**ALY6040 90248 Data Mining Applications SEC 01 Summer 2023 CPS [BOS-D-HY]**

**Module 6 Assignment — Final Project**

A black and white logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by**

Hanchan Xu [xu.hanc@northeastern.edu](mailto:xu.hanc@northeastern.edu)

Zihan Ma [ma.zihan1@northeastern.edu](mailto:ma.zihan1@northeastern.edu)

**Instructor**

 Prof. Kasun Samarasinghe

**Date**

08/16/2023

**Final Project**

**Introduction**

This report delves into an in-depth analysis of Airbnb listings data for Singapore. For hosts to competitively price their listings and for guests to ensure they receive value for money, understanding the determinants of Airbnb listing prices in Singapore becomes paramount. This business problem poses two critical questions:

1. What are the key factors or features of a listing that significantly influence its pricing?
2. How can potential hosts use this information to optimize their listing prices for maximum profitability without deterring potential guests?

Our primary aim is to decipher the intricate structure of the dataset and employ modeling techniques to discern the factors influencing the pricing of Singapore Airbnb listings.

**Tools and Techniques Used and Their Justification**

1. **Multiple Linear Regression**

* **Why Chosen**: Multiple Linear Regression (MLR) was selected as a foundational model to establish a baseline performance. It is a straightforward method to understand the relationships between multiple independent variables and the dependent variable. Given its simplicity, it provides a clear way to see which predictors are significant in predicting the outcome.

2. **Lasso Regression**

* **Why Chosen**: Lasso Regression was incorporated to manage potential multicollinearity and feature selection. As our dataset had numerous predictors, Lasso (L1 regularization) aids in shrinking the coefficients of less important features to zero, effectively performing feature elimination. This ensures a simpler and more interpretable model.

3. **Ridge Regression**

* **Why Chosen**: Similar to Lasso, Ridge Regression (L2 regularization) was utilized to address potential multicollinearity and overfitting. Instead of eliminating features like Lasso, Ridge shrinks all coefficients towards zero, ensuring that the model doesn't overly rely on any single predictor. This method was chosen as an alternative to Lasso to see if retaining all features (with reduced influence) improves performance.

4. **Gradient Boosting Machine (GBM)**

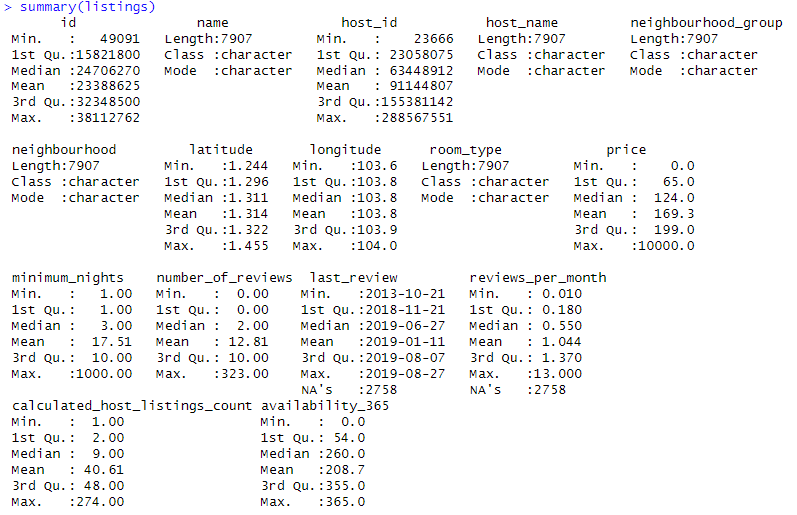
* **Why Chosen**: GBM was implemented as an ensemble method to enhance prediction accuracy. Given its capability to optimize on residuals and handle non-linear relationships, it was deemed valuable for our complex dataset where interactions between variables might not be captured adequately by linear models. Additionally, GBM's ability to assign importance scores to features provided insights into which predictors are most influential in our model.

