Анализ открытых источников данных (aka Mass Research)

Курс рекомендуется студентам 3 курса. Все материалы, используемые в курсе можно будет найти на страничке курса на Github: https://github.com/FUlyankin/massResearch_houses

Что будем делать

Мы будем делать большой проект. Вместе. Делать его будем на примере недвижимости. Скачаем данные, предобработаем их, обогатим, отработаем все те знания, которые получили на эконометрике. Проверим несколько гипотез, которые придут к нам в голову, поищем аномалии, попробуем прогнозировать цены и многое другое! В финале нашего проекта мы сделаем интерактивную веб-страничку и, если найдём что-то интересное, небольшую заметку на habr и medium. Весь код будем сохранять в общем репозитории на github в своих ветках. Самые крутые решения будем вливать в общую ветку.

У меня есть примерный план, ко которому пойдёт наше исследование, но при этом нет чёткого сценария. Для каждой пары я готовлюсь к нескольким вариантам развития событий и поощряю любую вашу инициативу. Что конкретно мы стараемся искать в данных выбираете вы. Цепляйтесь за любые даже самые мимолётные идеи и озвучивайте их, насколько бы глупыми они вам не казались.

В примерном плане исследования те темы, о которых я хотел бы поговорить. Вы можете сбивать меня и направлять в любую сторону. Даже в сторону разной математической подноготной моделей, если вам это вдруг стало интересно.

Понятное дело, что мы не успеем сделать всё, но мы попробуем выжать из нашего большого датасета максимум.

Примерный план исследования

1. Накидываем план исследования. Обсуждаем какие гипотезы про недвижимость нам хотелось бы проверить. Регистрируемся на github, присоединяемся к репозиторию, заводим свою ветку, учимся комитить и разрешать конфликты.

2. Начинаем собирать данные. Пишем на python парсер для CIAN. Набрасываем костяк и и основной цикл для сбора. Отправляемся домой дописывать.

3. Нас всех забанило. Учимся не злить сервер и модернизируем свои парсеры. Отправляемся домой собирать данные.

4. Строим первые визуализации. Смотрим разную описательную статистику и снова

обсуждаем гипотезы. Думаем как обогатить данные для их проверки. Смотрим на открытые данные и разные API. Отправляемся домой обогащать.

5. Смотрим кто как обогатил, забираем лучшие решения в мастер. Подключаемся к google maps и качаем географическую информацию про окрестности квартир. Смотрим на Selenium, обогащаем данные поисковыми запросами. Дома доделываем свои итоговые датасеты.

6. Приводим в порядок git, который вы скорее всего сломали. Предобработка и варка фичей. Основы работы с текстами (на CIAN есть описания). Дома варим всё, что придумаем. На следущей паре делимся находками.

7. Ищем в данных по недвижимости аномалии. Думаем как поступить с аномалиями.

8. Ищем разные кластерные структуры. Размышляем о фроде и фродерах.

9. Визуализируем всё, что только можно. Смотрим на plotly.

10. Отрабатываем всё, что в первом семестре узнали на эконометрике. Оцениваем предельные эффекты, проверяем интересные гипотезы. Дома проверяем ещё!

11. Пробуем нонстопом на небольших субсэмплах построить кучу моделей. От линейных и соседей до деревьев и бустинга. Дома пытаемся улучшить результаты.

12. Обсуждаем улучшения. Пытаемся улучшить вместе. Квантильная регрессия, блендинг и стекинг.

13. Flask. Создаём свою первую веб-страницу для презентации результатов.

14. Снова наводим порядок в проекте. Доводим до конца всё, над чем начали работать. Начинаем соединять лучшие решения в одном репозитории.