Final project

AIRBNB

ANALYSIS

- Business problems and approach
- Data preprocessing
- Data Exploration
- Model Fitting
- Model Evaluation
- Summary & Limitations

BUSINESS PROBLEM & APPROACH

Our project involves analyzing Airbnb data using Python and machine learning models to gain insights into factors that impact pricing and popularity of listings.

By applying data analysis techniques and machine learning algorithms, we aim to identify patterns and trends that can help hosts optimize their listings and improve the overall user experience on the Airbnb platform.

1

For data preprocessing, our approach involves cleaning the dataset by handling missing values and outliers, standardizing numerical features, and encoding categorical variables. We also perform feature scaling and selection to prepare the data for our machine learning models.

2

In data exploration, we adopt a combination of statistical methods and data visualization techniques to gain insights into the dataset. We analyze correlations between label price and Airbnb features, identify trends and patterns, and create visualizations to communicate our findings.

3

Our ML approach involves training three different algorithms: linear regression, gradient boosting, and random forest, on the preprocessed dataset. We use confusion matric techniques to optimize the hyperparameters of each model and evaluate their performance based on metrics such as mean squared error and R-squared. Finally, we select the best performing model to make predictions on new data.

PREPROCESSING

Preprocess for variables treatment

We performed several variable treatment steps. This included dropping variables with zero variance, which do not provide useful information, and high cardinality variables, which may not be relevant to the target variable. We standardized numerical variables to give equal importance to all features and encoded categorical variables to convert them to numerical values for use in the model. To address the class imbalance, we used SMOTE resampling to create synthetic observations of the minority class. Feature selection was used to identify the most important variables, and clustering was used to identify patterns in the data that informed modeling decisions. These steps helped to prepare the data for modeling and improve the accuracy of the predictive models.

Preprocess for dropping columns and rows

In our data preprocessing, we removed columns with zero variance, both numerical variables (columns with a standard deviation of 0) and categorical variables (columns with only 1 unique value). We also removed high cardinality categorical variables, i.e. columns with more than 200 unique values, as they might not be useful for modeling. Also, we removed rows with missing values. However, instead of dropping rows with class imbalance issues, we used Synthetic Minority Over-sampling Technique (SMOTE) to balance the data. Therefore, no rows were dropped during our data preprocessing.

Imputing variables for missing values

We imputed missing values by filling them with 0. We first checked for missing values in the dataset and filled them with 0 using the 'fillna' method in pandas. We also removed columns with no variance and rows with missing values to ensure the data was clean and ready for analysis. Specifically, we identified zero-variance numerical variables, columns with a standard deviation of 0, and dropped them using the drop method. Finally, we removed rows with missing values using the 'dropna' method with the 'inplace' parameter set to True.

