Predicting Airbnb Property Prices in NYC: A Regression Modeling Approach

Name: Kasif Hasnaen Zisan
Department: *Department of Computer Science And Engineering*
ID: 2020-1-60-027
Email: 2020-1-60-027@std.ewubd.edu

Name: Md.Maruf Shahriar
Department: *Department of Computer Science And Engineering*
ID: 2020-1-60-008
Email: 2020-1-60-008@std.ewubd.edu

Name: Rifa Tasnim Orin
Department: *Department of Computer Science And Engineering*
ID: 2020-1-60-103
Email: 2020-1-60-103@std.ewubd.edu

Name: Azman Ahmed
Department: *Department of Computer Science And Engineering*
ID: 2020-1-60-217
Email: 2020-1-60-217@std.ewubd.edu

***Abstract*— This paper involves the prediction of property prices for various Airbnb listings in New York City, using the 'New York City Airbnb Open Data' dataset. To accomplish this task, we employ an array of regression models, including Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. We then evaluate the models using key evaluation metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) score and then compare the results. Furthermore, to enhance the interpretability of our models, we integrate Explainable AI (XAI) techniques such as Feature Importance, Partial Dependence Plots (PDP), SHapley Additive exPlanations (SHAP). This study aims to provide accurate predictions while enhancing the interpretability of the underlying factors influencing Airbnb property prices.**

I. Background Study

This "Background Study" section serves as a foundational pillar, providing readers with essential knowledge on regression techniques used in this paper, namely Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. Additionally, it provides info about fundamental evaluation metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) score. Furthermore, this section delves into the realm of Explainable Artificial Intelligence (XAI), elucidating concepts like Feature Importance, Partial Dependence Plots (PDP), and SHapley Additive exPlanations (SHAP).

Regression is a statistical method used in machine learning and data analysis to model the relationship between a dependent variable (also known as the response or outcome variable) and one or more independent variables (predictors or features). The goal of regression analysis is to understand the nature of the relationship between the variables and to make predictions.

Linear regression assumes a linear relationship between the independent variables and the dependent variable. The model tries to find the best-fit straight line that minimizes the sum of the squared differences between the observed and predicted values.

Random Forest Regression is a versatile machine-learning technique used for predicting numerical values. It's particularly useful because it combines predictions of multiple decision trees to reduce overfitting and improve accuracy.

Gradient Boosting Regression is another powerful ensemble learning technique for regression tasks. Like Random Forest, Gradient Boosting builds a predictive model in the form of an ensemble of weak learners (typically decision trees). However, it constructs these trees sequentially, with each new tree attempting to correct the errors of the previous ones.

Root Mean Squared Error (RMSE) measures the average magnitude of errors between predicted and actual values, providing an insight into the model's overall accuracy.

Mean Absolute Error (MAE) quantifies the average absolute difference between predicted and actual values, offering a robust evaluation of prediction accuracy.

R-squared (R2) Score assesses the proportion of variance in the target variable explained by the model. A higher R2 score indicates better predictive performance.

XAI techniques like Feature Importance unveil the significance of input features, highlighting those with the most substantial impact on model predictions.

Partial Dependence Plots (PDP) illustrate the relationship between a specific feature and the model's predictions while holding other features constant, aiding in understanding feature interactions.

SHapley Additive exPlanations (SHAP) values provide a nuanced understanding of the contribution of each feature to an individual prediction, enhancing interpretability and transparency in model outputs.

II. Methodology

A. *Exploratory Data Analysis*

We imported the dataset and created a Data Frame called 'df'. We printed the first five and last five rows of the Data Frame to have a generalized idea about our dataset. We then printed out the info, shape, columns and the statistical description of the dataset.

From the info –

- The data types: float64(3), int64(7), object(6).

From the shape, we can see that we have 16 columns and 48,895 rows.

We can also see what our columns are from the df.columns.

Moreover, we can get statistical descriptions such as the mean, std. min. max and three quantiles of various columns from the describe() method.

