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

Исследовательская работа

Солодов Алексей Валерьевич

2018

Постановка задачи

Предметная область, компании, в которой я работаю – проведение торгов по государственным закупкам в электронной форме в соответствии с 44-ФЗ. Государственные и муниципальные организации обязаны производить закупку практически всех необходимых им для жизнедеятельности товаров (до 1 Января 2019 года, часть товаров может закупаться не в электронной форме, с 1 Января 2019 только в электронной форме) с использованием одной из отобранных электронных площадок. Наша площадка [www.rts-tender.ru](http://www.rts-tender.ru) является ведущей по объему размещения закупок (общей начальной максимальной цене объявленных закупок) и второй по количеству объявленных на ней процедур.

**Задача:** Составить прогноз размещения новых процедур (количество + сумма) на месяц вперед.

Бизнес обоснование. Заработок электронной площадки строится:

* До 1 Октября 2018 года. Для принятия участия в закупках, участнику необходимо внести обеспечение (сумма зависит от начальной максимальной цены контракта), которое служит гарантом серьезности намерений участника. Обеспечение необходимо только на время участия в закупки. По прохождении определенного этапа, в случае, если участник вел себя добросовестно, обеспечение возвращается участнику. Обеспечение размещается на счете площадки. Площадка получает процент от размещения денежных средств в банке на что и существует.
* С 1 Октября 2018 года. У участника отменяется необходимость перечисления денежных средств на площадку. Вводится система специальных счетов. Поставщик открывает специальный счет и средства в качестве обеспечения заявки блокируются на специальном счете. Но вводится плата, взимаемая с победителей закупки за заключение контракта.

Таким образом, заработок электронной площадки на прямую зависит от: Суммы начальных цен размещенных закупок (до 1 Октября 2018 года) и количества размещенных закупок (с 1 Октября 2018 года)

**Данные для построения прогноза**

Имеется историческая информация обо всех электронных процедурах, проведенных на площадке с момента функционирования (с 1 Октября 2010 года).

Но:

* До 1 Января 2014 года закупки проводились по другим правилам (до 31 Января 2013 года действовал другой закон – 94-ФЗ)
* С 1 Января 2016 года был полностью перестроен Общероссийский классификатор продукции по видам экономической деятельности. До 31 Декабря 2015 года действовал классификатор ОКПД, с 1 Января 2018 – ОКПД2.

В связи с этим был принято решение, для построения прогноза использовать данные о процедурах, размещенных с 1 Января 2016 года.

Была подготовлена выборка данных по процедурам. Для каждой процедуры выбирались следующие данные:

* 1. **Номер**. Будем использовать как уникальный идентификатор
  2. **Дата размещения**. Основной параметр, на основании которого производилось агрегирование данных
  3. **Организатор**. Организация, которая производит все необходимые действия по проведению процедуры
  4. **Уровень заказчика (организатора)**. Заказчики делятся на несколько уровней (федеральный, региональный, муниципальный). Отличаются тем, из какого бюджета выделяются средства на закупку товаров. Предполагаем, что данный показатель является существенным и может влиять на количество и сумму размещаемых закупок
  5. **Заказчик**. Организация, для которой осуществляется закупка. Если заказчиков в процедуре несколько (совместные торги), то заказчик не выбирался
  6. **Адрес поставки**. Будем использовать для агрегирования закупок по регионам поставки (к сожалению, другого более мелкого агрегирования на получилось, т.к., например, город есть далеко не во всех адресах)
  7. **Начальная максимальная цена контракта.** Параметр, который будем использовать при агрегировании
  8. **Наличие преференций для субъектов малого и среднего бизнеса**. Будем использовать этот признак для агрегирования. По закону заказчики должны производить от 10 до 15 процентов своих закупок у субъектов малого и среднего бизнеса. Предположим, что это будет существенно влиять.
  9. **Наличие преференций для товаров российского или белорусского происхождения.** При закупке некоторых видов товаров производители из России и Белоруссии имеют преимущества. Предположим, что данный фактор будет оказывать влияние
  10. **Наличие преференций организаций использующих труд инвалидов.** Предположим, что данный фактор тоже вносит существенное влияние на объем размещения
  11. **Наличие преференций для учреждений исправительной системы.** Возьмем этот признак, так же, как и предыдущий и предположим, что он существенно влияет на объем размещения
  12. **Классификатор (ОКПД2 – Общероссийский классификатор продукции по видам экономической деятельности).** Классификатор – один из основных элементов для агрегирования данных. Для эксперимента предлагается разбить данные на «отрасли», на основе ОКПД2. В качестве отрасли будем брать агрегирующие коды, состоящие из 2-ух цифр

**Выборка данных и их очистка**

Выборка данных осуществлялись средствами MS SQL. Была создана временная таблица с полями, описанными в разделе выше. С данными производились следующие действия:

* Из данных убирались, тестовые закупки, отмененные закупки, закупки с нарушенными параметрами (такие закупки считаются не действительными).
* Процедуры были разбиты на «отрасли» на основе классификатора ОКПД2. Если в составе процедуры закупались товары из нескольких разных групп, то такие процедуры относились к отдельной отрасли «составная», которая кодировалась кодом «100»
* Из адресов поставки выделялся код региона и кодировался двумя цифрами в соответствии с цифровыми кодами Российской Федерации
* Наличие преференций кодировалось флагом
* Для удобства агрегирования по нескольким параметрам производилось «склеивание» кодов, отвечающих за соответствующую сущность. Например, процедура для региона Москва, с составным типом закупки, размещенная организатором регионального уровня, имеющая преференции для товаров российского производства и никаких других имеет код 7710020100, где
  + 77 – код региона поставки
  + 100 – код товара отрасли «составная»
  + 2 – организатор имеет уровень «региональный»
  + 0100 – битовая маска преференций. Преференции для товаров российского производства идут на втором месте (отмечено 1), других преференций нет (отмечено 0)

Для построения прогнозов агрегируем данные по некоторым последовательным временным интервалам (по неделям или месяцам). Для агрегирования:

* По неделям будем использовать формулу: Остаток от деления года на 2016 умножить на 52 плюс данные о неделе с начала года
* По месяцам будем использовать формулу: Остаток от деления года на 2016 умножить на 12 плюс данные о месяце с начала года

Для построения прогнозов предлагается проделать следующие агрегации:

* 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 связь между закупками, проводимыми по различным регионам и отраслям
* По региону и уровню субъекта организатора. Более детальное разбиение по региону
* По региону, уровню субъекта, преференциям. Детализация предыдущих вариантов агрегации

**Анализ подходов к решению задачи**

Существует несколько подходов к решению предлагаемой задачи:

1. Рассмотрение каждой последовательности процедур, как отдельного временного ряда и отдельное прогнозирование каждого такого ряда
2. Объединение всех последовательностей в единый датасет и построение модели над всей совокупностью данных

На мой взгляд, первый подход очень трудоемкий, поскольку необходимо учитывать особенности каждого временного ряда в отдельности, что весьма накладно. Кроме того, при изменении принципов агрегации, все временные ряды изменяются и работу необходимо проводить заново. В рамках решения данной задачи использовался второй подход.