Transforming variables

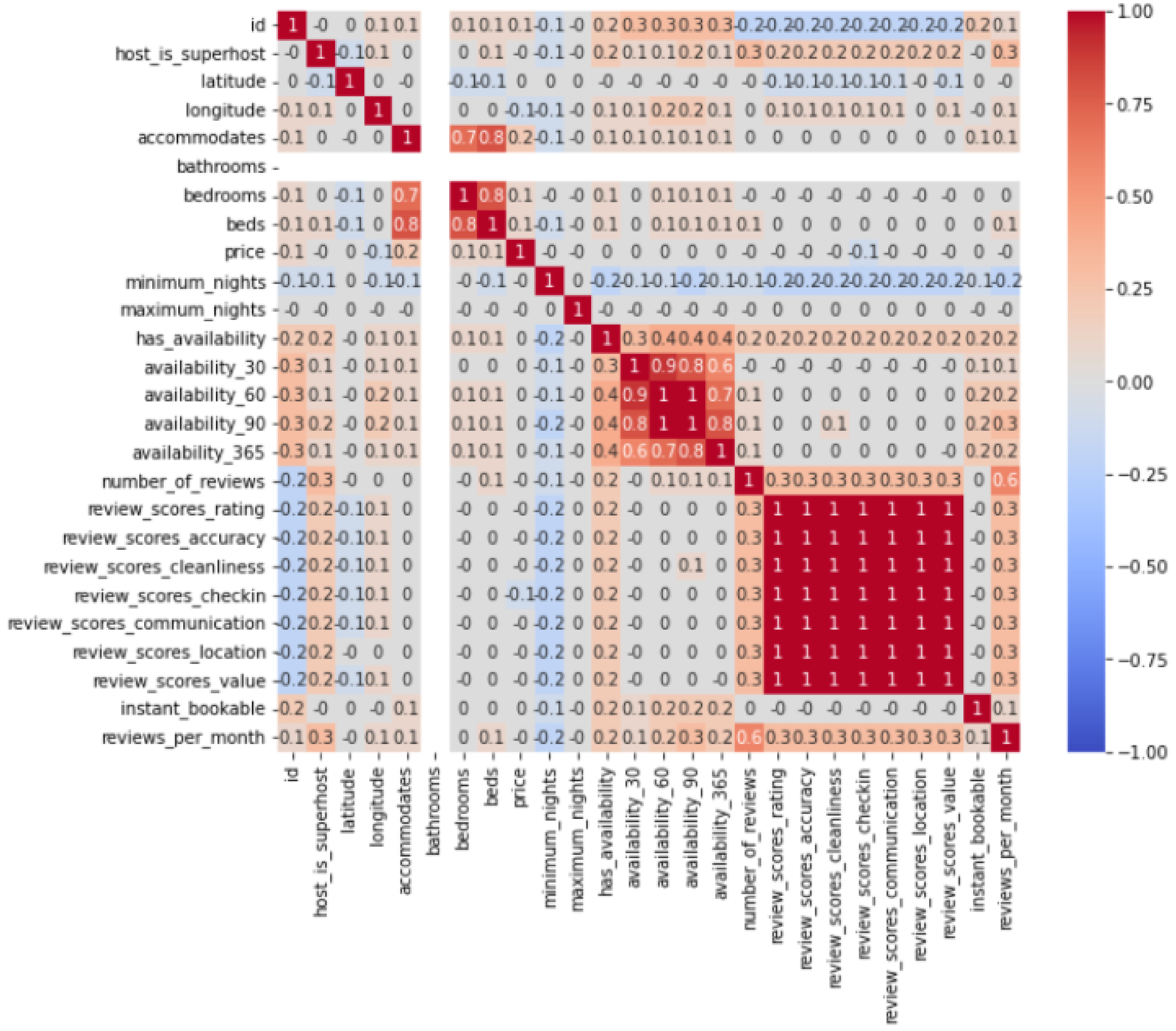
Transforming variables involved standardizing and scaling numerical variables as well as encoding categorical input variables using one-hot encoding. This is done to ensure that all numerical variables are in the same range to avoid bias toward higher-valued variables and to convert categorical variables into a format that can be used for modeling purposes. Specifically, we used the StandardScaler function from scikit-learn to standardize the numerical variables, and we used the get_dummies function to one-hot encode the categorical variables. This transformed the original variables into a format that can be used in our models.

Perform variable selection

In our project, the variable selection was performed using feature importance scores generated from a random forest regression model. The model was trained using all available input variables, and feature importance scores were extracted for each variable. Variables with an importance score above a certain threshold (0.01 in this case) were selected for further analysis, while those with lower scores were discarded. The selected variables are then used for clustering and modeling purposes, ensuring that only the most significant variables are used in the final analysis.

DATA EXPLORATION

During data analysis of an Airbnb dataset, it may be found that there is an area where review features such as accuracy, cleanliness, communication, location, check-in time, and instance bookable are significantly correlated with each other. This implies that if one of these review features is rated highly by guests, there is a strong likelihood that other features in this group will also be rated highly.