**Discussion on the Change of Direction:**

In our initial analysis, as presented in Module 3, our approach was rooted in predicting price. However, as the project evolved, we realized the complexities associated with predicting categorized prices, especially without domain expertise. It leads to results that seem less meaningful to match with the result in this report.

**Data Import and Basic Statistics**

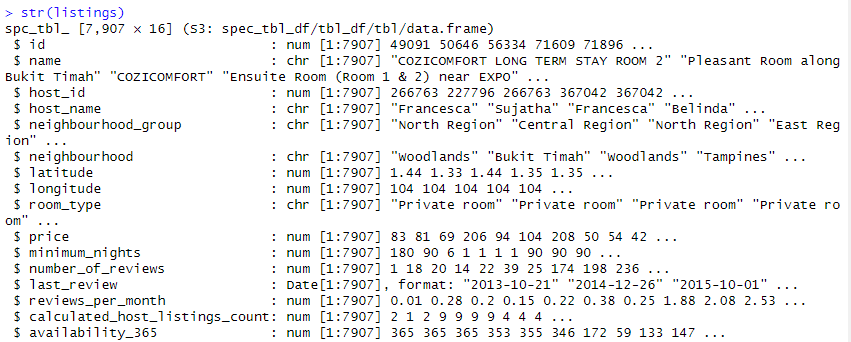
**Data Import**  
The dataset, sourced from a CSV file, comprises 7907 listings, each characterized by 16 attributes.



**Basic Statistics**

Our initial exploration of the dataset revealed:

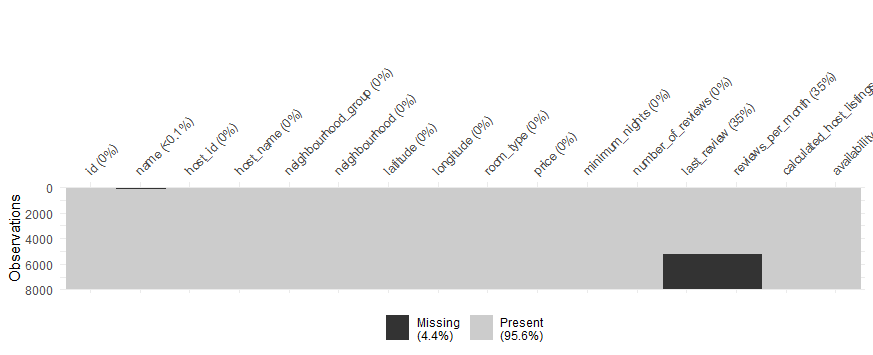
* Textual attributes such as 'name', 'host\_name', 'neighbourhood\_group', 'neighbourhood', and 'room\_type'.
* A date attribute represented by the 'last\_review' column.
* Several numerical attributes like 'id', 'latitude', 'longitude', and 'price' that offer quantitative insights into the listings.



**Data Cleaning**

**Missing Data**

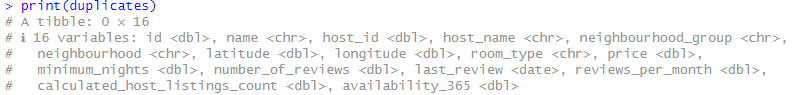
Our preliminary checks identified 2758 missing entries in the 'last\_review' and 'reviews\_per\_month' columns. A visualization (to be inserted) depicts the distribution of this missing data.



To address this, we imputed zeros in the 'reviews\_per\_month' column, rationalizing that listings without reviews would naturally have a monthly review count of zero.

**Duplicates**

Upon checking for duplicated rows in the dataset, it was found that there are no repeating data points.



**Outliers**

A boxplot analysis of the 'price' column highlighted several outliers, with prices reaching as high as 10,000.

A graph with black lines and dots

Description automatically generated with medium confidence

Given our model's focus on predicting average prices and the potential distortion caused by these outliers, we opted to exclude them.