**Методика построения прогноза**

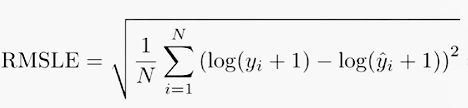
За основу методики построения прогноза выбран следующий подход:

1. Создаем некоторое количество временных рядов и объединяем их в один датасет
2. Делаем преобразование melt, после которого получаем «вытянутый» датасет состоящий из 3 колонок «код продукта», «временной интервал», значение
3. Строим фичи, на основе данных о предыдущих месяцах
4. Делаем модель основанную на полученных фичах используя обычные алгоритмы машинного обучения
5. Проводим обучение и смотрим результаты

Подробно данный подход описан в статье (см. раздел ссылки)

**Используемые метрики**

В качестве метрики, для оценки качества предлагается использовать Root Mean Squared Logarithmic Error (RMSLE)



Поскольку мы оцениваем общее количество заведенных процедур и их сумму, будем так же смотреть на показатели:

* Средняя абсолютная ошибка = сумма разницы между предсказанным значением и реальным деленная на количество экспериментов
* Средний процент ошибки = сумма разницы между предсказанным значением и реальным деленная на общую сумму реальных значений

При помощи RMSLE, мы будем следить, насколько спрогнозированные нами значения удаляются от некоторого «идеала». Данная характеристика очень тяжело интерпретируется, в наших измерениях мы будем считать, что чем RMSLE меньше, тем лучше. Средняя ошибка и средний процент ошибки очень легко интерпретируются человеком и практически любой человек понимает их суть. Данные показатели будем использовать для лучшего «человеческого» восприятия экспериментов.

**Алгоритмы машинного обучения, используемые для прогноза.**

Для построения прогноза попробуем использовать несколько различных алгоритмов машинного обучения и сравним получившиеся результаты:

1. Линейная регрессия – это самый простой алгоритм, попробуем использовать его. Он быстрый и в некоторых задачах показывает очень хорошие результаты
2. Случайный лес (random forest) – алгоритм основанный на использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество, но за счёт их большого количества достигается хороший результат
3. Градиентный бустинг – метод последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов. В нашем исследовании, будем пользоваться библиотекой lightGBM
4. Скомбинируем случайный лес и градиентный бустинг

**Прогнозирование:**

**Алгоритм получения прогноза:**

Для построения прогноза использовался следующий алгоритм. Выбирался набор агрегированных по какому-либо параметру данных. Первый столбец содержит код агрегации, все остальные столбцы – это сквозные номера временных интервалов (недели, месяцы) и агрегированные данные (количество или сумма закупок) за указанный период.

Выбранных набор проходил следующие стадии обработки:

1. Преобразование melt, после которого весь датасет превращался в длинную таблицу из трех столбцов «код агрегации», «временной интервал», «агрегированный столбец»
2. Построение базового прогноза. В качестве него использовался подход – В следующем месяце будет так же как в предыдущем
3. Добавление фичей. Таких этапов для каждого агрегированного датасета было несколько (использовались преобразования добавляющие фичи за предыдущий месяц, два предыдущих месяца, три предыдущих месяца, за 12 предыдущих месяцев, за 24 предыдущих месяца). На каждом из них производилось:
   1. Добавление новых фичей, путем добавления данных за предыдущие месяцы в новые столбцы.
   2. Обучение моделей и прогнозирование. Прогнозирование осуществлялось, в начале на 3 ближайших периода, затем, когда добавились данные за Август и Сентябрь 2018 по ним так же стали строить прогнозы, т.е. прогнозировались 5 интервалов
   3. Измерение результатов и сравнение с базовым прогнозом. Все характеристики усреднялись по количеству измерений
4. По окончании добавления фичей производился «подбор» фичей. Этот процесс производился итерационно. На первой итерации, в цикле выбиралось 2 фичи, прогнозирование на основании которых давало бы лучшие показатели, затем, к ним добавлялась еще одна фича, таким образом, чтобы результат максимально улучшался, затем следующая и так до тех пор, пока показатели качества продолжались улучшаться

**Агрегирование по неделям**

Первоначально использовался точно такой же подход, как описано в статье. Данные по закупкам были разбиты по неделям, произведена агрегация по «отраслям», регионам и совместная агрегация по «отраслям» и регионам.

В результате выполнения эксперимента выявлена глобальная проблема с интерпретацией данных:

1. Поскольку в году 52 недели, они не делятся ровно на 12 месяцев, т.е. существуют недели, которые одновременно относятся к двум разным месяцам
2. Поскольку у нас в стране принято очень многие вещи завязывать на окончание месяца (квартала, года) из-за возможной раздробленности недель, возникает некоторая неопределенность в данных, размываются известные циклы и т.п.
3. Из-за «плавающего» графика праздников, случаются провалы в работе в разные недели
4. В составляемом прогнозе трудно сделать акцент на месяце. 4 недели меньше чем месяц, 5 – больше. Трудно презентовать прогнозы другим людям, т.к. все привыкли мыслить «месяцами»

Из-за трудности корреляции данных в зависимости от указанных факторов было принято решение отказаться от агрегирования по неделям и работать с агрегированием по месяцам.