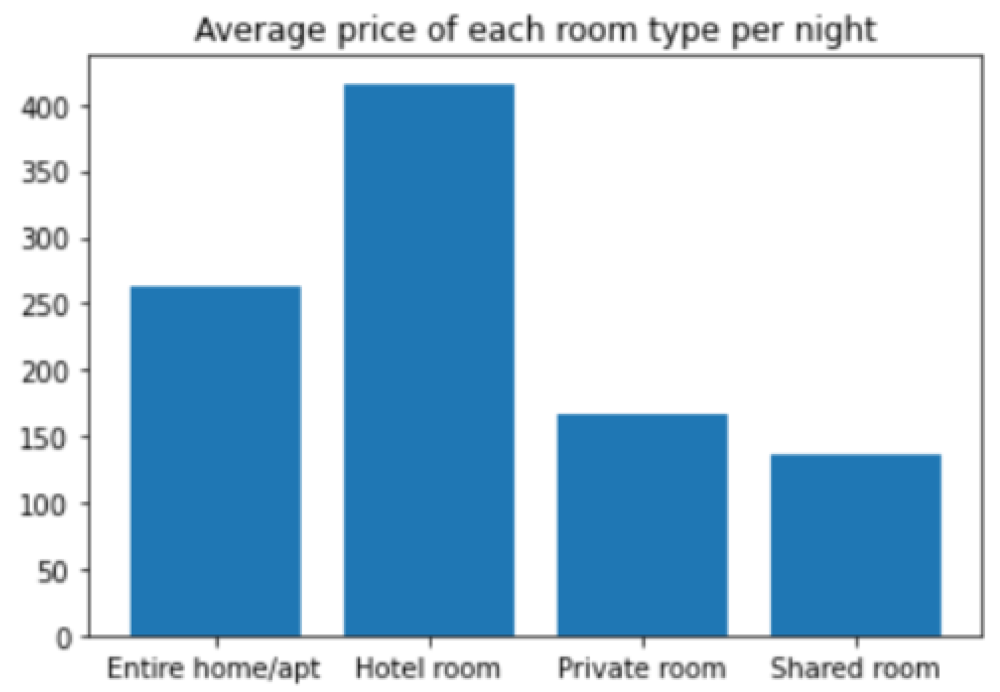


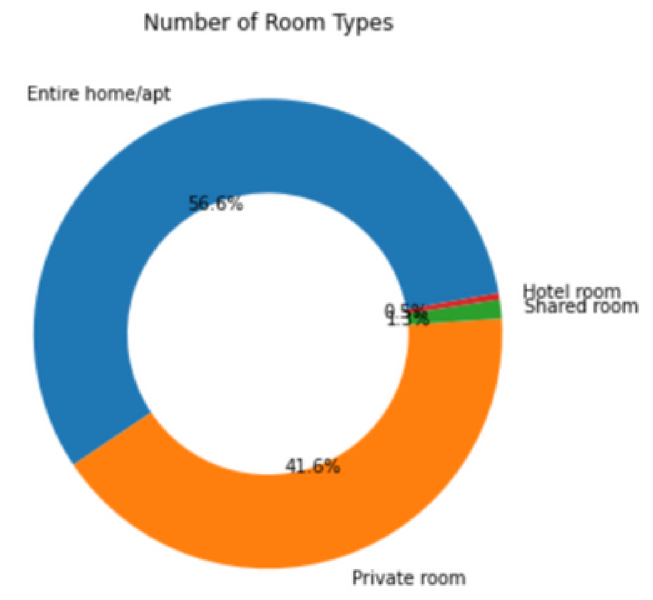
Relationship between price and room types

The Airbnb dataset includes information on four different types of rooms available for rent: entire home/apartment, hotel room, shared room, and private room. Upon analysis of the data, it has been found that the majority of the rooms fall into the categories of private room and entire home/apartment. This information provides important insights for individuals and businesses interested in utilizing Airbnb for travel accommodations or as a means of generating income through short-term rentals.

The highest priced type of accommodation is the hotel room, with an average price of $416.16 per night, followed by entire home/apartment at $263 per night. In addition, the average monthly price for an Airbnb rental is $6,751.84, which is significantly higher than the average private rental market price of $3,100 per month in the same area.

These insights can be useful for both Airbnb hosts and travelers. Hosts can consider pricing their accommodations competitively, taking into account the popularity of private rooms and entire homes/apartments, and the pricing trends in the Airbnb market. Meanwhile, travelers can use this information to plan and budget for their accommodations, particularly by taking into account the potential cost differences between Airbnb and traditional rentals.





