DATA 211 Midterm Project | Stella Chen, Yangge Xu

A. Overview and Motivation

Since its founding in 2008, Airbnb has emerged to be one of the biggest home-sharing platforms in the world. Not only does Airbnb provide tourists with a brand-new way of traveling, but more importantly, it allows the short-term leasing of vacant housing, which creates passive revenue for homeowners. In 2022, Airbnb had about 2,249,434 listings in the US alone with a revenue of $8.4 billion.[1]

The goal of this project is to accurately predict Airbnb prices in New York based on a variety of different parameters. Previous literature suggested that Airbnbs are more profitable compared to long-term listings, and therefore there has been an increased threat of displacement and a decrease in affordability for long-term residents(Yrigoy 2018)[2]. We hope to first understand the dataset itself and conduct some data cleaning as well as feature extraction to see which parameters are the most important in determining the price of the listings. After that, we will develop machine learning models(MLP, KNN, Regression Tree, Neural Network) and compare the accuracy of each model. By developing the predictive model for Airbnb prices, we will be able to provide both hosts and users with an estimate of the price of the listing to facilitate more transactions.

B. Data Source

The dataset used in our project comes from insideAirbnb.com[3], a website that collects data on Airbnb in different cities. The dataset gathers data on all the existing Airbnbs within a specific city and features of these places, ordered by month. Notable features include room type, occupancy, listings per host, etc. The data is collected by a group of volunteers who web-scraped the Airbnb website in order to empower local communities to understand, decide and control the role of renting residential homes to tourists. We decided to focus on the Airbnb data specifically for New York City up to March 2023.

There are a total of 42,931 listings across 250 different neighborhoods in New York City. Out of those, 56.6% of the listings are entire homes/apartments, 41.6% are private rooms, and the remaining are shared rooms and hotel rooms.

C. Data Cleaning

The data itself is actually quite messy due to how it was scraped. Therefore, we had to conduct our data cleaning and feature engineering processes. Some variables like price and bathrooms were originally in strings format, we parsed them and converted the number figures from text to float. Additionally, three columns: 'neighbourhood_cleansed', 'property_type', and 'room_type' are categorical variables, so we decided to utilize one-hot encoding on these columns in order to convert the categorical variables into 1s and 0s(we dropped the first column of one hot encoding to avoid over parameterization). About 1%(394) of the data has a price

[1] https://news.Airbnb.com/Airbnb-q4-2022-and-full-year-financial-results/

[2] Yrigoy, Ismael. "Rent gap reloaded: Airbnb and the shift from residential to touristic rental housing in the Palma Old Quarter in Mallorca, Spain." Urban Studies 56, no. 13 (2019): 2709-2726.

[3] "New York City." Inside Airbnb: Home. Accessed May 3, 2023. http://insideAirbnb.com/new-york-city/.

above $1000, so we decided to drop them as they might be outliers to our training data. After dropping nan and certain columns, we are left with 22,079 rows and 299 columns.

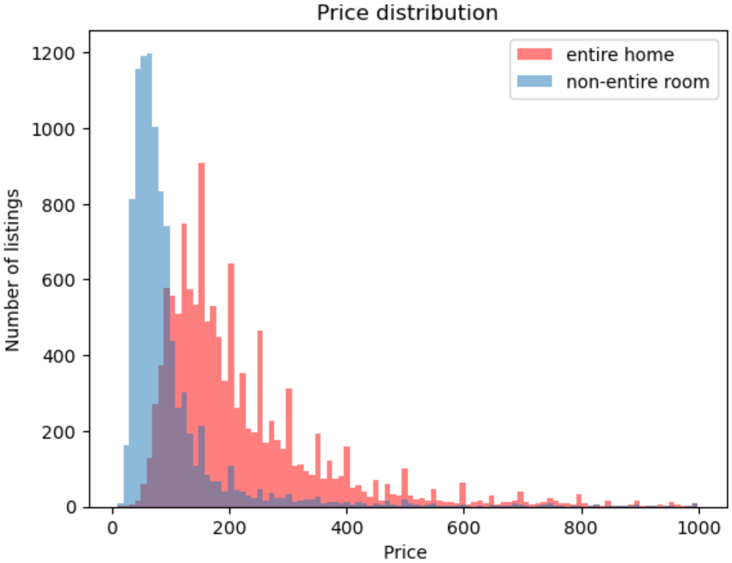
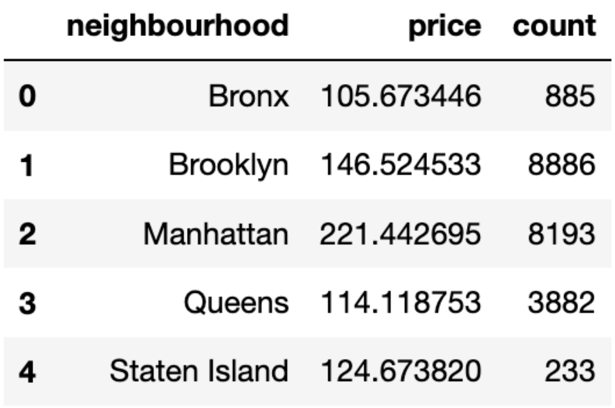