**Агрегирование по месяцам**

**Отрасли**

На первом этапе произведено прогнозирование на основании разбиения по отраслям. Поскольку в качестве отрасли брались коды ОКПД2 состоящие из двух цифр, то в наборе получилось 87 реальных отраслей + одна составная (в составную отрасль определялись процедуры в рамках которых производилась закупка из разных отраслей)

Для базового прогноза получились следующие показатели:

Средний RMSLE - 0.44740

Абсолютная средняя ошибка – 5704

Средний процент ошибки – 14.7%

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LR2 | RF2 | LGB2 | LR3 | RF3 | LGB3 | LR4 | RF4 | LGB4 |
| Ср. 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 |

В результате остальных экспериментов:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LR12 | RF12 | LGB12 | LR24 | RF24 | LGB24 | MRMSLE | MAE |
| Ср. RMSLE | 1.1669 | 0.5107 | 0.4486 | 1.2892 | 0.5104 | 0.4486 | 0.4363 | 0.6843 |
| Abs error | 7673 | 5721 | 7918 | 8117 | 5680 | 7918 | 4880 | 3265 |
| % | 19.77 | 14.74 | 20.40 | 20.91 | 14.63 | 20.40 | 12.57 | 8.41 |

**LR2** – линейная регрессия, построенная на данных за предыдущий месяц

**RF2** – случайный лес на данных за предыдущий месяц

**LGB2** – градиентный бустинг на данных за предыдущий месяц

**LR3** – линейная регрессия, построенная на данных за два предыдущих месяца

**RF3** – случайный лес на данных за два предыдущих месяца

**LGB3** – градиентный бустинг на данных за два предыдущих месяца

**LR4** – линейная регрессия, построенная на данных за три предыдущих месяца

**RF4** – случайный лес на данных за три предыдущих месяца

**LGB4** – градиентный бустинг на данных за три предыдущих месяца

**LR12** – линейная регрессия, построенная на данных за 12 предыдущих месяцев

**RF12** – случайный лес на данных за 12 предыдущих месяцев

**LG12** – градиентный бустинг на данных за 12 предыдущих месяцев

**LR24** – линейная регрессия, построенная на данных за 24 предыдущих месяца

**RF24** – случайный лес на данных за 24 предыдущих месяца

**LGB24** – градиентный бустинг на данных за 24 предыдущих месяца

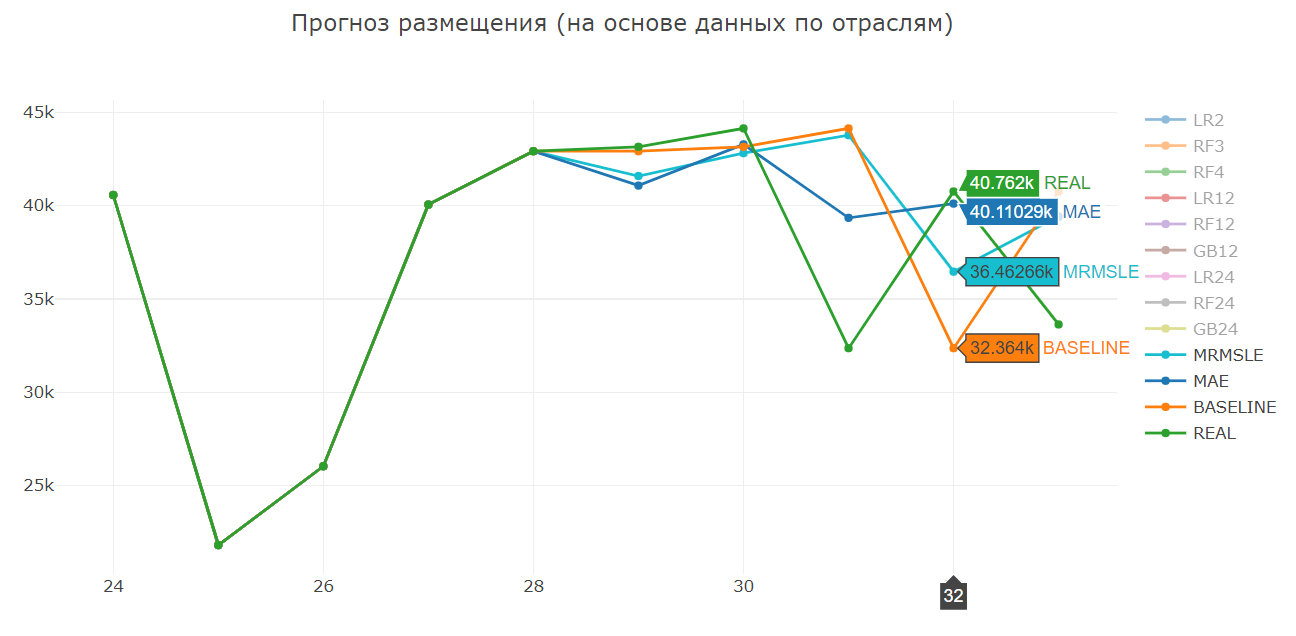
**MRMSLE** – прогноз с минимальной средней ошибкой RMSLE

**MAE** – прогноз с минимальной абсолютной ошибкой

Красным цветом в таблице отмечены значения, которые хуже базового показателя. Зеленым, те что лучше.

Как видно из таблицы, ни один алгоритм сходу не дал результат лучше базового. Лучшие результаты, были получен в результате подбора фич. Лучшего значения среднего RMSLE удалось добиться путем комбинирования фичи с отраслью и месяцем. Лучших абсолютных значений по данным о количестве процедур, проведенных на площадке за 2 месяца от текущего и за 12 месяцев. На основе этих данных можно построить лучший прогноз при данном разбиении. Если количество закупок проведенных через 12 месяцев выглядит логичным – через год возникает потребность в тех же товарах. Это выглядит как проявление сезонности. А вот количество закупок, размещенных за два месяца до текущего, с логической точки зрения объяснить трудно.

Графики прогнозов лучших прогнозов, а так же baseline и реальное размещение приведены на рисунке



**Регионы**

На этом этапе производилось прогнозирования на основании агрегирования данных только по региону. На площадке в разное время размещались закупки из 84 различных регионов. В качестве базового показателя так же использовалось значения предыдущего месяца

Для базового прогноза получились следующие показатели:

Средний RMSLE - 0.42341

Абсолютная средняя ошибка – 5704

Средний процент ошибки – 14.7%

В результате остальных экспериментов:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LR2 | RF2 | LGB2 | LR3 | RF3 | LGB3 | LR4 | RF4 | LGB4 |
| Ср. RMSLE | 0.8078 | 0.4536 | 0.4385 | 0.6854 | 0.4544 | 0.4586 | 0.7266 | 0.4509 | 0.4582 |
| Abs error | 5413 | 4790 | 5486 | 7045 | 5720 | 6253 | 7830 | 6213 | 6915 |
| % | 13.94 | 12.34 | 14.13 | 18.15 | 14.73 | 16.11 | 20.17 | 16.01 | 17.81 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LR12 | RF12 | LGB12 | LR24 | RF24 | LGB24 | MAE | MRMSLE |
| Ср. RMSLE | 0.7214 | 0.4521 | 0.4581 | 0.4514 | 0.4486 | 0.4581 | 0.6914 | 0.4502 |
| Abs error | 7814 | 6215 | 6736 | 7917 | 6244 | 6736 | 3353 | 4967 |
| % | 20.13 | 16.01 | 17.35 | 20.39 | 16.08 | 17.35 | 8.64 | 12.80 |