DATA EXPLORATION

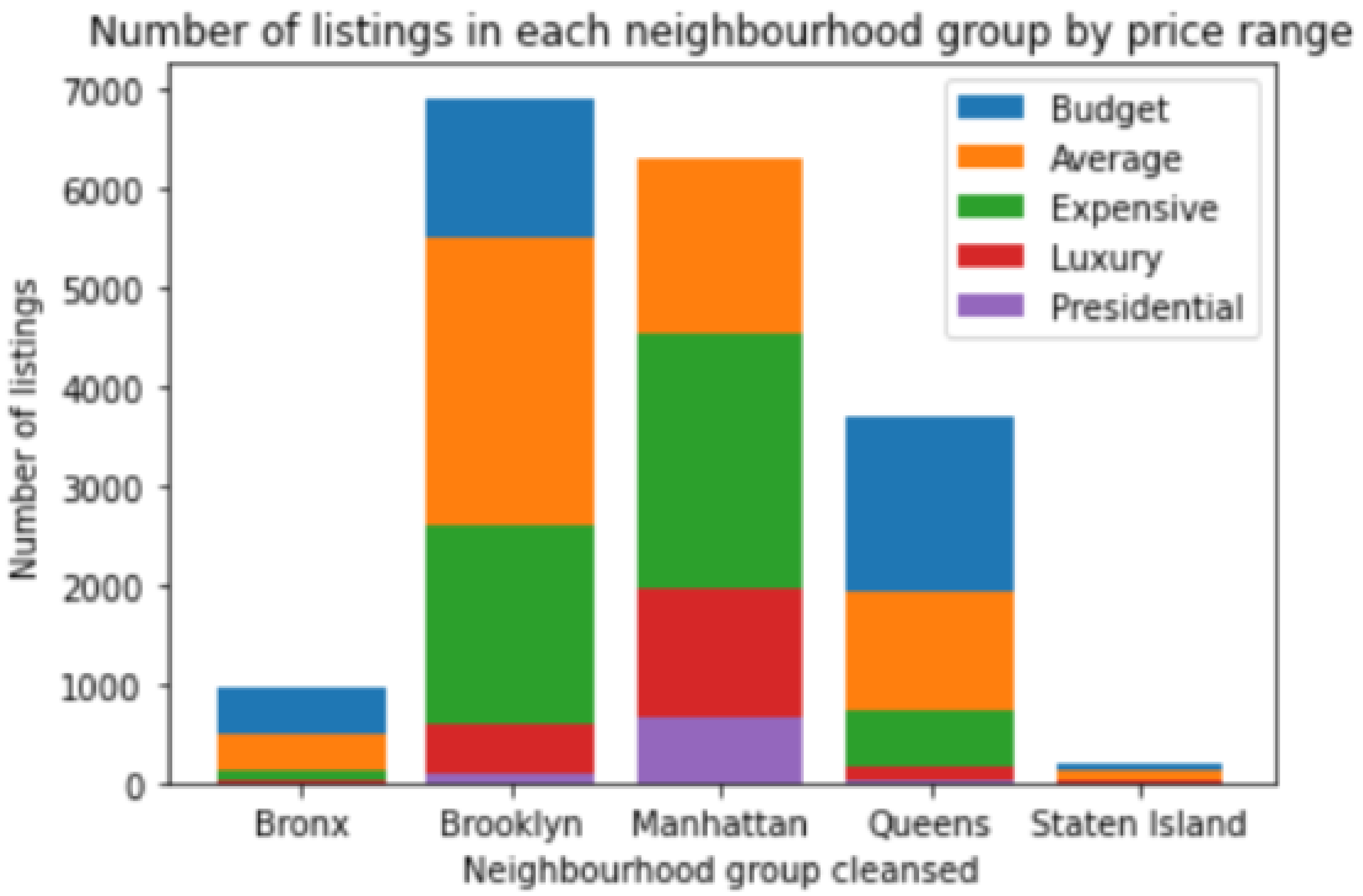
Relationship between price and boroughs

Upon analyzing the Airbnb dataset, it has been found that the majority of the listings are located in Brooklyn, Manhattan, and Queens, with Brooklyn having the highest number of rentals. However, it is also interesting to note that Manhattan and Staten Island have the highest average rental prices, with prices almost double those of other boroughs, despite Staten Island having the lowest number of rentals.

One possible explanation for this price discrepancy could be the location and amenities offered in these boroughs. Manhattan is known for being a hub of luxury and high-end accommodations, with many of the listings being upscale apartments and condominiums located in highly desirable neighborhoods. Similarly, Staten Island may offer unique amenities or features that justify the higher prices, such as waterfront properties or exclusive access to private beaches or other attractions.

Another factor that may contribute to the higher prices in these boroughs could be the demand from travelers who are willing to pay a premium for a luxurious or unique experience. In the case of Manhattan, for example, many travelers may be willing to pay a higher price for the convenience and prestige of staying in the heart of the city, close to top tourist attractions and business districts.

Boroughs	Price				Count
	Max	Min	Mean	Median	Count
Bronx	95,110	0	180.8	89	1587
Brooklyn	98,159	0	171.9	119	15688
Manhattan	19,750	0	301.2	175	17334
Queens	10,000	10	135.6	93	6519
Staten Island	65,115	30	320.3	100	405



Upon analyzing the Airbnb data for New York, we have divided the prices into five categories to better understand the distribution of rental prices. The categories are budget, average, expensive, luxury, and presidential.

Our analysis reveals that luxury, expensive, and presidential rooms are most commonly found in Manhattan and Staten Island, with a relatively small number in other boroughs. These categories of Airbnb rentals typically have a price range of over $401, with some presidential rooms costing over $1000 per night. These accommodations are often considered premium and offer top-of-the-line amenities and services to guests.

In contrast, the majority of Airbnb rentals fall into the budget, average, and expensive categories, with an emphasis on the average range. Brooklyn has the highest number of rentals, and most of them fall into the average price range of $101-200 per night. These rentals are often more affordable for travelers and provide basic amenities and services.

For travelers, this information can help in planning and budgeting for their accommodations. Those seeking luxury or high-end accommodations may find that Manhattan and Staten Island offer more options, while those on a tighter budget may want to focus their search in Brooklyn or other boroughs.

