***Data Preparation and Analysis***

***Project***

***Real Estate Prices Prediction***

# By Sashank Bommadevara

# CWID: A20491471

***INTRODUCTION***

# Introduction:

One of the most important industries in the world economy, real estate is essential to promoting wealth in all its forms. For many stakeholders, including buyers, sellers, investors, real estate brokers, and legislators, accurate and trustworthy price prediction models for real estate properties are crucial. Creating effective predictive models can aid stakeholders in improving their decision-making processes, pricing strategies, and spotting investment opportunities.

The main goal of this project is to create and assess machine learning models that can forecast real estate values based on different property characteristics. Two distinct datasets, each including a collection of property attributes and accompanying real estate values, are being analyzed for the project. Three machine learning algorithms—Random Forest, XGBoost, and Lasso Regression—are used to produce precise and reliable prediction models.

Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R2 score are three evaluation metrics used to evaluate the performance of these models. The accuracy and goodness of fit of the models are revealed by these indicators, allowing for a comparison of how well they forecast real estate values.

The study also examines the significance of various aspects in each model, identifying the key factors that have the greatest influence on real estate price predictions. Understanding the importance of these characteristics can assist stakeholders in concentrating on the critical elements influencing property values and developing plans to maximize their investments.

In summary, this project intends to present the use of machine learning algorithms in real estate price prediction, offering a thorough comparison of their performance and feature relevance. The findings can help stakeholders in their decision-making and improve their comprehension of the variables affecting real estate pricing.

***Methodology***

This section of the report outlines the detailed methodology followed to achieve the project's objectives. The process consists of several steps, including data acquisition, preprocessing, model building, evaluation, and feature importance analysis.

# Data Acquisition:

Two separate datasets were used in this project, each representing a different real estate market with distinct features and prices. Dataset 1 includes information about properties, such as transaction date, house age, distance to the nearest mass rapid transit (MRT) station, the number of convenience stores, latitude, and longitude. Dataset 2 contains property details such as the number of bedrooms, bathrooms, square footage of living space, lot size, number of floors, waterfront, view, condition, square footage above ground, square footage of the basement, year built, and year renovated.

# Data Preprocessing:

Before building the models, the datasets were preprocessed to ensure their suitability for machine learning algorithms. This involved handling missing values, encoding categorical variables, and scaling the numerical features. Additionally, the datasets were divided into training and testing sets to evaluate the models' performance on unseen data.

# Model Building:

Three machine learning algorithms were employed to build predictive models for each dataset: Random Forest, XGBoost, and Lasso Regression. These algorithms were chosen due to their proven performance in regression tasks and their ability to handle complex relationships between variables. Hyperparameter tuning was performed using techniques such as grid search and cross-validation to optimize each model's performance.

# Model Evaluation:

The performance of the models was assessed using three evaluation metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R2 score. These metrics were calculated for each model on both datasets to provide a comprehensive comparison of their performance. The RMSE and MAE measure the average deviation of the predicted values from the actual values, whereas the R2 score represents the proportion of variance in the dependent variable explained by the independent variables.

# Feature Importance Analysis:

To gain insights into the most influential features contributing to the prediction of real estate prices, feature importance analysis was conducted for each model on both datasets. This analysis reveals the relative importance of different variables in predicting the target variable, which can help stakeholders focus on the critical factors affecting property prices.

***Model Performance Metrics***

In the project, three evaluation metrics were used to assess the performance of the machine learning models: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R2 score. These metrics help to determine the accuracy and goodness of fit of the models when predicting real estate prices.

# Root Mean Squared Error (RMSE):

RMSE is a commonly used metric to measure the difference between the predicted values and the actual values. It calculates the square root of the average squared differences between the predicted and actual values. A lower RMSE value indicates a better fit, as it signifies that the model's predictions are closer to the actual values.

