Отчет Помозов Егор Игоревич

В работу были получены 4 набора данных:

* df\_transaction - транзакции магазинов;
* df\_competitors – информация о конкурентах;
* df\_cost – информация о себестоимости товаров;
* df\_weather – информация о погоде.

Первым шагом была проведена чистка данных от выбросов и пропусков. В датасете df\_transaction были отрицательные значения в поле price и amount (где их не должно было быть). Эти значения были заменены медианными по данному месту и данному товару.

После чего была проведена трансформация признаков для последующей подачи ее в модели машинного обучения (кодирование категориальных признаков).

После этого была произведена агрегация данных по полю date чтобы избавиться от поля time и работать с дневными данными.

Следующим шагом датасет был конкатенирован с датасетом о себестоимости. При конкатенации было получено много NaN значений, так как товары закупаются один раз на несколько дней вперед. Эти значения были заполнены значениями последней закупки данного товара в данном месте. Так же к датасету были добавлены данные о погоде.

После этого была произведена чистка данных от выбросов.

Анализ полученного датасета позволил сформировать следующие выводы:

* Поля price и amount есть выбросы, которые были успешно удалены;
* Почти в 45% дней присутствуют различные погодные явления;
* При присутствии этих погодных явлений число продаж вырастает примерно на 35 процентов;
* Эльфийская пыльца продается в меньшем количестве, но имеет сильно более высокую цену ввиду высокой себестоимости, в следствии этого, с этого продукта идет максимальная прибыль;
* Наивысшая прибыль наблюдается в Кеджистане;
* Количество продаж возрастает в четвертом квартале;
* Прибыль растет из года в год, как и себестоимость товаров;
* Количество продаж имеет годовую сезонную составляющую, описываемую аддитивной моделью, и боковой тренд;
* Продажи никак не зависят от дня недели;

После этого было решено начинать построение моделей. Изначально имеем только 2 признака - дата и цена (которую мы задаем), но нам точно нужны колонки количества проданного товара и себестоимости. Себестоимость нужно спрогнозировать на основе данных прошлых месяцев, так как на нее не влияют такие факторы как, например, цена, которую мы будем менять. Для количества проданного товара необходимо построить регрессионную модель. Помимо этого, будет полезно предсказывать погоду, так как есть зависимость по этому признаку. Также нужно спрогнозировать цены конкурентов, чтобы иметь понимание о всем рынке

Первым делом было решено спрогнозировать себестоимости товаров. Было подсчитано, как часто производятся закупки каждого товара:

* 54 закупки за 1000 дней для первого товара;
* 93 закупки за 1000 дней для второго товара;
* 76 закупки за 1000 дней для третьего товара;

Исходя из этих данных были заданы горизонты прогнозирования для каждого товара. Экспоненциальное сглаживание показало неплохие результаты для прогнозирования себестоимостей.

Следующим шагом были спрогнозированы цены у конкурентов. С этой задачей экспоненциальное сглаживание также справилось неплохо.

Следующей моделью стал xgboost для предсказания погоды в зависимости от даты и места. После гридсерча он показал точность в 98 процентов на тестовых данных.

После этого была построена полносвязная нейронная сеть для прогнозирования дневного количества продаж. Ошибка mae на валидационном наборе равна 6.6328, а r2\_score равен 0.6822.

На данном этапе мы готовы собирать наш итоговый датасет. Была создана оптимизационная функция для максимизации прибыли на основе различных вариантов цены товара. Цена предсказывалась сразу на 3 дня, так как чаще менять нельзя, при этом интервал прогнозирования [price+1; price-1]. Помимо этого, производился учет цен у конкурентов.

По итогу всех манипуляций была получена модель, которая, по предварительным оценкам, позволит получить прибыль в 860097 золотых, что на 332158 выше, чем у baseline. Всю интересующую вас информацию (код, выводы, комментарии, графики и более подробное описание) вы можете найти в файле predict\_price.ipynb.

Возможные доработки при наличии дополнительных ресурсов:

* Более детальный разведочный анализ;
* Использование более сложных моделей для прогнозирования (prophet или arima);
* Использование более сложной оптимизационной функции (например, метод Брента);
* Более качественное прогнозирование цен у конкурентов (работа с выбросами, очистка данных, использование более сложных моделей);
* Использование более сложной модели предсказания количества продаж;
* Более качественный код.