**Handling Time Data**

Our temporal analysis of reviews revealed:

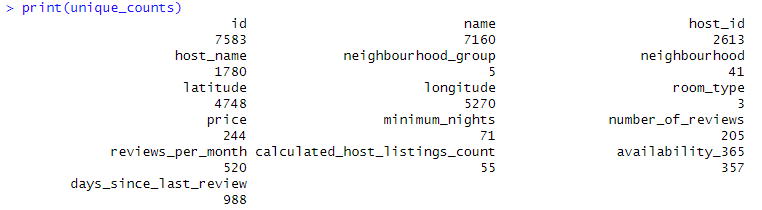
* The review timeline spans from 2013-10-21 to 2019-08-27.  
  A close up of numbers

  Description automatically generated
* Missing 'last\_review' entries were imputed with the earliest review date.
* We transformed the 'last\_review' column to depict the duration since the last review, using the most recent review as a reference. The original date column was subsequently discarded.

**Data Engineering and Transformation**

**Unique Value Assessment**

A quick assessment was conducted to understand the unique value count for each column.  
All column that has less than 6 unique values Have been printed out.  
Columns 'neighbourhood\_group' and 'room\_type' were particularly notable:



A screenshot of a computer

Description automatically generated

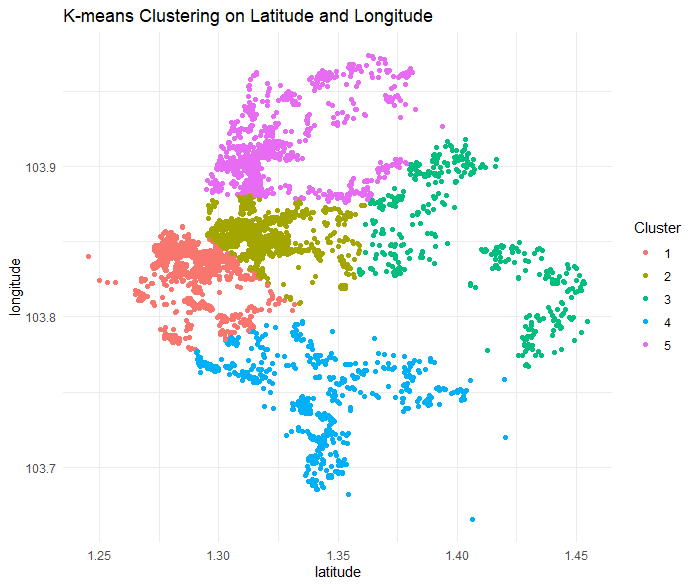
**One-Hot Encoding**

To make the dataset suitable for machine learning models:

* 'neighbourhood\_group' was one-hot encoded into separate columns: Central\_Region, East\_Region, North\_East\_Region, North\_Region, and West\_Region.
* 'room\_type' was similarly one-hot encoded into: Entire\_home\_apt, Private\_room, and Shared\_room. The original columns were removed post-encoding.

**K-means Clustering on Geographical Data**

We employed K-means clustering on the latitude and longitude columns to discern geographical patterns within the listings. This exercise yielded five distinct clusters, providing a geographical segmentation of Airbnb listings. A visualization showcased these clusters based on their geographical coordinates. While we contemplated clustering into 28 segments, mirroring Singapore's postal areas, we settled on five clusters to maintain a manageable column count.



**Finalizing the Data**

**Removing Redundant Columns**

Certain columns, deemed superfluous or challenging for modeling, were excluded. These include:

* 'name'
* 'host\_name'
* 'neighbourhood'
* 'host\_id'

**Data Scaling and Normalization**

To ensure that no variable disproportionately impacts the modeling due to its scale, numerical columns were subjected to standardization and normalization:

* Standardization: This involved transforming each numerical column to have a mean of 0 and a standard deviation of 1.
* Normalization: Each numerical column was then transformed again to have a minimum value of 0 and a maximum value of 1, using a custom normalization function.

Columns that underwent this transformation process are:

* 'price'
* 'minimum\_nights'
* 'number\_of\_reviews'
* 'reviews\_per\_month'
* 'calculated\_host\_listings\_count'
* 'availability\_365'
* 'days\_since\_last\_review'

**Data Split for Modeling**

For modeling purposes, the dataset was partitioned into:

* **Training set**: 80% of the data.
* **Testing set**: The residual 20% for model validation.