B. *Data Cleaning*

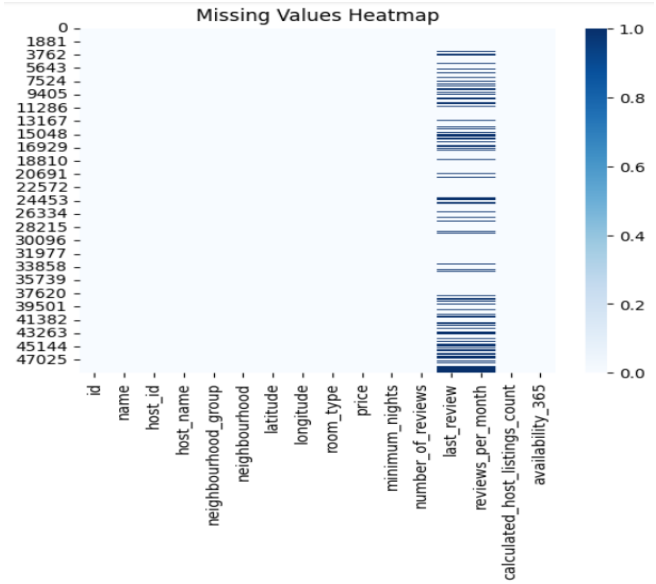
Identifying Duplicate Values:

In this section we ran the duplicated method and saw that there are no duplicate values in this dataset.

Identifying Missing Values:

Here we saw that the columns - name, host_name, last_review, reviews_per_month have missing values. We used various methods such as isna() and isna().sum() to find out the number of missing values in this dataset.

We also created a heatmap to visualize the missing values. As we can see that the number of missing values in the name and host_name columns are insignificant when compared to the missing values in the last_review and reviews_per_month column.



Handling Inconsistencies:

A focus is placed on instances where 'availability_365' is recorded as zero. These occurrences are rectified by substituting zero values with the mean availability, ensuring that no rows indicate zero availability after this correction.

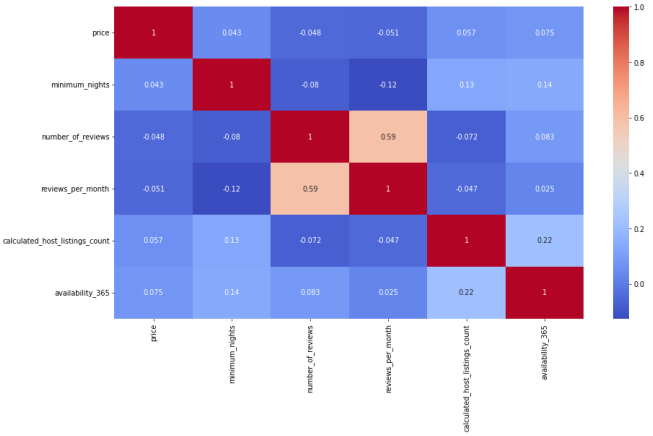
Handling Missing Values:

The 'last_review' column, containing date objects and featuring missing values, is subject to scrutiny. Recognizing the significance of temporal data, rather than discarding rows with missing 'last_review' values, the column is altogether excluded from the dataset. For the 'reviews_per_month' column, a comparative analysis is conducted between missing values and instances of zero reviews. Subsequently, missing values in 'reviews_per_month' are imputed with zero.

C. *Data Visualization*

Correlation Analysis:

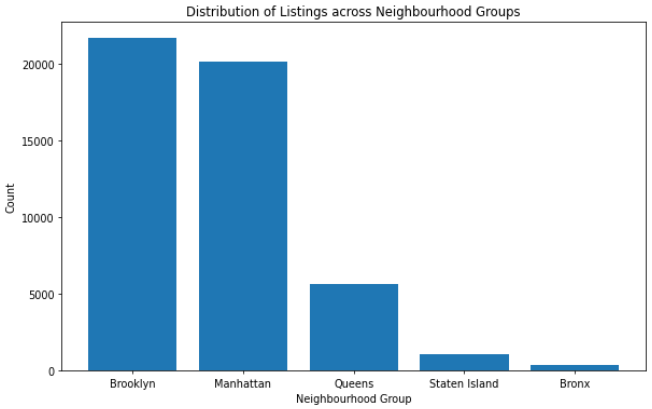
Correlation analysis is a statistical method that quantifies the degree of association between two variables, providing insights into the strength and direction of their relationship.



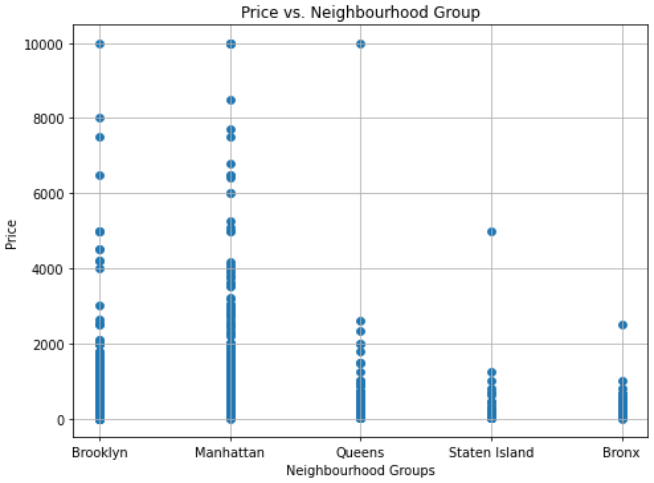
The correlation analysis reveals weak associations between price and various other features. Positive correlations exist with only minimum_nights, calculated_host_listings_count and availability_365 features. This hints that having a higher value of these features will slightly increase the price of the airbnb.

