Sberbank Russian Housing Market

Евгений Патеха (1 место)

Конкурс на Kaggle
Sberbank Russian Housing Market

- Задача – предсказание цен на квартиры в Москве на основе заявок на ипотеку, полученных Сбербанком
- Объем данных:
  - Train 08.2011-06.2015 – **30 471**
  - Test 07.2015-05.2016 – **7 662**
  - Macro – **2 485**
- Метрика – **RMSLE**

  public lb – 35%, private lb – **65**%

1	**alijs & Evgeny**	**0.30088**
2	data_mining2	0.30926
3	Computer says no	0.31033
4	Sher Dil	0.31073
5	Patrick_	0.31104
6	Leonid	0.31118
7	Parag, Adam, & BreakfastPirate	0.31119
8	Pervyj_Poslednij_Ne_Predlagat'	0.31120
9	Hello World	0.31128
10	STL	0.31163

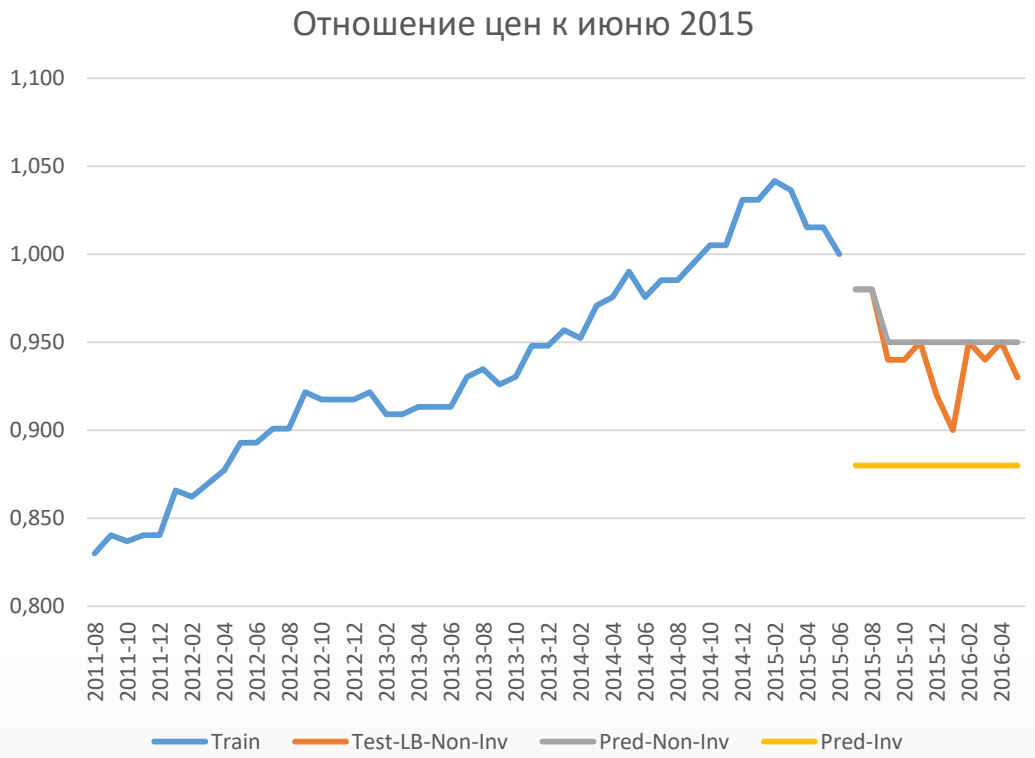
Две задачи в одном соревновании

На цену влияют разные факторы

- Макроэкономика, общее состояние рынка недвижимости
- Для новостроек дополнительный фактор – степень готовности дома и возможно фактор новой Москвы.
- Особенности квартиры (площадь, этаж, район, расстояние до метро и тд)

Макро-фактор

- Для исключения влияния макро-фактора были определены коэффициенты среднего изменения цены для каждого месяца тренировочных данных
- Далее цены для всех тренировочных периодов были приведены к уровню июня 2015г (последний месяц)
- Коэффициенты для обратной корректировки тестовых данных получены проверкой лидерборда.



