A close-up of logos

Description automatically generated

Module Code – IS6052

Module Title – Predictive Analytics

Name - Vaishnavi Mansingh Deshmukh

Student ID - 124112412

Submission Date – 02/12/2024

1. **Dataset Description and Initial Exploration**

The dataset employed in this study is labeled “Ireland House Price Final.csv”, to which the variables in the model correspond to attributes related to houses in Ireland. It contains 13,320 rows and 12 columns and every row represents a property whereas every column studies common characteristics of any property. The feature description accompanied with this dataset comprises of numerical and categorical data to the concepts such as type of property, its characteristics and price estimation. These key columns include ID (property identification number), property\_scope (specification or type of property like “Land Parcel”, “Constructed Space”), location (the geographical area of the property) and price-per-sqft-$ or the price per value of property per square feet..

Initial Exploration:

At first glance, it seemeed that the data collection is quite comprehensive, but, as always, there are a number of empty cells in several rows. That is, there is one missing observation for location, 16 observations are missing for size, while a considerable number of missing entries are found in bath (73 missing observations), balcony (609 missing observations), and the price per square feet in dollars (246 missing observations) features. All these missing values will have to be dealt with at a further stage of the analysis.

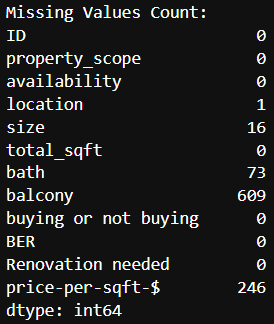


Fig 1: Missing Value Count

Statistical Summary

Certain observations can be made from looking at some basic statistical measures of the data obtained from the table. The price-per-sqft-$ varies from about $30 right up to more than $4.95 million, which would indicate that some sort of rent or price for property could be more affordable and it could also be very expensive. The bath feature has a visual mean of 2.7 bath however there is an extreme case of one or two houses having between 10-40 baths. Likewise, the balcony feature shows most properties have one or two balconies; however, some properties have more than two balconies. The total square footage (total\_sqft) distributive is quite positively skewed due to the presence of some exceptional properties among most small-sized properties.

A screenshot of a computer screen

Description automatically generated

Fig 2: Summary Statistics for Numerical Columns

Exploratory Data Analysis:

To give more information about the distribution of feature variables that are numerical and nominal, Exploratory Data Analysis (EDA) in the form of histograms and count plots is done. The analysis shows that the largest number of properties is located in the South Dublin and Fingal region; most dwellings have 2 to 3 bedrooms only. Housing renovations needed which if provided depict that a reasonable percentage of housing units demands renovation and this could cause their prices to alter.

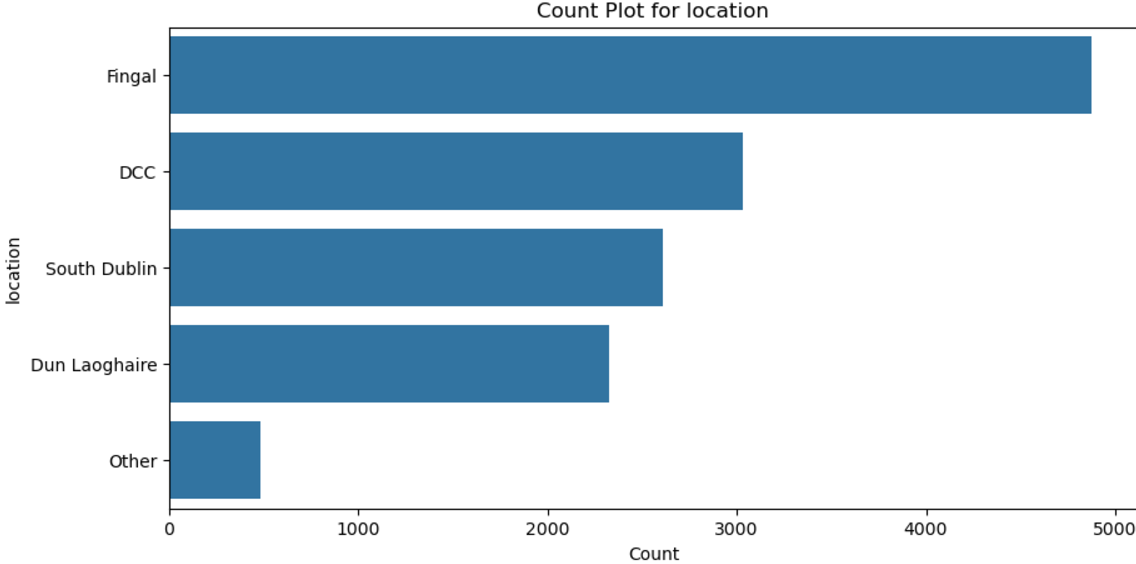


Fig 3: Count For Location

A graph with numbers and lines

Description automatically generated

Fig 4: Distributuion of Total\_sqft

Nevertheless, having the given dataset, one can infer a number of general propositions about the housing market in Ireland, and additional information clearing and data analysis are required for using them as a basis for predictive modeling.