Conversely, there are weak negative correlations with the number_of_reviews and the reviews_per_month features, hinting at a possible trend where properties with more reviews have slightly lower prices.

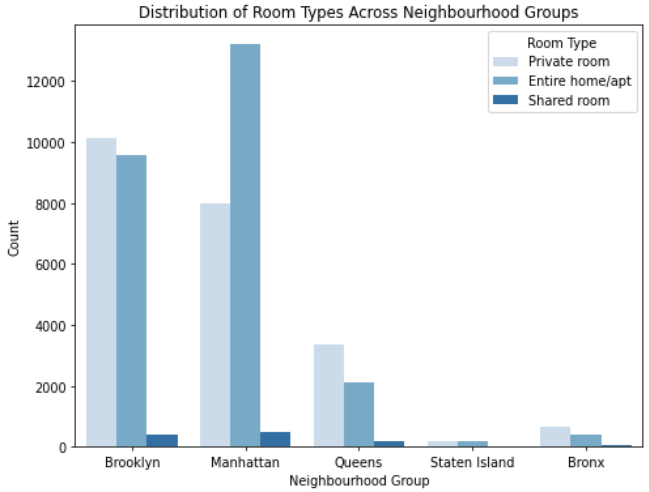
Visualizations of Relationships between different columns:



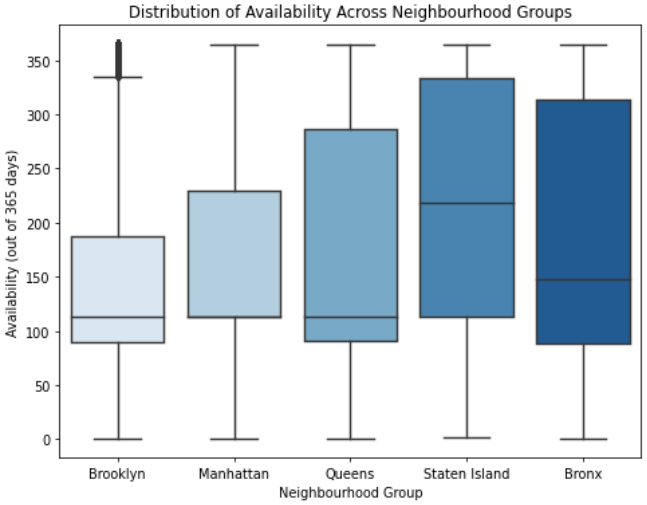
This bar plot shows the distribution of airbnb listings across neighborhood groups in the New York City neighbourhood groups of Brooklyn, Manhattan, Queens, Staten Island, and the Bronx. The y-axis represents the number of listings, while the x-axis represents the neighborhood group. Brooklyn has the highest number of listings, followed by Manhattan, Queens, the Bronx, and Staten Island.



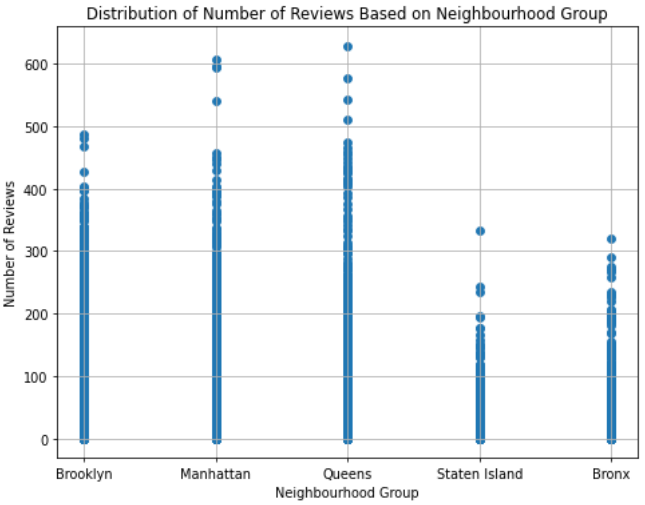
This scatter plot shows the distribution of airbnb prices in each neighborhood group in New York City. The x-axis represents the neighborhood_group feature, and the y-axis represents the price feature. We can see that airbnb's in Brooklyn and Manhattan have the highest distribution of prices whereas Staten Island and the Bronx have the lowest distribution of prices.