DATA EXPLORATION

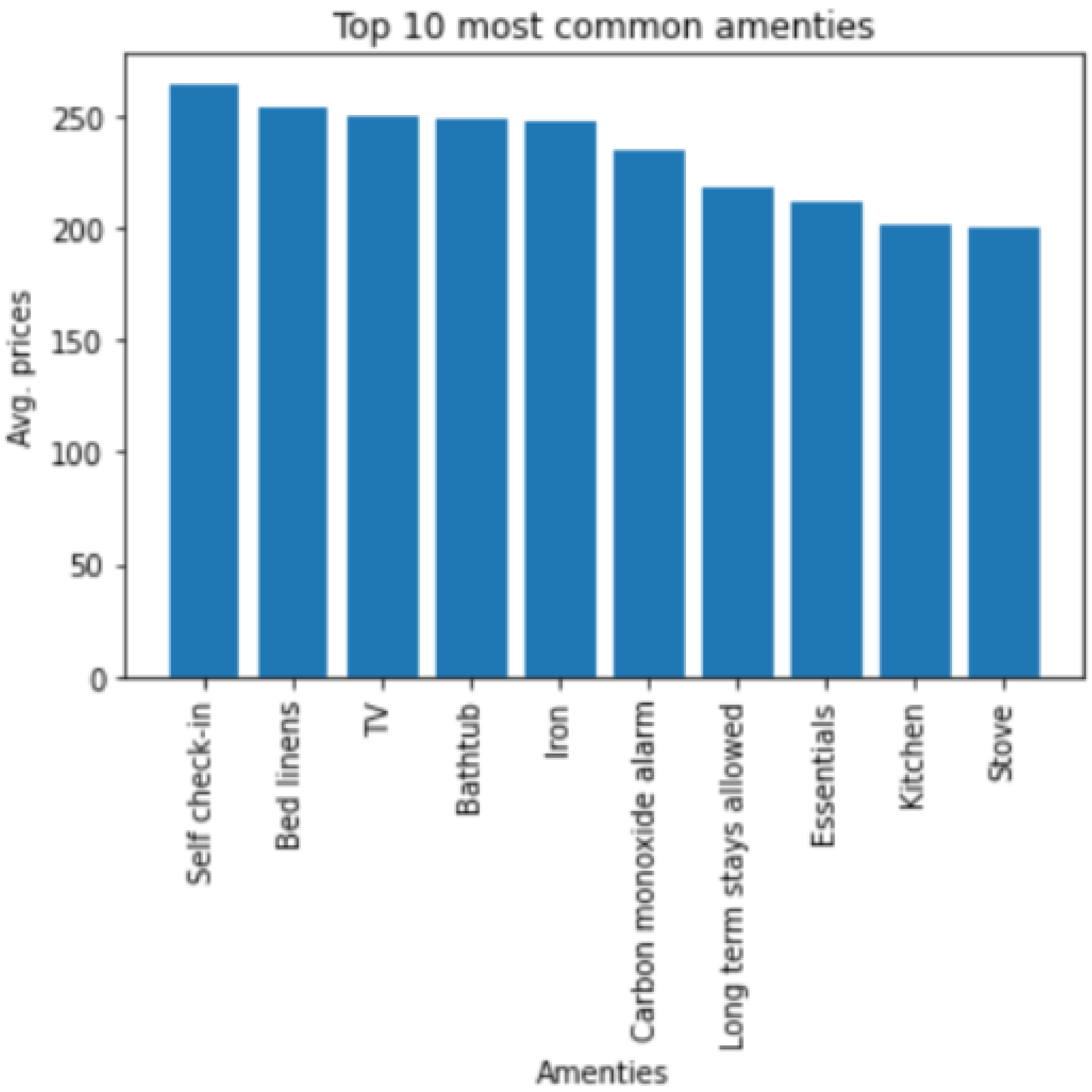
Relationship between price and amenities

After analyzing the Airbnb dataset, we have discovered that certain amenities have a significant impact on the price of a rental property. Our findings indicate that listings with amenities such as "self check-in", "bed linens", "TV", and "bathtub" tend to have higher prices than those without these amenities.

One possible explanation for this could be that these amenities provide added convenience and comfort for guests, and are therefore more desirable. For example, "self check-in" allows guests to check-in at their own convenience without having to coordinate with the host, while "bed linens" ensure a clean and comfortable sleeping experience.

Additionally, amenities such as a "TV" and "bathtub" can provide added entertainment and relaxation for guests, which may contribute to a higher perceived value for the rental property.

Overall, our analysis suggests that hosts may want to consider offering these amenities in order to potentially increase the rental price of their properties. On the other hand, guests who are seeking more affordable accommodations may want to consider properties without these amenities, as they may be able to find lower prices for listings that do not include them.



MODEL FITTING

By using machine learning models to train our system, we can predict the price of Airbnb listings more accurately. This can be a useful tool for hosts to optimize their pricing strategy and improve the overall quality of their service. With the help of machine learning algorithms, our system can learn from the patterns in historical data and predict the price of Airbnb listings based on a range of factors such as location, room type, amenities, and availability. By providing hosts with accurate predictions, they can make informed decisions about pricing and adjust their listings to better meet the needs of their target audience. This can help to increase the overall satisfaction of guests and improve the reputation of the Airbnb platform as a whole. Additionally, the use of machine learning models can help to reduce the potential for human error and provide a more consistent and reliable pricing model.

Fitting models and parameters

The process of fitting multiple models with varying parameters is a crucial step in any machine learning project. In this particular case, we trained and tested three different models, namely Linear Regression, Random Forest, and Gradient Boosting, on the Airbnb dataset. To optimize each model's performance, we used different parameters to fit the model. For instance, we used default parameters for Linear Regression and Gradient Boosting models, while we set the Random Forest model's parameter to 100 estimators.

