An Executive Report for Wazobia Real Estate Limited House Price Prediction.

Introduction:

The objective is to create a robust predictive model using the provided dataset. And to analyze various factors that impact house prices, identify meaningful patterns, and build a model that can generate reliable price predictions. With the data-driven solution, In order to empower Wazobia Real Estate Limited to make informed pricing decisions, enhance its competitiveness in the market, and deliver enhanced value to its customers.

Data Exploration:

During the data exploration phase, I employed various techniques to gain insights into the dataset. For the numerical variables, I utilized descriptive statistics, box plots, and distribution plots. Descriptive statistics helped me understand the central tendency, spread, and distribution of the numerical data. The box plots provided a visual representation of the data's quartiles, identifying potential outliers and the overall distribution. Additionally, distribution plots offered a clear visualization of how the data was distributed, indicating symmetry or skewness.

On the other hand, for the categorical data, I utilized value counts and bar plots. Value counts provided a summary of the frequency of each category in the categorical variables, offering an initial understanding of the data distribution. Bar plots visually displayed the distribution of the categories, enabling quick comparisons and identification of dominant categories.

By combining these exploratory techniques for both numerical and categorical variables, I gained a comprehensive understanding of the dataset's characteristics. These insights laid the foundation for further analysis, and model development in the subsequent stages of the project.

By conducting exploratory data analysis (EDA), I gain a deeper understanding of the dataset, leading to valuable insights that are useful for both model building and independent analysis.

During the exploration, I identified missing values in the columns: Location, Bedroom, Title, and Parking Space. To handle these missing values, I applied the following steps:

Location Column: Since the majority of the house locations are in Kaduna, I filled the missing values in the Location column with "Kaduna."

Bedroom Column: Observing that the majority of bedrooms are one bedroom, I filled the missing bedroom values with "1.0."

Title Column: As the majority of the house titles are "Flat," I filled the missing title values with "Flat."

Parking Space Column: For missing parking space values, I filled them with "4.0."

Bathroom Column: Additionally, I filled missing bathroom values with "1.0."

Furthermore, I encountered categorical variables in the House Title and Location columns. To handle these categorical variables, I utilized one-hot encoding, converting them into binary columns for each category. This allows me to transform the categorical data into a format that can be effectively used in the machine learning models.

By filling missing values and applying one-hot encoding to categorical variables, I ensure that the dataset is prepared and suitable for further analysis and model training. These steps contribute to the reliability and accuracy of the models, as well as providing valuable insights for independent exploration and decision-making in the real estate domain.

Methodology

Based on my comprehensive data exploration, I have identified that all the variables, including location, title, bedroom, bathroom, and parking space, play significant roles in influencing the housing prices in Nigeria.

To predict housing prices, I employed two models, namely Linear Regression and Random Forest. After thorough evaluation, I found that the Linear Regression model outperformed the Random Forest model in terms of accuracy and overall performance.

The decision to choose Linear Regression over Random Forest was driven by several factors. First, the Linearity Assumption was met, implying that the relationship between the independent variables and the target variable (housing prices) can be represented by a linear model. This aligns well with the nature of the data, making Linear Regression a suitable choice.

Secondly, there was no evidence of Multicollinearity among the variables present in the dataset. Fortunately, the data showed no such issues, ensuring the reliability of the Linear Regression model.

Lastly, I carefully examined the data for Outliers and Influential Points, which can exert a disproportionate impact on the model's performance. Thankfully, the dataset demonstrated the absence of such extreme values, further strengthening the validity of the Linear Regression model.

$$Y_{price} = \beta_{0_{intercept}} + \beta_{1_{loc\_kano}} + \beta_{2_{loc\_akwa\_ibom}} + \cdots + \beta_{48_{bathroom}} + \beta_{49_{parking\_space}}$$

Interpretation of Results:

Intercept: The intercept value is 2,134,999.12. the intercept represents the predicted house price when all other features are zero.

Location Coefficients: the impact of that location on the house price has Positive coefficients indicate that the house price tends to be higher in those location and the magnitude of the coefficient reflects the strength of the impact of the location on the house price.

Title Coefficients: Each "title_" feature represents a specific title type (e.g, Townhouse, Terrace duplex, etc.). The Positive coefficients imply higher house prices for those title types.

