**Разработка рекомендательной системы для инвестиций в коммерческую недвижимость**

Абдулин И.Н., 2023

**Введение**

1. **MVP алгоритма**

1.1 Данные

1.2 Алгоритм

1.3 Апробация

1.4 Выводы и рекомендации

1. **Proof of concept алгоритма**

2.1 User Value

2.2 Обзор методов оценки недвижимости

2.3 Модификации

2.4 Результаты разметки

2.5 Выводы и рекомендации

**Введение**

Рассматривается задача поиска интересных для инвестиций предложений коммерческой недвижимости г. Москва. В качестве зада

**1. MVP алгоритма**

**1.1 Данные**

С точки зрения моделирования апробация алгоритма демонстрирует адекватный результат – в топе находятся объекты чья стоимость

**1.2 Алгоритм**

С точки зрения моделирования апробация алгоритма демонстрирует адекватный результат – в топе находятся объекты чья стоимость

**1.3 Апробация**

С точки зрения моделирования апробация алгоритма демонстрирует адекватный результат – в топе находятся объекты чья стоимость

**1.4 Выводы и рекомендации**

С точки зрения моделирования апробация алгоритма демонстрирует адекватный результат – в топе находятся объекты чья стоимость за квадратный метр наиболее низкая. Однако стоимость всего сегмента этих объектов относительно низкая в целом. Поэтому одной из ключевых задач для повышения эффективности алгоритма является выделения сегмента интересного для инвестирования

**2. Proof of concept алгоритма**

**2.1 User Value**

С точки зрения моделирования апробация алгоритма демонстрирует адекватный результат – в топе находятся объекты чья стоимость

**2.2 Обзор методов оценки недвижимости**

С точки зрения моделирования апробация алгоритма демонстрирует адекватный результат – в топе находятся объекты чья стоимость

**2.3 Модификации**

С точки зрения моделирования апробация алгоритма демонстрирует адекватный результат – в топе находятся объекты чья стоимость

**2.4 Результаты разметки**

С точки зрения моделирования апробация алгоритма демонстрирует адекватный результат – в топе находятся объекты чья стоимость

**2.5 Выводы и рекомендации**

**Аналитический отчет по MVP (минимально жизнеспособному продукту) для поиска недооцененной коммерческой недвижимости методами машинного обучения (ML)**

1. Введение:

- Описание цели и задач проекта.

- Обзор данных и источника данных (сайт по продаже недвижимости).

- Обзор выбранного метода ML для поиска недооцененной недвижимости.

2. Предобработка данных:

- Описание процесса очистки и преобразования данных.

- Обработка пропущенных значений и выбросов.

- Преобразование категориальных признаков в числовые (при необходимости).

- Масштабирование признаков (при необходимости).

3. Feature Engineering:

- Описание созданных или использованных признаков.

- Объяснение выбора данных признаков и их роли в поиске недооцененной недвижимости.

4. Модель ML:

- Обзор выбранной модели ML для поиска недооцененной недвижимости (например, регрессионная модель).

- Описание архитектуры модели и ее параметров.

- Обучение модели на обучающих данных.

- Оценка производительности модели на валидационных данных (например, метрики качества модели).

5. Результаты и выводы:

- Анализ и интерпретация результатов модели.

- Описание выявленных недооцененных объектов недвижимости.

- Рекомендации для потенциальных инвесторов или покупателей, основанные на результате анализа.

6. Ограничения и дальнейшие исследования:

- Обсуждение ограничений и ограничивающих факторов модели.

