Real Estate Price Prediction Executive Summary

Indupama Herath, Sarasij Maitra, Rafatu Salis, Ersin Süer
Github: https://github.com/Sarasij93/Erdos-DataScience-May2024-RealEstate-Project/tree/FC-Trace-main

Overview: We predict house prices in King County, WA, USA based on a number of traditional (beds, bath, square feet, etc.) and non-traditional (school ratings, crime percentages) features. We chose King County as it is a densely populated location in the US and continues to attract more settlers because of highly popular companies such as Amazon, Google, Microsoft, etc. We use recent housing datasets from Redfin as well school and crime ratings from SchoolDigger and CrimeGrade.org. We implement various models and compare their performances against a baseline model. Finally, we select our best model and build a Streamlit app that predicts housing price based on user inputs.

Stakeholders:

- Families trying to move to King County
- Real estate agents trying to estimate housing prices

Key Performance Indicators (KPIs):

- Get a baseline model of predicting the mean real estate price and use root mean square errors (rmse) to observe the improvement upon this using our models.
- Get a good understanding of the key independent features that significantly affect house price.

Data Collection, Description and Cleaning: We used data sets that include prices and features values of properties. These were downloaded from the real estate property listing website Redfin. We used webscraping via *beautifulSoup* on SchoolDigger and CrimeGrade.org for each zipcode; then we used Bayesian Average of the school ratings and added crime and school columns to the housing data. After filling out missing location values using other location data such as city and zip codes and dropping rows with missing values, the cleaned data set includes 4700 rows and 19 columns with 5 categorical variables and 14 numerical variables.

Data Pre-Processing and Exploratory Data Analysis:

- Included new features such as Age and log_price. Age = 2024 - YEAR BUILT, Log_price = log(PRICE).
- Split the dataset into training and testing.
- Performed exploratory data analysis (eda) on training set to better understand the data set. Location and property type seemed to stand out slightly as features.
- Decided to use BEDS, BATHS, SQUARE FEET, LOT SIZE, LATITUDE, LONGITUDE, Bayes_RatingSchool, crime_percentage, Age, zipcode, Property type (5 classes) as our features.
- Deleted outlier and performed *One hot encoding* on the categorical variable property type.

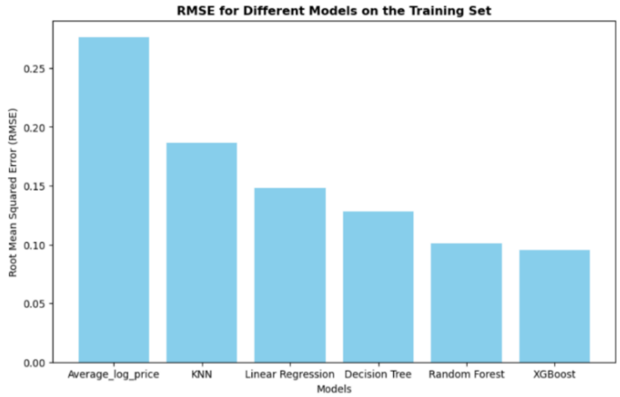
Modeling approach: We began with the baseline model of predicting the mean price of the training set. We compared this against the following models:

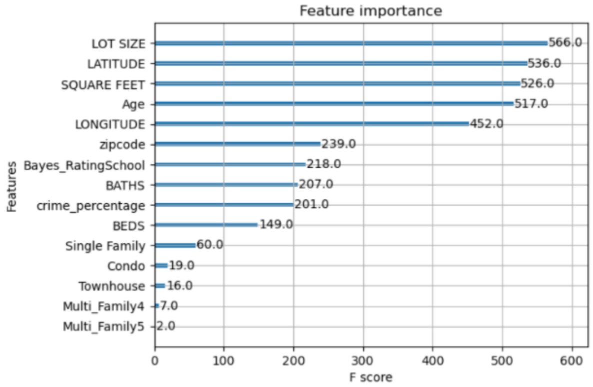
- k-Nearest Neighbors (kNN),
- Multiple Linear Regression,
- Random Forest Regressor,

- Gradient Boosting Regressor,
- XGBoost Regressor.

The kNN and the last three 'ensemble of trees' models were tuned via a gridsearch. To compare the models, we used cross-validation and calculated their average root mean squared error (rmse).

Results: Overall, all the models showed lower rmse compared to our baseline model; XGBoost model showed the best performance. We further explored the most significant features in this model.





Web Application: Using XGBoost, we built a simple web application on Streamlit that takes in user inputs (relevant to our model) and predicts the house price. The app is available at King County Price Prediction App 2024. We used modules such as *folium* to incorporate a map that shows nearby airports, big cities, etc. that hopefully gives a better idea of the desired location for the user.

Conclusions:

- Overall, all models did improve from the baseline model and did well on predicting the price.
- Contribution of non-traditional features are similar to some of the traditional features.

Future Work:

- Extend the study for other states.
- Incorporate more relevant features such as
  - whether the house has experienced flooding,
  - has mold issues,
  - the quality of construction materials,
  - the floor plan, and
  - whether fixtures and appliances have been recently updated.
- Get more data corresponding to various property types over a significant period of time to understand the effect of time as well on real estate.