нетология

Прогноз размещения (количество и сумма) новых процедур на ЭТП РТС-тендер в следующем месяце

Солодов Алексей Валерьевич

1. ПОСТАНОВКА ЗАДАЧИ

ТЕКСТ

Что за компания?

РТС-тендер – электронная площадка, отобранная Министерством финансов РФ и ФАС России для проведения закупок в электронной форме для государственных и муниципальных нужд в соответствии с 44-ФЗ.

ТЕКСТ

Что важно?

1. Обеспечение бесперебойной работы

2. Создание удобств для работы пользователей

3. Заработок компании

ТЕКСТ

Что поможет в достижении целей?

Знание ожидаемого количества проводимых процедур!

Зная количество можно:

1. Посчитать доходы (количество процедур * тариф)
2. Спрогнозировать нагрузку на:
    - Портал площадки
    - Колл-центр

ТЕКСТ

Как решать задачу?

Шаг 1. Собрать данные о ранее размещенных процедурах

Шаг 2. Агрегировать данные на основе факторов, которые предположительно оказывают влияние

Шаг 3. Создать модель и сделать прогноз

Шаг 4. Проверить получившиеся результаты

ТЕКСТ

Какие метрики будем смотреть?

- Средняя абсолютная погрешность

$$\Delta A = \frac{\sum_{k=1}^{N} (P_k - A_k)}{N}$$

- Средняя относительная погрешность

$$\Delta a = \frac{\sum_{k=1}^{N} (P_k - A_k)}{\sum_{k=1}^{N} A_k}$$

- Среднее значение

$$RMSLE = \sqrt{\sum_{i=1}^{N} \frac{1}{N} (log(y_i + 1) - log(\tilde{y}_i + 1))}$$

2. АНАЛИЗ

ТЕКСТ

Какие данные есть?

Данные о закупках проведенных на площадке с 01.10.2010
Особенности данных:

1. До 1.01.2014 закупки проводились в соответствии с другим законом. Сейчас 44-ФЗ, был 94-ФЗ

2. С 1.01.2016 полностью перестроен Общероссийский классификатор продукции по видам экономической деятельности

ТЕКСТ

Какие данные будем использовать?

Данные о процедурах размещенных с 1.01.2016.

Будем использовать следующие параметры:

- Номер
- Дата размещения
- Адрес поставки
- Цена контракта
- ОКПД2
- Данные о заказчике
- Данные о преференциях

ТЕКСТ

Как получим и агрегируем данные?

1. Создаем запрос в БД MS SQL, который содержит необходимые нам поля
2. Задаем условия в запросе, отделяющие некорректные данные (тестовые, отмененные и т.п.)
3. Агрегируем данные по выбранным параметрам и временным интервалам
4. Делаем выгрузку

3. МЕТОДИКА РЕШЕНИЯ

ТЕКСТ

Какую методику будем использовать?

Объединим все временные последовательности в один датасет и будем пользоваться обычными алгоритмами машинного обучения

Какие алгоритмы будем применять?

- Линейная регрессия
- Случайный лес
- Градиентный бустинг (библиотека lightGBM)

ТЕКСТ

Как реализуем?

Шаг 1 Загружаем данные в ноутбук

Шаг 2 Делаем преобразование melt, после которого получаем «вытянутый» датасет состоящий из 3 колонок «код сущности», «временной интервал», значение (количество или сумма)

Шаг 3 Генерируем фичи, на основе данных о предыдущих месяцах (размещение в предыдущем месяце, разница за месяц)

Шаг 4 Обучаем модель и делаем предсказание

Шаг 5 Смотрим результаты

ТЕКСТ

Какие гипотезы будем проверять?

Создаем датасеты по следующим критериям:

- По «отрасли». Предположительно, закупки товаров должны повторяться с некоторой периодичностью
- По региону. Предположительно, закупки в регионах так же должны иметь некоторый порядок
- По региону и отрасли. Предположительно, должна быть закономерность между закупками, проводимыми по различным регионам и отраслям
- По региону и уровню субъекта организатора. Более детальное разбиение по региону
- По региону, уровню субъекта, преференциям.