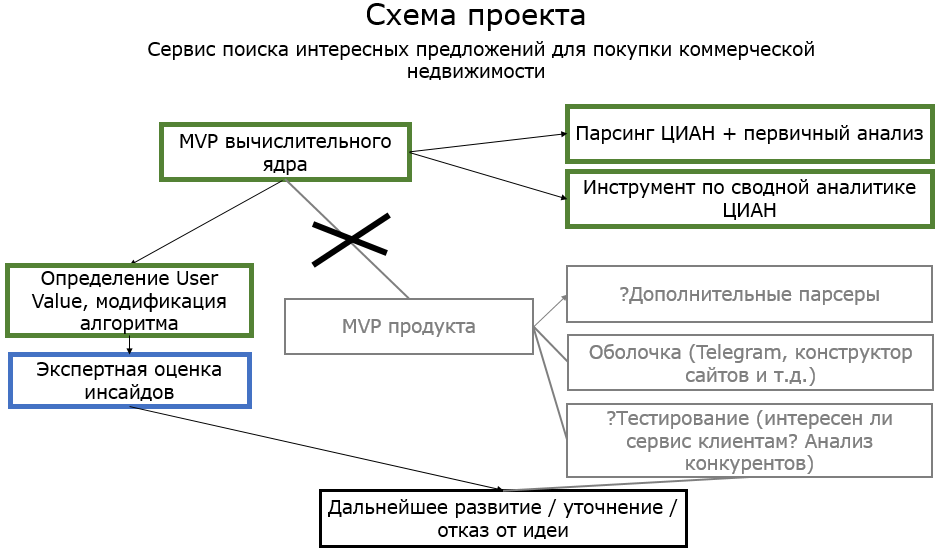
- Предложение идей для дальнейшего улучшения модели или расширения анализа.

7. Заключение:

- Подведение итогов исследования.

- Основные выводы и рекомендации.

Важно подчеркнуть, что конкретный вид и структура аналитического отчета может варьироваться в зависимости от требований проекта и целей исследования.



Спринт №1: **Парсинг ЦИАН + первичный анализ (14 апреля – 22 мая)**

**Цели:** выгрузить объявления по продаже коммерческой недвижимости третьего транспортного кольца Москвы, стоимость которых ниже рыночной оценки.

**Входные данные:** данные о продаже недвижимости в Москве с сайта ЦИАН.

**Проблемы:** наличие фейков.

**Задачи:**

1. Разработать парсер датасета с сайта ЦИАН (минимальный набор факторов: широта и долгота расположения дома, общая площадь, количество комнат, материал стен, этаж квартиры и этажность здания, адрес, округ, тип продажи);

2. Провести обработку пропусков и поиск аномалий;

3. Первичный анализ факторов, фича-инжиниринг исходя из бизнес логики (пример – расстояние до центра);

4. Провести вывод объявлений в .xlsx документ со стоимостью ниже выборочного среднего значения похожих объявлений.

**Методы и библиотеки:** Jupyter Notebook, Python 3, pandas, scikit-learn, matplotlib, geopy.

**Резюме результата спринта №1 (14 апреля – 22 мая)**

**Результат спринта:** разработан скрипт для выгрузки объявлений раз в 1-7 дней и алгоритм для оценки рыночной стоимости (вычислено по алгоритму А).

**Алгоритм А:**

1. Парсинг

1.1 Извлечение данных коммерческих объявлений по экспертным критериям - выгрузка HTML кода соответствующих страниц, результат - N страниц HTML c поискового запроса ЦИАН;

1.2 Скраббинг - выгрузка данных из HTML кода; Результат - датасет из N\*28 объявлений.

2. Подготовка данных

Очистка данных от нежелательных объектов и других шумов, преобразования факторов в нужный формат (например район в расстояние до географического центра), результат - датасет готовый для моделирования;

3. Моделирование

3.1 Сегментация данных - выделение подвыборок для обучения модели;

3.2 Обучаются предиктивные модели для оценки стоимости;

4. Интерпретация

**Результат: топ интересных объявлений по величине рыночная оценка-фактическая**.

**Обсуждение спринта №1**

**Возможные пути развития:**

* Главное - научиться отлавливать продажу бизнеса. Есть идея начать с поиска слова "аренд" из текста объявления, проблема - текст в HTML перепутан местами.
* Использовать k ближайших соседей (повысит интерпретируемость) и фактор материал стен.
* Разработка оболочки, первые наброски GUI - <http://researchmachine.pythonanywhere.com/>
* Расширение списка наблюдений.

**Ключевые идеи обсуждения:**

**Тигран:**

* Необходимо понять достаточно ли полученных факторов для извлечения практической пользы от данных.

**Алена:**

* Существуют агенства, которые следует отсеивать.

Спринт №2: **Анализ с использованием алгоритма А (Спринт 1) для извлечения User Value - эксперимент на основе доступных данных ЦИАН (22 мая – 5 июня)**

**Задачи:**

1. Устранить найденные в Спринте 1 недостатки в данных (добавить некоторые факторы, отсеять продажу бизнеса) и обновить модель. (готово 23.05)   
Подготовить результаты моделирования для оценки User Value в результатах алгоритма А. (готово 24.05)

2. Провести оценку User Value в результатах алгоритма А.

