|  |  |
| --- | --- |
|  | **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 \_\_\_\_\_\_\_\_Информатика и системы управления\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_\_\_\_\_\_\_Системы обработки информации и управления\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_\_\_\_\_\_\_\_\_\_предсказание покинет ли конкретный\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_клиент банк или нет \_\_\_\_\_\_\_\_\_\_\_\_\_***

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

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

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

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

Студент \_\_\_ИУ5-64Б\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_Гордеев Н.А.\_\_\_\_\_**

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

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Гапанюк Ю.Е.\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_ Гапанюк Ю.Е.\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2023 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 \_\_\_\_\_\_\_\_\_\_

(Индекс)

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

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_ предсказание покинет ли конкретный клиент банк или нет \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

Студент группы \_\_ИУ5-64Б\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Гордеев Никита Андреевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Исследоваткльская\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

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

\_\_\_\_\_Исследовать методы машинного обучения для решения задачи классификации\_\_\_\_\_\_

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

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

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

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

Расчетно-пояснительная записка на \_\_28\_\_ листах формата А4.

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

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

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

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

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

Дата выдачи задания « 13 » февраля 2023 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Гапанюк Ю.Е.\_\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Гордеев Н.А.\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

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

**Содержание**

[**Введение** 4](#_Toc133806658)

[**Постановка задачи** 6](#_Toc133806659)

[**Выполнение работы** 7](#_Toc133806660)

[**Создание веб-приложения** 29](#_Toc133806661)

[**Заключение** 31](#_Toc133806662)

[**Список использованной литературы** 32](#_Toc133806663)

# **Введение**

Проблема ухода клиентов из банков довольно актуальна в наши дни. Поэтому каждому банку выгодно не терять своих клиентов. Однако можно заранее узнать покинет ли банк тот или иной клиент.

В данной работе мы будем использовать данные, полученные от банков, чтобы построить модель машинного обучения, которая сможет предсказывать уход клиента из банка. Мы будем использовать алгоритмы классификации для определения факторов риска, включая возраст, пол, уровень заработка и страны проживания.

Целью данной работы является разработка эффективной модели, которая может помочь банкам понять собирается ли клиент уходить из банка или нет и при необходимости применить меры.

Для достижения поставленной цели были определены следующие этапы:

1. Поиск и выбор набора данных для построения моделей машинного обучения для решения задачи регрессии или классификации.
2. Проведение разведочного анализа данных.
3. Выбор признаков, подходящих для построения моделей.
4. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
5. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.
6. Выбор метрик для последующей оценки качества моделей.
7. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии.
8. Формирование обучающей и тестовой выборок на основе исходного набора данных.
9. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров и оценка качества моделей на основе тестовой выборки.
10. Подбор гиперпараметров для выбранных моделей. Построение оптимальных моделей.
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик.
12. Создание веб-приложение для демонстрации хотя бы одной модели машинного обучения.

# **Постановка задачи**

Данная работа по машинному обучению направлена на решение задачи классификации, а именно, предсказание ухода клиента из банка.

Имеются данные о клиентах, которые включают информацию о таких факторах, как кредитный рейтинг клиента, место нахождения заказчика, пол: мужской или женский, возраст заказчика, с какого года клиент находится в банке, средний остаток клиента, количество банковских продуктов, которыми пользуется клиент, имеет ли клиент кредитную карту, является ли клиент активным клиентом и предположительный заработок клиента. каждый клиент может быть классифицирован как покинет ли этот клиент банк или нет.

Целью задачи является создание модели машинного обучения, которая будет использовать имеющиеся данные для предсказания покинет ли этот клиент банк или нет. Для этого мы будем использовать различные алгоритмы классификации, такие как K ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг. Модель должна обучаться на тренировочных данных и проверяться на тестовых данных для оценки ее точности и эффективности.

Результатом работы должна быть модель, которая сможет помочь банкам понять собирается ли клиент уходить из банка или нет и при необходимости применить меры.

# **Выполнение работы**

Для решения задачи классификации был выбран набор данных содержащий информацию о клиентах, которые покинут этот банк.

В наборе данных присутствуют следующие столбцы:

1. RowNumber - Номер строки набора данны
2. CustomerId - Указан идентификатор клиента
3. Surname - Фамилия заказчика
4. CreditScore - Кредитный рейтинг клиента
5. Geography - Место нахождения заказчика
6. Gender - Пол: мужской или женский
7. Age - Возраст заказчика
8. Tenure - С какого года клиент находится в банке
9. Balance - Средний остаток клиента
10. NumOfProducts - Количество банковских продуктов, которыми пользуется клиент
11. HasCrCard - Имеет ли клиент кредитную карту
12. IsActiveMember - Является ли клиент активным клиентом
13. EstimatedSalary - Предположительный заработок клиента
14. Exited - Покинет ли этот клиент банк или нет

Данный датасет использован для решения задачи классификации - предсказания покинет ли этот клиент банк или нет.