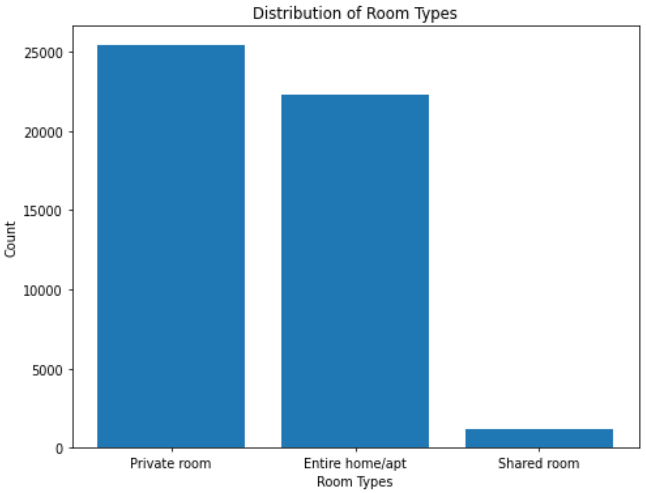
This countplot shows the number of different room types of airbnb in the different neighbourhood groups of New York. We can see that Manhattan has the highest number of Entire home/apt and Shared Room types, Brooklyn has the highest number of Private Room types.



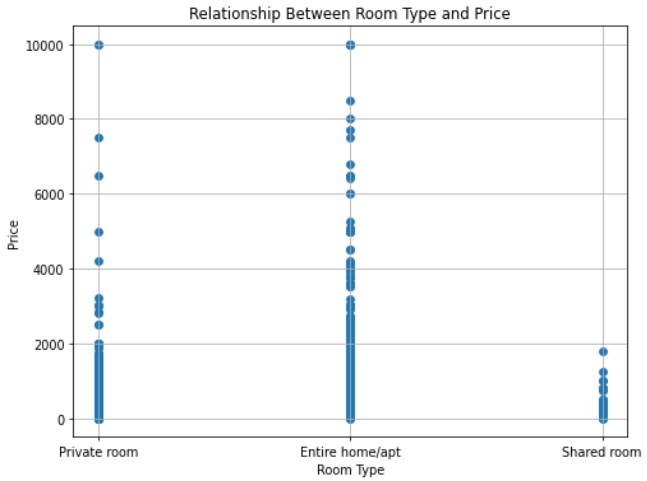
This boxplot shows a summary of the number of available days of all the airbnb in different neighbourhood groups of New York. A box plot gives us the idea about the mean, 25%, 75%, maximum and minimum range of the x-axis. In this case, we can see that Staten Island has the highest numbers of median available days among all of the neighbourhood groups. On the other hand, the Bronx and Queens have the lowest median available days.



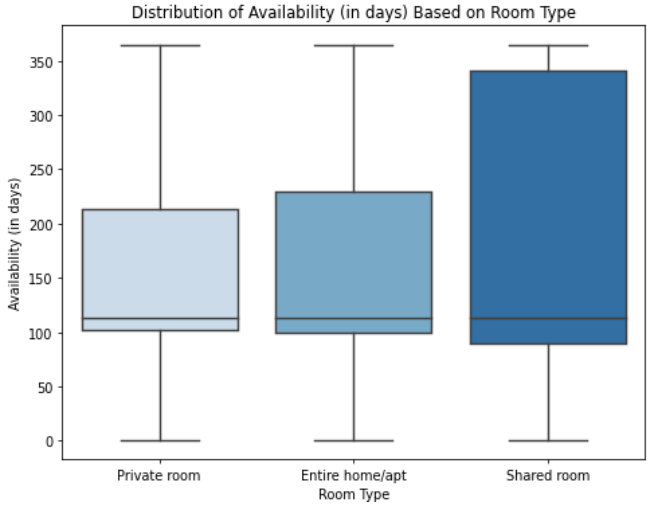
This scatter plot shows the distribution of the number of reviews of Airbnb in the different neighborhood groups of New York. We can see that the airbnb's in Queens and Manhattan have the most distribution and the most number of reviews whereas airbnb's in Staten Island has less distribution in the number of reviews.