3. Оценить успешность идеи: Оценить, насколько успешно идея с алгоритмом А может извлекать User Value из имеющихся данных и выяснить причины, по которым это может быть невозможно или затруднительно.

5. Провести анализ на возможность устранения недостатков при извлечении User Value на имеющихся данных.

6. Предложить рекомендации: На основе результатов эксперимента предложить рекомендации по дальнейшим действиям, основываясь на проверке идеи с алгоритмом А в извлечении User Value. В случае недостаточной успешности идеи, предложить альтернативные подходы или доработки алгоритма для достижения ожидаемого User Value.

Спринт №3: **Анализ с использованием алгоритма А (Спринт 1) для извлечения User Value – улучшения в соответствии с рекомендациями (5 июня – 19 июня/3 июля)**

**Задача:** Провести серию экспериментов для устранения возможных проблем в соответствии с User Value и данным по экспертной разметки от Алёны.

**Резюме результатов спринтов №2, 3 (22 мая – 16 июня/30 июня)**

**Результаты спринтов 2, 3:** усиление фильтров, фичаинжиниринг и расширение сегмента привели к значительному улучшению Топа, определены границы ранее выгруженных данных с ЦИАН для моделирования (только для задачи ранжирования внутри выборки).

**Подробнее:**

I. Текущий Value of User это список из N (от 20 до 70) объявлений коммерческой недвижимости Москвы на ЦИАН, которые выбиваются на текущий момент из общей массы с учетом геолокации и других факторов, указанных в объявлении. Для достижения именно этой цели принципиальных проблем по набору факторов нет. Это может помочь, например, для мониторинга нескольких городов на предмет объявлений, которые хотят быстро «скинуть».

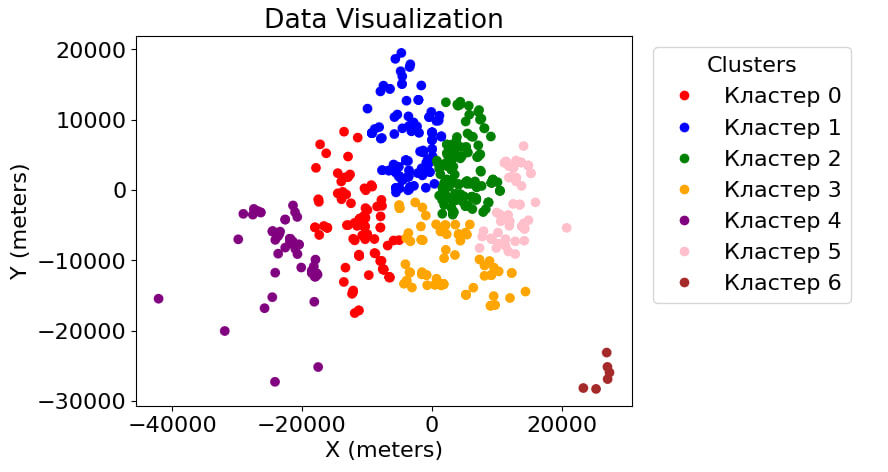
II. По профессиональным методикам оценки рыночной стоимости недвижимости нашел пособие Тепман Л.Н. (член-корр РАН) "Оценка недвижимости", 303 с., 2005 г (см. страницу 231). Хорошая книга для фичаинжиниринга. Главный вывод – имитировать профессионального оценщика нереально. Поэтому наш алгоритм выполняет только задачу ранжирования, использовать этот алгоритм для рекомендаций цены некорректно с точки зрения профессиональных методик оценки.

III. Усиление фильтров для выборки за 31.05 позволило отсеять более 40% нежелательных объявлений (111 объявлений из 253).

IV. Фичаинжиниринг позволил учитывать геоопозицию при вычислении средней стоимости. Это очень важно учитывать когда мы выделяем «средний» сегмент.

V. Алгоритм показал хорошие результаты на сегменте от 20 млн до 100млн. В топ попал одно целевой предложение с очевидной заниженной стоимостью

<https://www.cian.ru/sale/commercial/230805591/>



Другое: помимо фильтров из пункта III, попробовал убрать склады и автомойки (по описанию объявления), они сильно засоряют Топ и кажется что у них свои особенности формирования стоимости.

**Результаты оценок экспертов**

**Дополнительно: анализ влияния на рынок**