**Setting Up for Cross-Validation**

To bolster our model evaluation, we instituted a 10-fold cross-validation setup. This approach entails partitioning the training dataset into 10 subsets, training the model 10 times, and using 9 subsets for training and 1 for validation in each iteration. The cumulative performance across these iterations offers a robust metric of model efficacy.

**Multiple Linear Regression (MLR)**

**Model Implementation**

We employed Multiple Linear Regression (MLR) on the training dataset, designating 'price' as the dependent variable and all other attributes as independent variables.



**Challenges Encountered During Modeling**

During the modeling phase, we encountered warnings indicating potential multicollinearity within the dataset. Multicollinearity arises when two or more independent variables exhibit high correlation, complicating the model's ability to discern the individual impact of each variable on the dependent variable.

**Model Summary**

The summary of the MLR model is as follows:

* Residuals:
  + Minimum: -0.50757
  + 1st Quartile: -0.10569
  + Median: -0.03614
  + 3rd Quartile: 0.06848
  + Maximum: 0.82054
* Coefficients:
  + Several variables, including minimum\_nights, number\_of\_reviews, and availability\_365, exhibited significant influence on 'price' at the 0.001 level.
  + Certain variables, notably West\_Region, Shared\_room, and \5\, were undefined due to singularities, underscoring the multicollinearity concerns.
* Model Fit:
  + Residual Standard Error: 0.1614 on 6048 degrees of freedom
  + Multiple R-squared: 0.437
  + Adjusted R-squared: 0.4354

The R-squared value of 0.437 implies that the model explains approximately 43.7% of the 'price' variability. The Adjusted R-squared, which accounts for the number of predictors, corroborates this with a value of 0.4354.

**Insights**

While the MLR model offers a reasonable fit, the warnings about rank deficiencies highlight concerns regarding multicollinearity. To address this, future analyses might consider techniques like Variance Inflation Factor (VIF) assessment or regularization methods such as Ridge or Lasso regression.

**Data Preparation for Regression Models**

Before delving into regression models, it's crucial to format the data appropriately. We converted the training data to a model matrix format, facilitating the seamless conversion of categorical variables into dummy variables.

A close up of blue text

Description automatically generated

**Lasso Regression (L1 Regularization)**

**Model Implementation**

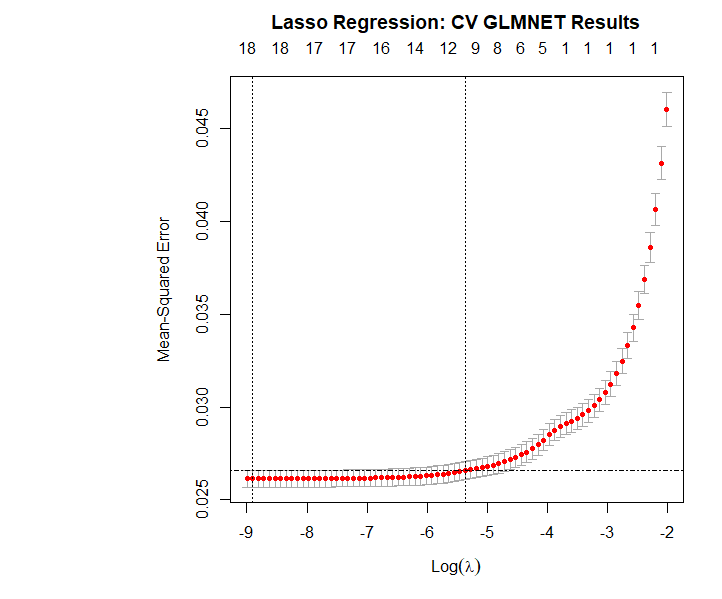
We applied Lasso Regression using cross-validation through the cv.glmnet function, setting alpha=1.