Загружаем данные, получаем обую информацию о датасете и делаем предположения о влиянии признаков на целевую переменную. В наборе данных содержится 10000 строк и 11 столбцов, из которых 7 типа int64 и 2 типа float64.

Меняем тип колонок Geography и Gender на ‘category’, так как они очевидно содержит категориальные данные и может внести несуществующие закономерности в данные.

Пропусков не обнаружено.

Дубликатов не обнаружено.

Строим график pairplot для визуализации распределения данных попарно для множества колонок.

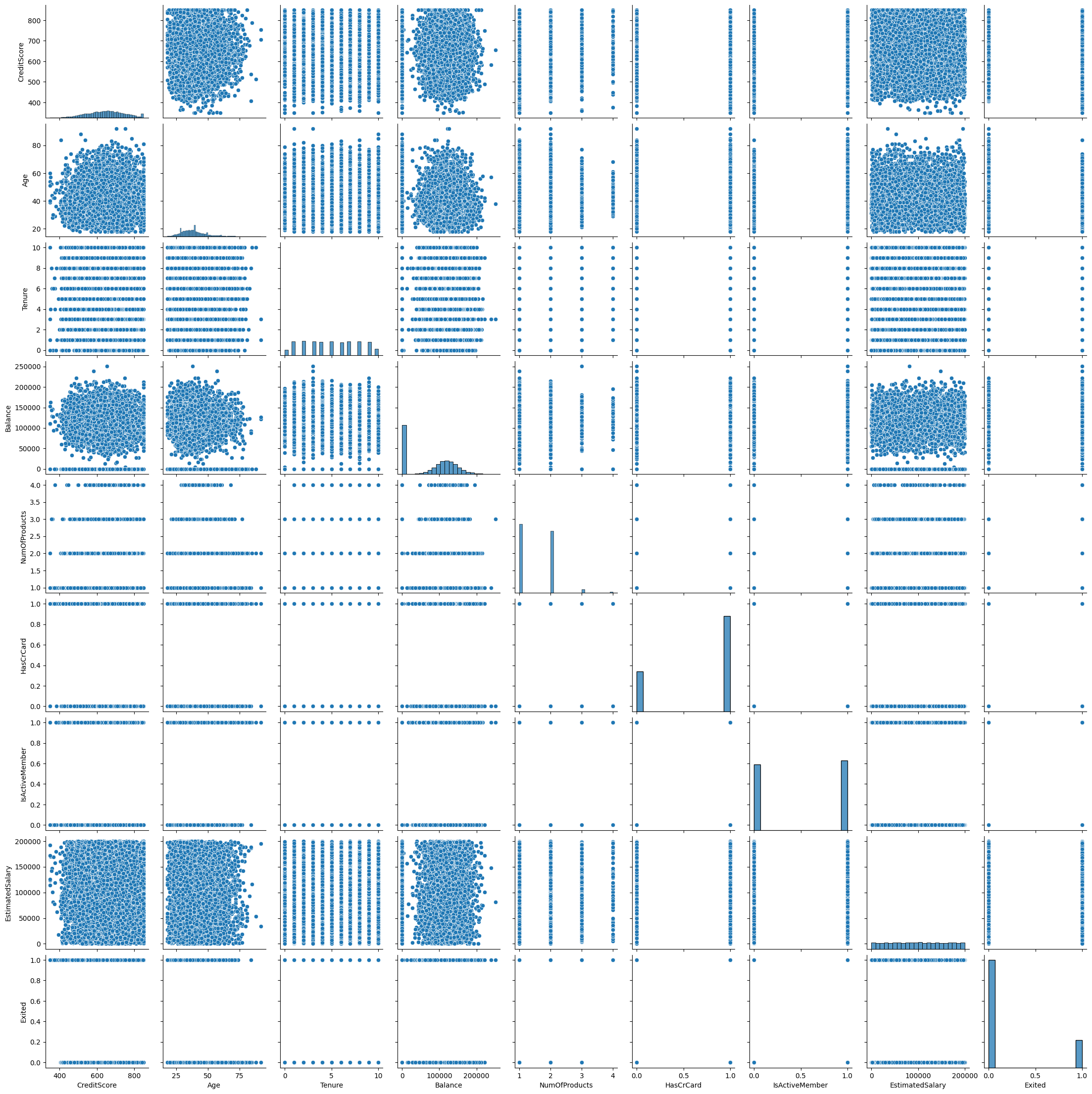


Рисунок 1 - Визуализация распределения данных попарно для множества колонок

Проверяем сбалансированы ли классы в нашем наборе данных. Получаем следующую гистограмму:

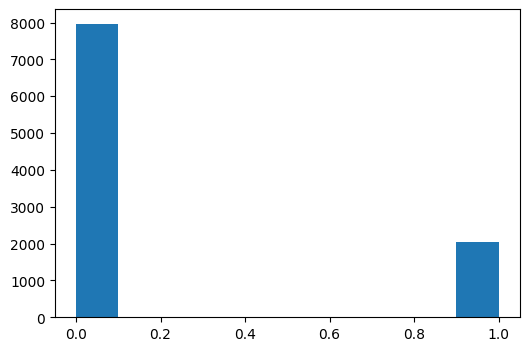


Рисунок 2 - Гистограмма классов

Видим, что классы не сбалансированы.

