BA 890 Professor Roy
Lyushen Song
Date: August 9th, 2024
Predicting Housing Prices in King County: A Comparative Analysis of Machine Learning Models
Collab Link: https://colab.research.google.com/drive/1jQt-53FUUOxNpk2pIYG3SUqIGRJ0FFh9?usp=sharing

Business Problem

The residential housing market has experienced significant growth in recent years, making it a crucial sector for various stakeholders, including buyers, sellers, agents, and investors. Accurate prediction of housing prices is essential for informed decision-making and strategy optimization. Traditionally, real estate professionals relied on their experience to determine housing prices, which could introduce bias and potentially misrepresent a property's true value. This research paper aims to address these limitations by employing 3 different machine learning models trained on the House Sales in King County dataset to provide reliable and objective housing price estimates.

Data Sourcing

This research paper utilizes the "House Sales in King County" dataset from Kaggle. The dataset contains information on residential home sales in King County, Washington, between May 2014 and May 2015. Key features include the number of bedrooms, bathrooms, living area square footage, zip code, and sale price. The dataset is publicly available and can be accessed through the Kaggle API. The dataset link is:
https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

Methodology

The methodology for this research involves several steps to ensure accurate housing price predictions:

- Data Preprocessing: The dataset was loaded using pandas and initially examined for missing values, data types, and basic statistics. No missing values were found, indicating a clean dataset. However, further preprocessing steps such as handling outliers and feature scaling are necessary.
- Exploratory Data Analysis (EDA): Initial EDA was conducted using pandas' describe() function to understand the distribution of numerical features. This provided insights into the range, mean, and standard deviation of key variables like price, square footage, and number of bedrooms/bathrooms. I also used visualization to understand the distribution of housing Price can be seen in the Collab file.
- Feature Engineering: Based on the initial analysis, several feature engineering steps were identified as potentially beneficial: (i) Converting the 'date' column to a datetime format and extracting relevant features like month and year.(ii)Creating interaction terms between related features (e.g., total square footage = sqft_living + sqft_basement) (iii)Adding a house_age variable by using year_sold – yr_built(iv)Using lambda function to create a new binary variable is_rennovated(v)Using Zip codes to identify the towns of each house, which is not included. The purpose of it is to reduce dimensions for the MLP. In addition, I also used VIF to check for multicollinearity and decided to drop sqft_living

and sqft_lot as they have high VIF number and keep total_sqft as it represent both sqft_living and sqft_lot.

- Model Selection and Training: Linear Regression: As a baseline model for its simplicity and interpretability. LightGBM: Good for efficiency and accuracy in handling large datasets and complex relationships. Multilayer Perceptron (MLP): To capture non-linear relationships and potentially improve prediction accuracy.
- Model Evaluation: The performance of each model will be evaluated using metrics Root Mean Squared Error (RMSE), and R-squared (R^2).

Analytical Roadmap:
The structure of this research paper follows these steps: (i). Develop Baseline Models (ii). Use Feature Selection techniques (iii) Use Hyper parameter tunning try to increase performance (iv)Deploy 'coefficients', Feature Importance, and LIME for all 3 models' interpretability.

Results
Three different models were evaluated for predicting housing prices:
1. Linear Regression Baseline Model: **RMSE: 179,955 R^2: 0.7858.** With Hyperparameter tuning RMSE: **179,956** R^2:**0.7858** 2. LightGBM: RMSE: **136098** R^2: **0.8774** With Hyperparameter tuning RMSE: **137800** R^2:**0.8744** 3. Multilayer Perceptron: RMSE: **158434** R^2: **0.8339**. The Linear Regression model served as a baseline, providing a reasonable starting point for prediction with an R^2 value of 0.7858. This indicates that the model explains approximately 78.58% of the variance in housing prices. However, the RMSE of 179,955 suggests that there is still significant room for improvement in prediction accuracy. LightGBM performed much better and have lowest RMSE **136098** and highest R-squared of **0.8774.** MLP's performance is a little better than linear regression yet not as competitive as LightGBM.

