ChatHDB

Group Number: 8

Introduction

Project Background

ChatHDB is designed to assist both buyers and sellers by providing critical information to facilitate property pricing decisions, with a primary focus on **price estimation**. In addition to pricing insights, ChatHDB offers key market summaries to enhance user understanding.

To achieve this, we develop a sophisticated pricing model by integrating multiple data sources, including:

- **HDB resale transaction data**
- **Unit-specific attributes (location, size, floor level, etc.)**
- **Rental market trends**
- **Sentiment analysis of news articles**
- **Google Trends data**

Additionally, ChatHDB offers mid-term price increase projections to assist investors in evaluating potential appreciation.

Beyond pricing insights, ChatHDB enhances information accessibility by providing:

- **Concise market summaries** available on a web platform
- **An AI-powered chatbot** that assists users in retrieving key market information
- **Real-time web search capabilities**, enabling access to the latest updates on:
    - Building attributes and conditions
    - Current sentiment about the area
    - News on urban planning and development
    - Nearby amenities

Market Research

The Singapore HDB resale market faces several challenges that hinder informed decision-making for buyers and sellers:

- **Valuation Uncertainty**: Determining fair market values is difficult due to the complex interplay of various price-influencing factors.
- **Information Asymmetry**: Limited access to comprehensive market data results in knowledge gaps among participants.
- **Decision Complexity**: Beyond price, property decisions are influenced by location quality, amenities, and future developments.
- **Market Volatility**: Rapid changes in market conditions make it challenging to stay updated on trends.

These challenges contribute to suboptimal decision-making, extended transaction times, and potential financial losses.

Project Scope

ChatHDB integrates three or more AI techniques as required by the assignment:

1. **Decision Automation:**
    - Automated fair pricing estimation based on structured market data and sentiment analysis.
2. **Business Resource Optimization:**
    - AI-powered chatbot for user-friendly and efficient information retrieval.
3. **Knowledge Discovery & Data Mining:**
    - Analysis of past HDB transactions and property characteristics.
    - Sentiment analysis from news articles to assess market perception.
    - Google Trends analysis for consumer interest insights.
4. **Cognitive System Design:**
    - **Web-based platform** (hosted on Vercel) to present data in an intuitive manner.
    - **Chatbot powered by GPT** for interactive property-related queries.

Data Collection and Preparation

Data sources:

- **HDB resale transaction data**
    - Source: data.gov.sg (Housing & Development Board dataset)
    - Features: town, flat type, block, street name, storey range, floor area, flat model, lease commencement date
    - Preprocessing: calculation of flat age based on lease commencement date and time of sale; one-hot encoding of categorical data such as town, flat type, storey range, flat model
- **Rental transaction data**
    - Source: data.gov.sg (Housing & Development Board dataset)
    - Features: town, flat type, block, street name, monthly rental amount
    - Preprocessing: aggregation of monthly rental amount based on town, flat type, block, street name
- **Population income levels**
    - Source: data.gov.sg (Ministry of Manpower dataset)
    - Features: annual median monthly income
    - Preprocessing: temporal disaggregation of higher frequency income
- **Planning area data**
    - Source: API calling onemap.gov.sg
    - Features: classification of URA town planning area
    - Preprocessing: extraction of planning area name from geojson; reclassification of town planning area based on classes within resale transaction dataset
- **News articles**
    - Source: web crawling of Google News RSS xml
    - Features: article title
    - Preprocessing: sentiment analysis based on article title; aggregation of lower frequency dataset
- **Search trends**
    - Source: web crawling of Google trends
    - Features: relative search interest of the respective town over time
    - Preprocessing: reformatting of month year