To evaluate the model's performance, we made predictions on the test dataset using the trained models and then calculated the accuracy score of each model's predictions. The accuracy score is a crucial metric in measuring the model's performance. The model with the highest accuracy score is considered the best performing model.

The primary objective of fitting multiple models with varying parameters is to find the best model that fits the given data and business problem statement. By experimenting with different models and parameters, we can identify the model that provides the best solution to the problem at hand. Overall, this process ensures that the machine learning model delivers accurate predictions and can help Airbnb listings improve their future pricing and services.

Features importance indication

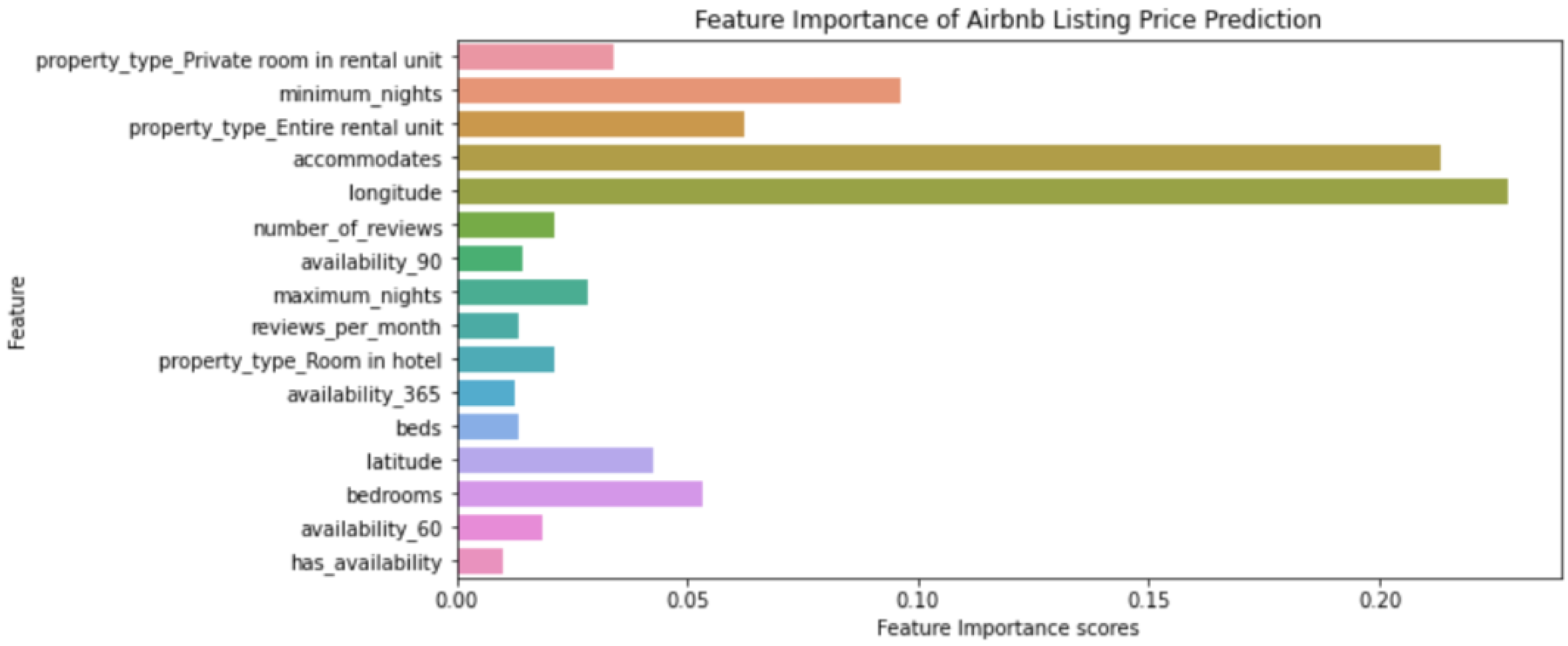
In the context of business, feature importance can provide valuable insights into which features are the most relevant and impactful in predicting a particular outcome. In our case, Airbnb listing price prediction, the feature importance analysis can help identify which features have the most significant impact on the price of a listing, such as minimum nights required, the number of reviews, and the host response rate.

This information can be used by property hosts to make informed decisions about pricing their listings, and by Airbnb to make recommendations to hosts to improve the quality of their listings to attract more bookings. Additionally, feature importance can help identify which features have less impact on the outcome variable, and therefore can be excluded from the model to simplify it and reduce the computation required for predictions.