Строим таблицу средних значений с группировкой по целевому признаку и делаем следующие предположения:

* у тех, кто покинет банк возраст выше
* у тех, кто покинет банк баланс на счете больше
* у тех, кто покинет банк активность ниже

Посмотрим влияет ли возраст клиента на целевой признак. Строим диаграмму, а также воспользуемся t-тестом, чтобы удостовериться что распределение не случайно. Получаем следующие значения:

t-statistic: -29.76681499437077, p-value: 1.2399313093427738e-186

P-значение называется вероятностью того, что результаты выборки данных произошли случайно. [1] Поэтому делаем вывод, что возраст клиента влияет на целевой признак.

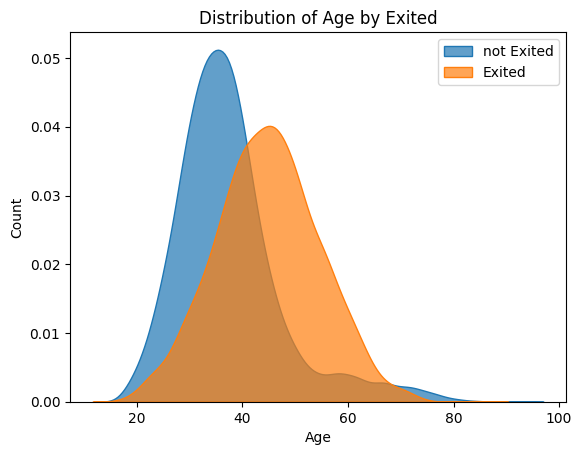


Рисунок 3 - График плотности возраста и целевого признака

Можно заметить, что те, кто покинут банк - старше, чем те, кто останется.

Построим график плотности баланса на счету от целевого признака и проверим t-статистику. Получаем следующие значения:

t-statistic: -11.936240300013841, p-value: 1.2399313093427738e-186

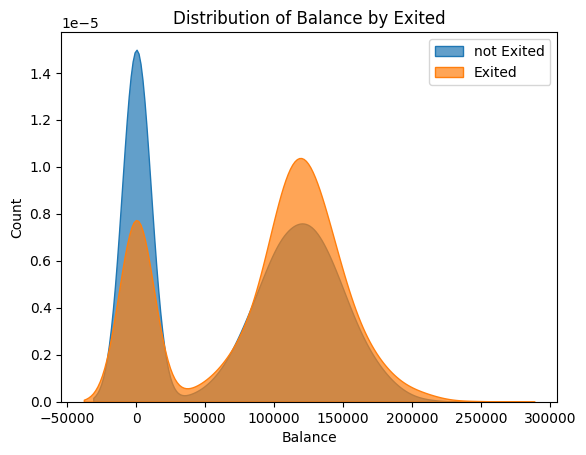


Рисунок 4 - график плотности баланса на счету от целевого признака

Можно заметить, что у тех, кто покинет банк средний баланс выше.

Смотрим на соотношение флага активного пользователя и целевого признака.

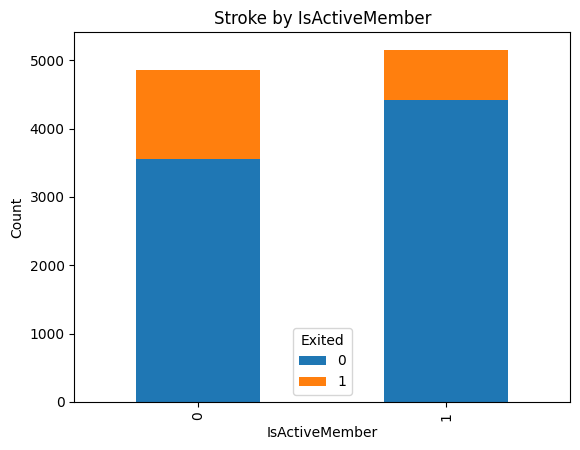


Рисунок 5 - Гистограмма зависимости флага активного пользователя и целевого признака

Можно заметить, что те, кто покинут банк чаще являются неактивными пользователям.

Далее приведем данные к нужному формату. Сначала масштабируем численные признаки методом Standard Scaler, который преобразует каждый признак таким образом, чтобы он имел среднее значение равное 0 и стандартное отклонение равное 1. Посмотрим на распределения колонок до и после масштабирования.

Таблица 1 - Распределение численных колонок до и после масштабирования

|  |  |
| --- | --- |
| До масштабирования | После масштабирования |
| Рисунок 6 - Распределение кредитного рейтинга до масштабирования | Рисунок 7 - Распределение кредитного рейтинга после масштабирования |
| Рисунок 8 - Распределение среднего остатка клиента до масштабирования | Рисунок 9 - Распределение среднего остатка клиента после масштабирования |
| Рисунок 10 - Распределение предположительного заработка клиента до масштабирования | Рисунок 11 – Распределение предположительного заработка клиента после масштабирования |

Распределение не изменилось.