**Lambda Selection**

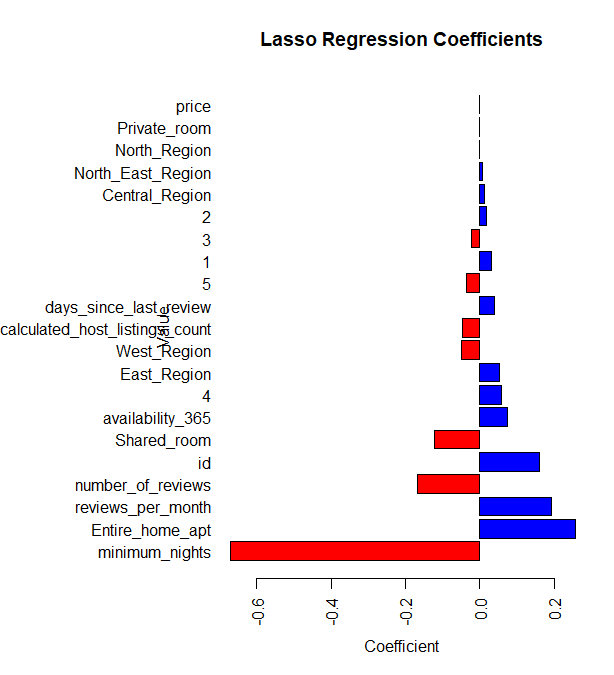
From the cv.glmnet function, two lambda values emerged as pivotal:

1. **lambda.min.lasso**: Corresponding to the smallest cross-validated error.
2. **lambda.1se.lasso**: The most regularized model with an error within one standard error of the minimum.

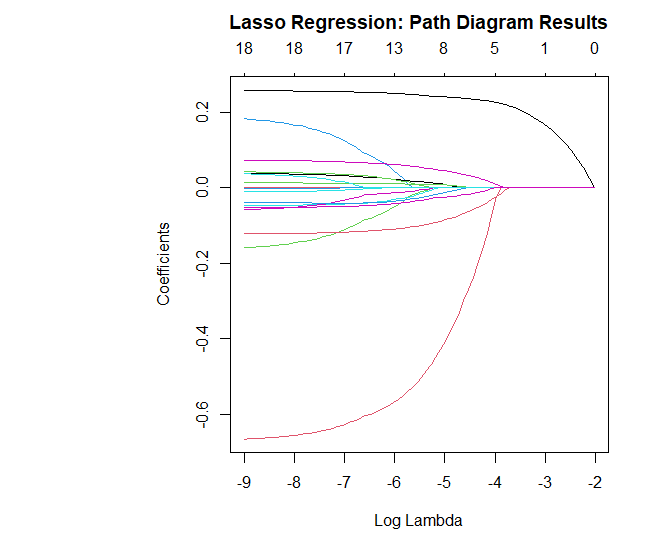


**Coefficient Analysis**

For the Lasso model at lambda.min.lasso, several coefficients were reduced to zero, while others retained non-zero values. Notably, variables like minimum\_nights, number\_of\_reviews, and availability\_365 exhibited significant coefficients.



The Lasso path diagram presents the trajectory of each predictor's coefficient as the penalty strength (lambda) changes. At high levels of lambda, all coefficients are at zero. As lambda decreases (moving to the left on the plot), more coefficients have non-zero values.



**Ridge Regression (L2 Regularization)**

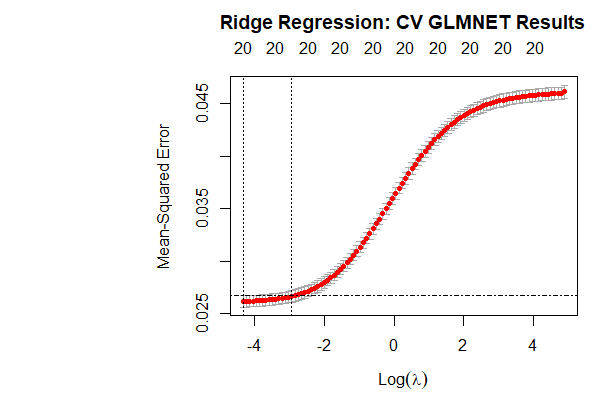
**Model Implementation**

Ridge regression, also known as L2 regularization, penalizes coefficient magnitude to deter overfitting. Unlike Lasso, Ridge doesn't force coefficients to zero but shrinks them towards it.