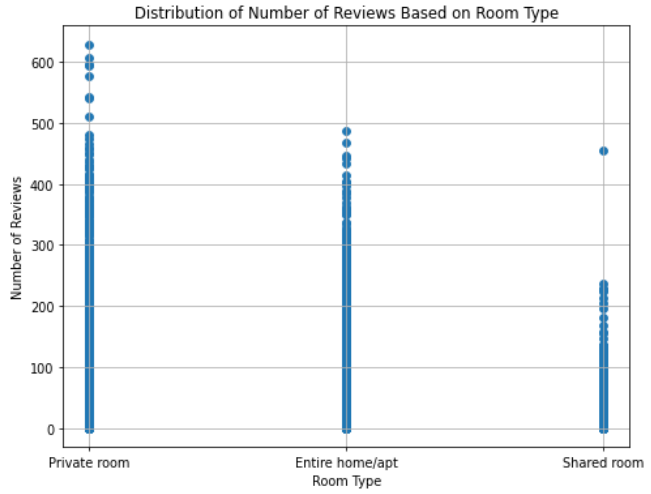
This bar plot shows the count of different room types in airbnbs across New York city. We can see that Private Room is the most common room type, followed by Entire home/apt and Shared Room.



This scatter plot shows the distribution between the different room types in airbnbs and their price. We can see that the Entire home/apt has a more distributed price ranging from low to high. It is also noticeable that some private rooms have really high prices. On the other hand, shared room prices are less distributed and on the lower end.



This boxplot shows a summary of the relationship between room type and availability of the airbnb. We can see that shared rooms are more available, followed by entire home/apt rooms and private rooms. Interestingly, the median availability value is almost the same across all the different room types.



This scatter plot shows the distribution of airbnb number of reviews based on room type in New York City. We can see that private rooms have the most reviews, followed by entire homes/apartments, and shared rooms.

D. Data Preprocessing

The comprehensive pre-processing methodology ensures that the dataset is meticulously curated, promoting the robustness and reliability of subsequent machine learning models.

Drop column:

We drop some unnecessary columns. Those columns don't need to train our model.

Divide data in X-axis Y-axis:

We divide our data in X-axis Y-axis. The Y- axis holds the price column and the other columns without price hold the X-axis.

E. Model Training

Linear Regression:

It begins by initializing a linear regression model and fitting it to the training data. Subsequently, the model predicts the target variable on the test set. Evaluation metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) score, are then computed to assess the model's performance. The actual and predicted values for the first five instances in the test set are displayed in a Data Frame for comparison. Additionally, a visual representation of prediction errors is generated through a bar plot, depicting the differences between actual and predicted values for the initial 25 instances in the test set. This comprehensive analysis provides insights into the accuracy and efficacy of the linear regression model in predicting Airbnb listing prices.

Random Forest Regression:

The RandomForestRegressor is configured with specific hyperparameters, including the number of estimators (200), maximum depth (50), minimum samples split (5), and minimum samples leaf (4). After fitting the model to the training data, predictions are generated on the test set. Key regression evaluation metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) score, are then computed to assess the model's performance. The actual and predicted values for the first five instances in the test set are displayed in a DataFrame for comparison. Additionally, a visual representation of prediction errors is created through a bar plot, illustrating the differences between actual and predicted values for the initial 25 instances in the test set. This thorough analysis provides insights into the accuracy and efficacy of the RandomForestRegressor in predicting Airbnb listing prices.

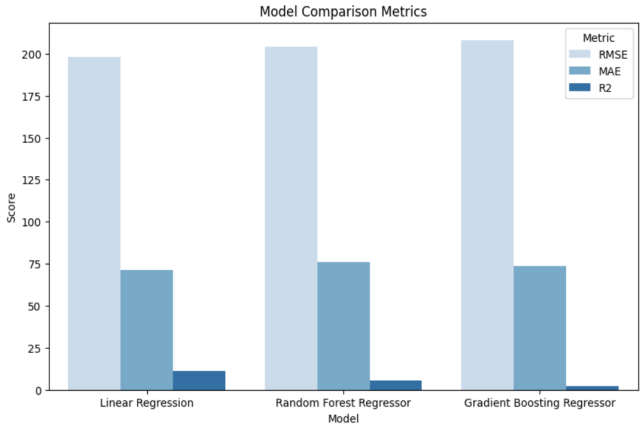
Gradient Boosting Regression:

The GradientBoostingRegressor is configured with specific hyperparameters, including the number of estimators (3000) and learning rate (0.01). After fitting the

model to the training data, predictions are generated on the test set. Key regression evaluation metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) score, are then computed to assess the model's performance. The actual and predicted values for the first five instances in the test set are displayed in a DataFrame for comparison. Additionally, a visual representation of prediction errors is created through a bar plot, illustrating the differences between actual and predicted values for the initial 25 instances in the test set. This comprehensive analysis provides insights into the accuracy and efficacy of the GradientBoostingRegressor in predicting Airbnb listing prices.

