# Итоговая работа на тему «Разработка и тестирование приложения для прогнозирования будущих событий с использованием машинного обучения»

## Введение

Подтема:

Прогнозирование оттока клиентов из банка.

Научный руководитель:

Корнеева Елена Игоревна

Цель:

Разработать приложение для прогнозирования событий о клиентском оттоке из банка.

Задачи:

Использовать нейронные сети для обучения модели прогнозирования событий.

Обзор проекта

Для реализации проекта был выбран язык программирования Python, так как данный язык имеет большое количество сторонних библиотек для разработки нейронных сетей и анализа данных. Он является самым передовым языком в области машинного обучения, при этом предоставляет простой синтаксис для работы

Используемые библиотеки:

NumPy - библиотека для вычислительных операций в Python. Она предоставляет массивы многомерных данных, а также множество функций для работы с ними. NumPy обеспечивает эффективное и быстрое выполнение математических операций, что важно для работы с данными и обучения моделей машинного обучения.

Pandas - библиотека для анализа и обработки данных в Python. Она предоставляет структуры данных, такие как DataFrame, которые упрощают работу с табличными данными. Pandas позволяет проводить очистку, преобразование и анализ данных, что важно при подготовке данных для обучения моделей.

Seaborn - библиотека для создания красивых и информативных статистических графиков в Python. Она обеспечивает высокоуровневый интерфейс для визуализации данных и упрощает создание графиков с использованием стандартных стилей.

Matplotlib - библиотека для создания разнообразных графиков и визуализации данных в Python. Она предоставляет широкие возможности для кастомизации графиков и создания практически любых видов визуализаций.

Collections – библиотека для реализации специализированных типов данных контейнеров, предоставляет альтернативу встроенным контейнерам общего назначения Python dict, list, set и tuple.

Imblearn - библиотека специально разработана для работы с несбалансированными наборами данных. Она предоставляет различные методы, такие как undersampling, oversampling и SMOTE, для обработки и устранения дисбаланса в наборе данных. В данной работе использовался метод oversampling.

Scikit-Learn - библиотека машинного обучения в Python, которая предоставляет широкий набор инструментов для обучения и оценки моделей. Она включает в себя множество алгоритмов машинного обучения, инструменты для подготовки данных, оценки производительности моделей и многое другое.

XGBoost – библиотека используемая в машинном обучении и предоставляющая функциональность для решения задач, связанных с регуляризацией градиентного бустинга.

Joblib - библиотека для эффективной сериализации Python-объектов, таких как обученные модели. Она упрощает сохранение и загрузку моделей, что важно в задачах, где нужно сохранить результаты обучения для долгосрочного использования.

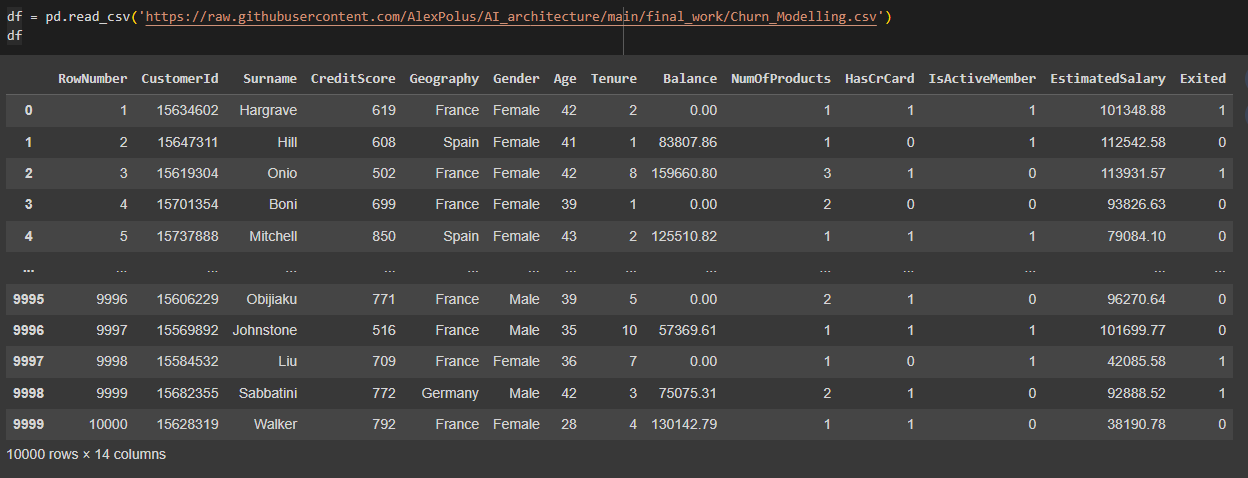
Каждая из этих библиотек является мощным инструментом в области анализа данных и машинного обучения, и выбор той или иной зависит от конкретной задачи и потребностей в проекте.