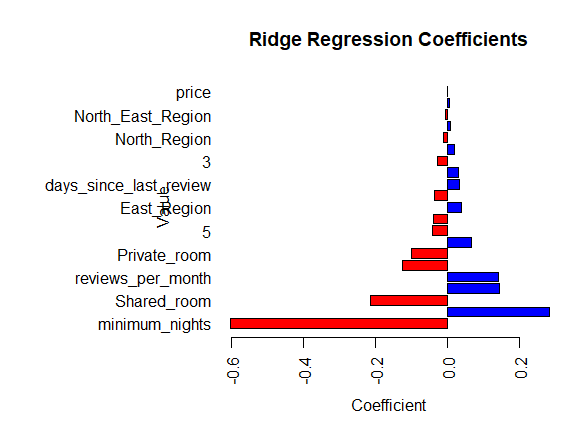
**Lambda Selection**

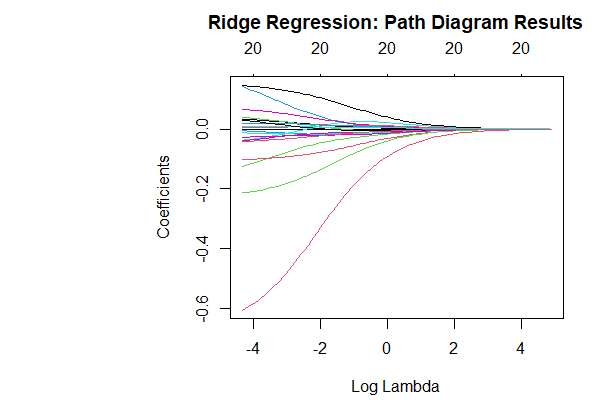
From the cv.glmnet function for Ridge regression:



**Coefficient Analysis**

In the Ridge regression model at lambda.min.ridge, no coefficients were set to zero. However, they were reduced towards zero. Variables such as minimum\_nights, number\_of\_reviews, and availability\_365 remained influential based on their coefficients.





**Summary**

Regularization techniques like Lasso and Ridge have demonstrated their utility in refining the MLR model, particularly in addressing multicollinearity. These methods simplify the model and counteract the overfitting observed in the base MLR model.

**GBM Model: Gradient Boosting Machine Analysis**

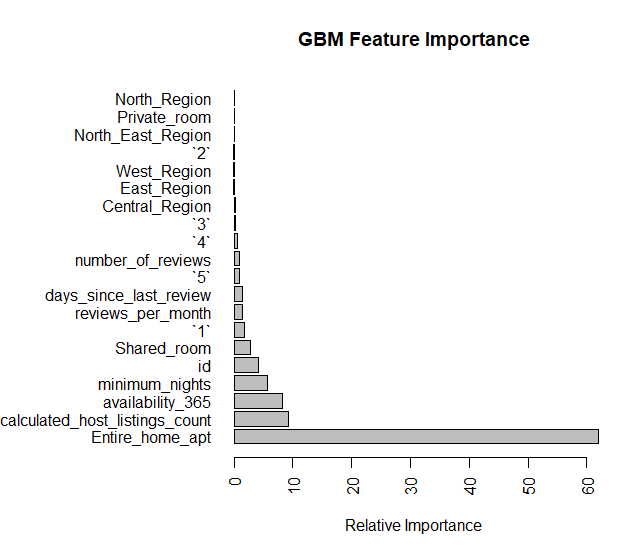
The Gradient Boosting Machine (GBM) stands as a robust ensemble learning algorithm adept at discerning intricate relationships and patterns within data. By iteratively adding trees, GBM aims to minimize errors and bolster predictive accuracy.

