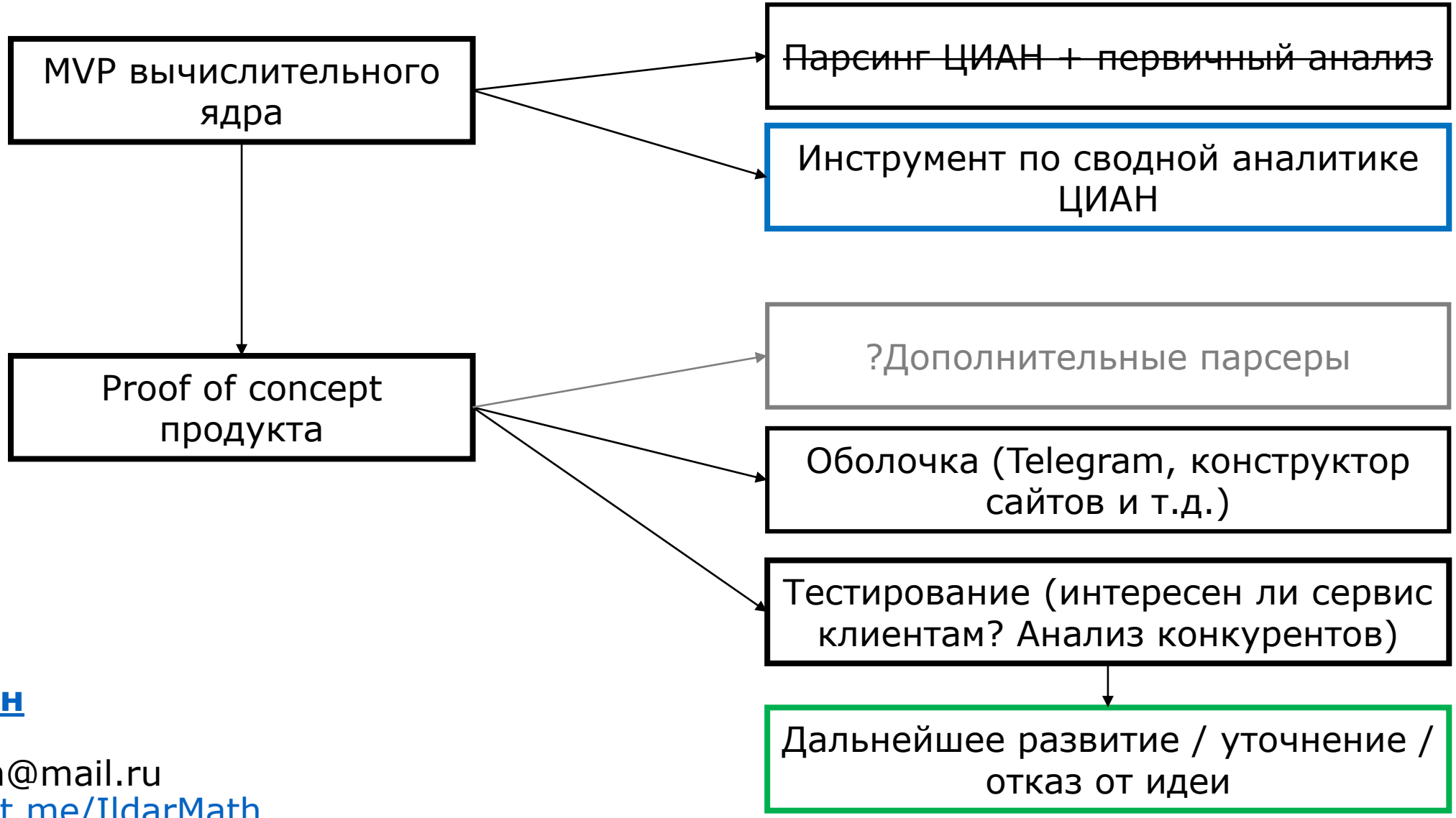
Схема проекта

Сервис поиска интересных предложений для покупки коммерческой недвижимости



Ильдар Абдулин
Контакты:
Email nakiullovich@mail.ru
Telegram https://t.me/IldarMath

Спринт №1: **Парсинг ЦИАН + первичный анализ**

Цели: выгрузить объявления по продаже жилой 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

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 с 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**

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.