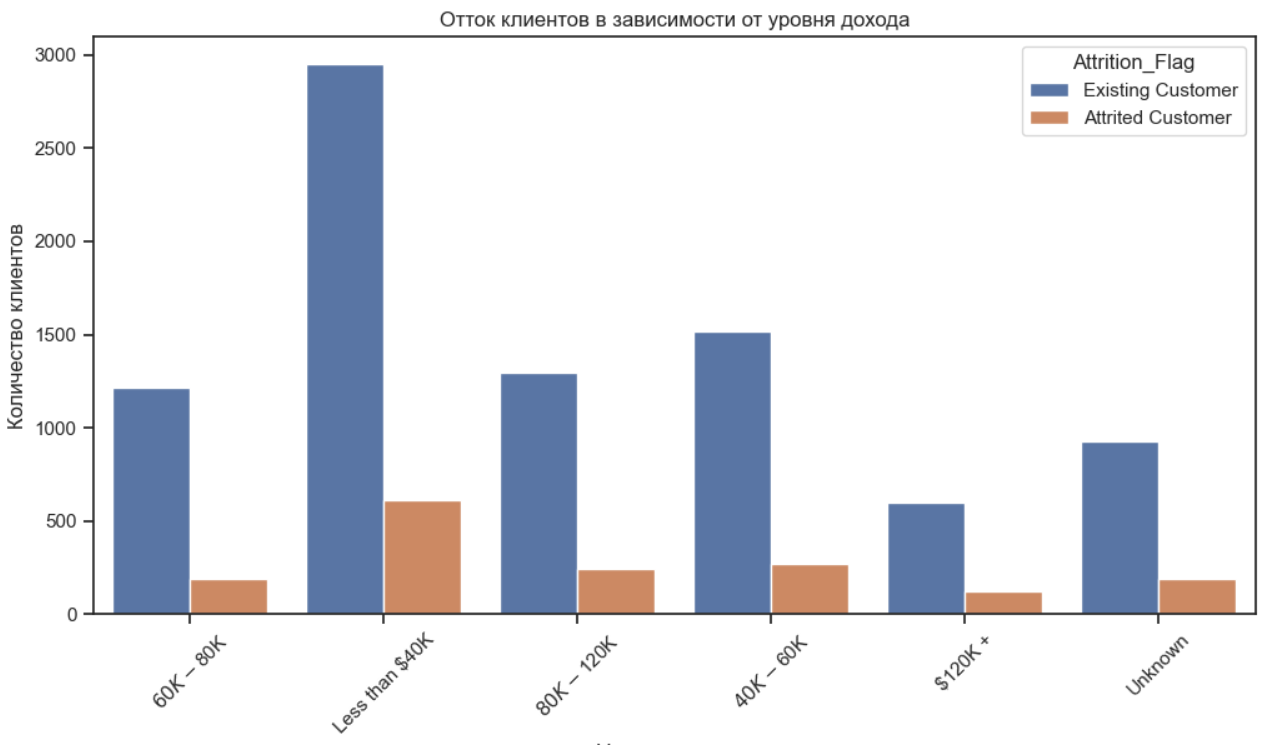
Пустых значений не обнаружено.