III. Result Analysis

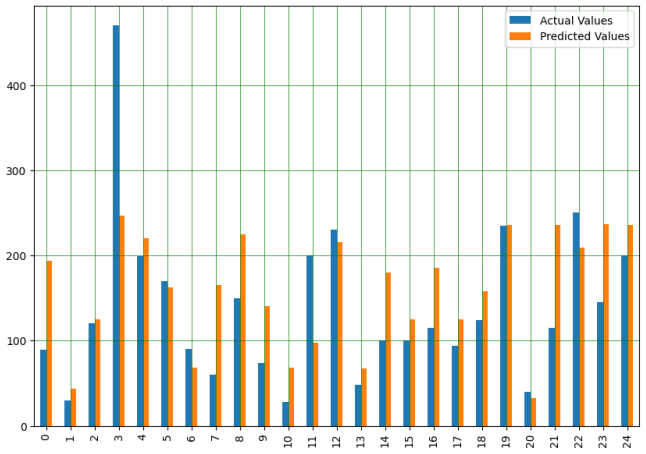
Analyzing the results of the three regression models: Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor based on the provided evaluation metrics (RMSE, MAE, and R2 Score).



Linear Regression

- The Root Mean Squared Error (RMSE) is a measure of the average magnitude of the errors between predicted and actual values. A lower RMSE is better, so 198.14 indicates a moderate level of accuracy.
- Mean Absolute Error (MAE) is another measure of prediction accuracy; 71.33 suggests a similar level of accuracy.
- The R2 Score measures the proportion of the variance in the dependent variable that is predictable from the independent variables. An R2 Score of 11.25% suggests that the model explains a small portion of the variance in the data.

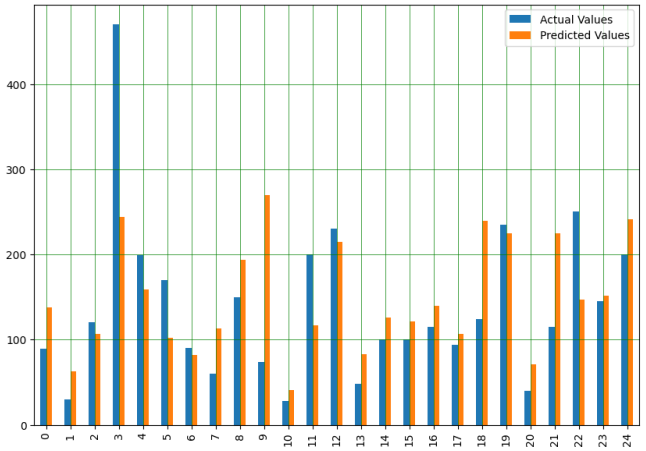
Here is the actual vs predicted value comparison for the linear regression model -



Random Forest Regression

- The Random Forest model has a slightly higher RMSE (204.27) and MAE (75.98) compared to Linear Regression, indicating a bit lower accuracy.
- The R2 Score of 5.68% suggests that Random Forest explains a larger portion of the variance compared to Linear Regression but still a relatively low amount.

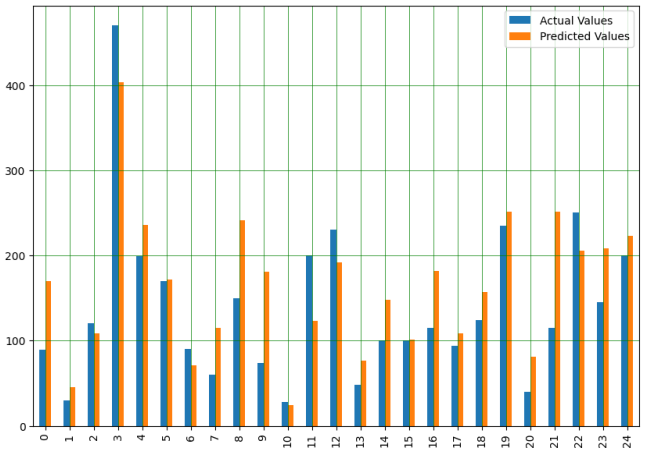
Here is the actual vs predicted value comparison for the random forest regressor model -



Gradient Boosting Regression

- The Gradient Boosting model has the highest RMSE (208.07), MAE(73.44), and the lowest R2(2.14%) Score among the three models, indicating the lowest level of accuracy and explained variance.