Figure 1



	neighbourhood	price	count
0	Bronx	105.673446	885
1	Brooklyn	146.524533	8886
2	Manhattan	221.442695	8193
3	Queens	114.118753	3882
4	Staten Island	124.673820	233

D. Exploratory Analysis

The next step in our analysis is to familiarize ourselves with the dataset itself. First, we looked at the distribution of our y variable(price) and found that most of the prices are concentrated within the 0-200 range. After the data cleaning processes, the current mean is 166.76 and the median is 125. Overall, Figure 1 shows that entire homes are more expensive than shared rooms/private rooms. We also investigated the geographic distribution of these listings and found that the majority of the listings are concentrated within the midtown Manhattan and Brooklyn region. Figures 2a and 2b indicate that the further away from these regions, the fewer listings there are. Additionally, there are price discrepancies between the

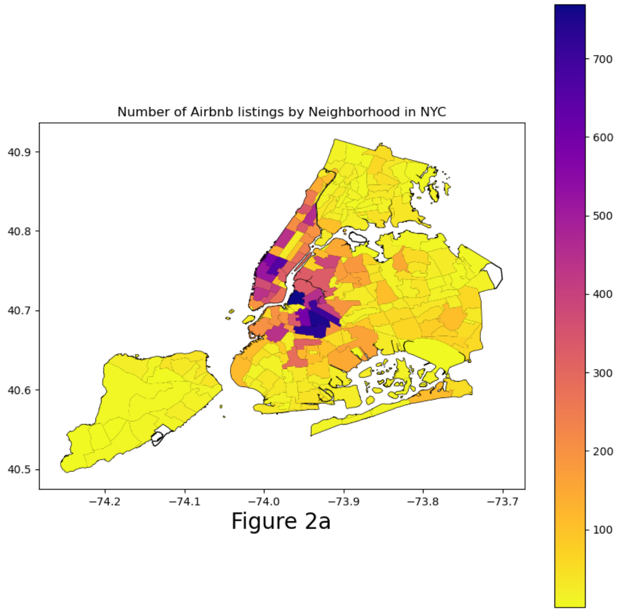


Figure 2a

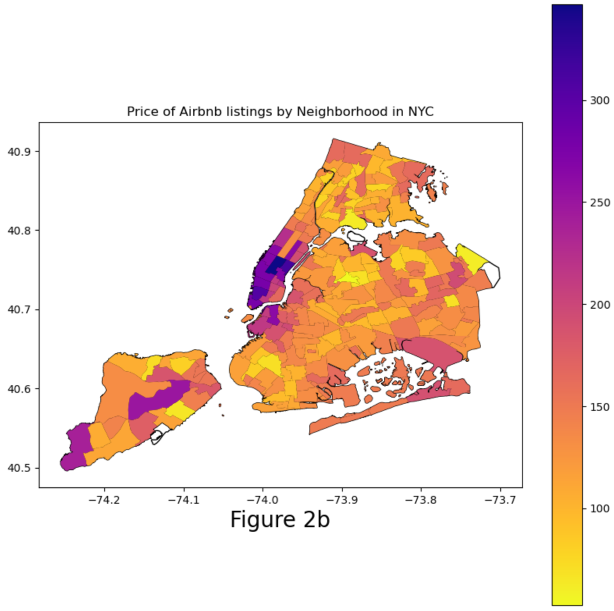


Figure 2b

different neighborhoods. Most of the listings with higher price are concentrated within the Manhattan area. The Bronx area has a price mean of $105.67, while the Manhattan area has a mean price of $221.44.