**Correlation Analysis:**

A correlation heatmap was generated to examine the relationships between numerical features. The analysis revealed:

* Price-per-sqft-$ shows a moderate positive correlation with total\_sqft and bath, suggesting that larger properties and those with more bathrooms tend to have a higher price per square foot.
* Balcony showed a weaker correlation with other features, indicating that it may have less impact on pricing compared to other attributes.

A screenshot of a graph

Description automatically generated

Fig 5: Correlation Analysis

**Overall Insights from Dataset and Initial Exploration**

1. Dataset Overview:  
   The dataset consist of 13320 real estate listings in Ireland and includes property type, location and area, availability of amenities and price per area. The dependent variable chosen for the purpose of prediction in this paper is price-per-sqft-$, that is, price per square feet in US dollars.
2. Missing Data:  
   Features such as location, size, bath, balcony, and price per square feet – dollar, respectively will be required to handle missing values for further analysis.
3. Numerical Data Distribution:  
   Features such as location, size, bath, balcony, and price per square feet – dollar, respectively will be required to handle missing values for further analysis.
4. Categorical Data Insights:  
   Categorical variables property\_scope, location and size contain diversity therefore differ, with some categories including Location and Size as the most repeated.
5. Correlation Analysis:  
   On looking at the heatmap generated it could be seen that total\_sqft had highest correlation with price-per-sqft-$ which matches our understanding that property size and price per sq ft should have direct. relationship as larger properties tend to be costly. All the other variables such as bath and balcony had comparatively low coefficients of determinant with the target variable.
6. **Data Preparation and Feature Engineering**

The data preparation process involved cleaning, standardizing, and enhancing the dataset to ensure it was ready for analysis.

Step 1: Data Standardization

The first process involved operating on the dataset and the first of such operations was normalization of several columns. If the property size was given in terms of different units such as acre, square meters, square yards etc they were converted into square feet (sqft) metric. In the size column numerous values were identified and extracted into pure numbers allowing similar entries to be maintained. Further, to make the comparison easier, the availability column was normalized to make all the text lower case and replace less informative date-like entries like ‘18-Apr’ with more informative form like ‘Available from April’.

A screen shot of a computer

Description automatically generated

Fig 6: Cleaned Data Stats

Step 2: Handling Missing Values and Outliers

Before going through imputation process, a boxplot graphical representation was generated to check for the presence of outliers across each of the numerical features. No outliers were omitted but Missing value imputation excluded outliers. The gaps in numerical variables were predicted using the mean value for most of the variables and median for particular variables such as “price/sqft”. For categorical data, the missing values were replaced where necessary with the mode (the most important value). This made the dataset accurate in the sense that any missing values in the response examination were filled, but no outliers affected the imputation steps.

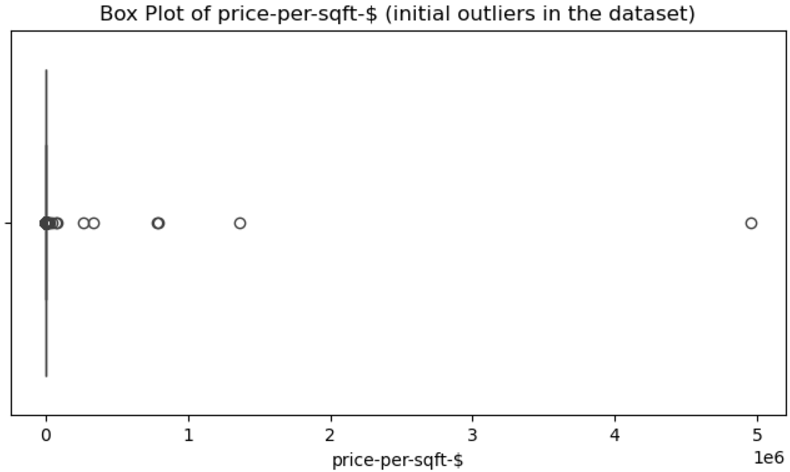


Fig 7: Box Plot Price-per-sqft (before removing outliers)

