**Information Technology Talent Program by ATON**

**Задачи для стажеров**

Выполнил

Василенко Даниил Сергеевич

Задание 1.

Пропуски в столбцах transaction\_id и client\_id заменить нечем, так как они являются уникальными идентификаторами, соответственно удаляем строки с пропусками в них. Пропуски и аномальные значения в столбце transaction\_date так же удалим. Эти же столбцы привел к единому формату. Пропуски и аномальные значения в amount заменил на среднее оставшихся значений. Остальные столбцы привел к стоковому типу, если иначе было нельзя, удалил.

Итоговый датасет насчитывает 8394 строки.

Задание 2.

Пользуясь встроенными функциями, выполнил поставленные задачи.

1. Топ 5 услуг: Инвестиционное консультирование, Управление активами, Финансовое планирование, Налоговое планирование, Структурирование капитала.
2. Результаты в файле с кодом.
3. Услугой с наибольшей выручкой оказалось Инвестиционное консультирование.
4. Банковский перевод - 34.643793185608764

Кредитная карта = 40.04050512270669

Криптовалюта = 5.2894924946390285

Наличные = 9.947581605908983

Остальные транзакции произведены неизвестным способом оплаты.

1. Суммарная выручка за последний месяц равна 149259475.06223527

Задание 3.

Сначала обработал данные аналогично первому файлу, затем пользуясь функцией pd.merge(), объединил два датафрейма.

Из результатов анализа понятно, что клиенты с высоким капиталом приносят в среднем наибольшую выручку.

Задание 4.

Пользуясь встроенными функциями, построил необходимые графики и диаграммы.

Задание 5.

Для прогнозирования спроса необходимо учитывать все транзакции, даже с некорректными данными, иначе удаление каких-либо строк приведет к гарантированному уменьшению прогнозируемой величины. Из-за этого обработанные ранее данные не подходят для анализа. Чтобы заново обработать данные я обработал столбец amount аналогичным способом. Обработать столбец дат транзакций было непросто, так как среди дат есть пропуски. Для решения этой проблемы я заменил их на равномерно распределенные даты по всему периоду. Таким образом я учел все транзакции, хоть и с достаточно высоким риском погрешности в данных. Все изначальные признаки, кроме даты бесполезны для прогнозирования, так как они не известны для дней вне исходных данных, поэтому я их отбросил.

Для прогнозирования на следующий месяц я сгруппировал данные по дням, чтобы остался только день и суммарная выручка за день. Планируется прогнозировать следующий день на основе известных данных о прошлом. Итого остался всего один признак, дата, но этого мало для хорошего прогноза, поэтому я пробовал добавлять следующие признаки:

1. Сумму транзакций за каждый из предыдущих дней (от 1 до 10);
2. 2-9 степень сумм транзакций за предыдущие дни
3. Средняя сумма транзакций за предыдущие дни (от 1 до 5);
4. Календарные показатели (день недели, выходной, день месяца).

Однако результат всех моделях на кросс валидации был ниже R2, что означало практически отсутствие результата. Среди рассмотренных моделей была выбрана модель ExtraTreesRegressor, так как R2 было хотя бы больше нуля, а точность составляла порядка 10%. Я подобрал для нее лучшие параметры с помощью GridSearchCV и запустил. Так же на тесте моделей, при тестовой выборке в 30 дней хороший результат показала линейная регрессия, если в данные добавить квадраты и кубы сумм транзакций за предыдущие дни (R2 = 0.12, MAPE = 0.098). Я решил так же запустить линейную регрессию для предсказания. Однако модель дала неудовлетворительный результат, возможно, из-за переобучения были получены нормальные средние оценки точности.

В результате, при использовании ExtraTreesRegressor, предсказанная сумма выручки за следующий месяц = 184611499