In order to further investigate the relationship between variables, we created the correlation heatmap(Figure 3) between the numerical columns. There are some correlations between price and accommodates(0.524), price and bedrooms(0.4225), and price and beds(0.4279). However, none of the correlations is above 0.6.

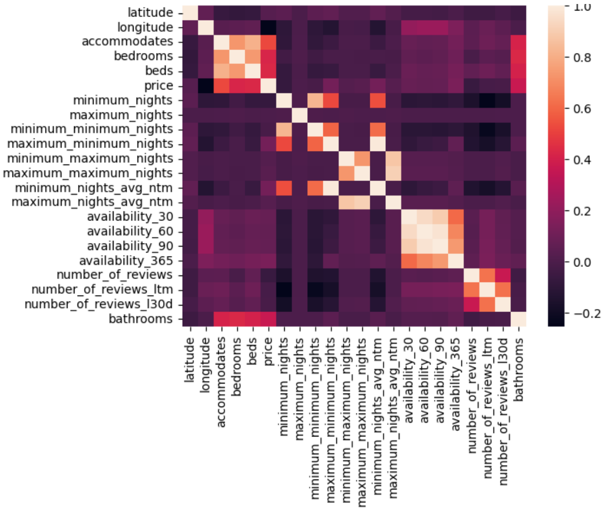
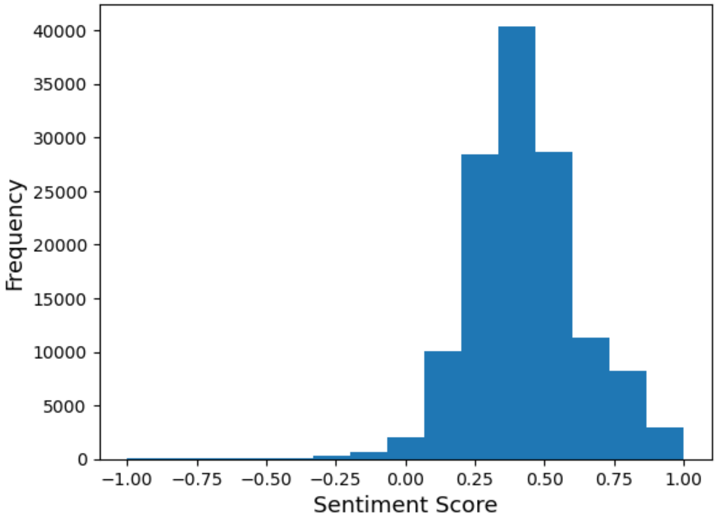
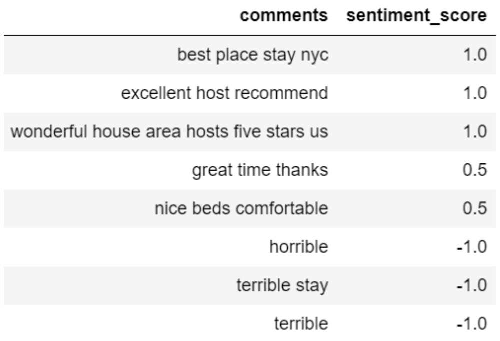


Figure 3

E. Sentiment Analysis

Figure 4. Distribution of Sentiment Scores





comments	sentiment_score
best place stay nyc	1.0
excellent host recommend	1.0
wonderful house area hosts five stars us	1.0
great time thanks	0.5
nice beds comfortable	0.5
horrible	-1.0
terrible stay	-1.0
terrible	-1.0

To investigate the relationship between comments and ratings in the Airbnb dataset, sentiment analysis was conducted on the listing comments. This involved combining the review dataset with the listing dataset using the common identifier of listing ID. To ensure the efficiency and representativeness of the model, a random sample of 1000 listings with at least 100 reviews was used for training.

The comment data underwent preprocessing, which involved removing punctuation, stop words, and converting all text to lowercase. Stop words such as "a" and "again," which do not contribute much to sentiment expression, were removed from the dataset using the nltk.corpus. A pre-trained sentiment analysis model from the TextBlob library was then applied to each review comment to calculate its sentiment score. The sentiment score is on a scale ranging from -1 to 1, with a variance of approximately 0.043. Negative comments scored -1 often contain strong negative words like "horrible" and "terrible," while positive comments scored 1 contain words like "best," "excellent," and "awesome." Comments with scores of 0.5 or other intermediate values tend to have more neutral words like "nice" or "lovely." Finally, the sentiment scores were aggregated by computing the average score for each listing ID.

