Автор отчета Пышногуб А.А.

Что представляет собой этот отчет?

В данном отчете кратко описаны пройденные этапы при выполнении задания. Подробное пошаговое описание действий и развернутые комментарии к коду находятся в Jupyter-ноутбуке.

Что представляет из себя задача?

Задача представляет из себя создание системы динамического ценообразования на основе прогноза временного ряда с экзогенными признаками.

Что было сделано на этапе EDA?

На этом этапе был проведен обзорный анализ предоставленных таблиц. Были рассмотрены размеры таблиц, содержащиеся в них данные, наличие пропусков, ошибок и аномалий. Сделаны основные следующие выводы:

- В таблице транзакций представлены данные по каждой из 875 тыс. транзакций по всем городам и продуктам за период времени с 2216-01-02 до 2218-09-27 с точностью до минуты.
- Наиболее зашумленной является таблица транзакций – в столбцах price и amount присутствуют отрицательные значения, причем для price отрицательные значения сильно сдвигают среднее значение, что говорит о присутствии сильной аномалии. Также в таблице присутствуют 432 строки с отсутствующим значением в столбце place.
- Остальные данные менее зашумлены, но собраны с разной периодичностью, требуется привести все данные к ежедневной периодичности.

Информация о транзакциях была обработана. Из таблицы были удалены данные с отрицательными значениями цены. На основании данных о распределениях также было принято решение удалить записи с отрицательным значением объема продаж. Таймстемпы с точностью до

минуты были округлены до дней, после чего данные были сгруппированы и аггрегированы по сумме объема продаж, после чего стала заметная сильная сезонность.

Итак, данные были успешно обработаны и сгруппированы (от 875 тыс. единичных транзакций к 15 тыс. ежедневных записей), на их основе можно создать модель динамического ценообразования.

Какая была предложена архитектура решения задачи?

На основе имеющихся данных (день, цена, время, цены конкурентов, погода и др.) обучаем модель (одну на все комбинации товар/город, либо 15 разных) прогнозирования временного ряда с экзогенными признаками, таргет - объем продаж.

По данным, получаемым с инференса модели, рассчитывается критерий качества. На основе каких-либо правил создаем массив цен-кандидатов (цены-кандидаты должны удовлетворять заданным ограничениям) либо выполняем поиск лучшей цены численными методами оптимизации с ограничениями. Прогнозируем все экзогенные признаки, на основании которых прогнозируем объем продаж.

На основе массива цен рассчитываем критерий качества, выбираем лучшего кандидата, устанавливаем цену на товар на следущие n дней.

Что было сделано на этапе поиска зависимостей и создания признаков?

Основной задачей на моменте углубленного анализа данных было понять, зависит ли спрос на наши товары от цены, которую мы устанавливаем. После анализа корреляции признаков друг с другом, а также их взаимного влияния без учета инфляции стало понятно, что спрос на наш товар от цены не зависит.

Это значит, что наш товар неэластичен по цене (что подтверждает то, что мы продаем товары для здоровья), поэтому было необходимо найти другие признаки, по которым можно было бы сделать прогнозирование.

Очевидно, что цена на наш товар зависит от цены конкурентов, но прямой корреляции найдено не было. Тогда были рассчитаны новые признаки, относительная разница между нашей ценой и максимальными, минимальными и средними ценами конкурентов, которые показали сильную корреляцию с целевой переменной.

Также сильную корреляцию с объемом продаж показывала погода. Это логично, так как в плохую погоду спрос на товары для здоровья будет больше. Но, после анализа имеющихся данных, был сделан вывод, что погоду с достаточной точностью предсказать невозможно, и из кандидатов в признаки погода была исключена.

Как были сделаны предсказания признаков, по которым прогнозируется спрос?

Следующей задачей было понять, можем ли мы предсказывать цены конкурентов и затраты на производство товара. Здесь был использован Prophet – пакет для прогноза временных рядов на основе преобразования Фурье. Для цены лучше результаты показало прогнозирование только линии тренда, когда как для цен было принято решение предсказывать не по каждому конкуренту, а максимум, минимум и среднее значение цен, что может увеливает точность прогноза на длительном отрезке. После обретения уверенности в возможности предсказания требуемых признаков и создания новых, которые лучше коррелируют с целевой переменной, была создана модель.