Таким образом, в проекте был использован широкий спектр библиотек для работы с данными, моделью и визуализацией.

Описание процесса решения:

Подбор данных:

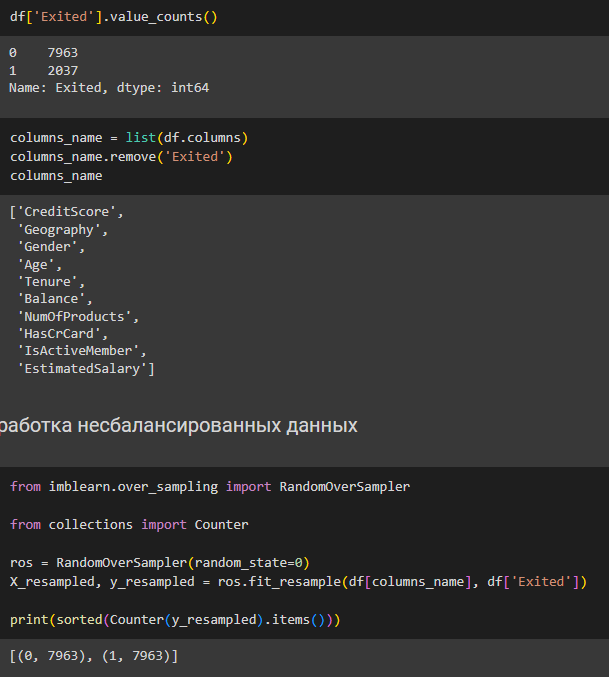
Для решения задачи прогнозирования был найден подходящий датасет. Модель была обучена на данных, включающих различные параметры связанными с клиентами банков, а именно кредитный рейтинг, пол, возраст, срок пребывания клиента в банке, количество кредитных продуктов и другие.



Подготовка данных:

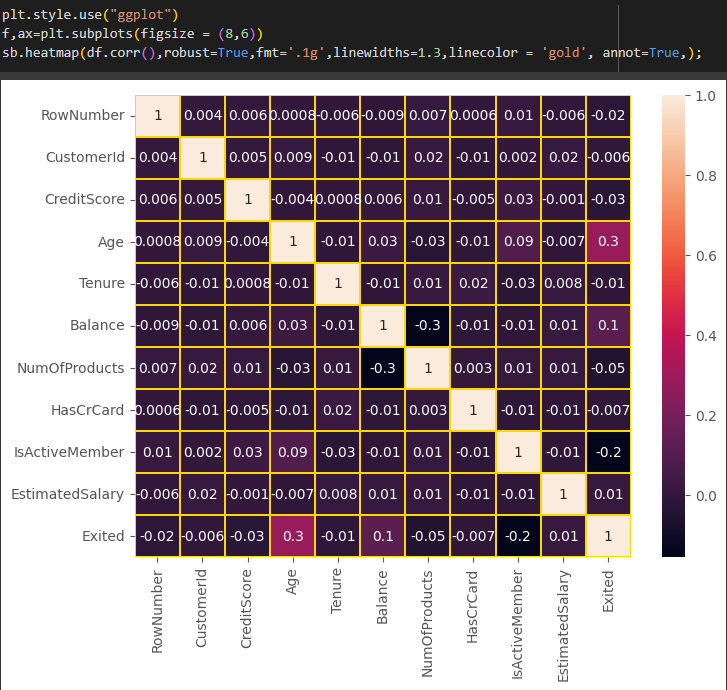
Вначале был проведен визуальный анализ данных, определение типов данных в различных столбцах, проверка на пустые значения, удаление лишних данных, которые не имеют ценности для датасета, такие как номер строки, имя и идентификатор клиента.