The data was split into training and testing sets, and a multi-linear regression was performed with sentiment score as a predictor and rating as the target variable. The model was evaluated using mean squared relative error (MSRE), mean squared error (MSE) and mean absolute error (MAE), with a MSRE of 11.8%, a MSE of 0.0254 and MAE of 0.122 obtained. These results indicate that the sentiment analysis model provides a relatively accurate prediction of ratings given the dataset's maximum rating of 5.0, a standard deviation of 0.167, and a variance of 0.028.

However, when exploring the correlation between price and ratings, we found that the correlation coefficient was only 0.07. This indicates a weak positive correlation between the two variables. In other words, there is only a slight tendency for higher-rated listings to have higher prices, but the relationship is not strong.

It is important to note that while sentiment analysis provides insights into the sentiment expressed in the comments, it may not directly correlate with price. Other factors such as location, amenities, and property characteristics might have a more significant impact on pricing decisions. Therefore, it is necessary to consider additional variables and build machine learning models that encompass a broader range of features to accurately predict Airbnb prices.

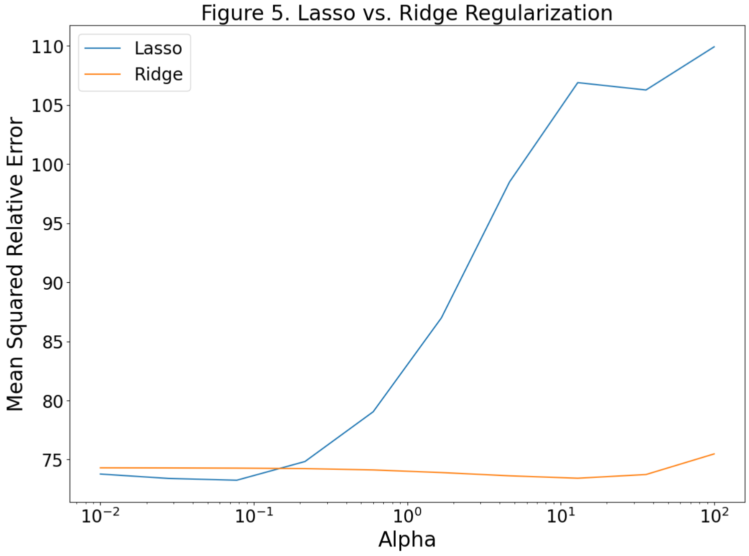
In conclusion, the sentiment analysis results suggest that sentiment can be a useful tool for predicting ratings in the Airbnb dataset. However, the correlation between sentiment and price is relatively weak. To further understand and predict Airbnb prices, it is crucial to explore and incorporate additional variables that are more directly related to pricing decisions.

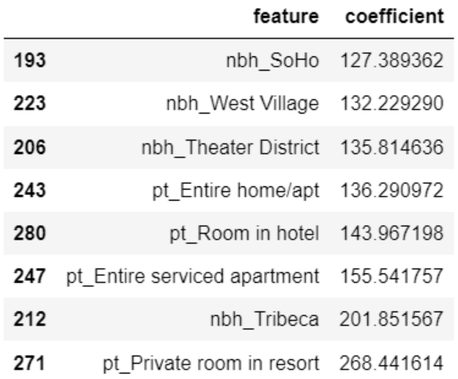
F. Predictive Methods

In order to predict the prices of Airbnb listings, we have deployed several machine learning algorithms including: multi-linear regression, regression tree, K nearest neighbor, and neural network.

i) Multi-linear regression

Figure 5. Lasso vs. Ridge Regularization





	feature	coefficient
193	nbh_SoHo	127.389362
223	nbh_West Village	132.229290
206	nbh_Theater District	135.814636
243	pt_Entire home/apt	136.290972
280	pt_Room in hotel	143.967198
247	pt_Entire serviced apartment	155.541757
212	nbh_Tribeca	201.851567
271	pt_Private room in resort	268.441614

For the Lasso and Ridge models, a range of regularization strength was tested, and the model with the lowest mean absolute error score was selected as the best model. The best Lasso model had an MAE of $73.25, a mean squared relative error of 5.20, and a regularization strength of 0.0774, while the best Ridge model had an MAE of $73.42, a mean squared relative error of 3.88 and a regularization strength value of 12.92.

