



Дипломный проект на тему: «Анализ рынка недвижимости»

Слушатели:

Куров Александр Александрович

Актуальность темы и ее проблематика

На рынок недвижимости влияет множество внешних и внутренних факторов: политический кризис, нарушение цепочек поставок, спекулятивные действия отдельных игроков, и т.д.).

В этих условиях возрастает актуальность исследования и моделирования процессов, связанных с рынком недвижимости.

Цель проекта

В рамках данной работы была выбрана задача прогнозирования стоимости квартиры на основании совокупности характеристик квартиры и дома. А так как рынок динамично меняется, была поставлена цель – реализовать ETL процесс для данных по недвижимости и автоматизировать построение моделей на этих данных, а для доступа к результатам моделирования организовать витрину данных.

Задачи

- Разработка скриптов для инициализирующей и накопительной загрузки данных с публичных страниц сайта CIAN.RU
- Организация структуры хранения данных: сырой слой, промежуточный слой, слой витрины данных
- Создание baseline модели. Анализ доступных фичей и выделение наиболее значимых
- Реализация механизма для непрерывного дообучения модели на поступающих данных
- Реализация механизма для оценки стоимости квартир для тестового датасета

План реализации

Подготовка

- Формирование требований к проекту
- Первичный анализ данных

Проектирование

- Разработка блок-схемы архитектуры решения
- Проектирование структуры хранения данных
- Выбор технологического стека

Реализация

- Создание хранилища данных
- Разработка функций загрузки и обработки данных
- Прототипирование модели и анализ фичей
- Подготовка промышленной модели
- Создание docker контейнеров

Результаты

- Анализ полученных результатов и дальнейшего развития проекта

Технологии

Python

Является одним из самых популярных инструментов для работы с данными. Имеет богатый набор библиотек и интеграции с другими инструментами

CatBoost

Открытая программная библиотека, разработанная компанией Яндекс и реализующая уникальный патентованный алгоритм построения моделей машинного обучения, использующий одну из оригинальных схем градиентного бустинга. Основное API для работы с библиотекой реализовано для языка Python. Отличительной особенностью библиотеки является возможность использования в модели категориальных признаков без предварительного кодирования.

PostgreSQL

PostgreSQL отвечает требованиям надежности, скорости и удобства доступа к данным. Есть готовые образы для Docker, библиотеки для Python. Является бесплатным