Есть категориальные признаки, которые нужно закодировать

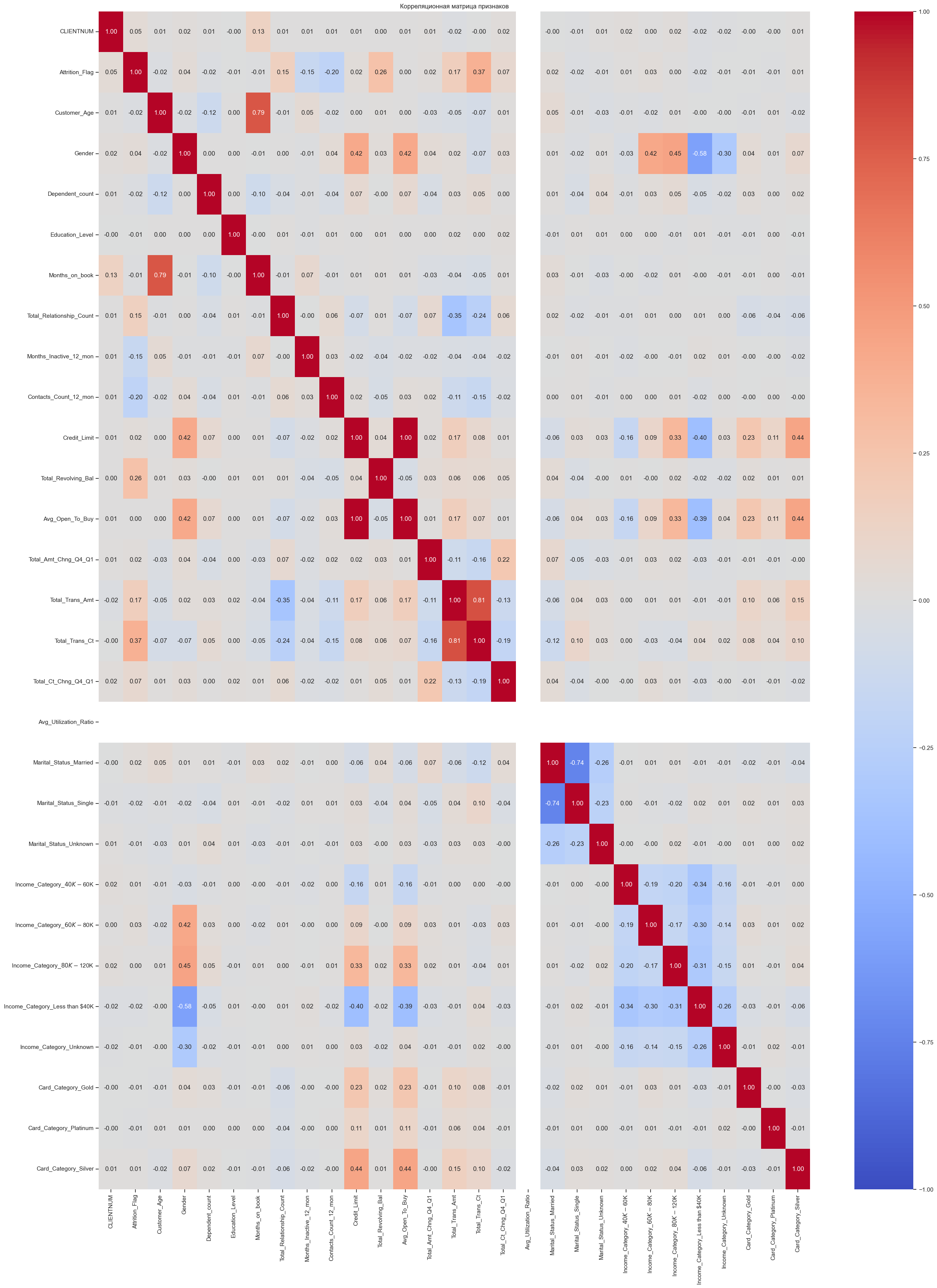
Графики, которые позволяют понять распределение и структуру данных:







1. Подготовка Данных: В процессе подготовки данных были выполнены следующие действия:
   * Масштабирование числовых признаков;
   * Кодирование категориальных признаков.
2. Вычисление и визуализация корреляционной матрицы



На основе корреляционной матрицы можно сделать следующие выводы:

* Флаг оттока (Attrition\_Flag)
  + Общее количество транзакций (Total\_Trans\_Ct): Наивысшая положительная корреляция (0.37), что указывает на то, что клиенты с большим количеством транзакций менее склонны к оттоку.
  + Изменение количества транзакций за квартал (Total\_Ct\_Chng\_Q4\_Q1): Сильная положительная корреляция (0.29), что означает, что клиенты с большим изменением количества транзакций между четвертым кварталом и первым менее склонны к оттоку.
  + Общий оборотный баланс (Total\_Revolving\_Bal): Положительная корреляция (0.26), что предполагает, что клиенты с большими оборотом менее склонны к оттоку.
  + Количество контактов за 12 месяцев (Contacts\_Count\_12\_mon): Отрицательная корреляция (-0.20), что означает, что большее количество контактов с банком может увеличить вероятность оттока.
  + Месяцы неактивности за 12 месяцев (Months\_Inactive\_12\_mon): Отрицательная корреляция (-0.15), указывающая на то, что клиенты, которые были неактивны в течение большего количества месяцев за последний год, более склонны к оттоку.
* Возраст клиента (Customer\_Age)
  + Количество месяцев на учете (Months\_on\_book): Очень сильная положительная корреляция (0.79), что указывает на то, что более старшие клиенты дольше находятся на учете в банке.
  + Изменение суммы за квартал (Total\_Amt\_Chng\_Q4\_Q1): Слабая отрицательная корреляция (-0.06), что указывает на небольшую тенденцию, что у более старших клиентов меньше изменений в суммах транзакций между четвертым и первым кварталом.
  + Общее количество транзакций (Total\_Trans\_Ct): Слабая отрицательная корреляция (-0.07), что предполагает, что у более старших клиентов может быть меньше транзакций.
* Общее количество транзакций (Total\_Trans\_Ct)
  + Изменение количества транзакций за квартал (Total\_Ct\_Chng\_Q4\_Q1): Сильная положительная корреляция (0.71), что указывает на то, что клиенты с большим количеством транзакций склонны к большему изменению количества транзакций между кварталами.
  + Общая сумма транзакций (Total\_Trans\_Amt): Сильная положительная корреляция (0.80), что означает, что клиенты с большим количеством транзакций также имеют большие общие суммы транзакций.
  + Общее количество отношений (Total\_Relationship\_Count): Умеренная отрицательная корреляция (-0.24), что указывает на то, что у клиентов с большим количеством транзакций может быть меньше общих отношений.
  + Среднее использование лимита (Avg\_Utilization\_Ratio): Умеренная отрицательная корреляция (-0.36), что предполагает, что у клиентов с большим количеством транзакций, как правило, ниже средний коэффициент использования лимита.
* Среднее использование лимита (Avg\_Utilization\_Ratio)
  + Кредитный лимит (Credit\_Limit): Сильная отрицательная корреляция (-0.71), что указывает на то, что более высокие кредитные лимиты связаны с более низким коэффициентом использования лимита.
  + Общий оборотный баланс (Total\_Revolving\_Bal): Сильная положительная корреляция (0.89), что предполагает, что более высокие оборотный балансы ведут к более высоким коэффициентам использования лимита.
  + Средняя доступная сумма (Avg\_Open\_To\_Buy): Сильная отрицательная корреляция (-0.76), что указывает на то, что большие доступные суммы связаны с более низким коэффициентом использования лимита.

1. Выбор метрик для последующей оценки качества моделей.

Было принято решение использовать следующие метрики: accuracy, precision и recall

1. Выбор наиболее подходящих моделей для решения задачи регрессии.