**LR2** – линейная регрессия, построенная на данных за предыдущий месяц

**RF2** – случайный лес на данных за предыдущий месяц

**LGB2** – градиентный бустинг на данных за предыдущий месяц

**LR3** – линейная регрессия, построенная на данных за два предыдущих месяца

**RF3** – случайный лес на данных за два предыдущих месяца

**LGB3** – градиентный бустинг на данных за два предыдущих месяца

**LR4** – линейная регрессия, построенная на данных за три предыдущих месяца

**RF4** – случайный лес на данных за три предыдущих месяца

**LGB4** – градиентный бустинг на данных за три предыдущих месяца

**LR12** – линейная регрессия, построенная на данных за 12 предыдущих месяцев

**RF12** – случайный лес на данных за 12 предыдущих месяцев

**LG12** – градиентный бустинг на данных за 12 предыдущих месяцев

**LR24** – линейная регрессия, построенная на данных за 24 предыдущих месяца

**RF24** – случайный лес на данных за 24 предыдущих месяца

**LGB24** – градиентный бустинг на данных за 24 предыдущих месяца

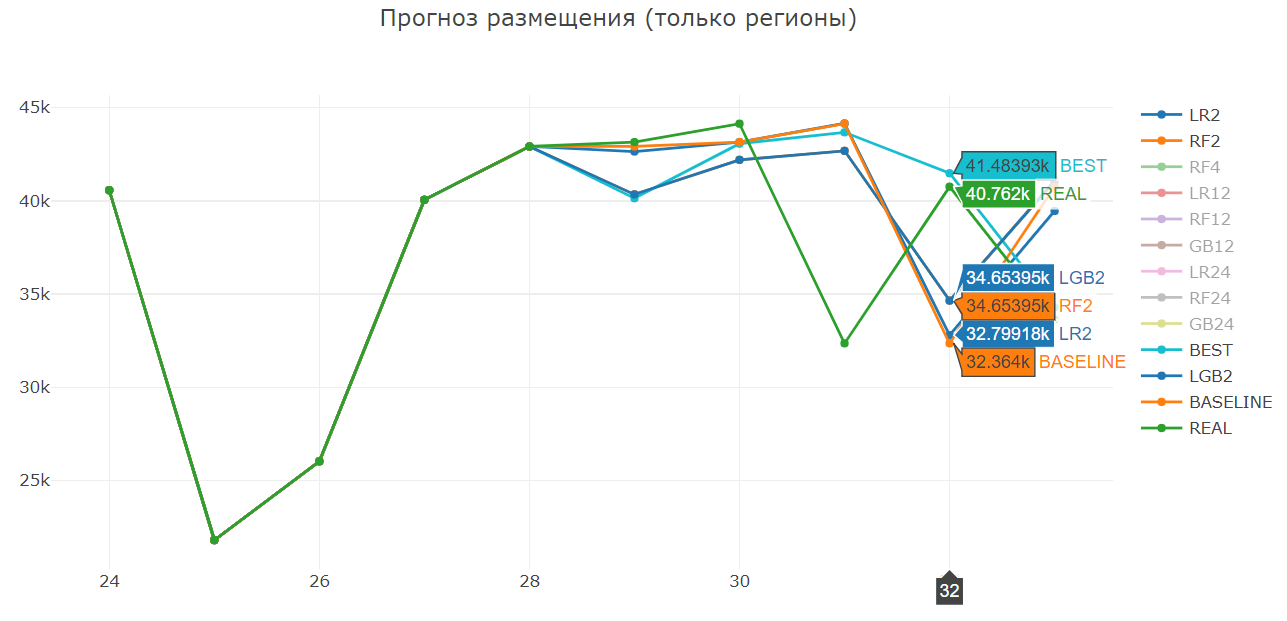
**MRMSLE** – прогноз с минимальной средней ошибкой RMSLE

**MAE** – прогноз с минимальной абсолютной ошибкой

Красным цветом в таблице отмечены значения, которые хуже базового показателя. Зеленым, те что лучше.

Так же как и для прогноза по отраслям, лучшие показатели среднего RMSLE получились при использовании кода региона и номера месяца. Наименьшей ошибки удалось достичь при использовании данных о процедурах размещенных за два месяца и за двенадцать месяцев от даты прогноза

На рисунке ниже отображаются лучшие прогнозы, полученные в результате эксперимента



**Отрасли + регионы**

В этом разделе производилось совместное агрегирование по отраслям и регионам. Если перемножить 84 региона на 88 отраслей, то должно получиться 7392 категории, однако в части регионов никогда не закупали продукцию некоторых отраслей. Такие отрасли не попали в выборку. В результате комбинирования получилось 5310 различных категорий.

Для базового прогноза получились следующие показатели:

Средний RMSLE - 0.6384

Абсолютная средняя ошибка – 5704

Средний процент ошибки – 14.7%

В результате остальных экспериментов:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LR2 | RF2 | LGB2 | CE2 | LR3 | RF3 | LGB3 | CE3 | LR4 | RF4 | LGB4 |
| Ср. RMSLE | 0.6970 | 0.6651 | 0.5665 | 0.6142 | 0.6684 | 0.6652 | 0.5601 | 0.6071 | 0.7035 | 0.6637 | 0.5556 |
| Abs error | 5198 | 8130 | 6818 | 5852 | 6276 | 8436 | 7166.5 | 5888 | 6596 | 8155 | 7362 |
| % | 13.39 | 20.94 | 17.56 | 15.07 | 16.17 | 21.73 | 18.46 | 15.17 | 16.99 | 21.01 | 18.97 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CE4 | LR12 | RF12 | LGB12 | CE12 | LR24 | RF24 | LGB24 | CE24 |
| Ср. RMSLE | 0.6145 | 0.7034 | 0.6645 | 0.5556 | 0.6148 | 0.7034 | 0.6645 | 0.5556 | 0.6148 |
| Abs error | 6192 | 6596 | 8131 | 7362 | 6192 | 6596 | 8131 | 7362 | 6192 |
| % | 15.95 | 16.99 | 21.01 | 18.97 | 15.95 | 16.99 | 21.01 | 18.97 | 15.95 |