Реальные данные – фактор нестабильности. Низкие цены

- Значительное количество записей с ценами существенно ниже рынка. Основная гипотеза – цены занижены для ухода продавцов от налогов
- Ошибка организаторов – оставить заниженные цены в тестовой части, поскольку точное предсказание заниженных цен бессмысленно для бизнеса банка, но при этом они значительно влияли на итоговый результат (0.15-0.20 при среднем результате в районе 0.31)
- Решение – исключить эти данные из тренировочной выборки, поскольку нет надежного способа их предсказать по имеющимся данным, а их исключение повышает стабильность модели и качество локальной валидации
- Простой способ – предсказать цены в тренировочном датасете с помощью базовой модели и исключить записи с большим разрывом с предсказанными

Реальные данные – фактор нестабильности. Ошибки в адресах. Пропуски

- Из-за ошибок в исходных адресах, заполненных клиентами, более 700 записей после геокодирования оказались на Красной площади и были исправлены организаторами по ходу соревнования
- В процессе сопоставления наших результатов были обнаружены несколько новостроек с неверным расположением, которые значительно влияли на результат (Басманное, Кунцево)
- Корректные адреса были восстановлены через сопоставление площадей квартир (в тесте точность 2 знака после запятой)
- Довольно много пропусков из-за ошибок при заполнении документов клиентами. Попытка восстановления по аналогичным квартирам.