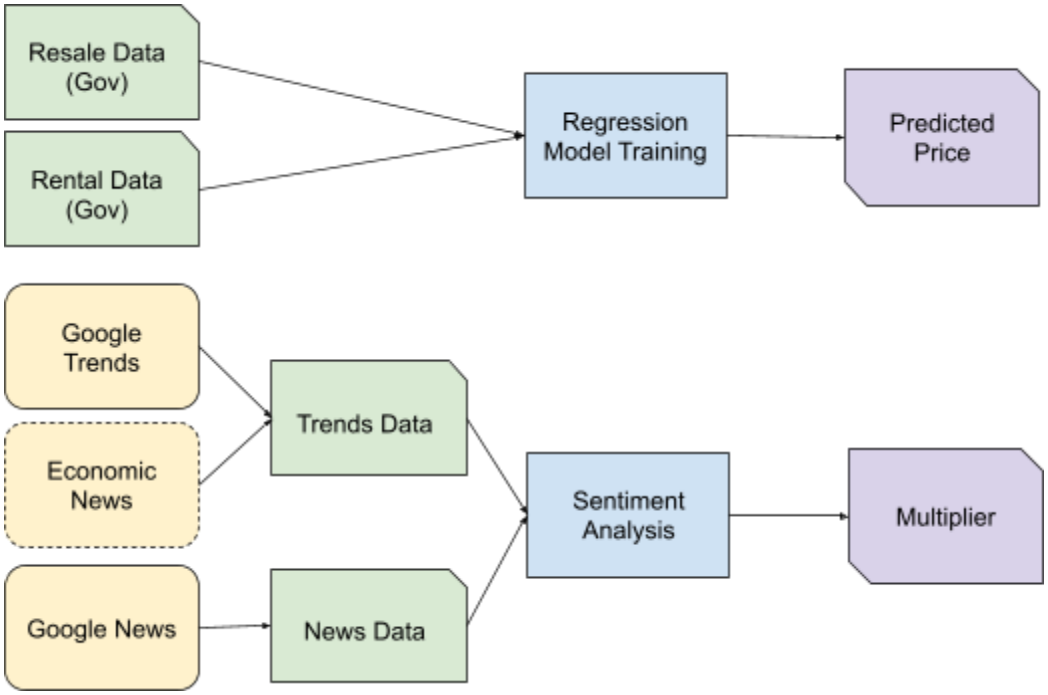
Challenges faced:

- Dataset like the news article and search trends are not readily available and require web scraping and parsing of the raw response into usable data.
- Alternative datasets are in varying levels of granularity which requires aggregation and disaggregation to map to resale transaction dataset.
- Alternative datasets are sparse and do not extend to the entire time frame of the resale transaction dataset.

System Design

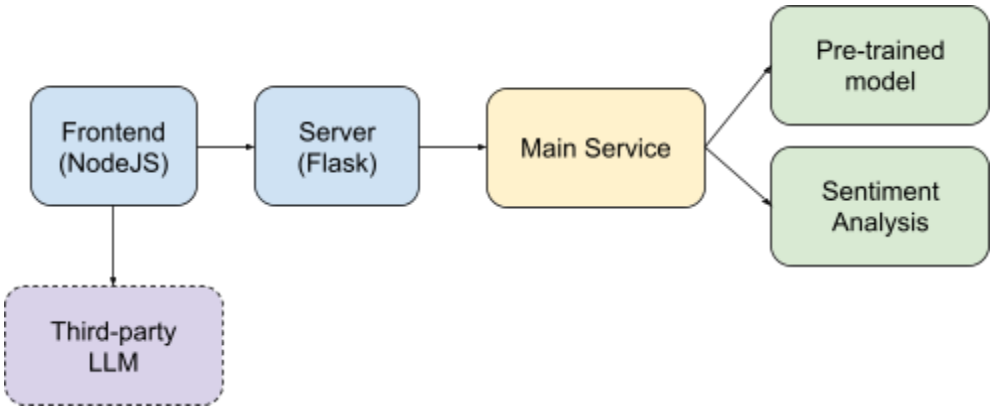
Considering the market requirements for the project, we want to focus on several aspects of the application: Price prediction, estimation of a price multiplier, as well as an LLM chatbot.

Price Prediction and Multiplier



Application Overall Architecture

We will run the frontend on Vercel using the NextJS framework, and the backend will be running on Google Cloud Run using Flask.