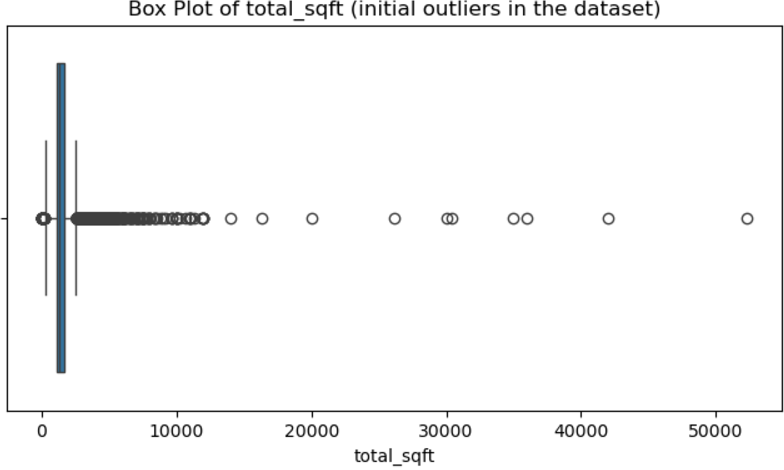


Fig 8: Box Plot Total\_sqft (before removing outliers)

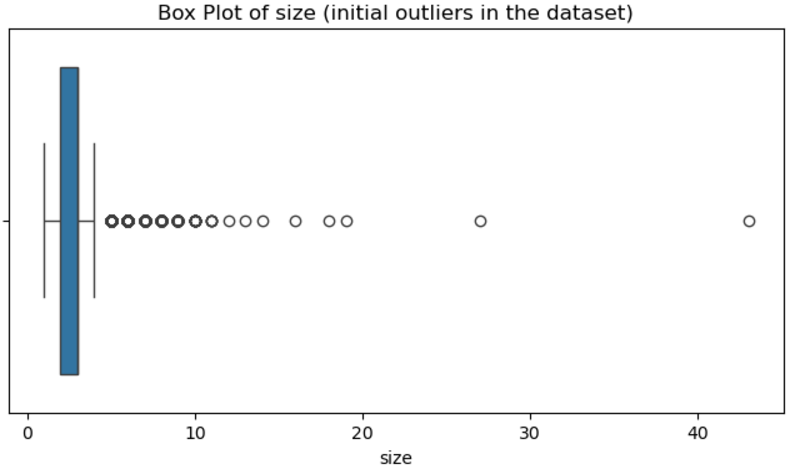


Fig 9: Box Plot of Size (before removing outliers)

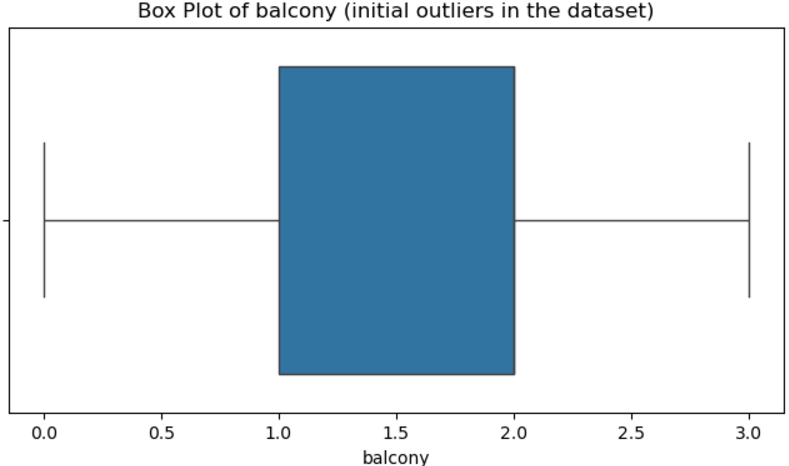


Fig 10: Box Plot of balcony (before removing outliers)