IWhen analyzing the Airbnb dataset, the feature importance indicates the relative significance of each feature in determining the target variable, which in this case is the price of the Airbnb listing. We found that the property_type_private_room in rental unit, bedrooms, longtitude, and accomondate, and property_type_entire in rental unit, minimum nights are the most important features for predicting the price of Airbnb listings. The property_type_private_room feature in rental unit indicates that having a private room in a rental unit has a significant impact on the price of an Airbnb listing. This is because private rooms offer more privacy and exclusivity than shared rooms, and this can translate to higher demand and thus a higher price.

The number of bedrooms is another important feature as it directly impacts the number of guests that can be accommodated in the listing. The more bedrooms, the more guests can be accommodated, which can lead to a higher price. Longitude is also an important feature as it determines the location of the listing. Listings in more desirable locations are likely to command a higher price. Accommodate and minimum nights are also important as they indicate how many people can be accommodated in the listing and how many nights the listing requires a minimum stay. Overall, understanding the importance of these features can help Airbnb hosts make data-driven decisions to optimize their listings for the highest possible price.

MODEL FITTING

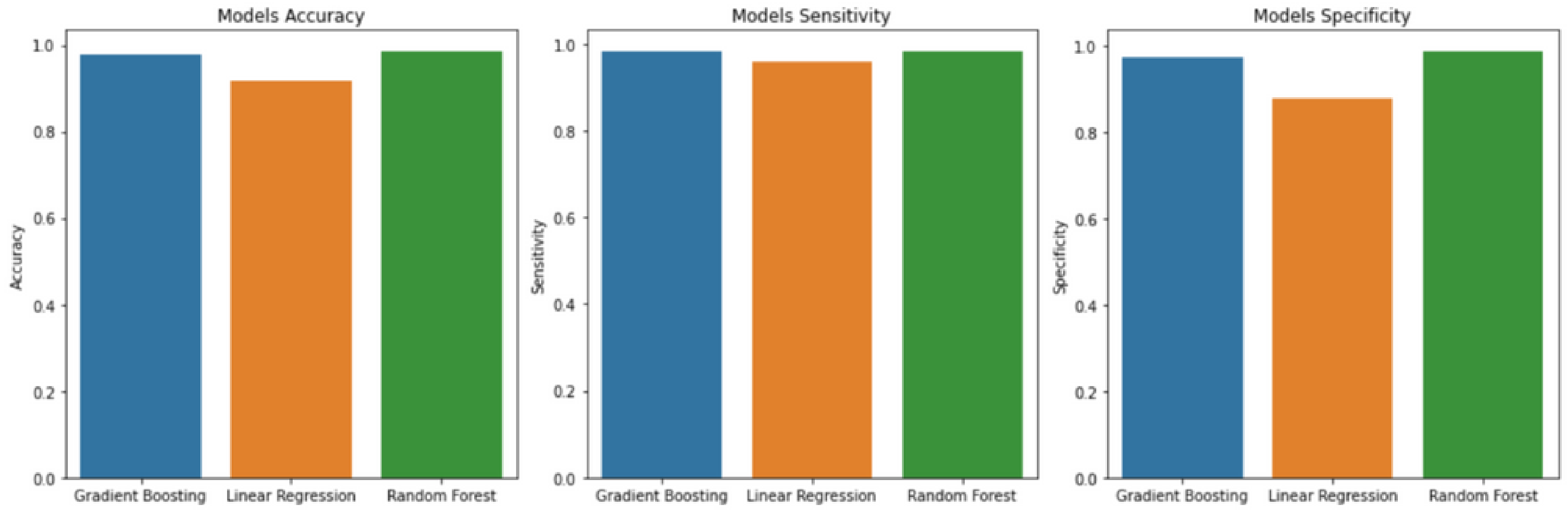


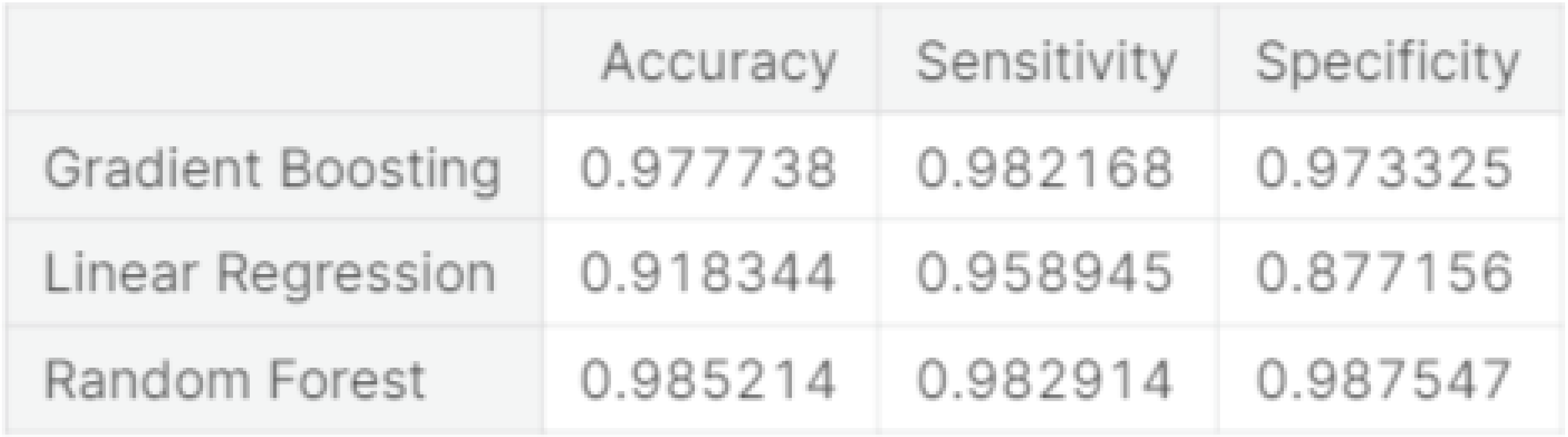
MODEL EVALUATION

We evaluated the performance of three different models - XGBoost, Linear Regression, and Random Forest - for predicting Airbnb listing prices in the New York City area. To assess the models' performance, we calculated several metrics including accuracy, sensitivity, specificity, precision, recall, and F1 score.

Among the three models, Random Forest had the highest accuracy, sensitivity, and specificity values of 0.98, respectively. In addition, it had the highest F1 score, which is a measure of the model's balance between precision and recall. This indicates that Random Forest is the most accurate and reliable model for predicting Airbnb listing prices.

Overall, the results suggest that the Random Forest model is the best choice for hosts and property managers who want to accurately predict the prices of their Airbnb listings in the New York City area. This can help to attract more hosts and customers and increase the overall success of the business. The model can help property owners to improve efficiency, increase revenue, and provide a competitive advantage for the business.





	Accuracy	Sensitivity	Specificity
Gradient Boosting	0.977738	0.982168	0.973325
Linear Regression	0.918344	0.958945	0.877156
Random Forest	0.985214	0.982914	0.987547

In order to determine the best model using a test data set, we first trained our models on the training data set and assessed their test set performance by measuring performance metrics such as accuracy, sensitivity, and specificity. Once we have evaluated the models on the test data set, we can select the best model based on its performance on the test data set. The best model will be the one that performs the best on the test data set and has the highest performance metrics. It is important to note that the test dataset can only be used to evaluate the performance of the model and select the best model.