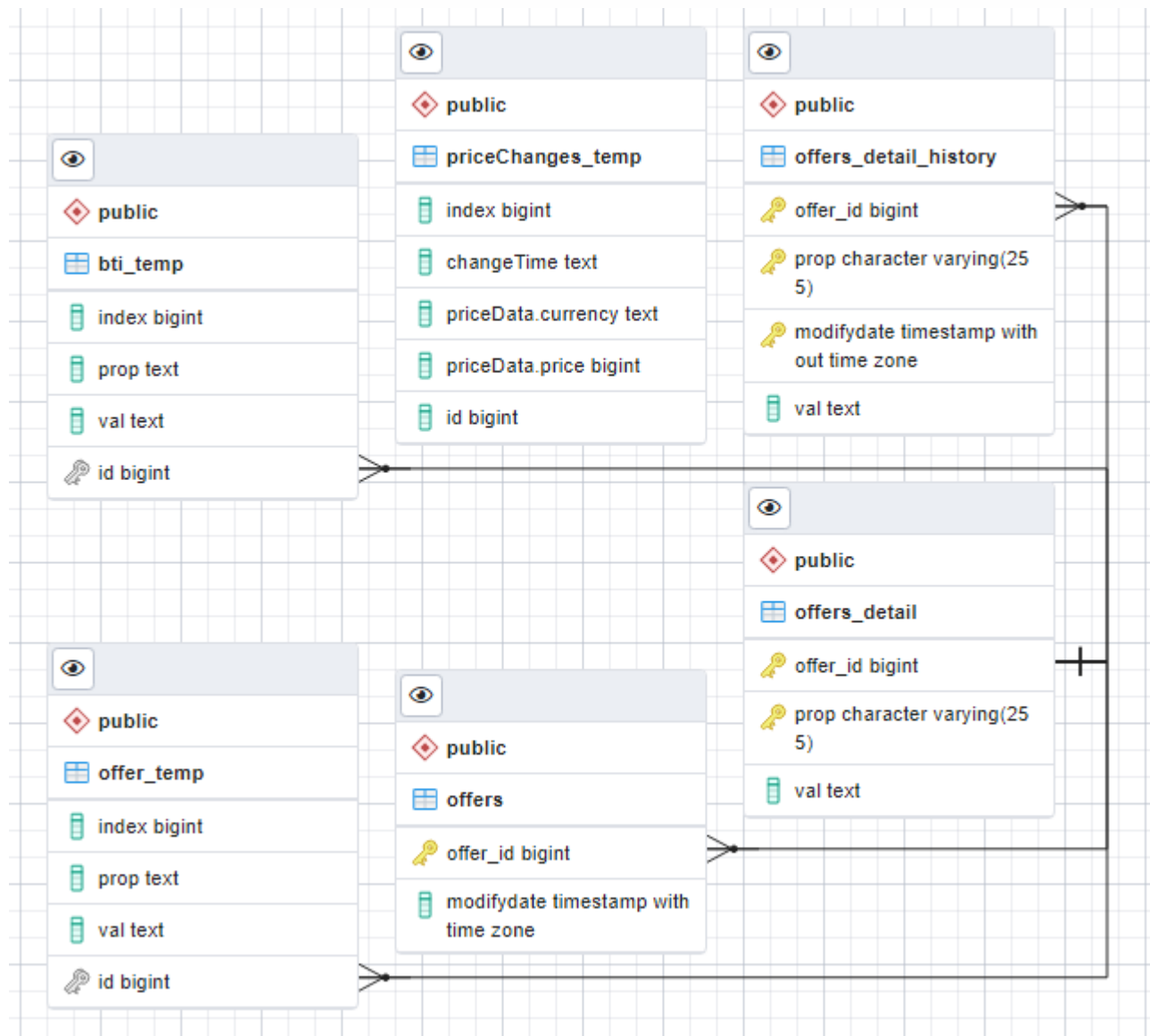
Apache Airflow

Открытое ПО для создания, выполнения, мониторинга и оркестровки потоков операций по обработке данных

Docker

Одно из наиболее популярных и доступных средств контейнеризации







Структура данных



Сущности с суффиксом «_temp» создаются динамически на основании предустановленных правил парсинга данных сайта cian.ru

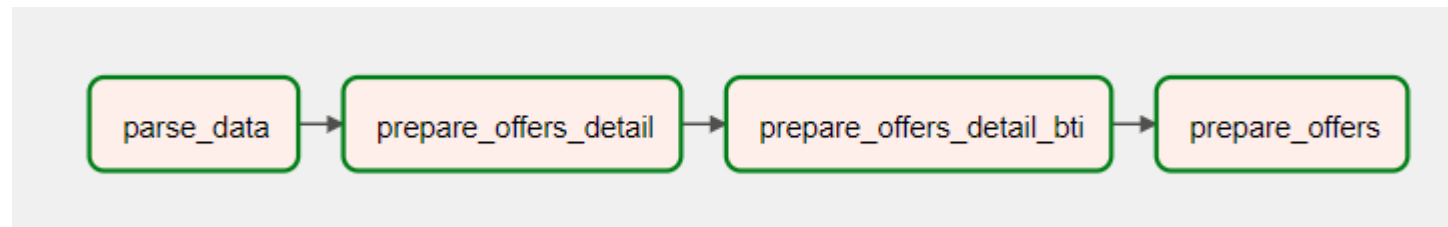
При загрузке обновленных данных по существующей квартире, автоматически пополняется таблица offers_detail_history