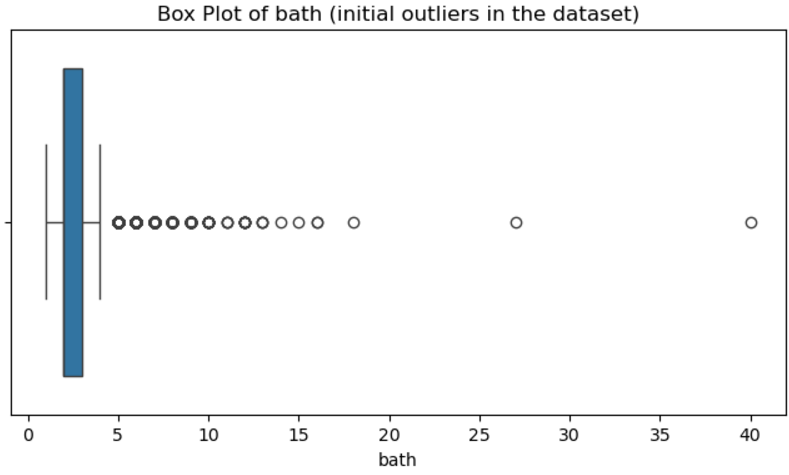


Fig 11: Box Plot of Bath (before removing outliers)

A screen shot of a computer

Description automatically generated

Fig 12: Missing Values Summary After Filling

Step 3: Feature Engineering

By doing the feature engineering we have been able to enrich the dataset in terms of a higher predictive value. A new total\_price feature was derived by simply multiplying the total\_sqft column by the price by the square-feet-/dollars. Thus, this new feature can be used as the variable for the potential models that would estimate the price of properties. By completing these steps, the dataset was thoroughly prepared: The deleted values were managed, by removing them during the imputation process, Outliers were also omitted for imputation, while new features were introduced that would add meaning. This made it possible for the data to always be ready for more refined results and more efficient data modeling and analytics, increasing the relevance of the results for prediction.

A screen shot of a computer

Description automatically generated

Fig 13: New Features Created

1. **Outlier Detection and Handling**

Outlier occurs when there is a significant difference between a data point and the real value for a particular test or assessment instrument This is applicable in learning institutions where a student’s score is extremely different from the actual score for a particular course Thus the conversation Outlier detection involves recognition of the point which deviates from the typical results In handling the same it means how to go about identifying the point which has greatly differed from the normal results of the test or examination Outliers in the dataset were also checked to improve result quality and filter out distortions when building models for making predictions. Widely known facts like ‘gross outliers’ which are extremely valuable data points considerably different from most other values comprise a major threat towards ‘skewing’ predictive model precision if not handled suitably. To identify these instances, the authors used the most commonly applied technique known as Interquartile Range (IQR).

Identifying Outliers:

Once the outliers were identified I like to calculate the first quartile (Q1) and third quartile (Q3) for each numeric column for the dataset at hand, and compute the interquartile range (IQR) that is Q3 – Q1. In cases where data in analyzer a, b, c above was less than Q1 - 1.5 IQR, or greater than Q3 + 1.5 IQR, thesaid data was considered an outlier.The following is a summary of the results of analyzing the above outlined variables: This method is useful in making an early estimate of the outcome of the analysis since outliers can so much affect the result.

Handling Outliers:

I took them out from the performances and the dataset I used. This was done for all the numeric columns basically size, total\_sqft, bath, balcony and price per sqft. Exclusion of the outliers was very useful to reduce the effect of unreliable observations on the overall data analysis.

Post-Removal Evaluation:

Upon the deletion of the outliers, I plotted a box plot for each of the numeric features. The means, median, and box plots let me to verify that the outliers were omitted, it was clear that the values were more reasonable. Also, the summary statistics of the dataset displayed more reasonable ranges suggesting effectiveness of the method that was used to remove outliers.