**LR2** – линейная регрессия, построенная на данных за предыдущий месяц

**RF2** – случайный лес на данных за предыдущий месяц

**LGB2** – градиентный бустинг на данных за предыдущий месяц

**CE2** – комбинированный прогноз на основе прогноза полученного RF2 + LGB2

**LR3** – линейная регрессия, построенная на данных за два предыдущих месяца

**RF3** – случайный лес на данных за два предыдущих месяца

**LGB3** – градиентный бустинг на данных за два предыдущих месяца

**CE3** – комбинированный прогноз на основе прогноза полученного RF3 + LGB3

**LR4** – линейная регрессия, построенная на данных за три предыдущих месяца

**RF4** – случайный лес на данных за три предыдущих месяца

**LGB4** – градиентный бустинг на данных за три предыдущих месяца

**CE4** – комбинированный прогноз на основе прогноза полученного RF3 + LGB3

**LR12** – линейная регрессия, построенная на данных за 12 предыдущих месяцев

**RF12** – случайный лес на данных за 12 предыдущих месяцев

**LG12** – градиентный бустинг на данных за 12 предыдущих месяцев

**CE12** – комбинированный прогноз на основе прогноза полученного RF4 + LGB4

**LR24** – линейная регрессия, построенная на данных за 24 предыдущих месяца

**RF24** – случайный лес на данных за 24 предыдущих месяца

**LGB24** – градиентный бустинг на данных за 24 предыдущих месяца

**CE24** – комбинированный прогноз на основе прогноза полученного RF4 + LGB4

Красным цветом в таблице отмечены значения, которые хуже базового показателя. Зеленым, те что лучше.

Эксперименты показали, что наилучшие результаты по общей ошибке показывает линейная регрессия, построенная на данных за предыдущий месяц, но по RMSLE ее показатели хуже baseline. Показать лучшее чем baseline среднюю RMSLE удалось градиентному бустингу и совместному алгоритму, построенному на предсказаниях случайного леса и бустинга. Совместный алгоритм показал самые сбалансированные показания, но все равно по лучшим значениям он не превосходит baseline

**Регионы + уровень субъекта**

В этом разделе производилось дополнительное агрегирования организаторов по уровням заказчиков. Всего существует три уровня заказчика: Федеральный, Региональный, Муниципальный. Кроме некоторых специфических субъектов, заказчики каждого уровня присутствуют в каждом регионе. В результате агрегирования по данному принципу получилось 243 категории.

Для базового прогноза получились следующие показатели:

Средний RMSLE - 0.59873

Абсолютная средняя ошибка – 5704

Средний процент ошибки – 14.7%

В результате остальных экспериментов:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LR2 | RF2 | LGB2 | CE2 | LR3 | RF3 | LGB3 | CE3 | LR4 | RF4 | LGB4 |
| Ср. RMSLE | 0.9748 | 0.6320 | 0.6320 | 0.7463 | 0.8550 | 0.6207 | 0.6207 | 0.6801 | 0.8934 | 0.6062 | 0.6062 |
| Abs error | 5433 | 5987 | 6500 | 5627 | 6848 | 5632 | 6719 | 6315 | 7714 | 6116 | 6768 |
| % | 13.99 | 15.42 | 16.74 | 14.49 | 17.64 | 14.51 | 17.31 | 16.27 | 19.87 | 15.75 | 17.43 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CE4 | LR12 | RF12 | LGB12 | CE12 | LR24 | RF24 | LGB24 | CE24 | МАЕ | MRSLE |
| Ср. RMSLE | 0.6912 | 0.8922 | 0.6055 | 0.6055 | 0.6905 | 0.8849 | 0.6057 | 0.6057 | 0.6869 | 0.5816 | 0.5769 |
| Abs error | 6866 | 7693 | 6150 | 6768 | 6870 | 7565 | 6160 | 6768 | 6831 | 2774 | 5154 |
| % | 17.69 | 19.82 | 15.84 | 17.43 | 17.70 | 19.49 | 15.87 | 17.43 | 17.60 | 7.15 | 13.28 |