**Training and** **Results**

The GBM model was trained using various features to predict the target variable price. The summary of the model revealed the relative importance of these features in making predictions.



A horizontal barplot was created to visualize the feature importance. This plot aids in quickly identifying the variables that the GBM model deems important for the prediction of price.



The 'Entire\_home\_apt' variable emerged as the dominant predictor, accounting for a substantial 61.995% of the model's importance. It was closely trailed by 'calculated\_host\_listings\_count' and 'availability\_365', which held relative importance values of 9.201% and 8.215%, respectively.

While features like 'Shared\_room', variables labeled '1', 'reviews\_per\_month', and 'days\_since\_last\_review' made notable contributions, they paled in comparison to the top three variables.

Interestingly, geographic indicators such as 'Central\_Region', 'East\_Region', and 'West\_Region' held diminished importance values. This suggests that property-specific attributes might wield more influence over pricing than mere geographical locations.

**Summary**

The GBM model adeptly ranked features based on their relative importance in predicting the target variable. The findings accentuate the pivotal roles of property type, host's listing count, and annual availability in shaping listing prices. Moreover, it's evident that certain geographical regions might not be as influential in price determination as property-centric attributes.

**Interpretations**

From the results of the models and their predictions on the test dataset, we can derive the following interpretations:

A screenshot of a computer code

Description automatically generated

* **Model Performance**:
  + Of the models evaluated, the GBM model surpassed both Lasso and Ridge regressions in performance metrics.
  + he GBM model's MAE stood at a mere 0.1063, signaling its minimal average prediction error. Its R-squared value of 0.5548 indicates that it accounts for a significant portion of the variance in the test set compared to other models. The model's minimal RMSE further cements its position as the top-performing model.
* **Significance of Variables**:
  + The 'Entire\_home\_apt' variable's paramount importance in price prediction suggests that listings labeled as 'Entire Home/Apartment' significantly sway their pricing, potentially commanding a premium.

**Recommendations & Conclusions**

**Recommendations**

1. Stakeholders could reap substantial benefits by channeling marketing efforts towards entire homes and apartments, given their pronounced influence on pricing.
2. Future analytical endeavors should consider integrating more property-centric features, especially since the GBM model underscored their importance over location-based attributes.
3. In light of the GBM model's stellar performance, subsequent modeling attempts might lean towards boosting methods or delve into advanced ensemble techniques.

**Next Steps**

1. Delve into the nexus between 'Entire\_home\_apt' and price to ascertain if specific factors (e.g., location, size, amenities) amplify its premium.
2. Integrate interaction terms or polynomial features to unearth potential non-linear relationships.
3. Given the prominence of certain property features, procuring more detailed property data could unveil nuanced insights.

**Methods and Algorithms for Future**

1. Neural Networks: Given the right computational resources and data preprocessing, deep learning methods might capture complex relationships in the data.
2. Stacking or Blending: Combining predictions from multiple models can often lead to better accuracy.
3. Time-Series Analysis: If the data has a temporal component, exploring time-series models can be advantageous.

**Final Conclusions**

* Property-specific attributes significantly influenced price predictions, overshadowing geographical regions.
* 'Entire\_home\_apt' emerged as the most influential predictor, with host-related attributes following suit.
* The GBM feature importance bar plot and the numerical metrics collectively provided a comprehensive overview of each model's accuracy, with GBM outshining in all facets, from MAE to R^2 and RMSE.

Our comprehensive analysis of Singapore's Airbnb listings data has illuminated several critical determinants influencing listing prices. The essence of this report was not only to identify these key factors but to empower potential hosts with actionable insights for price optimization. By understanding the weightage of elements like the 'Entire\_home\_apt' and its importance in the market, hosts can strategically price their properties. This ensures not just higher profitability for them but also guarantees value for their guests. In a competitive market such as Singapore's Airbnb scene, harnessing this data-driven knowledge becomes the linchpin for both hosts seeking profitability and guests looking for value.