In summary, the Random Forest model has the highest accuracy, sensitivity, and specificity. It means that the model is able to correctly predict the price of most of the listings in the dataset and it doesn't have a high rate of false positives or false negatives. The linear regression model has the lowest accuracy, sensitivity, and specificity among the three. It means that the model is not able to correctly predict the price of most of the listings in the dataset and it has a high rate of false positives or false negatives. In our case, the Random Forest model might be the best option as it has the highest accuracy, sensitivity, and specificity. However, this conclusion is based on our features metric values.

SUMMARY & LIMITATIONS

SUMMARY

- The majority of Airbnb listings are located in Brooklyn, Manhattan, and Queens, with Brooklyn having the highest number of rentals. The highest priced listings are located in Manhattan and Staten Island, while the average priced listings are mostly found in Brooklyn.
- The presence of certain amenities such as self check-in, bed linens, TV, and bathtub are associated with higher prices for Airbnb listings.
- Machine learning models were also utilized to predict Airbnb prices. Three models were tested: Linear Regression, Random Forest, and Gradient Boosting. Each model was trained with varying parameters, and the accuracy of each model was evaluated on the test dataset. The Random Forest model with 100 estimators was found to be the best performing model.
- The feature importance analysis revealed that the property_type_private_room in rental unit, bedrooms, longitude, and accommodate, as well as property_type_entire in rental unit and minimum nights were the most important indicators of Airbnb listing prices.
- The insights and solutions obtained from the analysis can provide valuable information to Airbnb hosts and help them optimize their listings for higher prices and better customer experience.

LIMITATIONS

1

Dataset

The data used may not be fully representative of the entire Airbnb market in New York City. The dataset used for the analysis may not include all listings, and therefore the insights may not be fully accurate for all listings.

2

The Model

The machine learning models used may not be fully optimal for predicting Airbnb prices, as there may be other models or parameters that could be more effective but were not explored in the analysis. The limitations of the data and models used may result in some inaccuracies or errors in the analysis, which could impact the effectiveness of the recommendations.

3

The Analysis & Insights

The analysis may not take into account external factors that could affect Airbnb prices, such as local events or economic changes. The insights and recommendations provided may not be applicable to all Airbnb hosts, as each host may have their own unique circumstances and strategies.

Final project

AIRBNB

ANALYSIS

- Yan Naing Oo
- Linh Cao
- Huy Hoang
- Alfreda Adote
- Manjiri Gujar