Как была реализована модель?

Наилучшие результаты показала комбинация двух моделей градиентного бустинга CatBoost. Обе модели были помещены внутрь ForecasterAutoreg, который позволяет легко адаптировать различные модели для задач предсказания временных рядом. Первая модель, на вход которой подавалось 365 лагов, была обучена только на них, в результате чего она хорошо была адаптирована под сезонность данных. Вторая модель получала на вход самые важные лаги первой, а также остальные экзогенные признаки, что уже обеспечивало хорошее отображение как сезонности, так и зависимости от цен конкурентов.

Как была выбрана рекомендуемая цена?

Следующей задачей была максимизации прибыли. Для этого была выполнена оптимизация критериальной функции, которая представляла собой прибыль со штрафами за выход за ограничения. Был создан алгоритм для оптимизации на основе метода Нелдера-Милда, который позволял получить соответствующую требованиям и максимизирующую прибыль цену.

Затем все описанные до этого шаги были повторены уже для последующего инференса, без разделения на тестовые и обучающие данные. На всем объеме данных была обучена модель, сделаны предсказания признаков, прогнозы целевой переменной, и с помощью численной оптимизации была получена рекомендованная цена на каждый из пяти видов товаров в каждом из трех городов.

Как прогноз, основанный на рекомендованной цене, отличается от baseline?

В конце ноутбука были проведены расчеты и сравнение прибыли.

product	place	profit_base	profit_rec
Целебные травы	Анор Лондо	19329	20298
Эльфийская пыльца	Анор Лондо	31169	33337
Эстус	Анор Лондо	18004	21607
Целебные травы	Врата Балдура	11541	12592
Эльфийская пыльца	Врата Балдура	16225	17208
Эстус	Врата Балдура	22015	24353
Целебные травы	Кеджистан	27208	31759
Эльфийская пыльца	Кеджистан	46433	47629
Эстус	Кеджистан	48377	50813
Целебные травы	Нокрон	16382	17647
Эльфийская пыльца	Нокрон	21863	23642
Эстус	Нокрон	16692	18786
Целебные травы	Фалькония	19962	23110
Эльфийская пыльца	Фалькония	29073	29561
Эстус	Фалькония	20250	20445

По нашим прогнозам, при установке рекомендованных цен на товары мы получим на 28264 зол. Больше (или на 7.75 %) что сравнимо с открытием еще одной торговой точки. **Файл с рекомендуемыми ценами называется df_answer.parquet**

Какие глобальные выводы можно сделать о исходных данных?

- Скорее всего, перед нами реальные, но специально измененные в угоду безопасности данные.
- Данные почти наверняка "сжаты/растянуты" по оси времени, так как сезонность не синхронизирована с количеством дней в году, а составляет по грубой оценке графиков и анализу самых важных лагов в моделях примерно 250 дней.
- Автор контеста большой фанат темного фэнтези.

Вопросы, оставшиеся без ответа?

- Почему в данных отсутствует ровно 2 дня продаж для Эльфийской пыльцы во Вратах Балдура?
- Почему распределение распределение объема продаж практически равномерное для всех товаров?
- Какой товар скрывался за этими названиями на самом деле? (скорее всего, что-то вроде бензина)

Что можно было сделать лучше, если бы было больше времени?

- Провести более обширный grid search/random search параметров во всех моделях.
- Проверить данные о погоде и попытаться найти закономерости / восстановить если они специально испорчены, так как погода обладает сильной (относительно других имеющихся признаков) корреляцией со спросом.
- Проверить и, возможно, использовать в качестве признаков цены на другие типа товаров. Наши товары могут быть взаимозаменяемы.
- Модифицировать алгоритм, так чтобы цену не обязательно было менять раз в 4 дня, а например, можно было держать цену 6 дней подряд, а потом поменять на 7 день, мы можем упускать прибыль.
- Либо предложить другой, более эффективный алгоритм, который бы быстрее или точнее справлялся с поиском оптимума, либо изменить подход на генерацию цен по какому-либо правилу и последующий выбор самых результативных и подходящих требованию цен-кандидатов.
- Проверить и при необходимости отредактировать код на соответствие PEP8, через, например, pylint.