Here is the actual vs predicted value comparison for the random forest regressor model -



Overall Comparison

- The Linear Regression model outperforms both Random Forest and Gradient Boosting in terms of RMSE, MAE, and R2 Score.
- Random Forest performs better than Gradient Boosting but is still less accurate than Linear Regression.
- The R2 Scores for all three models are relatively low, suggesting that the models do not explain a significant portion of the variance in the data.

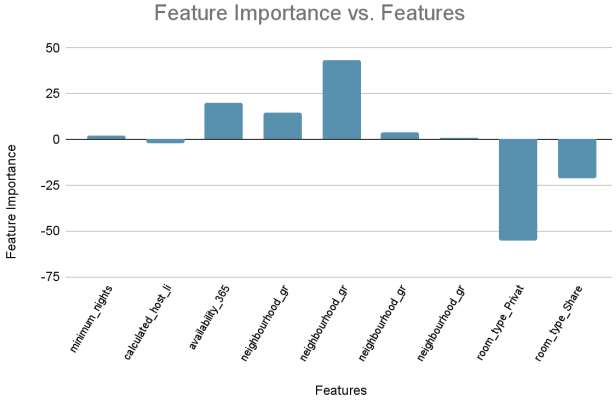
IV. Result Explanation with Explainable AI

Explainable Artificial Intelligence (XAI) is a critical facet of AI that seeks to demystify and interpret the decisions made by complex machine learning models. In contrast to black-box models, XAI techniques aim to make the decision-making process of AI systems more transparent and understandable for human users.

In our project, we used three XAI techniques - Feature Importance for Linear Regression Models, Partial Dependence Plots (PDP) for Random Forest Regressors and SHapley Additive exPlanations (SHAP) for Gradient Boosting Regressors.

A. Feature Importance:

For feature importance for each feature we got these values- [1.86660138, -1.91258022, 20.03734728, 14.54507099, 43.12310789, 3.62733687, 0.72614803 -55.27329158, -21.41252409]

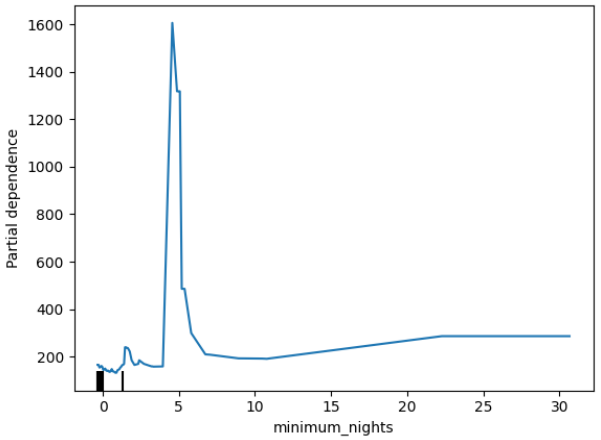


This feature importance array interprets the amount of importance each feature has in predicting the target variable which in this case is 'price'. Higher values mean that those features are more important.

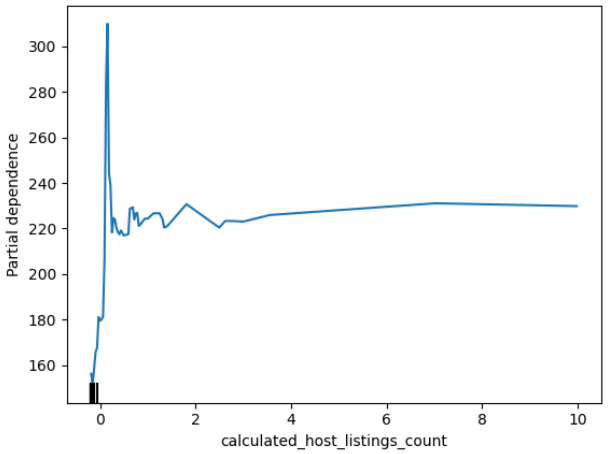
For example, the feature 'availability_365' has a feature importance score of 20.03734728 interpreting a strong positive impact on price. Higher availability throughout the year is associated with significantly higher predicted prices.

On the other hand, the feature 'room_type_Private room' has a feature importance score of -55.27329158 interpreting a strong negative impact on price. Private rooms are associated with significantly lower predicted prices than the reference group.