Затем используем One Hot encoding для кодирования колонок 'Geography', 'Gender'. В этом случае каждое уникальное значение признака становится новым отдельным признаком. [4]

Проводим корреляционный анализ данных. Строим тепловую карту корреляций.

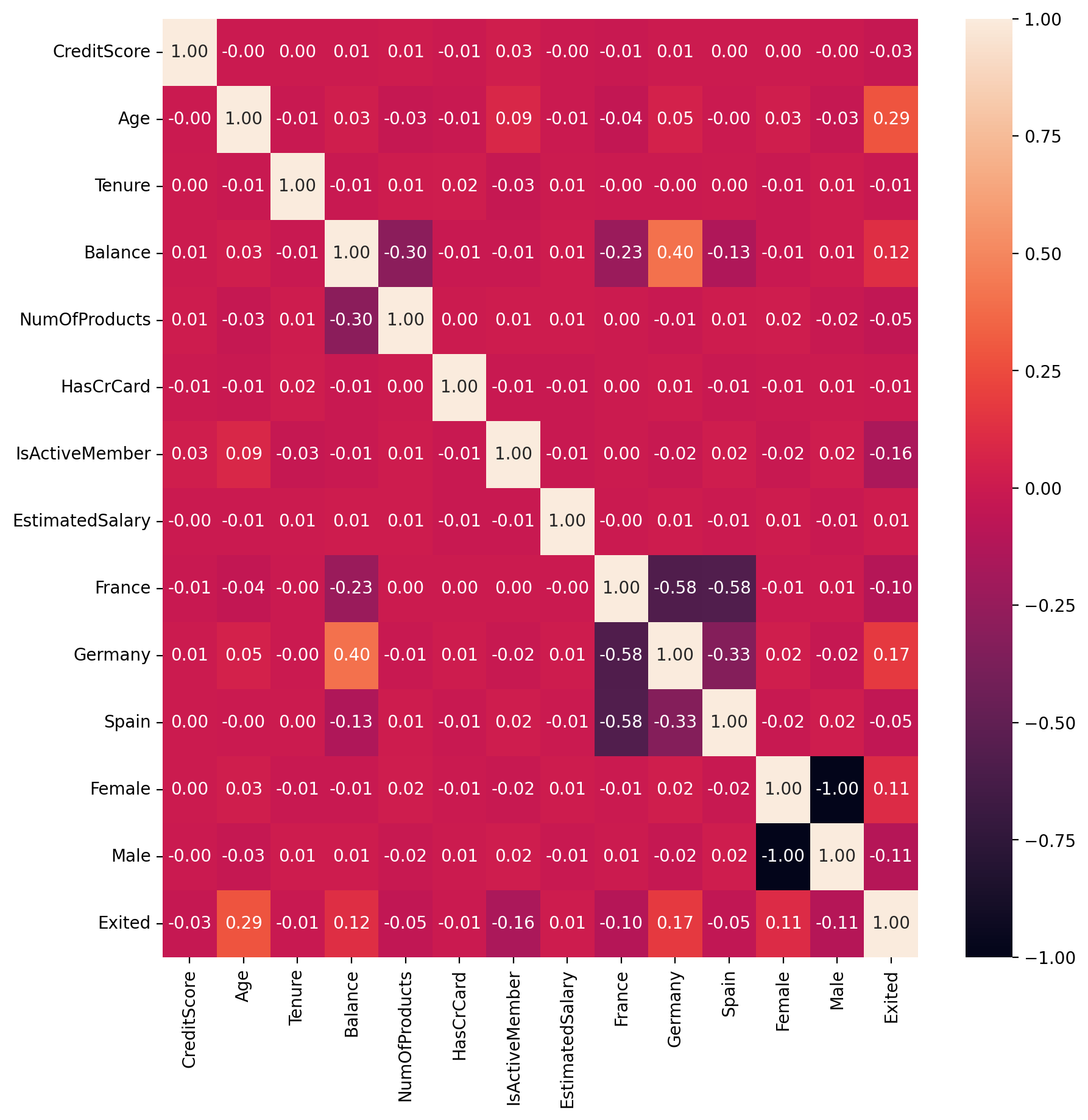


Рисунок 16 - Тепловая карта корреляций

Выводы:

* целевой признак Exited больше всего коррелирует с возрастом (0.29), проживанием в Германии (0.17) и активным клиентом (-0.16).
* столбец со средним балансом сильно коррелирует с проживанием в Германии. Оставим его.
* слабокоррелирующие столбцы France и Male также оставим. Female не берем так как это значение является полным противоположностью Male
* Все остальные столбцы можно не учитывать для упрощения обучения

Выберем метрики для оценки качества модели:

* - показывает, какую долю объектов, которые модель предсказала как положительные, действительно являются положительными.
* - показывает, какую долю положительных объектов модель способна обнаружить.
* - среднее гармоническое precision и recall. Другими словами, это средневзвешенное значение точности и отзыва. [2]
* - oснована на вычислении следующих характеристик: - True Positive Rate, откладывается по оси ординат. Совпадает с recall. - False Positive Rate, откладывается по оси абсцисс. Показывает какую долю из объектов отрицательного класса алгоритм предсказал неверно. Идеальная ROC-кривая проходит через точки (0,0)-(0,1)-(1,1), то есть через верхний левый угол графика. Чем сильнее отклоняется кривая от верхнего левого угла графика, тем хуже качество классификации. [3]

Выберем модели для решения задачи классификации:

* KNN;
* SVC;
* Дерево решений;
* Случайный лес;
* Градиентный бустинг.