Затем был проанализирован датасет на несбалансированность данных по выходному параметру. По итогу анализа было принято решение сбалансировать данные с помощью RandomOverSampler. В следствии чего количество данных возросло с 10000 записей до 15926.



Далее данные имеющие строковые значения были преобразованы в категориальные данные.

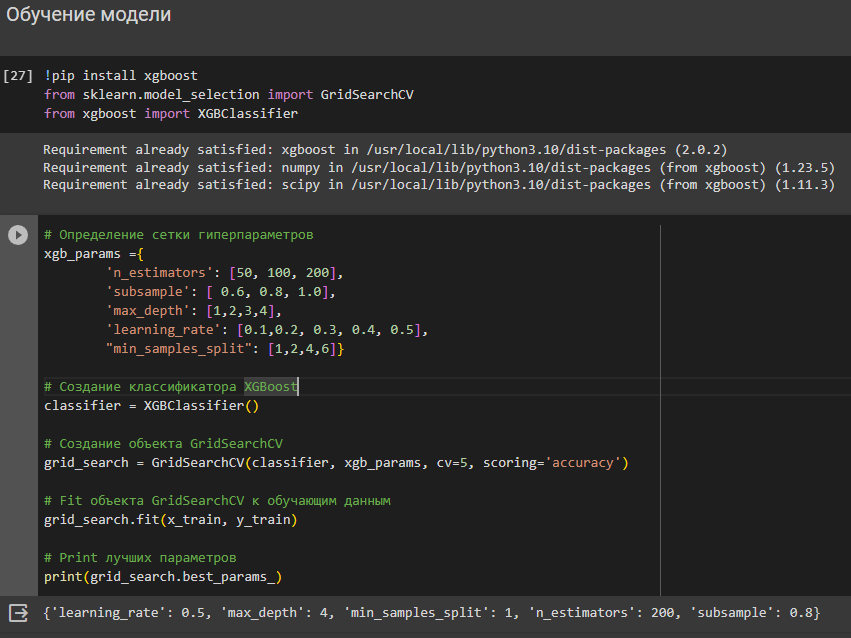
Затем были простроены heatmap и некоторые графики зависимостей, для лучшего понимания зависимости различных столбцов друг от друга.

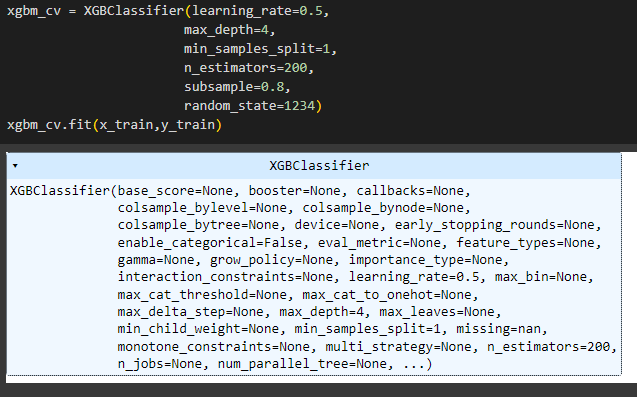


Также перед обучение модели были выделены тренировочные и тестовые сплиты с параметрами train\_size= 0.3 и random\_state=1234

Обучение модели:

Были выявлены гиперпараметры оптимальные для обучения и непосредственно само обучение.