* Метод ближайших соседей
* Метод опорных векторов
* Дерево решений
* Случайный лес
* Градиентный бустинг

1. Формирование обучающей и тестовой выборок на основе исходного набора данных.
2. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров.

accuracy precision recall

KNN 0.833333 0.243243 0.022333

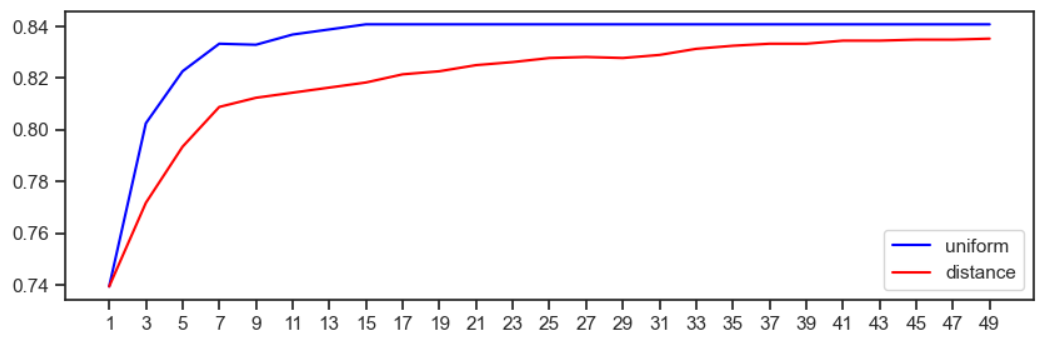
SVM 0.840837 0.000000 0.000000

Decision Tree 0.937204 0.793269 0.818859

Random Forest 0.954976 0.921283 0.784119

Gradient Boosting 0.956951 0.910615 0.808933

1. Подбор гиперпараметров для выбранных моделей. Анализ результатов



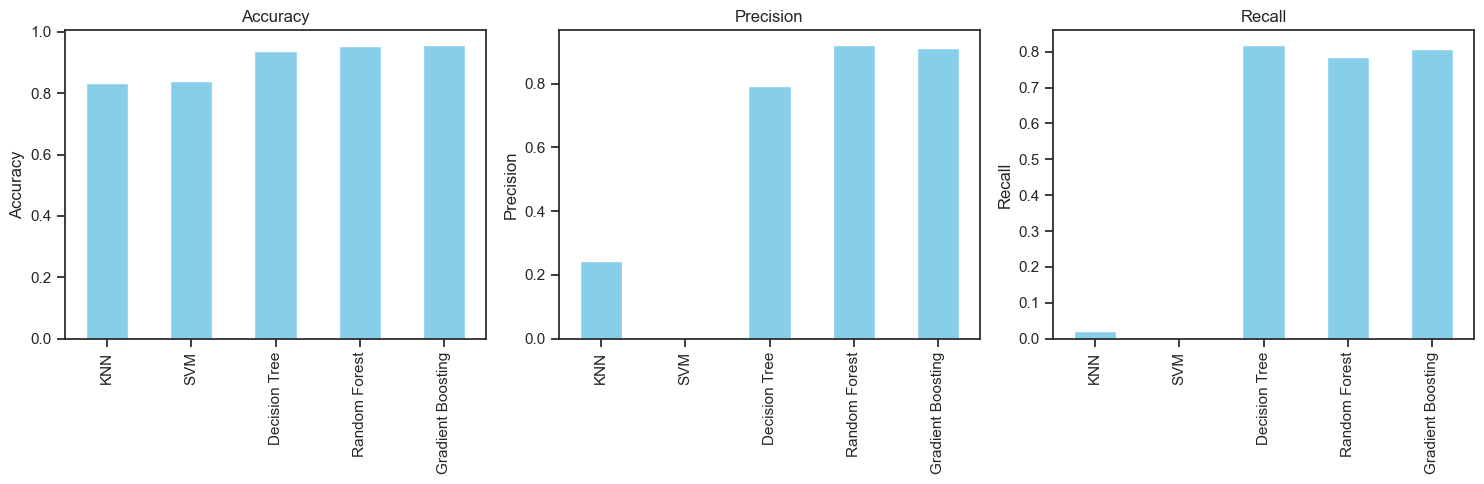
Для метода ближайших соседей значение n = 19 оказалось наилучшим.