Формируем обучающую и тестовую выборку в соотношении 8:2. Удаляем колонки 'CreditScore', 'Tenure', 'NumOfProducts', 'HasCrCard', 'EstimatedSalary','Spain' и 'Female', т.к. они почти не влияют на целевой признак.

Строим базовое решения, выводим значениями метрик и ROC-кривую.

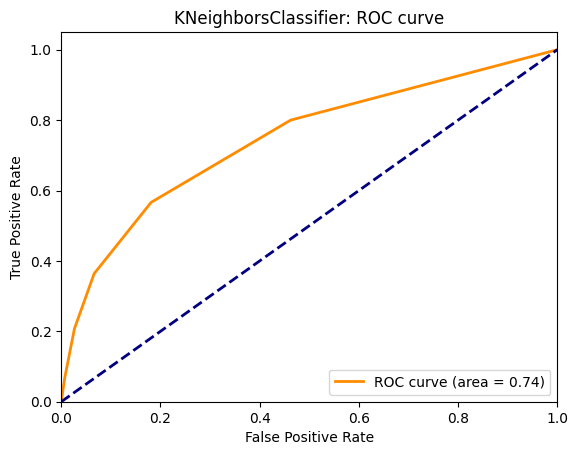


Рисунок 17 - ROC-кривая базовой модели KNN

KNeighborsClassifier:

Precision: 0.59

Recall: 0.36

F1-score: 0.45

ROC AUC score: 0.7436821101440463

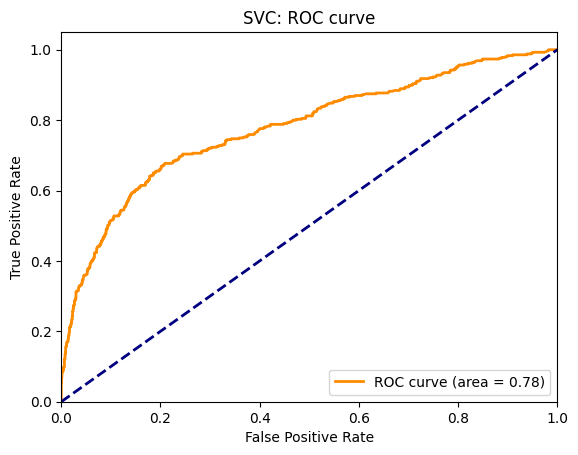


Рисунок 18- ROC-кривая базовой модели SVC

SVC:

Precision: 0.0

Recall: 0.0

F1-score: 0.0

ROC AUC score: 0.777870092356809

(Нулевые значения из-за того, что внутри функции обучения происходит ошибка деления на ноль и она автоматически возвращает значение 0)

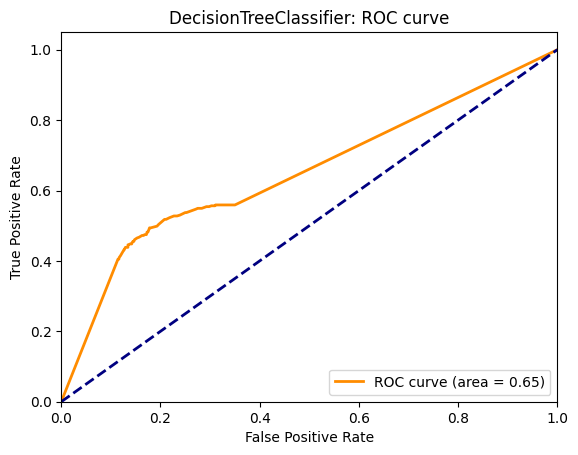


Рисунок 19 - ROC-кривая базовой модели Decision Tree

DecisionTreeClassifier:

Precision: 0.48

Recall: 0.41

F1-score: 0.44

ROC AUC score: 0.6511337463418342

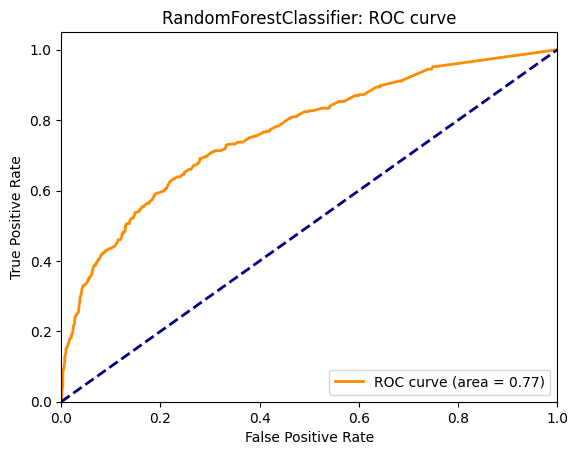


Рисунок 20 - ROC-кривая базовой модели Random Forest

RandomForestClassifier:

Precision: 0.57

Recall: 0.42

F1-score: 0.49

ROC AUC score: 0.7658743491315417

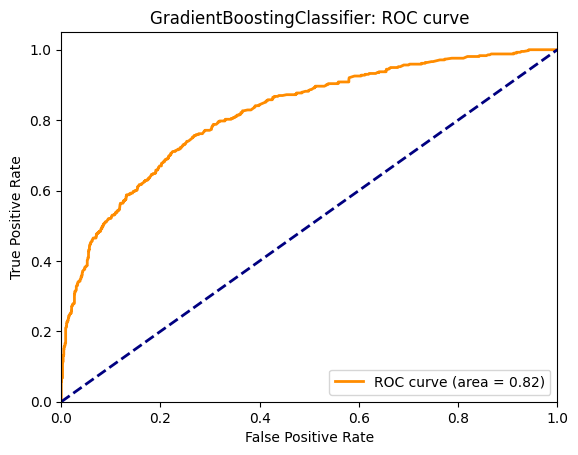


Рисунок 21 - ROC-кривая базовой модели Gradient Boosting

GradientBoostingClassifier:

Precision: 0.71

Recall: 0.34

F1-score: 0.46

ROC AUC score: 0.8178921363688192

Используем GridSearch для поиска оптимальных гиперпараметров для каждой модели.

KNeighboursClassifier:

Best hyperparameters: {'algorithm': 'auto', 'n\_neighbors': 10, 'weights': 'uniform'}

SVC:

Best hyperparameters: {'C': 1, 'degree': 4, 'gamma': 'auto', 'kernel': 'rbf'}

DecisionTreeClassifier:

Best hyperparameters: {'criterion': 'entropy', 'max\_depth': 5, 'max\_features': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2}

RandomForestClassifier:

Best hyperparameters: {'max\_depth': 5, 'max\_features': 'sqrt', 'min\_samples\_leaf': 1, 'min\_samples\_split': 5, 'n\_estimators': 100}

GradientBoostingClassifier:

Best hyperparameters: {'learning\_rate': 0.05, 'max\_depth': 5, 'max\_features': 'sqrt', 'min\_samples\_leaf': 4, 'min\_samples\_split': 10}

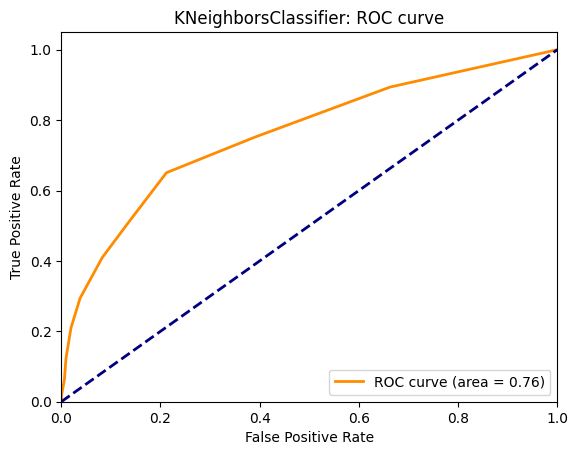


Рисунок 22 - ROC-кривая модели KNN после поиска гиперпараметров

KNeighborsClassifier:

Precision: 0.67

Recall: 0.29

F1-score: 0.41

ROC AUC score: 0.7597514347611265

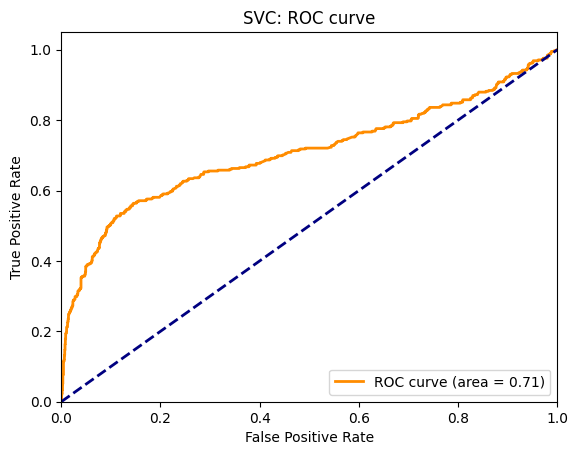


Рисунок 23 - ROC-кривая модели SVC после поиска гиперпараметров

SVC:

Precision: 0.75

Recall: 0.29

F1-score: 0.42

ROC AUC score: 0.7076576336893314

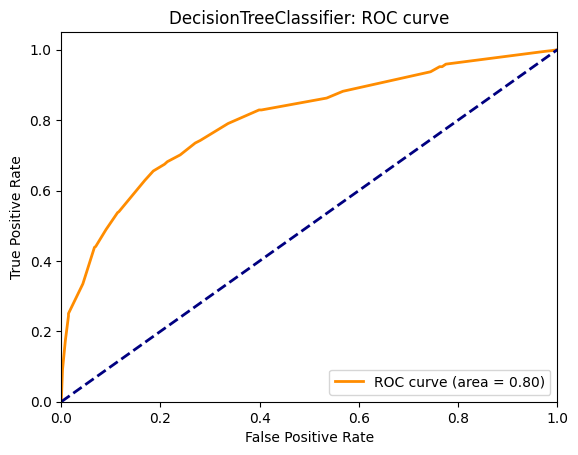


Рисунок 24 - ROC-кривая модели Decision Tree после поиска гиперпараметров

DecisionTreeClassifier:

Precision: 0.82

Recall: 0.25

F1-score: 0.38

ROC AUC score: 0.7988202652882825

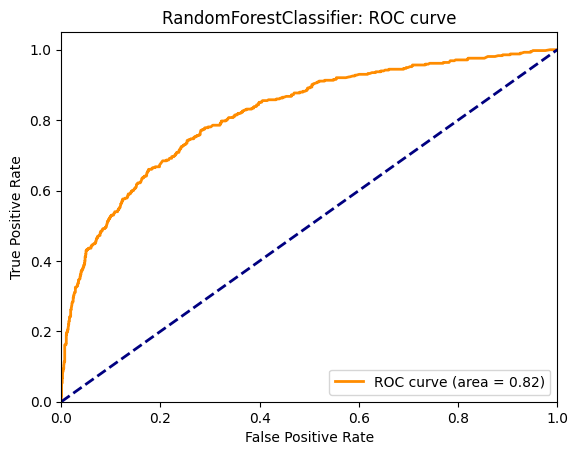


Рисунок 25 - ROC-кривая модели Random Forest после поиска гиперпараметров

RandomForestClassifier:

Precision: 0.76

Recall: 0.3

F1-score: 0.43

ROC AUC score: 0.8176253278096614

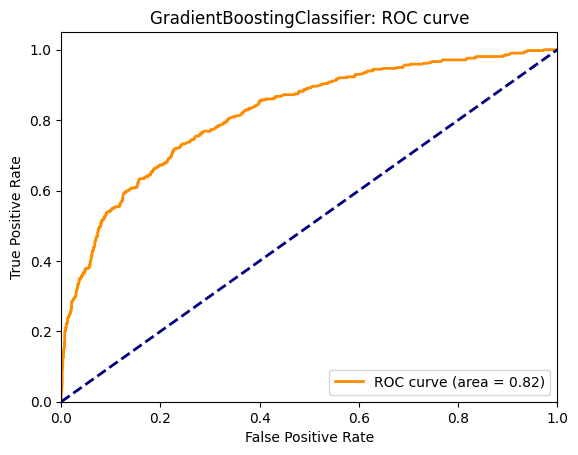


Рисунок 26 - ROC-кривая модели Gradient Boosting после поиска гиперпараметров

GradientBoostingClassifier:

Precision: 0.71

Recall: 0.34

F1-score: 0.46

ROC AUC score: 0.8189388468701303

Таблица 2 - Сравнение базовых моделей с моделями после подбора гиперпараметров по 4 метрикам

|  |  |  |
| --- | --- | --- |
| Модель | Baseline | GridSearch() |
| KNN | Precision: 0.59  Recall: 0.36  F1-score: 0.45  ROC AUC score: 0.7436821101440463 | Precision: 0.67  Recall: 0.29  F1-score: 0.41  ROC AUC score: 0.7597514347611265 |
| SVC | Precision: 0.0  Recall: 0.0  F1-score: 0.0  ROC AUC score: 0.777870092356809 | Precision: 0.75  Recall: 0.29  F1-score: 0.42  ROC AUC score: 0.7076576336893314 |
| Decision Tree | Precision: 0.48  Recall: 0.41  F1-score: 0.44  ROC AUC score: 0.6511337463418342 | Precision: 0.82  Recall: 0.25  F1-score: 0.38  ROC AUC score: 0.7988202652882825 |
| Random forest | Precision: 0.57  Recall: 0.42  F1-score: 0.49  ROC AUC score: 0.7658743491315417 | Precision: 0.76  Recall: 0.3  F1-score: 0.43  ROC AUC score: 0.8176253278096614 |
| Gradient Boosting | Precision: 0.71  Recall: 0.34  F1-score: 0.46  ROC AUC score: 0.8178921363688192 | Precision: 0.71  Recall: 0.34  F1-score: 0.46  ROC AUC score: 0.8189388468701303 |

На основании трех метрик из четырех лучшими для решения данной задачи классификации оказались модели градиентного бустинга и метод случайного леса.

# **Создание веб-приложения**

В качестве модели выбираем решающее дерево.