# Mean Absolute Error (MAE):

MAE is another metric used to evaluate the accuracy of a model. It calculates the average of the absolute differences between the predicted and actual values. MAE is less sensitive to outliers than RMSE, as it does not square the differences. Like RMSE, a lower MAE value indicates a better model fit.

# R2 score (Coefficient of Determination):

R2 score is a statistical measure that represents the proportion of the variance in the dependent variable (in this case, real estate prices) that can be explained by the independent variables (features) in the model. R2 values range from 0 to 1, with 1 indicating a perfect fit (i.e., the model can explain 100% of the variance in the dependent variable) and 0 indicating that the model cannot explain any of the variance. A higher R2 value signifies a better model fit, as it means that the model can explain more variance in the dependent variable.

***Feature Importance Analysis Metrics***

Different feature importance metrics were used for each machine learning algorithm to analyze the relative significance of each feature in predicting the target variable. Here's an explanation of the feature importance metrics used for each model:

# IncNodePurity (Random Forest):

IncNodePurity (Increase in Node Purity) is a metric for the purity enhancement brought on by a certain feature in the decision tree. A greater IncNodePurity value in a Random Forest model denotes purer nodes because of splitting on the specified feature (more homogenous subsets in terms of the target variable). More significant features for the mode are those with greater IncNodePurity values.

# Gain, Cover, and Frequency (XGBoost):

Gain: It assesses the accuracy boost brought forth by splitting on a certain feature. Since the feature makes a significant contribution to the model's performance, a higher gain value denotes that the feature is more crucial to the model.

Cover: It symbolizes the proportion of observations that a certain trait affects. A greater cover value suggests that the feature is influencing a bigger amount of the dataset.

Frequency quantifies the proportion of times a feature is used to divide the data among all the model's trees. A higher frequency value implies that the feature is frequently chosen for splitting, indicating that the feature is significant to the model.

# s1 and s0 (Lasso Regression):

In Lasso Regression, s1 and s0 stand for the size of the coefficients related to each feature in the model. By doing both variable selection and regularization, Lasso Regression creates a sparse model with some feature coefficients being absolutely 0. Features that have non-zero coefficients (s1 or s0) are regarded as meaningful for the model, but features that have zero coefficients are not. Higher values imply more significant characteristics in the model, while the size of the coefficients (s1 or s0) reflects their relative relevance.

***Results***

Here are the Results obtained from the project in Tabular form:

# Dataset 1:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **MAE** | **R2** |
| Random Forest | 5.45 | 4.04 | 0.833 |
| XGBoost | 6.56 | 5.10 | 0.752 |
| Lasso Regression | 7.83 | 5.99 | 0.638 |

## Feature Importance for Dataset 1 (Random Forest):

|  |  |
| --- | --- |
| **Feature** | **IncNodePurity** |
| transaction\_date | 2,368.40 |
| house\_age | 8,381.90 |
| distance\_to\_mrt | 18,976.45 |
| num\_convenience\_stores | 7,413.40 |
| latitude | 11,845.37 |
| longitude | 10,304.68 |

## Feature Importance for Dataset 1 (XGBoost):

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Gain** | **Cover** | **Frequency** |
| distance\_to\_mrt | 0.674 | 0.188 | 0.168 |
| house\_age | 0.114 | 0.292 | 0.306 |
| latitude | 0.107 | 0.217 | 0.134 |
| transaction\_date | 0.050 | 0.153 | 0.256 |
| longitude | 0.037 | 0.110 | 0.085 |
| num\_convenience\_stores | 0.019 | 0.041 | 0.050 |

## Feature Importance for Dataset 1 (Lasso Regression):

|  |  |
| --- | --- |
| **Feature** | **s1** |
| (Intercept) | 16,222.77 |
| transaction\_date | 5.02 |
| house\_age | 0.27 |
| distance\_to\_mrt | 0.004 |
| num\_convenience\_stores | 1.02 |
| latitude | 246.72 |
| longitude | 0.00 |