Coefficients results, Feature Importance and MLP Model Interpretability (LIME)
Taking a couple of examples from the regression results, grade13 (coefficient: 1.975e+06), which means properties with a grade_13 have a predicted price that is approximately $1,975,000 higher than properties without this grade, assuming all other variables are held constant, and it is statistically significant as p values is less than 0.05. sqft_above(coefficient: 1.376e+06) has strong positive impact to price and it is statistically significant. And City_kenmore(coefficient: -1.484e+05), Vashon(coefficient: -1.437e+05) have strong negative impact to pricing and are statically significant.
Based on the results from feature importance, variables like square footage, locations, view, number of bathrooms and bedrooms, Waterfront status, and conditions of the houses are significant factors for predicting the price of a house. These features align with common intuitions about factors that influence housing prices. To enhance the interpretability of the MLP model, LIME (Local Interpretable Model-agnostic Explanations) been used. LIME as it is model-agnostic and provide a good individual prediction. Furthermore, LIME is quite simple and fast to use. From the result of first row, city_medina and others have negative impact to the pricing. And city_medina contributes to decreasing the predicted price by about 748321.

Business Applications:
This research can help various stakeholders in the King County real estate market by providing accurate and objective housing price predictions: (i) Buyers and sellers can make more informed decisions about property values. (ii)Real estate agents can offer data-driven advice to their

clients, help their values propositions and credibility. (iii)Investors can better understand potential returns on investment properties by having a more accurate estimate of market values.

Conclusion

As MLP is a more complex model, it does perform better than linear regression yet not as good as LightGBM for this dataset. The reason why MLP is not the best performer might be the size of dataset is quite small and number of variables in the dataset are somewhat limited. Based on the graph from model comparison, I would conclude that LightGBM perform the best between 3 models as it has the lowest RMSE of just over 130000 and more than 87% of R-squared, which means that LightGBM model can capture more than 87% of variance in housing prices. If I have access to a bigger dataset and include more variables, I think that MLP would perform much better.

Limitations and Future Work

Limitations: Historical Data: The dataset only covers the period from May 2014 to May 2015. Given the significant changes in the real estate market since then, particularly due to the COVID-19 pandemic, the model's applicability to the current market may be limited. Therefore, if there are more recent data available, practically, data from post-COVID will greatly increase the performance of the dataset. Geographic Specificity: The model is specific to King County, Washington, and may not be generalizable to other real estate markets without further adaptation and training on local data. Limitation on dataset and computational strains: The dataset didn’t provide more variables that can be used to improve models’ performance, some of those variables can be distance to school, distance to nearest public transportation, and others. Therefore, the 3 machine learning models could not predict housing prices accurately despite the 3 models all have decent R-squared values. Because of computational constrains and data availability, I cannot use more advanced deep learning models for prediction and even the Multiple layer perceptron and LightGBM with hyperparameter took too long to execute.

Future Work:

(i)Update the Dataset: Incorporate more recent data, especially post-COVID data to improve the model's relevance to current market conditions. (ii)Expand Geographic Coverage: Acquiring more data from other counties or regions to create a more comprehensive real estate prediction model. (iii)Incorporate Additional Features: Include macroeconomic indicators, local amenities, and neighborhood characteristics to potentially improve prediction accuracy. (iv)Explore Advanced Techniques: Due to computational constrains, I did not use too advanced models like transformers and if possible, I want to investigate the use of ensemble methods or more sophisticated deep learning architectures to further improve prediction accuracy.By addressing these limitations and pursuing these avenues for future work, this research project can evolve into a powerful tool for understanding and predicting housing prices, benefiting a wide range of stakeholders in the real estate market.

Reference:

LIME: prashant111. “Explain Your Model Predictions with Lime.” Kaggle, February 13, 2020. https://www.kaggle.com/code/prashant111/explain-your-model-predictions-with-lime.

1. “How to Get Lime Predictions vs Actual Predictions in a Dataframe?” Stack Overflow, September 1, 2022. https://stackoverflow.com/questions/71646792/how-to-get-lime-predictions-vs-actual-predictions-in-a-dataframe.