**LR2** – линейная регрессия, построенная на данных за предыдущий месяц

**RF2** – случайный лес на данных за предыдущий месяц

**LGB2** – градиентный бустинг на данных за предыдущий месяц

**CE2** – комбинированный прогноз на основе прогноза полученного RF2 + LGB2

**LR3** – линейная регрессия, построенная на данных за два предыдущих месяца

**RF3** – случайный лес на данных за два предыдущих месяца

**LGB3** – градиентный бустинг на данных за два предыдущих месяца

**CE3** – комбинированный прогноз на основе прогноза полученного RF3 + LGB3

**LR4** – линейная регрессия, построенная на данных за три предыдущих месяца

**RF4** – случайный лес на данных за три предыдущих месяца

**LGB4** – градиентный бустинг на данных за три предыдущих месяца

**CE4** – комбинированный прогноз на основе прогноза полученного RF3 + LGB3

**LR12** – линейная регрессия, построенная на данных за 12 предыдущих месяцев

**RF12** – случайный лес на данных за 12 предыдущих месяцев

**LG12** – градиентный бустинг на данных за 12 предыдущих месяцев

**CE12** – комбинированный прогноз на основе прогноза полученного RF4 + LGB4

**LR24** – линейная регрессия, построенная на данных за 24 предыдущих месяца

**RF24** – случайный лес на данных за 24 предыдущих месяца

**LGB24** – градиентный бустинг на данных за 24 предыдущих месяца

**CE24** – комбинированный прогноз на основе прогноза полученного RF4 + LGB4

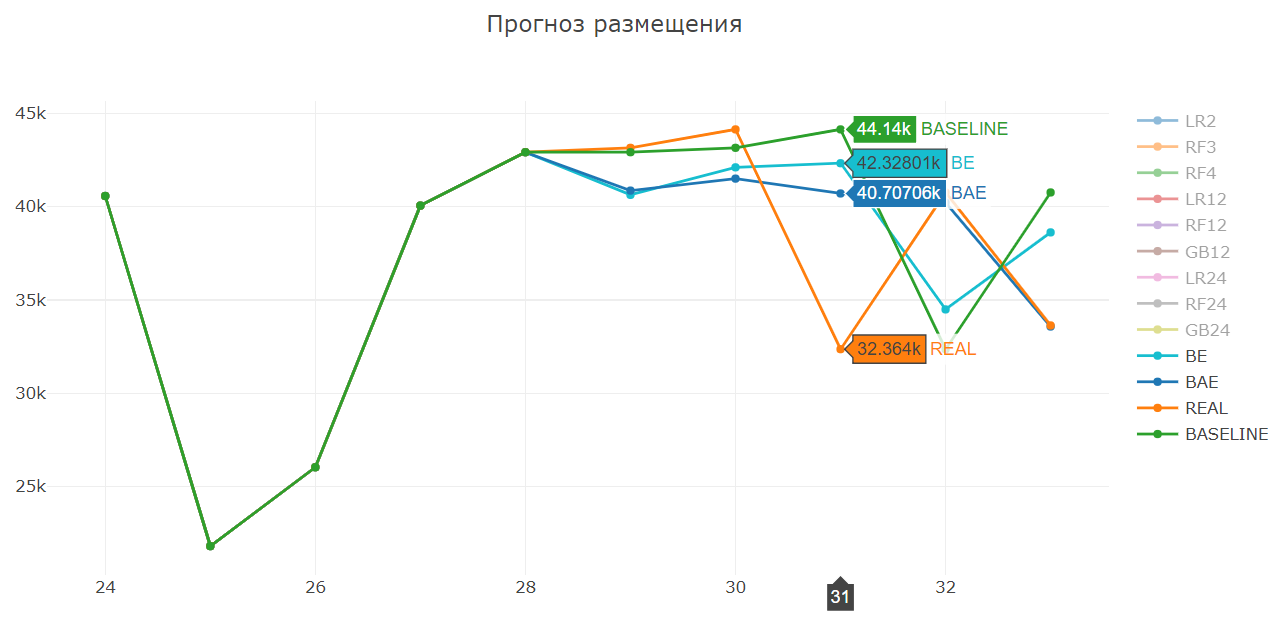
**MRMSLE** – прогноз с минимальной средней ошибкой RMSLE

**MAE** – прогноз с минимальной абсолютной ошибкой

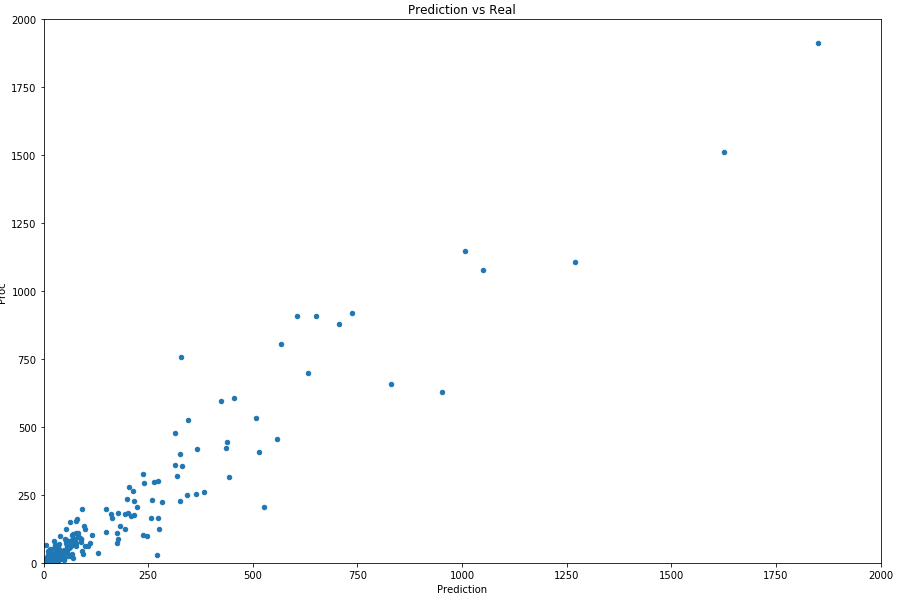
Красным цветом в таблице отмечены значения, которые хуже базового показателя. Зеленым, те что лучше.

Эксперименты показали, что наилучшие результаты по общей ошибке показывает линейная регрессия, построенная на данных за предыдущий месяц, но по RMSLE ее показатели значительно хуже baseline. Наилучших показателей удалось добиться комбинированием. Для данного датасета, лучшей комбинацией стал набор из кода, данных о закупках за позапрошлый месяц, данные о закупках, размещенных 12 месяцев назад и данные об изменениях в закупках размещенных 3 месяца назад. Прогноз на данном наборе показал **среднее значение RMSLE 0.5769, Среднее значение абсолютной ошибки 2774, что составило 7.14%** от общего объема размещения. Это лучшие показатели, достигнутые в процессе эксперимента. На мой взгляд, показатели, полученные в результате работы с данной выборкой могут служить основой для ежемесячных предсказаний.

На рисунке представлены графики реального размещения, baseline, BAE- минимальная абсолютная ошибка и BE – минимальная ошибка по средней RMSLE



На рисунке отображено распределение прогноз-значение для предсказания с набором лучших параметров (код, размещение в позапрошлом месяце, размещение год назад, изменение в размещении 3 месяца назад)



**Регионы + преференции + уровень субъекта**

После успеха, достигнутого по результатам предыдущего эксперимента. Захотелось попробовать его закрепить, добавив разбиение по наличию преференций. От заказчиков закон требует выполнения определенных требования по отношению к поставщикам, возможно, это отражается на размещении. По результатам агрегирования получилось разбиение на

Для базового прогноза получились следующие показатели:

Средний RMSLE - 0.62653

Абсолютная средняя ошибка – 5704

Средний процент ошибки – 14.7%

В результате остальных экспериментов:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LR2 | RF2 | LGB2 | CE2 | LR3 | RF3 | LGB3 | CE3 | LR4 | RF4 | LGB4 |
| Ср. RMSLE | 0.8597 | 0.6674 | 0.5711 | 0.6081 | 0.7759 | 0.6661 | 0.5653 | 0.6041 | 0.7841 | 0.6581 | 0.5600 |
| Abs error | 5561 | 7125 | 6373 | 6101 | 6746 | 6926 | 6305 | 5771 | 7372 | 5978 | 6573 |
| % | 14.32 | 18.35 | 16.42 | 15.71 | 17.38 | 17.84 | 16.24 | 14.86 | 18.99 | 15.40 | 16.93 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CE4 | LR12 | RF12 | LGB12 | CE12 | LR24 | RF24 | LGB24 | CE24 | МАЕ | MRSLE |
| Ср. 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 |

