Predicting House Prices - Erdos Institute Data Science Boot Camp Summer 2024

Group: Sarasij Maitra, Indupama Herath, Rafatu Salis, Ersin SÜER

1. **Project Overview and goals:**
   Our primary goal is to predict housing prices for King County, WA, USA based on a number of features using machine learning methods. Apart from traditional features like number of beds, baths, and square feet we intend to incorporate other non-traditional features such as ratings of schools nearby and crime rate of the location that could affect the prices of the houses.
   After identifying the best model for predicting prices, we plan to develop a web application to help our stakeholders gain insights into the housing market in King County.

2. **Stakeholders:**
   Families looking to settle down in King County, WA
   Real estate agents who are trying to give estimates to housing prices based on customer inputs like sqft, number of beds, baths, safety of neighborhood, etc.

3. **KPIs**
   Get the mean price of houses in King County, WA using the data and use our models to compare with that. Root mean squared error of the training and testing sets for each model we use. To further evaluate the model we can look at qq-plots and residual plots.

4. **Data set:**
   Data Collection - We used a data set that includes prices and features values of properties in King county, Washington, USA. The data sets were downloaded from the real estate property listing website Redfin. In order to add school and crime ratings we used the websites SchoolDigger and CrimeGrade.org.
   Data Description - 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.

5. **Data Pre-Processing:**
   Split the dataset into training and testing
   Perform exploratory data analysis on training set to better understand the data set. After this analysis, outliers were deleted based on price column, performed one hot encoding on PROPERTY TYPE categorical variable.
   Identify features to be used in the model. The final feature set that were used in the models is `BEDS, BATHS, SQUARE FEET, LOT SIZE, AGE, LATITUDE, LONGITUDE, Bayes_RatingSchool, crime_percentage, Age, zipcode, Property type (5 classes)`

6. **Modeling approach and Results**
   For the baseline model, we used the average log prices of properties. Five additional models were implemented. After tuning the models using a grid search with 5-fold cross validation RMSEs for the training sets were recorded.

Model	RMSE
Average log_price	0.2766
KNN	0.1865
Multiple Linear Regression	0.1483
Decision Tree Regressor	0.1277
Random Forest Regressor	0.1010
XGBoost Regressor	0.0954

The XGBoost model was selected as the final model due to its best performance. It was tested on the testing set, yielding an RMSE of 0.1003, which is close to the training error, as desired. To further evaluate the XGBoost model qq-plot and residual plots were used.

7. Web application

We also developed a simple web application that uses the XGBoost model to display predicted prices based on user input feature values.

8. Conclusion

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.

9. Future improvement

- 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.