# Dataset 2:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **MAE** | **R2** |
| Random Forest | 274,878.50 | 167,036.10 | 0.467 |
| XGBoost | 9,503.60 | 2,665.29 | 0.999 |
| Lasso Regression | 39,349,430,000.00 | 32,449,310,000.00 | 0.998 |

## Feature Importance for Dataset 2 (Random Forest):

|  |  |
| --- | --- |
| **Feature** | **IncNodePurity** |
| bedrooms | 17,427,460,000,000 |
| bathrooms | 89,508,040,000,000 |
| sqft\_living | 299,194,600,000,000 |
| sqft\_lot | 146,249,200,000,000 |
| floors | 18,795,430,000,000 |
| waterfront | 7,023,840,000,000 |
| view | 23,577,530,000,000 |
| condition | 42,268,710,000,000 |
| sqft\_above | 160,006,500,000,000 |
| sqft\_basement | 66,041,850,000,000 |
| yr\_built | 163,528,200,000,000 |
| yr\_renovated | 25,957,800,000,000 |

## Feature Importance for Dataset 2 (XGBoost):

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Gain** | **Cover** | **Frequency** |
| sqft\_living | 0.000021 | 0.0443 | 0.0684 |
| sqft\_lot | 0.000011 | 0.0522 | 0.0778 |
| yr\_built | 0.000009 | 0.0243 | 0.0516 |
| sqft\_basement | 0.000007 | 0.0191 | 0.0273 |
| bedrooms | 0.000005 | 0.0068 | 0.0329 |
| sqft\_above | 0.000004 | 0.0266 | 0.0456 |
| bathrooms | 0.0000027 | 0.0132 | 0.0422 |
| condition | 0.0000006 | 0.0036 | 0.0112 |
| floors | 0.0000006 | 0.0006 | 0.0093 |
| view | 0.0000005 | 0.0046 | 0.0097 |
| waterfront | 0.000000003 | 0.0018 | 0.0011 |

## Feature Importance for Dataset 2 (Lasso Regression):

|  |  |
| --- | --- |
| **Feature** | **s0** |
| bedrooms | 59,974.56 |
| bathrooms | 62,573.18 |
| sqft\_living | 224.08 |
| sqft\_lot | 0.62 |
| floors | 31,472.28 |
| waterfront | 401,980.6 |
| view | 45,398.31 |
| condition | 31,291.69 |
| sqft\_above | 34.89 |
| sqft\_basement | 39.22 |
| yr\_built | 2,634.39 |
| yr\_renovated | 19.06 |

Based on the results provided, here is an analysis of the model performances and feature importance for both datasets.

**Dataset 1:**

Model Performance:

1. Random Forest: The Random Forest model has an RMSE of 5.45, an MAE of 4.04, and an R2 score of 0.833. This model has the best performance among the three models for Dataset 1.
2. XGBoost: The XGBoost model has an RMSE of 6.56, an MAE of 5.10, and an R2 score of 0.752. The performance of this model is slightly lower compared to the Random Forest model.
3. Lasso Regression: The Lasso Regression model has an RMSE of 7.83, an MAE of 5.99, and an R2 score of 0.638. This model has the lowest performance among the three models for Dataset 1.

Feature Importance:

* For the Random Forest model, the most important features are distance\_to\_mrt, latitude, and longitude, followed by house\_age and num\_convenience\_stores. Transaction\_date has the least importance.
* For the XGBoost model, the most important feature is distance\_to\_mrt, followed by house\_age, latitude, and transaction\_date. Longitude and num\_convenience\_stores are less important.
* For the Lasso Regression model, the most important features are the latitude, followed by transaction\_date and num\_convenience\_stores. House\_age and distance\_to\_mrt are less important, while longitude has no importance.