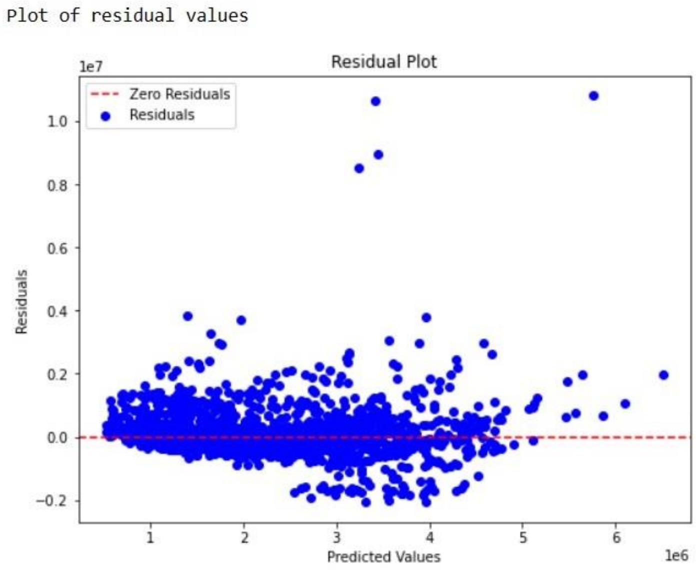
Bedroom, Bathroom, and Parking Space Coefficients: These coefficients represent the impact of the number of bedrooms, bathrooms, and parking spaces on the house price. The Positive coefficients indicate that an increase in the number of bedrooms, bathrooms, or parking spaces is associated with higher house prices.

Model Training and Evaluation:

The model's performance was evaluated using a range of performance metrics, including The Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Model Accuracy, R-squared (R^2), and the residual plot.

By employing these performance metrics, I was able to comprehensively evaluate the model's effectiveness in making accurate predictions. These metrics provided valuable insights into the model's strengths and potential areas for improvement, guiding me in refining the model and optimizing its performance.

```
Root Mean Square Error (RMSE): 660822.29
Model Accuracy: 0.65
R-squared (R2): 0.65
Mean house price: 2135673.35
Baseline MAE: 776993.9646515128
Training MAE: 339551.4945787647
```



Plot of residual values

Training and Evaluation results,

The Root Mean Square Error (RMSE) is a measure of the average prediction error of a regression model. In the house pricing prediction model, the RMSE of 660,822.29 means that, on average, the difference between the predicted house prices and the actual house prices is approximately NGN 660,822.29.

Model Accuracy: 0.65. The model's accuracy is 65%, indicating that it correctly predicts outcomes with a probability of 65%.

R-squared (R2): 0.65. The variance in the target variable is explained by the independent variables included in the model. In other words, the model captures a reasonable amount of the variability in the target variable and provides a moderately good fit to the data. However, there is still about 34.77% of the variance that remains unexplained by the model.

Baseline MAE: 776993.9646515128

Training MAE: 339740.9537250625

The lower MAE for the training set indicates that the model is performing better than the baseline model on the training data.

The residual plot is scattered randomly around the horizontal line at y = 0, it generally indicates that the regression model is a good fit for the data. This pattern suggests that the model's predictions are on average very close to the actual observed values, and the residuals (the differences between the observed and predicted values) are distributed evenly across the range of the predictor variable(s).

indicating that the regression model is performing well and providing accurate predictions for the data at hand.

Business Impact and Recommendations:

- This house pricing predictions enable real estate investors to make well-informed decisions when purchasing or selling properties. They can identify undervalued properties and invest in areas with potential growth, maximizing their return on investment.

- Real estate agents and property sellers can use the house price predictions to set competitive listing prices. By pricing properties more accurately, they can attract more potential buyers and close deals faster.
- Financial institutions can use house price predictions for risk assessment in mortgage lending. Accurate predictions help identify properties that may be overvalued, reducing the risk of loan defaults.
- House price predictions provide valuable insights into local real estate markets, including demand patterns, market trends, and price fluctuations. This information can be used to adapt business strategies to changing market conditions.
- Real estate marketers can tailor their marketing efforts based on predicted house prices in different areas. This allows them to target specific buyer segments effectively and optimize their marketing campaigns.
- House price predictions can serve as a valuable tool during price negotiations. Buyers and sellers can use this information to support their bargaining positions.

Recommendations:

- Regularly update and refine the predictive model to ensure it captures changing market dynamics.

- Collaborate with domain experts to include additional features or variables that may improve prediction accuracy.

- Consider incorporating external data sources, such as economic indicators or local development plans, to enhance the model's performance.