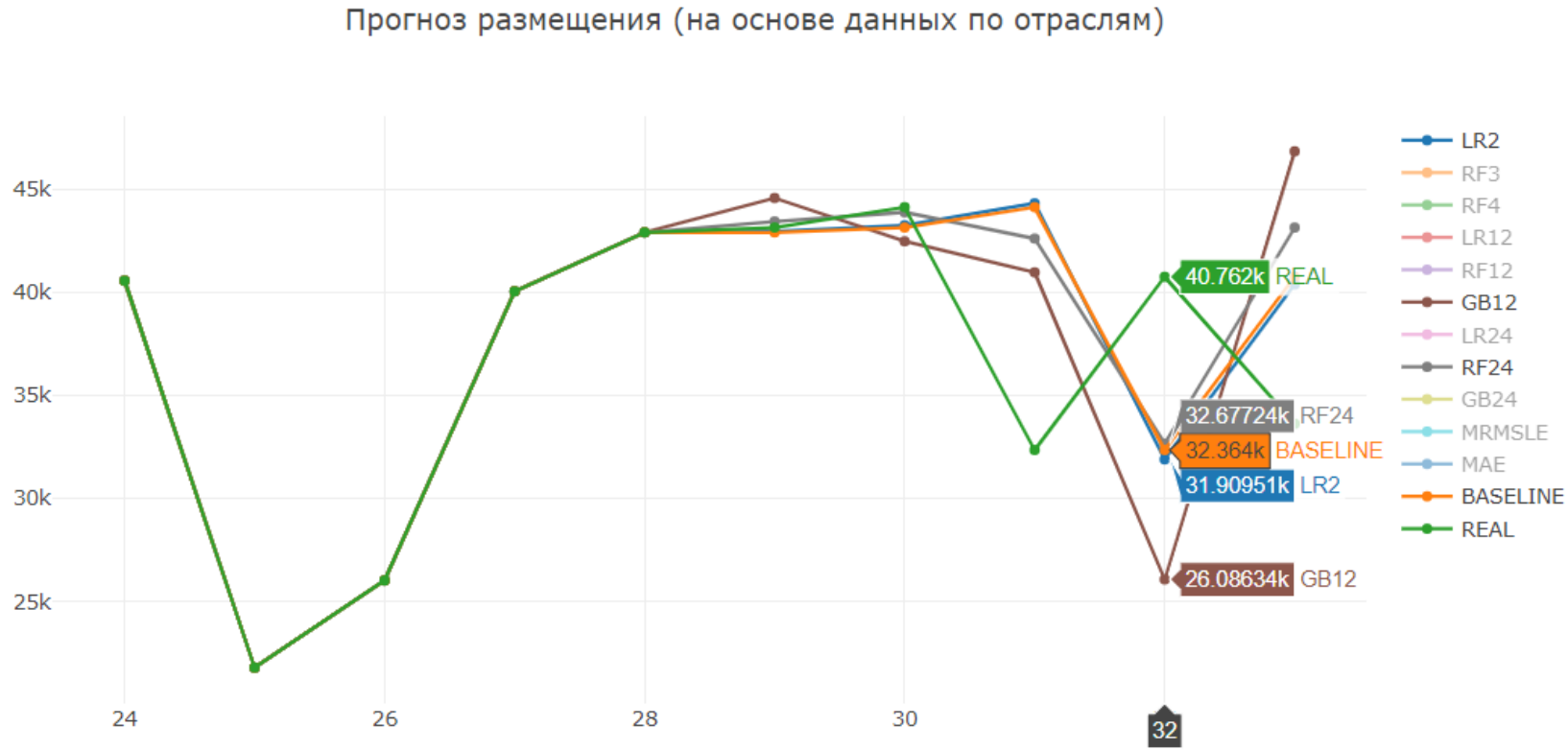
ТЕКСТ

Что в итоге проверяли?