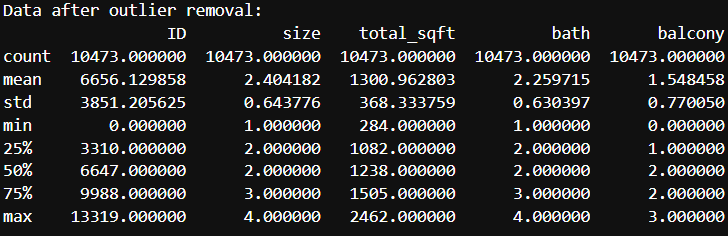


Fig 14: After removing outliers Stats

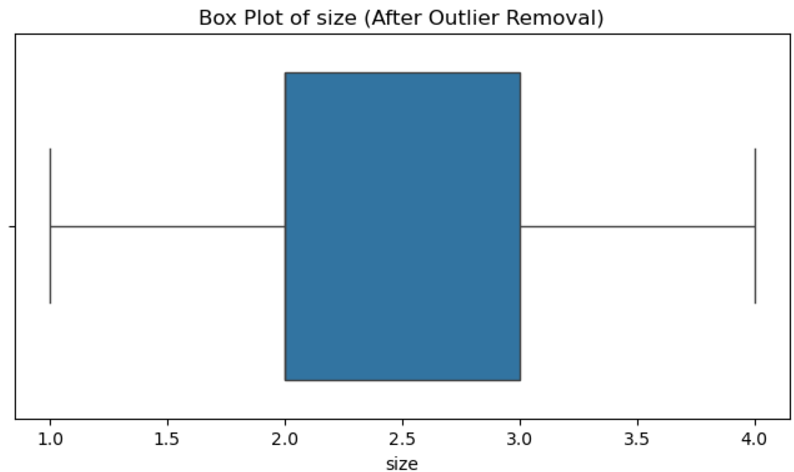


Fig 15: Box Plot Size (After removing outliers)

A blue rectangular bar graph

Description automatically generated

Fig 16: Box plot of Total\_sqft (After removing outliers)

A blue rectangular object with white text

Description automatically generated

Fig 17: Box plot of Bath (After removing outliers)

A blue rectangular object with white text

Description automatically generated

Fig 18: Box plot of Balcony (After removing outliers)

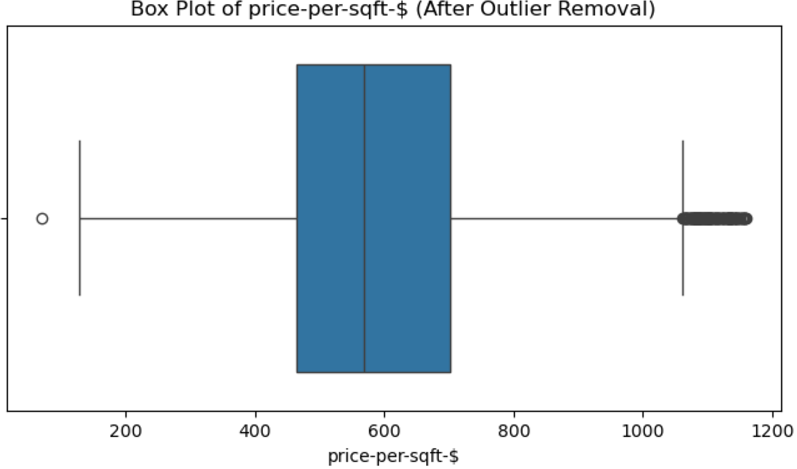


Fig 19: Box Plot of Price-per-sqft (After removing outliers)

Impact on Model Performance:

Outliers reduce the accuracy of a dataset for predictive modeling and therefore it was easier for analysts to make those datasets more conducive to their use. If the dataset contains outliers this can influence the results, and the model will perform a poorly fit prediction for these sets of values. These values were useful in preprocessing data but excluding them eased the development of more dependable models from the data. Therefore, there was an expectation of an enhanced performance of the consequent analyses and the resulting approximate models, enabling the provision of more accurate results.

**Pearson correlation:**

Subsequently, since the outliers were eliminated from the data, the Pearson correlation test was applied to identify the nature of linear dependencies of the primary numeric variables. Pearson correlation is useful for determining in which extent two variables are correlated, and runs from -1 (correlated negatively) to 1/0 (correlated positively). It ranges between -1 and 1; the former showing no linear relationship at all.

Pearson Correlation Insights:

1. Total\_sqft and Size: The three rows suggest a very positive relationship between the total\_sqft and the size. This makes sense if you consider that higher order properties have tendency to consist more square footage as is observed in the case of housing statistics.
2. Price-per-sqft and Size: The next trend we assessed on the list was the value of case by size which was positively correlated at a moderate level with price-per-sqft. Prices of the properties are also found to depend on the size of the property; however, the correlation between size and price per square foot is not very strong, so it can be assumed there may be other factors affecting the price more, like location, amenities, or state of the property.
3. Price-per-sqft and Total\_sqft: It shows a fairly poor coefficient of determination between price-per-sqft and total\_sqft, meaning that total area of the property does not have much bearing on price per square foot. This also means other attribute (such as location or property type) may have a stronger influence on the price per square foot.
4. Price-per-sqft with Bathrooms and Balcony: The coefficient for number of bathrooms and balconies present very low correlation with price per sqft so it means that irrespective of the number of bathrooms that may be there or even if the property comes with a balcony, there isn’t much variation in the price per sqft in the dataset used here. This could be an indication that these features are less sensitive to price determination compared to say the size of the property.