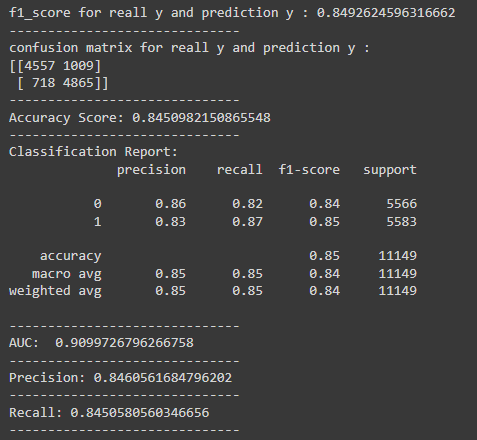


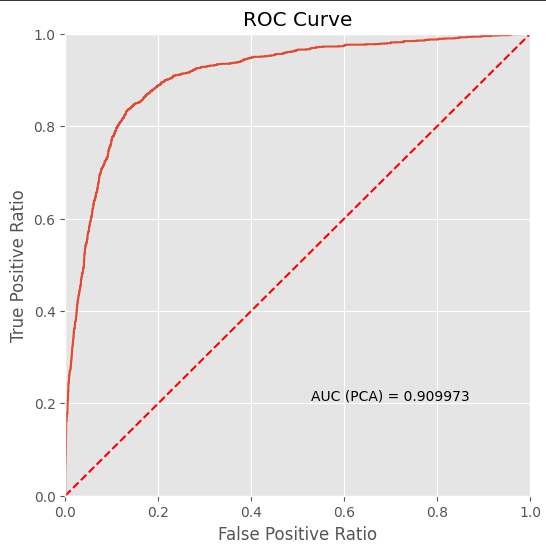


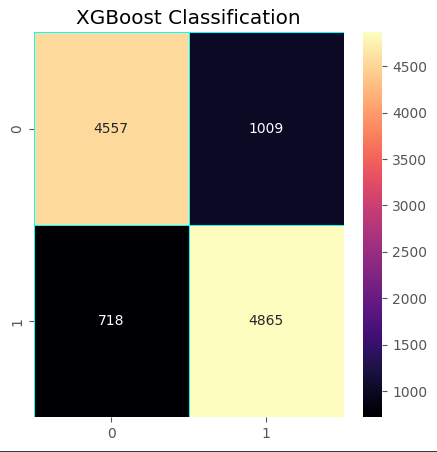
Результаты

Производительность модели:

Обученная модель демонстрирует высокую точность на тестовом наборе данных, которая составила около 84,5%. Это говорит о способности модели правильно классифицировать данные, что является критическим аспектом для задачи прогнозирования.







Итоги:

Полученные результаты подтверждают, что нейронные сети эффективны в задачах прогнозирования событий. Данное решение может быть важным для банковских приложений, в которых точность распознавания играет важную роль.

Вывод

В настоящий момент, разработка и тестирование приложений для прогнозирования событий с использованием машинного обучения стали ключевой темой и весьма актуальными в банковской и других сферах. Прогнозирование имеет широкий спектр использования в различных отраслях работы банков, от понимания лояльности клиентов к банку до анализа динамики инвестиционных продуктов.

Технологический прогресс и развитие машинного обучения придают этой области огромный потенциал. Применение машинного обучения улучшает эффективность рабочих процессов, понижает отток клиентов и укорачивает клиентский путь.

С развитием машинного обучения и нейронных сетей, точность и надежность систем прогнозирования значительно возросли. Это сделало их более надежными инструментами в различных секторах экономики. Однако, существует потенциал для дальнейшего улучшения и оптимизации этих систем.

Таким образом, разработка и тестирование приложений для прогнозирования событий с использованием машинного обучения имеют огромное значение.