Для каждого полученного датасета:

1. Создали отдельные наборы фичей (данные за 1 месяц, 2 месяца, 3 месяца, 12 месяцев, 24 месяца)
2. Определили baseline и сняли для него метрики
3. Для каждого набора фичей провели обучение и сделали прогноз на 5 месяцев с использованием LR, RF, LGB
4. Сняли метрики и сравнили с baseline

ТЕКСТ


Прогноз размещения (на основе данных по отраслям)
45k
40k
35k
30k
25k
24
26
28
30
32
40.762k REAL
32.67724k RF24
32.364k BASELINE
31.90951k LR2
26.08634k GB12
LR2
RF3
RF4
LR12
RF12
GB12
LR24
RF24
GB24
MRMSLE
MAE
BASELINE
REAL

4. РЕЗУЛЬТАТЫ

Что получилось?

В большинстве случаев, алгоритмы дают точность ниже baseline!

Для разбиения по «отраслям»:

Baseline - RMSLE =0.4474, $\Delta A = 5704$, $\Delta a = 14.7\%$

	LR2	RF2	LGB2	LR3	RF3	LGB3	LR4	RF4	LGB4
RMSLE	1.1115	0.5019	0.4276	0.6854	0.4544	0.4586	1.1815	0.5094	0.4486
Abs error	5726	6621	7602	7045	5720	6253	7633	5738	7918
%	14.75	17.05	19.58	18.15	14.73	16.11	19.66	14.78	20.40

	LR12	RF12	LGB12	LR24	RF24	LGB24
RMSLE	1.1669	0.5107	0.4486	1.2892	0.5104	0.4486
Abs error	7673	5721	7918	8117	5680	7918
%	19.77	14.74	20.40	20.91	14.63	20.40

Что можно попробовать?

1. Попробовать ограниченный набор фич
   - Перебираем в цикле все комбинации из двух фич и ищем лучшую по характеристикам
   - В цикле добавляем третью фичу и смотрим, улучшились ли характеристики
   - Продолжаем добавлять характеристики, пока качество улучшается
2. Комбинация моделей
   - Пробуем скомбинировать модели. В рамках проекта просто брал среднее арифметическое предсказаний

Что получилось в итоге?

- Совместная модель дает более сбалансированный результат (особенно ощутимо проявляется на датасетах с большим количеством строк) Но в большинстве случаев и она хуже baseline

- Попытка отобрать отдельные фичи для прогноза, практически всегда дает улучшение результата, но не всегда улучшает baseline

Baseline - RMSLE = 0.62653, $\Delta A = 5704$, $\Delta a = 14.7\%$

	CE4	LR12	RF12	LGB12	CE12	LR24	RF24	LGB24	CE24	MAE	MRSLE
RMSLE	0.5965	0.7841	0.6592	0.5600	0.5970	0.7841	0.6594	0.5600	0.5971	0.6948	0.6008
Abs error	6040	7372	5978	6573	6057	7372	5981	6573	6056	3788	5029
%	15.56	18.99	15.40	16.93	15.60	18.99	15.41	16.93	15.60	9.76	12.96

5. ЗАКЛЮЧЕНИЕ

ТЕКСТ

ВЫВОДЫ

1. Простейшую базовую модель улучшить очень не просто. Из **399** измерений параметров в ходе работы лучше baseline было только **81**!
2. Удалось найти модель которая по всех характеристикам улучшила baseline. **RMSLE = 0.5769, ΔA = 2774, Δa = 7.14%**
3. Комбинирование результатов делает модель более сбалансированной и зачастую дает выигрыш в результатах над одиночными моделями

ТЕКСТ

Что дальше?

- Представить модель руководству, как базовый вариант
- Попробовать построить модель на данных размещения на всех площадках
- Попробовать различные комбинации имеющихся моделей для повышения качества
- Попробовать использовать временные ряды
- Посмотреть на более низкий уровень сегментации. Построить модели для предсказания количества по отраслям. Данный прогноз можно использовать в маркетинговых целях, рассматривая площадку ни как единое целое, а как совокупность небольших отраслевых площадок

нетология

Спасибо за внимание!