Overall, both Lasso and Ridge models performed similarly, with Lasso model slightly outperforming Ridge model in terms of MAE. Also, since the Lasso model also performs feature selection, it may be preferred if the goal is to identify the most important features for predicting car prices. In fact, 103 out of 300 features are selected within the Lasso model. The table shows the top features with the largest absolute values of coefficients, which give us insight into the factors that are most strongly associated with price in the dataset, and suggest that location and type of property are key drivers of price in the Airbnb market.

Figure 7. Regression Tree: predicted price vs actual price

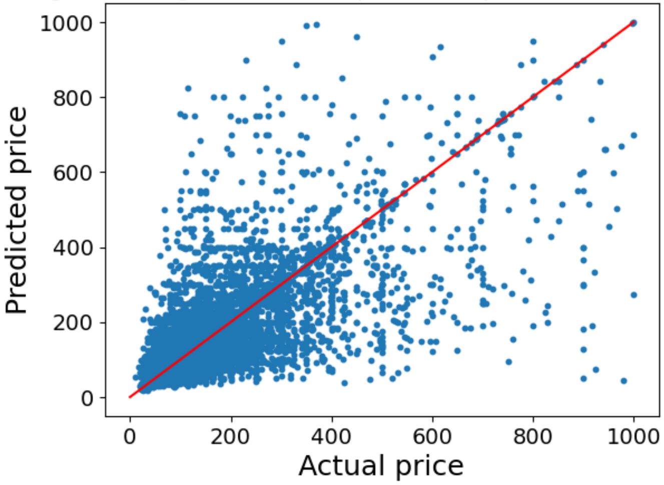
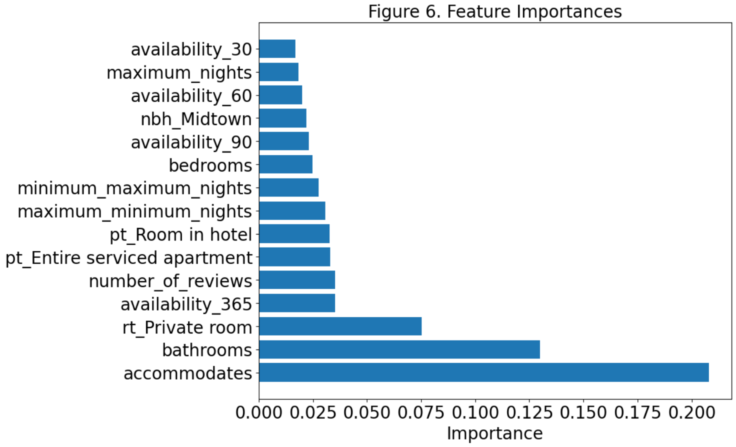


Figure 6. Feature Importances



ii) Regression Tree

For the Regression Tree model, we used the DecisionTreeRegressor function to predict the price of Airbnb listings. The model was able to relatively accurately predict the price of Airbnb listings, achieving a MAE of $63.04 and mean squared relative error of 33%.

The feature importance analysis showed that the most important variables for predicting the price of Airbnb listings in New York City were the number of accommodates, the number of bathrooms, and the type of room (private room or entire apartment). Location-based variables such as the neighborhood and proximity to popular tourist attractions were also important factors in predicting the price of Airbnb listings.

The scatter plot of predicted prices vs. actual prices showed a strong positive correlation with even error distribution between the predicted and actual prices, indicating that the model was able to accurately predict the price of Airbnb listings.

iii) K nearest neighbor Regression

For the KNN model, we decided to try out a variety of k values in order to test out the best one. We standardized the columns themselves and tried k values from 10 to 35. Figure 8a shows that when k = 22, the mean absolute error is the lowest around 61.5(lowest mean relative squared error is about 42% of median price). Additionally, when we look at Figure 8b, it seems like the actual price tends to be somewhat higher than the prediction. And it is the most accurate between 0 to 200, probably due to the fact that there are more data points in that range.