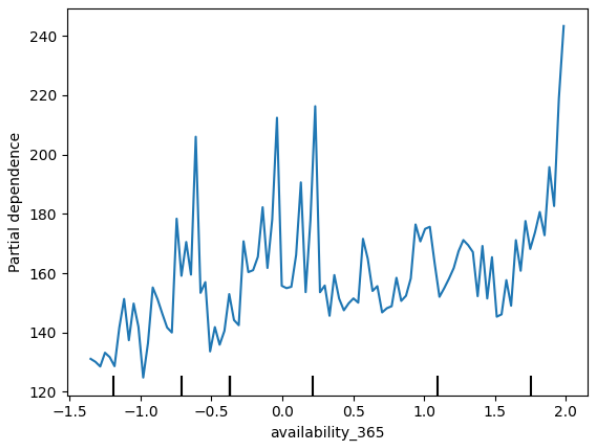
B. Partial Dependence Plots (PDP) -



This plot interprets that increase of minimum_nights leads to a sharp increase in price but that is only for about 4 to 8 minimum_nights. After that the price falls only to increase slightly between 20 to 30 minimum nights. Also by this curve, we can see that minimum_nights and price have a nonlinear relationship.



This plot interprets the change of price in accordance with the calculated_host_listings_count feature. We can see that the price takes a sharp increase with the increase of calculated_host_listings_count feature but soon it normalizes and becomes almost a horizontal line.

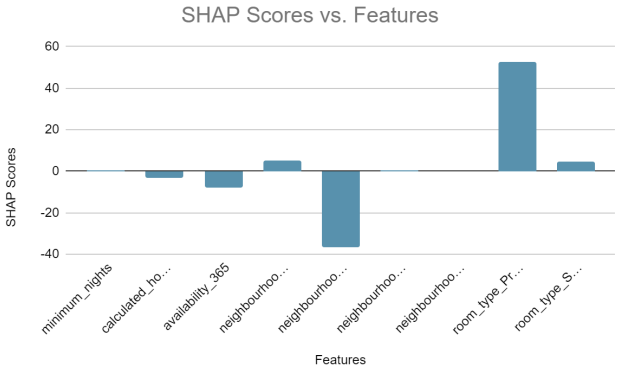


This plot interprets the change of price in accordance with the availability_365 feature. We can see that the increase in availability_365 affects the price randomly.

C. *SHapley Additive exPlanations (SHAP)*

We get the SHAP score values as: 4.33355425×10^{-1}, -3.18825882, -7.62275757, 4.99945511, -3.64342114×10^{1}, 6.41238635×10^{-1}, 1.57231646×10^{-2}, 5.27810294×10^{1}, 4.61106031

If we plot these values into a graph against the features we get -



The array represents the SHAP values for each feature in the model for a specific instance (data point). The values indicate the contribution of each feature to the predicted price for this particular listing. Positive values signify a positive contribution, meaning the feature increases the predicted price. Negative values indicate a negative contribution, meaning the feature decreases the predicted price.

If we want to interpret these scores, for example if we take the SHAP score of 'minimum_nights' we can see that it has a score of 0.433 which means that this feature increases the predicted price by 0.433 units.

On the other hand, the SHAP score of 'neighbourhood_group Manhattan' is -36.43 meaning this feature decreases the predicted price by 36.43 units.

V. Conclusion

In this study, we harnessed the power of data and machine learning models to predict Airbnb prices in New York City. Our journey involved meticulous data cleaning, insightful visualizations, model training, and comparison. We demystified our models using Explainable AI, uncovering critical features and relationships. Through this holistic approach, our study not only enhances our understanding of Airbnb pricing dynamics and regression models but also underscores the importance of transparent and accountable AI practices using Explainable AI. This work contributes to the broader discourse on responsible AI deployment in real-world scenarios.

References

[1] "About Linear Regression | IBM." https://www.ibm.com/topics/linear-regression

[2] Shi, A. a. K. (2023, March 28). Decision Tree Regressor — A Visual Guide with Scikit Learn. Medium. https://towardsdatascience.com/decision-tree-regressor-a-visual-guide-with-scikit-learn-2aa9e01f5d7f

[3] Regression analysis using gradient boosting regression tree: Aurora articles | NEC. (n.d.). https://www.nec.com/en/global/solutions/hpc/articles/tech14.html

[4] Chugh, A. (2022, March 16). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? Medium. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

[5] What is explainable AI? | IBM. (n.d.). https://www.ibm.com/topics/explainable-ai

[6] 4.1. Partial Dependence and Individual Conditional Expectation plots. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/partial_dependence.html

[7] Trevisan, V. (2022, July 5). Using SHAP values to explain how your machine learning model works. Medium. https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137