**LR2** – линейная регрессия, построенная на данных за предыдущий месяц

**RF2** – случайный лес на данных за предыдущий месяц

**LGB2** – градиентный бустинг на данных за предыдущий месяц

**CE2** – комбинированный прогноз на основе прогноза полученного RF2 + LGB2

**LR3** – линейная регрессия, построенная на данных за два предыдущих месяца

**RF3** – случайный лес на данных за два предыдущих месяца

**LGB3** – градиентный бустинг на данных за два предыдущих месяца

**CE3** – комбинированный прогноз на основе прогноза полученного RF3 + LGB3

**LR4** – линейная регрессия, построенная на данных за три предыдущих месяца

**RF4** – случайный лес на данных за три предыдущих месяца

**LGB4** – градиентный бустинг на данных за три предыдущих месяца

**CE4** – комбинированный прогноз на основе прогноза полученного RF3 + LGB3

**LR12** – линейная регрессия, построенная на данных за 12 предыдущих месяцев

**RF12** – случайный лес на данных за 12 предыдущих месяцев

**LG12** – градиентный бустинг на данных за 12 предыдущих месяцев

**CE12** – комбинированный прогноз на основе прогноза полученного RF4 + LGB4

**LR24** – линейная регрессия, построенная на данных за 24 предыдущих месяца

**RF24** – случайный лес на данных за 24 предыдущих месяца

**LGB24** – градиентный бустинг на данных за 24 предыдущих месяца

**CE24** – комбинированный прогноз на основе прогноза полученного RF4 + LGB4

**MRMSLE** – прогноз с минимальной средней ошибкой RMSLE

**MAE** – прогноз с минимальной абсолютной ошибкой

Красным цветом в таблице отмечены значения, которые хуже базового показателя. Зеленым, те что лучше.

По данной выборке градиентный бустинг и комбинация из бустинга и случайного леса давала лучший результат, чем baseline, но абсолютные значения ошибки были хуже. При выборе полей, дала комбинация из кода и показателей по закупкам за прошлый месяц. Хотя показатели улучшили baseline, но они оказались хуже, чем показатели, полученные на наборе «регион + уровень субъекта».

**Прогноз по суммам размещения**

Для базового прогноза получились следующие показатели:

Средний RMSLE – 2.24694

Абсолютная средняя ошибка – 24232млн.

Средний процент ошибки – 17.88%

В результате остальных экспериментов:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | LR2 | RF2 | RF3 | RF4 | LGB4 | LR12 | RF12 |
| Ср. RMSLE | 4.7380 | 3.2847 | 3.3197 | 3.2465 | 1.9938 | 4.5296 | 3.2573 |
| Abs error(млн.) | 24828 | 37818 | 35940 | 34550 | 25981 | 27356 | 34376 |
| % | 18.32 | 27.90 | 26.52 | 25.49 | 19.17 | 20.18 | 25.36 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | LGB12 | LR24 | RF24 | LGB24 | MRMSLE | MAE |
| Ср. RMSLE | 1.9937 | 4.4973 | 3.2789 | 1.9937 | 2.9039 | 3.1316 |
| Abs error | 25981 | 25230 | 34369 | 25981 | 37704 | 17315 |
| % | 19.17 | 18.61 | 25.36 | 19.17 | 27.82 | 12.77 |

**LR2** – линейная регрессия, построенная на данных за предыдущий месяц

**RF2** – случайный лес на данных за предыдущий месяц

**RF3** – случайный лес на данных за два предыдущих месяца

**RF4** – случайный лес на данных за три предыдущих месяца

**LGB4** – градиентный бустинг на данных за три предыдущих месяца

**LR12** – линейная регрессия, построенная на данных за 12 предыдущих месяцев

**RF12** – случайный лес на данных за 12 предыдущих месяцев

**LG12** – градиентный бустинг на данных за 12 предыдущих месяцев

**LR24** – линейная регрессия, построенная на данных за 24 предыдущих месяца

**RF24** – случайный лес на данных за 24 предыдущих месяца

**LGB24** – градиентный бустинг на данных за 24 предыдущих месяца

**MRMSLE** – прогноз с минимальной средней ошибкой RMSLE

**MAE** – прогноз с минимальной абсолютной ошибкой

Красным цветом в таблице отмечены значения, которые хуже базового показателя. Зеленым, те что лучше.

Как видно из получившихся результатов, среднее значение RMSLE baseline не удалось улучшить ни одному алгоритму. В результате подбора столбцов удалось уменьшить значение по общей ошибке, но RMSLE в этом варианте сильно превышает baseline

**Регионы + уровень субъекта**

Лучшие результаты в прогнозе количества закупок получились при использовании агрегации по регионам и уровням субъекта, попробуем провести такой же эксперимент для предсказания сумм.

Для базового прогноза получились следующие показатели:

Средний RMSLE – 2.49364

Абсолютная средняя ошибка – 24232млн.

Средний процент ошибки – 17.88%

В результате остальных экспериментов:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LR2 | RF2 | LGB2 | CE2 | LR3 | RF3 | LGB3 | CE3 | LR4 | RF4 | LGB4 |
| Ср. RMSLE | 5.5599 | 3.8372 | 2.3186 | 5.1841 | 5.4797 | 3.8638 | 2.3436 | 5.1113 | 5.3905 | 3.6987 | 2.3397 |
| Abs error(млн.) | 19675 | 32472 | 45876 | 15871 | 21862 | 37124 | 36219 | 18626 | 24426 | 38420 | 37762 |
| % | 14.51 | 23.96 | 33.85 | 11.71 | 16.13 | 27.39 | 26.72 | 13.74 | 18.02 | 28.35 | 27.86 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CE4 | LR12 | RF12 | LGB12 | CE12 | LR24 | RF24 | LGB24 | CE24 | MRMSLE | MAE |
| Ср. RMSLE | 5.0295 | 5.3718 | 3.6852 | 2.3397 | 5.0135 | 5.4015 | 3.6564 | 2.3397 | 5.0373 | 3.8462 | 4.9347 |
| Abs error | 21290 | 24394 | 38283 | 37762 | 21192 | 24747 | 38019 | 37762 | 21075 | 13468 | 12373 |
| % | 15.71 | 18.00 | 28.25 | 27.86 | 15.63 | 18.26 | 28.05 | 27.86 | 15.55 | 9.94 | 9.13 |