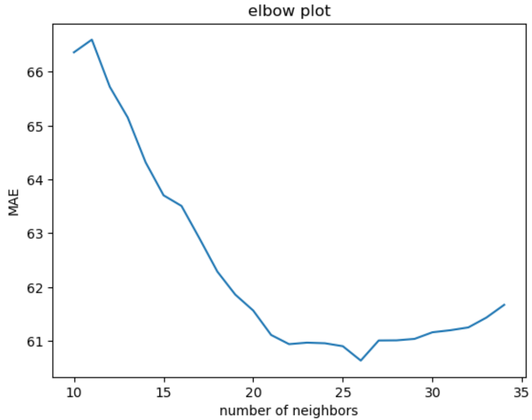


Figure 8a

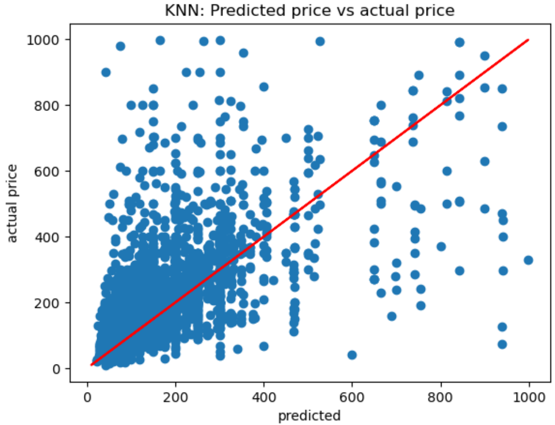


Figure 8b

iv)Neural Network

Finally, we also trained a sequential neural network model using tensorflow. Our neural network consists of four layers. The first layer is an input layer that takes as input the features we selected to predict Airbnb prices. This layer has a single node for each feature we included in our model. The second layer has 128 nodes. This layer takes the input from the input layer and applies a linear transformation and ReLU activation function, which allows the network to learn complex patterns and relationships between the features. The third layer has the same structure as the previous layer, but with 64 nodes. This layer further refines the features learned by the network in the previous layer, making it more sensitive to complex patterns in the data. The final layer is an output layer with a single node, which predicts the Airbnb price. This layer uses a dropout regularization technique to avoid overfitting. We split the dataset into 60% for training, 20% for validation, and 20% for testing.

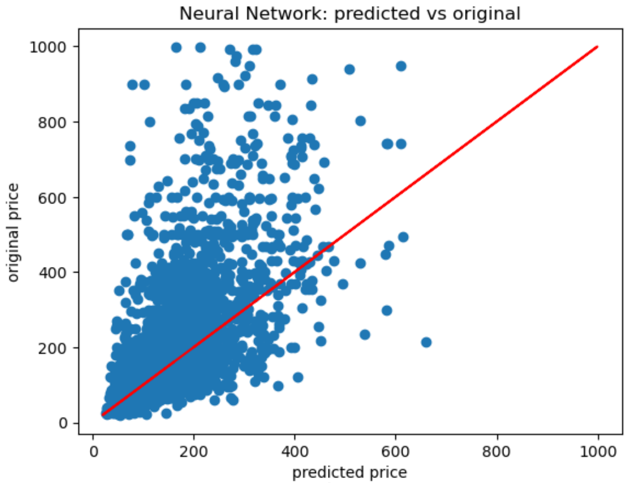
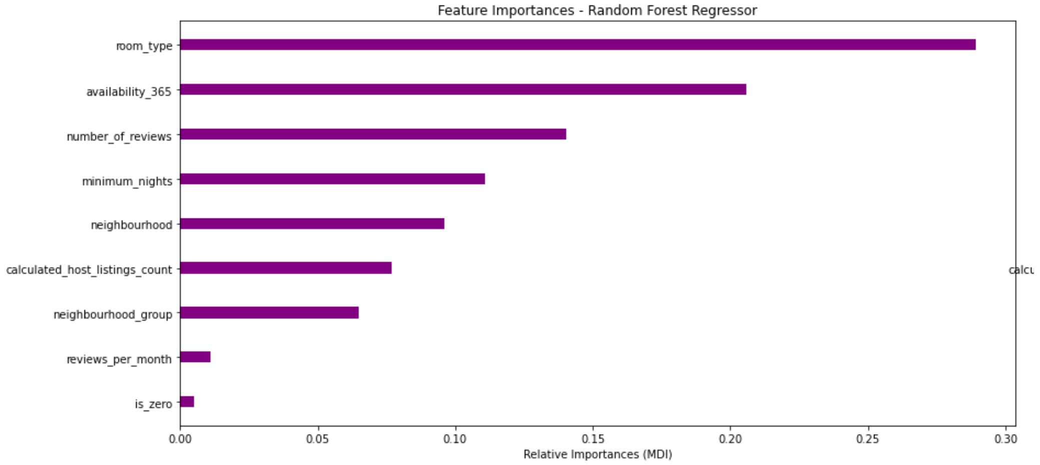