Сначала импортируем необходимые библиотеки, загружаем набор данных, обрабатываем данные: удаляем дубликаты, заполняем пропущенные значения, кодируем категориальные данные с помощью OneHotEncoder.

Добавляем функцию train\_model, которая принимает два аргумента: 'max\_depth', определяющий максимальную глубину дерева решений, и 'test\_size', определяющий размер тестового набора данных.

Внутри функции происходит разделение данных на обучающий и тестовый наборы, используя функцию 'train\_test\_split' из библиотеки sklearn. Затем создается экземпляр класса DecisionTreeClassifier из библиотеки sklearn, устанавливая параметр максимальной глубины, переданный в качестве аргумента. Модель обучается на обучающем наборе данных, вызывая метод 'fit'.

Затем модель используется для предсказания на тестовом наборе данных, вызывая метод 'predict', и вычисляются значения recall и precision, используя функции 'recall\_score' и 'precision\_score' из библиотеки sklearn.

Функция возвращает обученную модель решающего дерева, а также значения recall и precision в виде кортежа (model, recall, precision).

Функция 'app' создает интерфейс приложения на Streamlit.

С помощью функции 'st.title' устанавливается заголовок страницы. Затем создаются два слайдера для настройки максимальной глубины дерева и размера тестового набора данных с помощью функции.

Затем вызывается функция 'train\_model’, и результаты ее выполнения сохраняются в переменные model, recall и precision. Далее выводятся значения recall и precision.

С помощью функции 'export\_text' создается текстовое представление дерева решений, которое затем выводится с помощью функции.

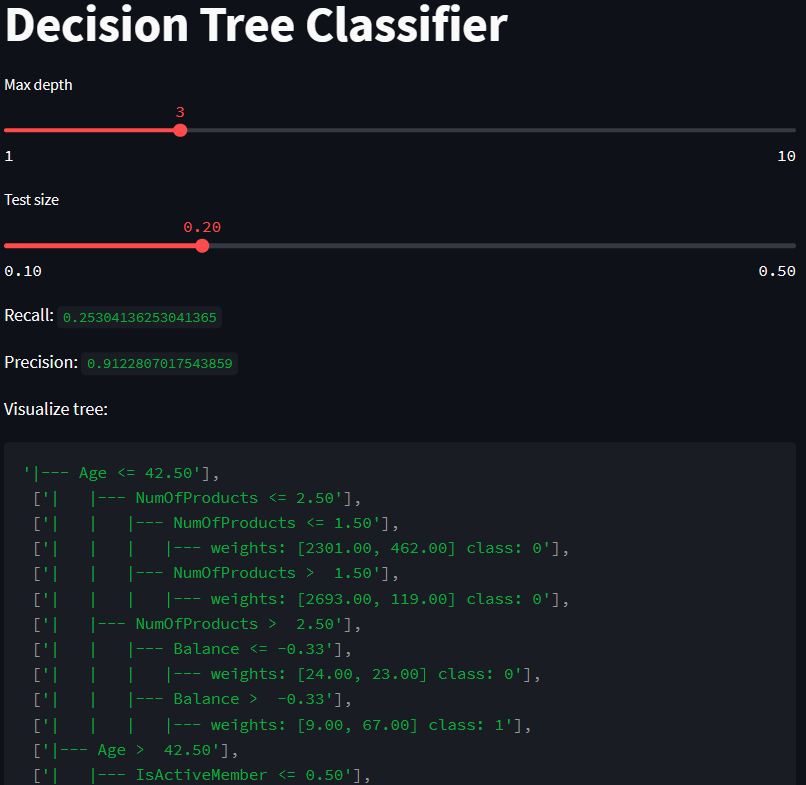


Рисунок 27 - Скриншот веб-приложения

# **Заключение**

Проблема ухода клиентов из банков довольно актуальна в наши дни. Поэтому каждому банку выгодно не терять своих клиентов. Однако можно заранее узнать покинет ли банк тот или иной клиент.

В рамках НИР была рассмотрена задача классификации случаев ухода клиентов из банка с помощью методов машинного обучения. Данные были проанализированы, визуализированы и подготовлены к обучению. Были применены различные алгоритмы, такие как метод ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг.

В результате исследования было показано, что большинство использованных методов могут достичь хороших результатов, но самыми точными на основании трех метрик из четырех оказались модели градиентного бустинга и метод случайного леса.

# **Список использованной литературы**

1. T-test на Python для проверки и получения t-статистики // Помощник Python URL: https://pythonpip.ru/osnovy/t-test-na-python
2. Machine Learning Metrics in simple terms // Medium URL: https://medium.com/analytics-vidhya/machine-learning-metrics-in-simple-terms-d58a9c85f9f6
3. Опорный пример для выполнения проекта по анализу данных. // Jupyter nbviewer URL: https://scikit-learn.org/stable/tutorial/statistical\_inference/supervised\_learning.html
4. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. // GitHub URL: https://github.com/ugapanyuk/courses\_current/wiki/COURSE\_TMO\_SPRING\_2023/