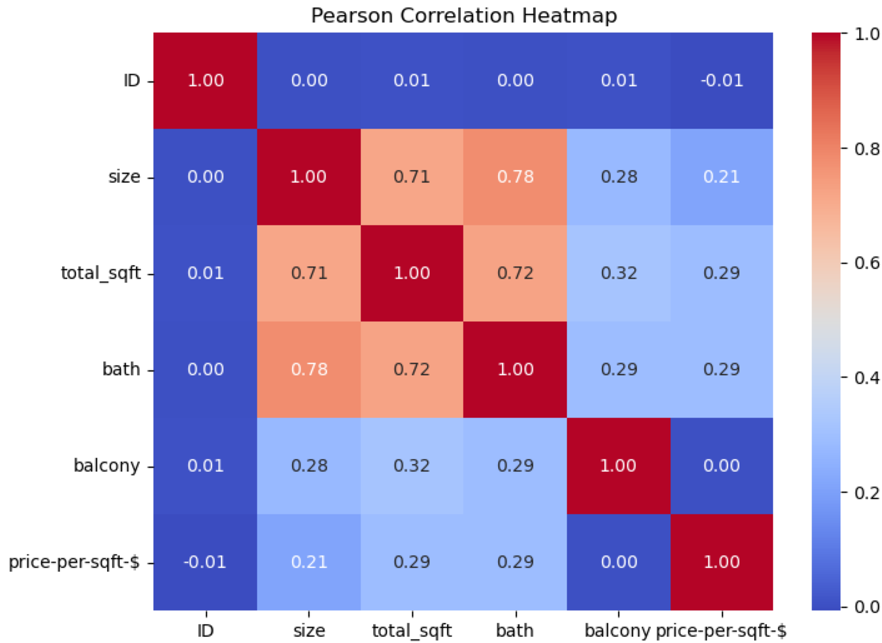


Fig 20: Pearson Correlation after Filling Missing Values

**Key Insights:**

By doing so, the Pearson correlation analysis can give useful information about the interdependency of some of the properties of the property. Although, there is clear association between total\_sqft and size predictors, the influence of all features including; number of bathrooms and balcony on price\_per\_sqft is not significant. Such research would imply that other variables which affect the price include locality and state of properties, among other considerations. It is particularly important for the heuristic model—not to be confused with a predictive model—by eliminating outliers and making the data more understandable, one can achieve better or more accurate results for an actual predictive model.

1. **Predictive Analysis:**

For the predictive analysis, I used three models such as Random forest, Linear Regression and Support Vector Machine on different technique to make exact price of property.

1. Random Forest

It was established that Random Forest model had the highest performance. Firstly, I employed the histogram bar chart to analyze the accuracy of all the three models. Random Forest gave an accuracy of 80.03% and the lowest RMSE of 160.33 making it the best model of the lot. This model proved the ability of the given approach by capturing the data characteristics well and mitigating overfitting caused by multiple features. To rectify this, I used RandomizedSearchCV to fine tune the model for other parameters which include the number of trees and the depth of trees.

2. Linear Regression

Linear Regression which was used as a simpler baseline gave inferior results as compared to Random Forest. Even though the model was not as accurate as the Random Forest in terms of accuracy, which was 77.70%, or the RMSE of 168.45, it deteriorated as compared to the Random Forest model. Linear Regression was used yet it began to perform poorly mainly because of skewness of the data with a log transform used on the dependent variable. It gave fairly reasonable predictions, however, as a basic model, its efficiency suffered in comparison to such models as, for instance, Random Forest.

3. Support Vector Machine (SVM)

The results shown by SVM were comparable to those achieved by Linear Regression with 77.73% of accuracy and 169.14 RMSE. With the default setting, the model was not that impressive so, I used random search to optimize its parameters a bit more. Still, it was not as efficient as Random Forest resulted from this evaluation.