**LR2** – линейная регрессия, построенная на данных за предыдущий месяц

**RF2** – случайный лес на данных за предыдущий месяц

**LGB2** – градиентный бустинг на данных за предыдущий месяц

**CE2** – комбинированный прогноз на основе прогноза полученного RF2 + LGB2

**LR3** – линейная регрессия, построенная на данных за два предыдущих месяца

**RF3** – случайный лес на данных за два предыдущих месяца

**LGB3** – градиентный бустинг на данных за два предыдущих месяца

**CE3** – комбинированный прогноз на основе прогноза полученного RF3 + LGB3

**LR4** – линейная регрессия, построенная на данных за три предыдущих месяца

**RF4** – случайный лес на данных за три предыдущих месяца

**LGB4** – градиентный бустинг на данных за три предыдущих месяца

**CE4** – комбинированный прогноз на основе прогноза полученного RF3 + LGB3

**LR12** – линейная регрессия, построенная на данных за 12 предыдущих месяцев

**RF12** – случайный лес на данных за 12 предыдущих месяцев

**LG12** – градиентный бустинг на данных за 12 предыдущих месяцев

**CE12** – комбинированный прогноз на основе прогноза полученного RF4 + LGB4

**LR24** – линейная регрессия, построенная на данных за 24 предыдущих месяца

**RF24** – случайный лес на данных за 24 предыдущих месяца

**LGB24** – градиентный бустинг на данных за 24 предыдущих месяца

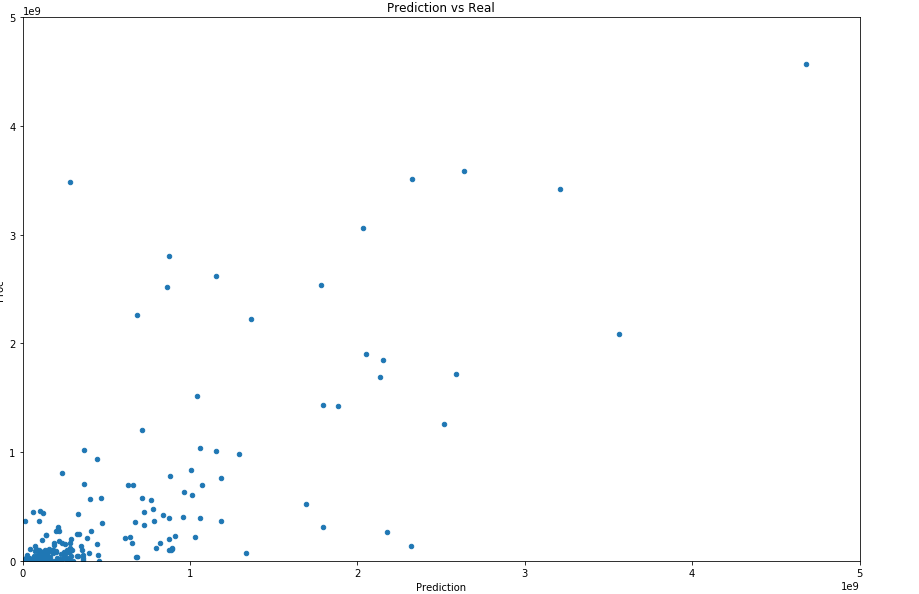
**CE24** – комбинированный прогноз на основе прогноза полученного RF4 + LGB4

**MRMSLE** – прогноз с минимальной средней ошибкой RMSLE

**MAE** – прогноз с минимальной абсолютной ошибкой

Красным цветом в таблице отмечены значения, которые хуже базового показателя. Зеленым, те что лучше.

По результатам хорошо видно, что градиентный бустинг показывает лучшие результаты по среднему значению RMSLE, однако предсказания общей суммы у него получается плохо. Лучшие результаты по абсолютной ошибке показывает совместный алгоритм, скомбинированный на показаниях градиентного бустинга и случайного леса. Самых лучших показателей по абсолютному значению удалось добиться комбинированием столбцов с суммами по процедурам проведенными 8 и 15 месяцев назад. Несмотря на то, что удалось значительно снизить абсолютную ошибку, относительно базовой, снизить RMSLE не удалось. Распределение прогноз-предсказание показывает, что на предсказание сильно влияют крупные размещения



**ВЫВОДЫ**

В результате проведенных исследования были получены следующие результаты:

1. Для прогнозирования количества размещенных закупок выбрана модель. Модель строится на данных по размещению процедур, агрегированных по региону и уровню субъекта заказчика размещающего процедуру. Модель строится на данных:
   1. Уникальный код (регион + уровень субъекта заказчика)
   2. Количество процедур, размещенных на площадке 2 месяца назад
   3. Разница между количеством процедур, размещенных 2 месяца назад и 3 месяца назад
   4. Количество процедур, размещенных на площадке 12 месяцев назад

В процессе проведения эксперимента указанная модель показала следующие результаты. Среднее значение **RMSLE 0.5769, Среднее значение абсолютной ошибки 2774, что составило 7.14%**

1. В большинстве случаев, при разных вариантах агрегирования, базовый вариант – «в следующем месяце будет так же как в предыдущем» показывает лучший результат по сравнению с машинными алгоритмами
2. В случае работы с датасетами, которые содержат большое количество строк, хорошие результаты начинает показывать комбинированный алгоритм, совмещающий значения, полученные градиентным бустингом и случайным лесом. На датасетах с небольшим набором строк, его результаты обычно не очень хорошие
3. Построить прогноз, превосходящий baseline не удалось. На общую суммы сильное влияние оказывают крупные закупки, которые могут составлять до нескольких процентов от общей суммы. Для улучшения прогноза, можно попробовать отдельно прогнозировать значения для крупных закупок (понятие крупных, требует отдельного исследования) и всех остальных.

**ВАРИАНТЫ РАЗВИТИЯ**

1. Представить модель руководству, для внедрения
2. Попробовать построить модель на данных размещения на всех площадках
3. Попробовать различные комбинации имеющихся моделей для повышения качества
4. Попробовать использовать временные ряды
5. 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

**Ссылки**

Mario Filho

How To Predict Multiple Time Series With Scikit-Learn (With a Sales Forecasting Example) <http://mariofilho.com/how-to-predict-multiple-time-series-with-scikit-learn-with-sales-forecasting-example/>