Схема обработки

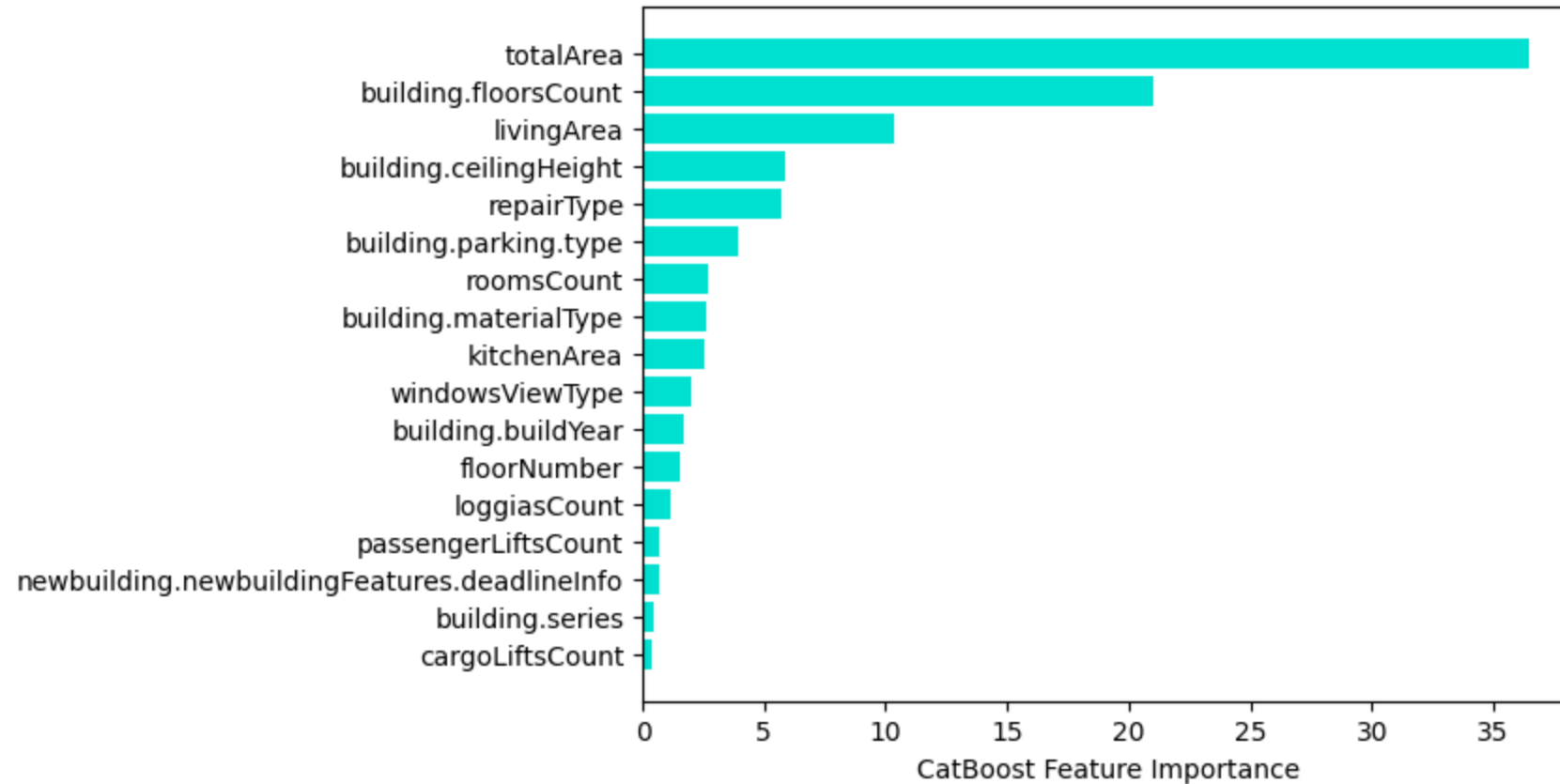
DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks
<input checked="" type="checkbox"/> full_download	airflow		<code>*/60 *</code>	2023-03-09, 02:00:00	2023-03-09, 03:00:00	
<input checked="" type="checkbox"/> model	airflow		<code>*/120 *</code>	2023-03-09, 03:06:12	2023-03-09, 03:00:00	
<input checked="" type="checkbox"/> process	airflow		<code>*/10 *</code>	2023-03-09, 03:04:45	2023-03-09, 03:10:00	

Выделено три независимых потока

- Full_download – первичная загрузка данных страниц с объявлениями
- Process (рис. ниже) – парсинг загруженных данных и заполнение витрин, впоследствии используемых для моделей
- Model – предварительная подготовка данных, построение модели, проверка качества, сохранение результатов моделирования



Модель



Для предсказания стоимости квартиры была выбрана модель **CatBoostRegressor**

Функция ошибок **RMSE**

Метрика для оценки качества модели **R2**

Baseline модель на небольшом наборе данных (~8000 наблюдений с выделением 20% на тестовый датасет) показала неплохое качество даже без работы с фичами **R2=.86**

Особенности

- Проект реализован в виде пакета `docker_compose`
- Сбор данных, парсинг и обработка уже собранных данных и обновление модели выполняются в независимых потоках
- Данные о каждой построенной модели автоматически сохраняются в `\src\model` в папке проекта
- Все настройки загрузки, обработки данных и построения модели хранятся в конфигурационных файлах. Таким образом можно менять параметры выборки, набор фичей, параметры обучения модели без изменения исходного кода

Выводы

В процессе подготовки и выполнения работы были изучены:

- Методики парсинга сайтов, особенности работы библиотеки request, принципы обхода WAF
- Методики организации ETL процессов с использованием Apache Airflow
- Принципы контейнеризации и работы с Python внутри контейнеров
- Базовые возможности библиотеки CatBoost

Итогом работы является построенная baseline модель с хорошими показателями качества ($R^2=.86$), а также механизм, реализующий автоматизированное дообучение этой модели на новых данных

Планы:

- Добавить дополнительные метрики контроля качества
- Добавить витрину с фактическими и предсказанными значения таргета
- Добавить уведомление о случаях, когда предсказанное значение таргета значительно меньше фактического. Это может быть признаком того, что приобретение данного объекта недвижимости выгодно при текущих рыночных условиях

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. <https://catboost.ai/>
2. <https://stackoverflow.com/>
3. <https://pandas.pydata.org/>
4. <https://www.zenrows.com/blog/stealth-web-scraping-in-python-avoid-blocking-like-a-ninja>