Всем привет! На протяжении недели я выполнял задания кейс-чемпионата. Компанией были поставлены 3 задачи: во-первых, спрогнозировать продажи каждого вида товаров на 3 месяца, во-вторых, найти на тот же период наилучшую цену и в-третьих обосновать точность полученных прогнозов. Результаты выполнения задач находятся в таблице «prediction.csv» («prediction.tsv») и Jupyter-Notebook «gpn\_champ.ipynb» с подробными комментариями. Ключевые моменты отражены в этом кратком отчете о проделанной работе.

В первом задании необходимо спрогнозировать количество продаж товаров. Это означает, что целевым признаком является натуральное число. Мне стало понятно, что передо мной задача линейной регрессии обучения модели «с учителем». Для решения этой задачи я видел два подхода:

1. обучать модель конкретно для каждого SKU

2. обучать модель «по всему портфелю» целиком

Недостатком первого способа была невозможность оценить влияние на продажи таких категориальных признаков, как категория продукта, брэнд, начинка и т. д. Поэтому я выбрал второй подход.

Для выполнения заданий были предоставлены сведения о поставках товаров и о категориях покупателей за последние несколько лет. Первым делом нужно было произвести оценку данных и объединить все сведения в одну таблицу. На этом этапе стало понятно, что мы имеем дело с рядом сложностей:

- «битые» данные (отсутствовали цены на некоторые продукты),

- отсутствуют сведения о продажах продуктов в некоторых локациях,

- не было информации о покупателях в локации Village.

Так как разброс цен по «битым» товарам был незначительным, то первую проблему я решил вставками средних значений цен по данному продукту. Это не должно сильно исказить результаты прогнозирования и позволит сохранить и использовать другие признаки. Вторая и третья проблемы были существеннее. Так как сведения сильно отличались в зависимости от локаций я посчитал нецелесообразным генерировать приблизительные данные самостоятельно. Это сильно исказило бы прогноз по «чистым» данным. В результате из-за отсутствия сведений я не смог сделать прогноз по всем товарам в локации Village и по нескольким товарам в некоторых локациях (Moon - 15, 33, 38, 49, 42, 67, 78, 94, 106, Pirate Bay — 98, 55, 34, 37, 27, 43, 44, 72, 81, 82, 83, 84, 7, 11, 17, 3, 5, Alpha City - 98, 55, 34, 37, 27, 43, 44, 72, 81, 82, 83, 84, 7, 11, 17, 3, 5, 105, 60).

Далее я выполнил анализ целевого признака, отбор и исключение коррелирующих вещественных и категориальных признаков, нормализацию значений и one-hot-coding.

Для обучения модели я попытался использовать написанный мной алгоритм мини-батч градиентного спуска, однако стандартный инструмент библиотеки sklearn на тестовой выборке оказался немного лучше. Поэтому в дальнейшем я решил пользоваться им. Метрика SMAPE для всего датасета составила 15%. Результаты прогнозов обученной модели находятся в таблице «prediction.csv».

Для выполнения второго задания я решил гранулировать анализ для каждого SKU отдельно. На мой взгляд данный подход позволит модели сильнее уловить связь между ценой и объемом продаж. В результате я не смог воспользоваться моделью из первого задания, так как категориальные признаки становятся неизменными и соответственно избыточными. Логично представлять, что с увеличением цены не всегда будет происходить рост выручки, так как спрос на товар уменьшится. В такой ситуации нашей задачей становится определить некий баланс между ценой и продажами. В качестве решения,мной написана функция, которая, итерируясь по каждому типу товара, обучает отдельную модель. Изменяя цену в диапазоне от -15% до +15% (с шагом 5%), я пересчитывал значения объема продаж. Исходя из значений прогнозируемой выручки были определены «лучшие» цены. Результаты для каждого товара представлены в таблице «prediction.csv».

Для оценки точности прогноза выручки при наилучшей цене мной рассчитаны границы 95%-доверительного интервала. То есть мы можем утверждать, что истиная выручка с 95% вероятностью будет находится в рассчитанном интервале значений. Результаты для каждого товара представлены в таблице «prediction.csv».

Комментируя имеющиеся данные, хотелось бы отметить их репрезентативность, несмотря на описанные выше недостатки. Однако, получение полных «чистых» данных позволило бы увеличить точность прогноза. Так как мы имели дело с достаточно популярными кондитерскими изделиями, то считаю, что дополнительная информация о маркетинге (вложения производителя в рекламу, расположение товара на полках и т. д.) также позволила бы существенно увеличить точность прогнозов по продажам.

В качестве заключения, мне бы хотелось выразить благодарность всем организаторам кейс-чемпионата за приобритенный опыт. Мне было очень интересно решать настоящую задачу с «живыми» данными и применить на практике теорию из университета. Глупо отрицать карьерные и финансовые перспективы стажировки в такой организации как «Газпром». В сегодняшней общественной формации все мы так или иначе продаем свое рабочее время, руки и голову и, как и большинство мужчин, я не лишен здравых амбиций. Однако, для меня стажировка - это в первую очередь возможность погрузиться в мир науки о данных, отрасль, которая способна изменить наш мир так сильно, насколько мы сегодня до конца не можем представить! Моя цель стать не только востребованным специалистом в своей сфере, но и иметь возможность развивать и внедрять новые технологии и, в целом, делать общество и мир чуточку лучше. Ведь истинная ценность человеческой жизни определяется тем, насколько ты был полезен окружающим.