INITIAL MODEL ACCURACY

A screenshot of a computer screen

Description automatically generated

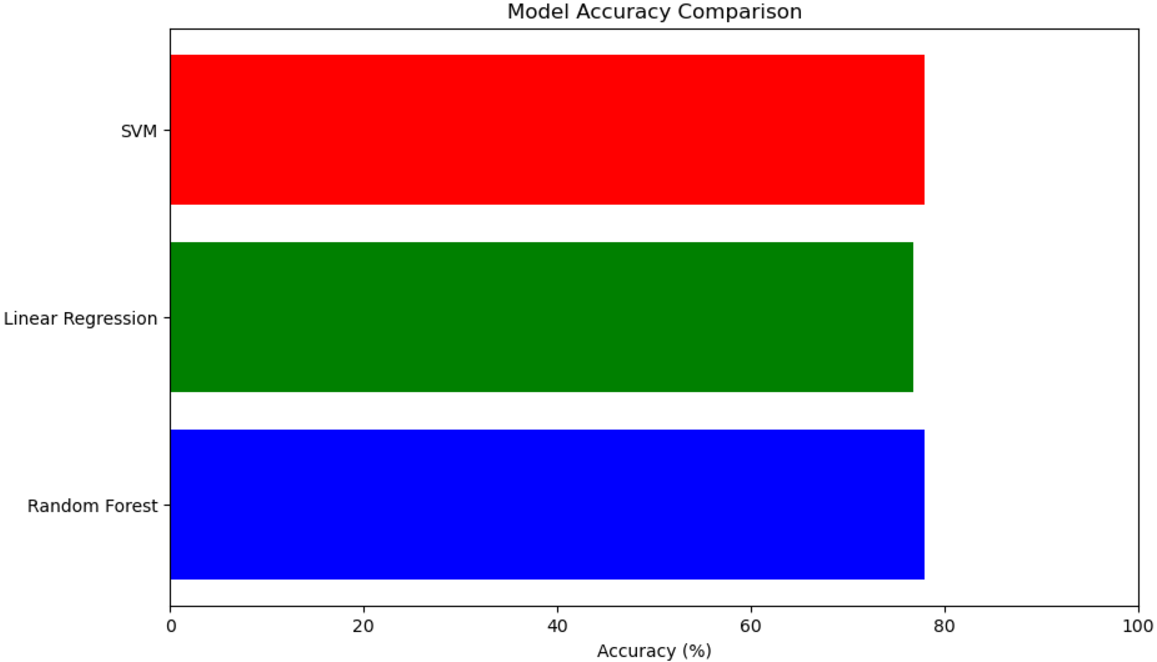


Fig 21 : Accuracy Comparison of Models

**Model Training and Accuracy:**

In the remaining of the analysis, the first step was to splitting the dataset into train and test, then standardizing the features and finally applying Principal component analysis (PCA) for dimensionality reduction in order to enhance the efficiency of the models. For hyper-parameter tuning, RandomizedSearchCV was used to ensure the best was chosen to increase model accuracy.

1. Random Forest: This featuring model achieved an average accuracy of 79.03% after adjusting other parameters such as n\_estimators, and max\_depth. The values of the cross-validation sum of squares were 258988 and 522241 for the training and testing datasets, respectively Remote mosaic score error RMSE = 160.33, good generalization and minimal overfitting can be observed..
2. Linear Regression: Linear regression achieved an accuracy of 77.70% But, RMSE is little bit high, which was 168.45 of this and it indicate this model was performed reasonable but not able capture all the interactions of the data as much as the Random Forest.
3. SVM: The accuracy of this model was 77.73% with an RMSE of 169.14 as was seen with Linear Regression. SVM had slightly lower accuracy simply because of its sensitivity towards parameters and the chosen kernel.

A screen shot of a black background

Description automatically generated

Fig 22 : Accuracy Of Models After Tunning

**Comparison and Insights:**

Random Forest Model was the most effective of the analyzed techniques both in terms of accuracy and Root Mean Square Error. Where PCA contributed to the amelioration of the situation was through enhancing the efficiency of the models With reference to the improvement of the accuracy of all models HV was responsible. Analyzing the model performances through a bar chart, it was seen that Random Forest the best one for this data set.

A red green and blue stripes

Description automatically generated

Fig 23 : Accuracy Comparison of Models after Tunning

**Conclusion on predictive techniques:**

Analyzing this project makes it clear that techniques like Random Forest are useful when solving practical problems in the real estate market. Linear and SVM models offered good results, yet it was not better than Random Forest model. The proper usage of dimensionality reduction and hyperparameter optimization drastically increased the modelling predictiveness and Random Forest was identified as the most suitable for this dataset model.

1. **Results, Evaluation, and Discussion**

Presentation of Results

The findings from the predictive analysis were assessed using various performance measures headed by RMSE and MAPE. I employ histogram bar chart to visually compare the models and the chart indicates the accuracy of the models. Random forest again showed the highest level of accuracy standing at 79.03% followed by Linear Regression accuracy that stood at 77.70% and SVM at 77.73%. Similarly, examining the values of RMSE,

Critical Discussion of Model Effectiveness

When comparing the results of the two models to identify the best performing model, Random Forest model was found to the best choice since it can detect both linear and non-linear relationships within the data. Its major disadvantage is that it is less clear when trying to interpret it than other methods are. While Linear Regression is accurate when the connection between variables is direct and relatively uncomplicated—the model doesn’t perform well with multi-variable and nonlinear data as this one. This is true because when tested with Self