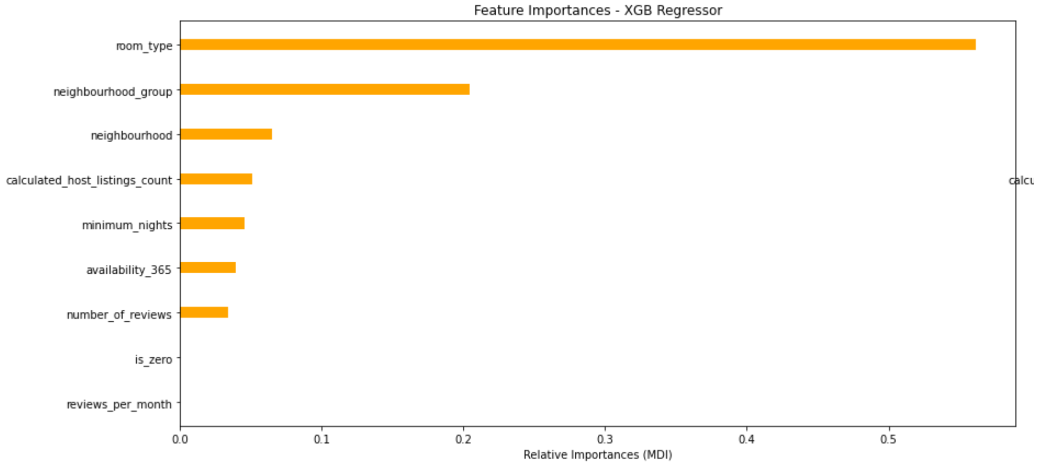
Figure 9

The MAE is 61 while the median absolute error is 31, and the mean squared relative error is 41(20% of the mean price). This is possibly due to the fact that there are a few falsely predicted small values that skewed the mean. R^2 score is 0.4, which means the variance of less than half of the data can be explained by the model(some more work needs to be done here to select features more carefully).

Figure 9 shows that the predicted price tends to be lower than the actual price, and the model doesn't produce predictions above 600. It would be interesting to find a few more $600-$1000 data points to balance the model.

Summary of data analysis conducted by others





On Kaggle, there are a few EDA and feature engineering notebooks. User “Jominjae” employed RandomForestRegressor and XGB Regressor, and the best MAE he got was about 46. He also found that room_type, availability_365, and neighbourhood_group are three of the most important features that contribute to the models. I think we used slightly different dataset(his is the cleaned up version so there’s no feature named accommodate), but we had some disagreements over which feature mattered the most since we found that accommodates is one of the more important features. [4] Additionally, Lawani et al.[5] conducted a sentiment analysis on Boston Airbnb data and found that sentiment analysis and disaggregated quality measures are better indicators of quality than single review scores, but they can serve as a proxy for rooms’ qualities, which agrees with what we found(Lawani 2019).

G. Implications and Future Improvements[6]

Based on the multiple approaches employed to predict listing price, we think the most accurate model is neural networks, followed by Regression Tree and KNN, and the least accurate is Multiple linear regression. This is due to the fact that a lot of variables employed in the model do not demonstrate a significant correlation with price, and also the relationship between price and various parameters is not necessarily linear.

As a next step in our data exploration, features should be more carefully examined and engineered in order to maximize explainability and accuracy. We should also make use of categorical variables such as neighborhoods and explore their corresponding wealth/education/crime rate status to inquire if they have an impact on listing prices.

[4] Jominjae. “Airbnb Prediction-RF,GB,XGB,LGB.” Kaggle. Kaggle, August 2, 2022. https://www.kaggle.com/code/jominjae/Airbnb-prediction-rf-gb-xgb-lgb#MODEL-LEARNING.

[5] Lawani, Abdelaziz, Michael R. Reed, Tyler Mark, and Yuqing Zheng. "Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston." Regional Science and Urban Economics 75 (2019): 22-34.

[6] Work division:
Stella: data cleaning, exploratory analysis, KNN, Neural Network
Yangge: Sentiment Analysis, Multilinear regression, regression tree