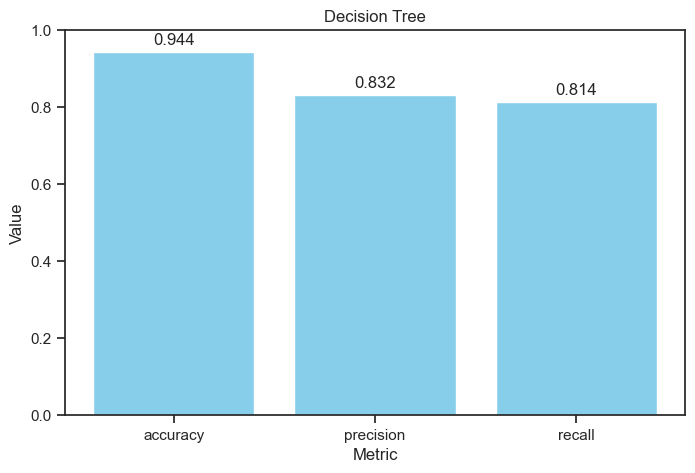
Для дерева решений наилучшая глубина - 7

1. Сравнение значений базового и с гиперпараметром

Для модели ближайших соседей точность (accuracy) стала равна 84%, а для дерева решений 94,44%. Также у дерева решений улучшилась метрики precision (точность) и recall (полнота), 83% и 81% соответственно.

1. Формулирование выводов.





Из полученных данных видно, что модели "Decision Tree", "Random Forest" и "Gradient Boosting" демонстрируют более высокие значения метрик accuracy, precision и recall по сравнению с моделями "KNN" и "SVM".

Модель "Decision Tree" показывает наилучшие результаты с точки зрения всех трех метрик: accuracy составляет 0.937204, precision - 0.793269, а recall - 0.818859. Эти результаты указывают на то, что модель "Decision Tree" обладает хорошей способностью к правильной классификации и обнаружению положительных случаев, а также способна минимизировать ложно-положительные результаты.

Модели "Random Forest" и "Gradient Boosting" также показывают высокую производительность, превосходя модели "KNN" и "SVM" по всем трем метрикам.

Модель "KNN" демонстрирует наихудшие результаты с точки зрения precision и recall, что может указывать на то, что она имеет трудности в правильном классифицировании положительных случаев и их обнаружении.

Модель "SVM" имеет нулевое значение precision и recall, что может свидетельствовать о том, что она не смогла корректно предсказать положительные случаи.

Таким образом, на основе этих данных можно сделать вывод, что модели "Decision Tree", "Random Forest" и "Gradient Boosting" наилучшим образом подходят для данной задачи классификации.

**Заключение**

В ходе проведенного исследования по прогнозированию оттока клиентов банка были выполнены следующие ключевые этапы:

1. Сбор и предварительная обработка данных:

Данные были загружены и предварительно обработаны, включая удаление ненужных столбцов, кодирование категориальных признаков и масштабирование числовых данных.

Датасет был разделен на обучающую и тестовую выборки для построения и оценки моделей.

1. Анализ взаимосвязей между признаками:

Были выявлены важные зависимости между транзакционными и демографическими характеристиками клиентов и их склонностью к оттоку.

Наиболее значимыми факторами, влияющими на отток клиентов, оказались количество транзакций, количество контактов с банком за последний год и месяцы неактивности.

1. Построение и оценка базовых моделей:

Были построены и оценены базовые модели, такие как метод k-ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг.

Метрики accuracy, precision и recall были использованы для оценки качества моделей.

1. Подбор гиперпараметров:

Проведен подбор гиперпараметров с использованием метода GridSearchCV для улучшения качества моделей.

Были найдены оптимальные значения гиперпараметров для каждой модели, что позволило повысить точность прогнозирования.

1. Сравнение и выбор наилучшей модели:

Модели с подобранными гиперпараметрами были повторно оценены и сравнины с базовыми моделями.

Модель случайного леса показала наилучшие результаты по всем метрикам и была выбрана как наиболее подходящая для прогнозирования оттока клиентов.