Хакатон “Что найдет сервер?”

Команда “Белый пес”

Немного о команде

Анастасия Батхина, PhD

Академический директор в НИУ ВШЭ, основатель и CEO международного технологического стартапа InMind, New York

Виталий Мальцев

Data Scientist, BCS Data Unicorn (коммерциализация больших данных)

Задача - мониторинг цен

Основные сложности:

- мало данных, по некоторым названиям товаров единицы записей
- данные разбиты по времени, по некоторым названиям товаров последние заказы были несколько лет назад

Ссылка на гитхаб с кодом:

Выбор подхода

Из-за особенности в данных, оптимально будет предсказывать цену на каждое наименование товаров по отдельности. Первые (самые старые исторически) ¾ записей будем использовать для обучения, последнюю четверть для тестирования качества моделей.

Выбор метрики

В качестве метрик качества будем использовать MAPE (mean absolute percentage error, ошибка в процентах).

Будем смотреть ошибку по каждому наименованию товаров отдельно и в качестве общей метрики берем среднее MAPE по всем товарам.

Базовый подход

Если всегда брать в качестве стоимости товара стоимость предыдущей покупки этого товара, то в таком случае среднее MAPE получается 9.45.

На отдельных товарах ошибка доходит до 38%.

Линейная модель

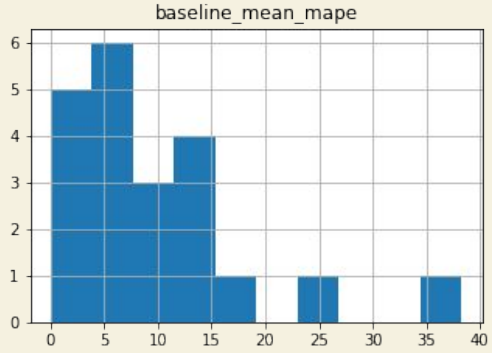
В общем случае, ошибка линейной модели получается около 29%. В редких случаях результат лучше, чем у базовой версии, на отдельных примерах ошибка очень большая.

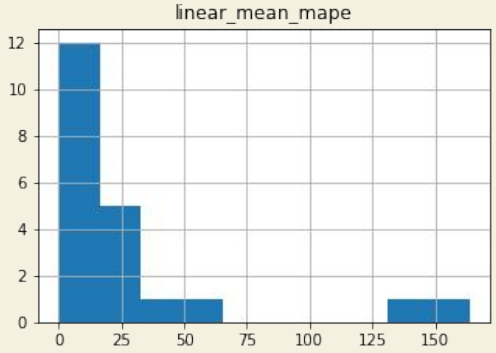
Модель *Catboost* с дополнительными данными

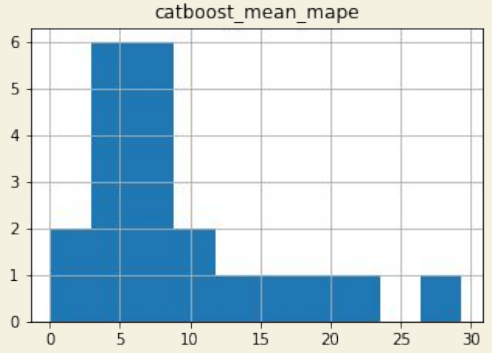
Какие данные мы добавили?

- *внешние:* разные признаки, основанные на стоимости доллара (средняя, максимальная, минимальная стоимость $ последние 30-100-350 дней до даты заявки), скачанные при помощи *yahoo finance api*.
- *внутренние:* обобщенные характеристики по товарам и поставщикам (стоимость товара в прошлые заказы, стоимость заказов у данного поставщика и т.д.)

Ошибка модели 8.9MAPE, что лучше базовой модели.







Для товаров с небольшим количеством данных нужно использовать иной подход определения стоимости.

Задачи на хакатон

1. Дополнительно обогатить данные стоимостью различных материалов (золота, металла и т.д.) на бирже в периоды времени.

Посмотреть корреляцию и влияние стоимостей на цену различных товаров.

2. Реализовать нормальную валидацию модели бустинга. Там, где достаточно много семплов, можно сделать кросс-валидацию.

Есть идея, при обучении валидироваться на других наиболее похожих названиях товаров.

Задачи на хакатон

3. Реализовать наглядный пример использования инструмента по предсказанию цен на товары.

4. Для тех товаров, по которым мало исторических данных, придумать иной способ предсказания.

Например, предсказание предыдущей покупки + корректировка на инфляцию или линейная модель.