Evaluation of Predictive Techniques

* Random Forest:
  + a) Yes, it is great at capturing complex relationships, making it suitable for datasets with non-linear features.
  + b) It’s harder to interpret compared to Linear Regression since it uses many decision trees.
  + c) Random Forest handles large datasets well but can become slower with too many trees.
  + d) While explaining the model is not easy, we can gain insights from feature importance.
  + e) Random Forest works well in various scenarios, especially with complex and non-linear relationships.
  + f) It requires careful tuning of hyperparameters, like the number of trees and depth, to avoid overfitting and to achieve optimal performance.
* Linear Regression:
  + a) It is limited to linear relationships, which can make it less useful for complex datasets like this.
  + b) The results are easy to interpret because the coefficients directly show the impact of each feature on the target variable.
  + c) Linear Regression is very efficient with large datasets, as it doesn’t require much computation.
  + d) Communicating results is straightforward since the relationships are clear and simple.
  + e) It struggles with complex or non-linear data and is best suited for linear relationships.
  + f) Linear Regression is sensitive to outliers and multicollinearity, so preprocessing steps like outlier handling and feature selection are necessary for better performance.
* SVM:
  + a) SVM can model non-linear relationships but tends to be computationally expensive.
  + b) It is harder to interpret than Linear Regression, though it’s possible to visualize decision boundaries in simple cases.
  + c) SVM can handle large datasets, but its performance might suffer with very large or noisy data.
  + d) Explaining SVM’s results is more difficult due to its complexity, though techniques like support vectors help in understanding it.
  + e) SVM is strong with non-linear relationships but might not perform as well with large or noisy datasets.
  + f) SVM requires proper parameter tuning (e.g., kernel choice, C value) and scaling of features to perform optimally, as it is sensitive to these aspects.

5. Feature Engineering and Outlier Handling Contribution

Feature engineering and outlier handling significantly improved model performance. By transforming the target variable (price) and removing extreme outliers, the models became more robust and accurate. Without these steps, the models would have been heavily affected by skewed data and outliers, resulting in less reliable predictions. This demonstrates the importance of proper data preparation in predictive modeling.

**Real-World Application**

As seen at the previous section, Random Forest was the most appropriate model for real-world application such as property prices, which demonstrate a complex and non-linear nature similar to real estate scenarios. In the mean time, Linear Regression is good when the problem is simple and the relationship between data points are linear; however, this is not the case when the data increases in complexity. SVM can be helpful, especially when working with high-dimensional data but as the number of data points rises, for instance property data, SVM might not do as well. Therefore, all being equal, Random Forest is the least biased towards this dataset and offers all the versatility that other models do wrt data patterns and is most accurate as opposed to the contrary.

**Conclusion**

The analysis highlighted the effectiveness of Random Forest for predicting property prices, with Linear Regression and SVM providing useful insights in specific cases. Feature engineering and outlier handling were critical in improving model accuracy. The results suggest that for complex, non-linear datasets like real estate pricing, Random Forest is the most effective technique, though Linear Regression and SVM can still be useful for simpler tasks or specific scenarios.

References:

1. Tutorialspoint, n.d. *Machine Learning - Supervised Learning*. <https://www.tutorialspoint.com/machine_learning/machine_learning_supervised.htm>
2. Tutorialspoint, n.d. *Machine Learning - Data Visualization*. <https://www.tutorialspoint.com/machine_learning/machine_learning_data_visualization.htm>
3. Tutorialspoint, n.d. *Machine Learning - Box and Whisker Plots*. <https://www.tutorialspoint.com/machine_learning/machine_learning_box_and_whisker_plots.htm>
4. Tutorialspoint, n.d. *Machine Learning - Linear Regression*. <https://www.tutorialspoint.com/machine_learning/machine_learning_linear_regression.htm>
5. Tutorialspoint, n.d. *Machine Learning - Correlation Matrix Plot*. <https://www.tutorialspoint.com/machine_learning/machine_learning_correlation_matrix_plot.htm>
6. Tutorialspoint, n.d. *Machine Learning - Random Forest Classification*. <https://www.tutorialspoint.com/machine_learning/machine_learning_random_forest_classification.htm>
7. Tutorialspoint, n.d. *Machine Learning - Performance Metrics*. <https://www.tutorialspoint.com/machine_learning/machine_learning_performance_metrics.htm>

Top of Form

Bottom of Form