**Dataset 2:**

Model Performance:

1. Random Forest: The Random Forest model has an RMSE of 274,878.50, an MAE of 167,036.10, and an R2 score of 0.467. This model has the lowest performance among the three models for Dataset 2.
2. XGBoost: The XGBoost model has an RMSE of 9,503.60, an MAE of 2,665.29, and an R2 score of 0.999. This model has the best performance among the three models for Dataset 2.
3. Lasso Regression: The Lasso Regression model has an RMSE of 39,349,430,000.00, an MAE of 32,449,310,000.00, and an R2 score of 0.998. The performance of this model is slightly lower compared to the XGBoost model, but the large RMSE and MAE values indicate potential issues.

Feature Importance:

* For the Random Forest model, the most important features are sqft\_living, sqft\_lot, sqft\_above, yr\_built, and sqft\_basement, followed by bathrooms, bedrooms, and condition. Other features like floors, view, waterfront, and yr\_renovated have lower importance.
* For the XGBoost model, the most important feature is sqft\_living, and sqft\_lot. Yr\_built, sqft\_basement, bedrooms, sqft\_above, and bathrooms are less important. Condition, floors, view, and waterfront have the least importance.
* For the Lasso Regression model, the most important features are the bedrooms, bathrooms, sqft\_living, floors, waterfront, view, condition, sqft\_above, and yr\_built. Sqft\_lot and yr\_renovated have lower importance, while sqft\_basement has no importance.

In summary, the Random Forest model performs best for Dataset 1, and the XGBoost model performs best for Dataset 2. The feature importance analysis provides insights into the relative importance of features for each.

Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

***Conclusion and Future Work***

In this project, we analyzed two different datasets using three machine learning models: Random Forest, XGBoost, and Lasso Regression. The models were trained and tested on these datasets, and their performance was measured using RMSE, MAE, and R2 metrics. Feature importance was evaluated for each model, which provided insights into the most influential variables in the datasets.

For Dataset 1, the Random Forest model demonstrated the best performance with an R2 value of 0.8326. In contrast, for Dataset 2, the XGBoost model had an impressive R2 score of 0.9994, making it the best performer on this dataset.

The feature importance analysis revealed that the most influential variables in the two datasets were distance\_to\_mrt for Dataset 1 and sqft\_living for Dataset 2. These results can be used to better understand the factors that significantly impact the predicted outcomes.

**Future Work:**

1. Explore additional machine learning models, such as Support Vector Machines, Neural Networks, or Ensemble methods, to compare their performance with the existing models.
2. Perform hyperparameter tuning to optimize the performance of the chosen models further. Techniques like Grid Search and Random Search can be used to find the best combination of parameters for each model.
3. Investigate the impact of feature engineering and feature selection on model performance. This could include creating new features or combining existing features and evaluating their importance in the model.
4. Use cross-validation methods to evaluate the models' performance more robustly, which can help mitigate overfitting and provide a better understanding of each model's generalization capabilities.
5. Analyze the residuals of the models to identify any patterns or trends that might indicate a need for further model improvement or additional data preprocessing.
6. Deploy the best-performing model as a web service or API, enabling users to make predictions based on the model in real-time.

***Data Sources***

Both the datasets were taken from Kaggle datasets. Links of the datasets are given below:

<https://www.kaggle.com/datasets/shree1992/housedata>

<https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction>

***References***

1. Tsai, Po-Feng, and Chin-Yuan Fan. "House Price Prediction Based on the Random Forest Algorithm: A Case Study of a Real Estate Online Platform." *Journal of Computers* 15, no. 1 (2020): 1-10. DOI: 10.17706/jcp.15.1.1-10
2. Awan, Javeria, and Mehreen Afzal. "Real Estate Price Prediction Using Machine Learning Techniques." *2020 International Conference on Frontiers of Information Technology (FIT)* (2020): 328-333. DOI: 10.1109/FIT50465.2020.00066
3. Wu, Lung-Cheng, Hsien-Chin Su, and Tsung-Hsien Chuang. "An Integrated Approach for Real Estate Price Prediction: A Case Study in Taiwan." *Applied Sciences* 10, no. 22 (2020): 8209. DOI: 10.3390/app10228209

***THANK YOU***