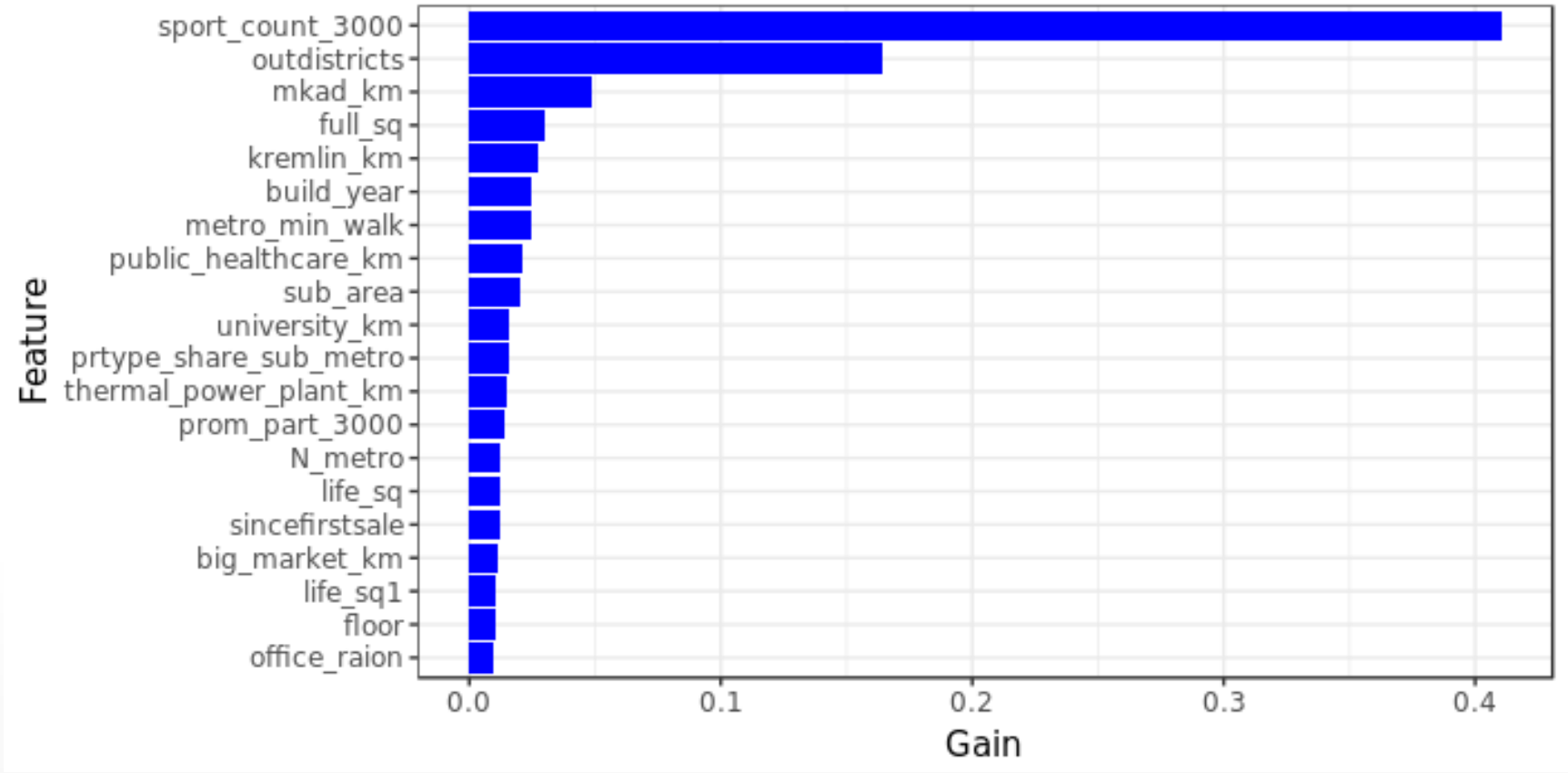
Валидация

- Валидация – разбиение по времени. Почему работала плохо as is:
  - В разные годы цены значительно отличались и макрофактор оказывал существенное влияние на скор при валидации
  - Заниженные цены распределены между периодами неравномерно и непредсказуемо влияли на валидацию, маскируя реальные эффекты
- Решение - выровнять цены между периодами через макро-коэффициенты, исключить заниженные цены. В результате для инвестиционной части SD составил .007-.008 при разбиении на 8 фолдов по полгода
- Альтернативный вариант от alijs – для неинвестиционной части нормально отработало stratified разбиение на фолды.

Признаки

- В моделях было 40-50 признаков
- Полезные новые
  - расстояние до МКАД - изменен знак на отрицательный для квартир внутри МКАД
  - внешние районы – разметил районы за МКАД как ближние (с метро, такие как Митино, Бутово) и дальние (в основном новая Москва)
  - время с момента первой продажи по данному адресу (для новостроек)
  - доля новостроек (type==“OwnerOccupier”) по району, метро
- Исходные (квартира) - full_sq, life_sq, build_year, floor
- Исходные (район) - sport_count_3000, kremlin_km, metro_min_walk, prom_part_3000, public_healthcare_km, university_km, office_raion

Признаки. Топ-20 для инвестмодели



Модели

- Отдельные модели для новостроек (OwnerOccupier) и вторичного рынка (Investment)
- Для новостроек дополнительно использовались коэффициенты готовности дома (рассчитаны как изменения медианных цен относительно апреля-июня 2015г)
- Для новостроек 2х этажный вариант – обучение на полных данных, затем обучение только на новостройках с использованием предсказаний 1й модели
- Alijs – развитие паблик скриптов, отдельные модели, использование предсказаний модели новостроек как признак для модели инвестиций, постпроцессинг – усреднение с последними известными ценами для аналогичной по площади квартиры по данному адресу, взвешенное по времени

Перспективы

- Возможно существенное улучшение точности моделей за счет следующих факторов:
  - На основе точных адресов по открытым данным (таким как reformagkh.ru) можно восстановить значительную часть пропущенных данных о квартире и доме
  - С помощью открытых данных по точным адресам можно дообогатить датасет новыми признаками (управляющая компания-ТСЖ, ставки на обслуживание, общая площадь дома)
  - Построение моделей предсказания макровлияния на рынок недвижимости и применение моделей на более близких периодах
  - Сбор дополнительной статистики по динамике цен новостроек в зависимости от стадии готовности дома
- Недалекая перспектива – переход от экспертной оценки стоимости квартир при рассмотрении ипотечных заявок к автоматическому алгоритму

Вычислительные ресурсы. ПО

- Ноутбук - 2 ядра, 12 GB
- R пакеты data.table, lightGBM

Спасибо за внимание!