Implementation

We will begin by selecting suitable models and then proceed with practical steps to optimize their performance.

Since our objective is to predict both **current resale prices** and **future resale prices**, we are dealing with a **regression problem**. As a baseline model, we will use **linear regression** to establish a simple benchmark.

For our primary model selection, we have researched regression algorithms that perform well on large datasets. The three promising candidates are:

- **XGBoost** - known for its efficiency and accuracy in structured data,
- **Random Forest** - a robust ensemble method that handles non-linearity well, and
- **Neural Networks** - capable of capturing complex patterns in the data.

To determine the best model, we will conduct comparative testing across these algorithms, evaluating them based on key performance metrics: **R^2 (coefficient of determination) and MSRE (Mean Squared Root Error).**

Beyond model selection, we will implement additional steps to enhance performance:

1. **Data Preprocessing** – Handling missing values, scaling numerical features, and applying one-hot encoding for categorical variables.
2. **Model Fine-Tuning** – Optimizing hyperparameters by testing multiple configurations to achieve the best performance.

This structured approach ensures that we select and refine the most effective model for resale price prediction.

Prediction Models

Model	Description
XGBoost	XGBoost, short for extreme gradient boosting, is an optimised distributed **gradient boosting** library designed to be highly efficient, flexible and portable. It is designed for efficiency and uses decision trees for its base model. Each new tree is trained to correct errors made by the previous tree, and it has built-in parallel processing to train models on large datasets quickly.
Random Forest	Random Forest algorithm also uses decision trees to make predictions based on **ensemble learning**. It uses many decision trees to do a consolidated voting to make a prediction, where each tree is trained by

	taking a random different part of the training dataset.
Neural Networks	Recurrent neural networks (RNNs) such as **Long Short-Term Memory LSTM)** can be used to predict future housing pricing. It captures long-term contextual information in sequential data such as time-series data, which would be helpful in our project.

Multiplier Estimation

On top of using related variable data to predict pricing, we are also going to scrape and consolidate Google News for each area of housing (e.g., Woodlands,Tampines) and split these news by month and year. We will then transform this data into a time-series data and perform **sentiment analysis** on each area for a given month and year. Correlating this with the results from the price prediction model, we can estimate and check for any correlation between the Google News sentiments and the price fluctuations.

Based on the correlation between the news sentiment results and resale price fluctuations, we expect to predict a multiplier value which estimates the impact on the changes in resale price based on overall societal sentiments. We would like to test this hypothesis and see if any accurate relation can be drawn here.

However, a challenge posed here would be the scraping and obtaining of the news. Currently, we are using Google News to obtain raw data and clean them up. However, some areas in Singapore could be largely underrepresented in the news data, as they could be too new or for other reasons. We might have to extrapolate based on nearby areas if necessary.

Conclusion

Conclusion
ChatHDB leverages AI-driven decision automation, data mining, and cognitive user interaction techniques to address challenges in the HDB resale market. By integrating real-time and historical data, it empowers users with actionable insights, ultimately improving decision-making efficiency in Singapore's property market.

AI Disclaimer

This project proposal has been refined using AI for grammar correction and coherence. We also plan to utilize AI tools to assist with code debugging and brainstorming model approaches.

Dump for final proposal

1. XGBoost is chosen for current price
    a. It produces better R2 at a much faster rate than NN & Random Forest, making it easier for multiple training
2. For future price, we are testing XGBoost first before testing others
3. Optuna library is used for optimisation
4. XGBoost
    a. XGBoost – What Is It and Why Does It Matter?
    b. Keyword:
        i. Extreme gradient boosting
        ii. Boosting is method to combine weak models to create a good model
        iii. Gradient - using gradient descent to correct mistake
        iv. Extreme - in terms of speed using pruning and parallelisation
5. Optuna
    a. Bayesian Sorcery for Hyperparameter Optimization using Optuna | by Becaye Baldé | Medium
    b. Keyword: Bayesian optimisation, testing several strategy and automatically optimise based on the most promising one using statistical probability