BDA 594 Big Data Science and Analytics Platforms

Group Project Report

Topic: Analysis of Airbnb Lodging Service in San Diego

<u>Group-9 Team Members</u>

Jiacheng Sun

Wenjin Wang

Wilson Gu

Mu-Ting Huang

Table of Contents

1. Problem Statement

Airbnb is an online rental marketplace that allows hosts to post their properties for short-term rental. Airbnb logging service has revolutionized the hotel industry and played a key role in our sharing economy. Unlike hotels that incorporate their system to set up a nightly rate, Airbnb hosts are fully responsible for pricing their rental properties. It becomes a challenge for hosts to decide on a price that could maximize their profits. New hosts would have to spend considerable time researching comparable listings and what features are attractive to the customers. Existing hosts also have to keep monitoring the market, so they do not set the prices too high or too low. Therefore, hosts need tools to help them price their property more conveniently and sufficiently.

San Diego is famous for its beaches and tourist sites. It hosts nearly "35.1 million visitors each year" and is a top U.S. travel destination. Our project intends to use data visualization and machine learning models to build tools for hosts in San Diego to price their properties. Our project would also help renters choose their ideal listings by providing knowledge of San Diego's market and the value of the property.

We would first conduct data analysis using data visualization on important features such as location and date. It would provide viewers a visual understanding of the overall Airbnb market in San Diego. It would be complementary to the price prediction tool built using machine learning models for hosts to make the most informed decisions on pricing their rental properties. We would build 5 different machine models (Simple linear regression, multiple linear regression model, Polynomial Linear regression model, decision tree regressor, Random Forest regression model, and XGBoost model), and compare which models can best predict the price.

2. Literature Review

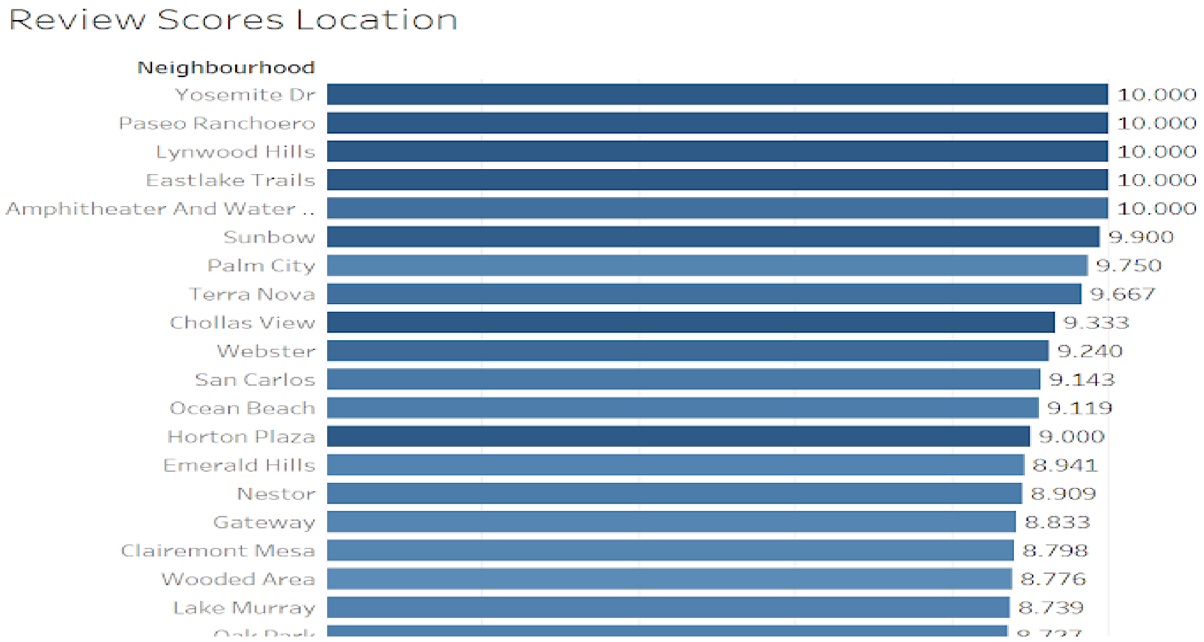
There have been multiple studies on Airbnb price prediction using machine learning in other cities. In a 2019 study (Kalehbast), authors employed linear regression, tree-based models, support-vector regression, k-means clustering, and neural networks to create the prediction models. The study found an abundance of features leads to high variance and weak performance of the model on the validation set compared to the train set. Among the models tested, Support Vector Regression (SVR) performed the best and produced an R2 score of 69% and an MSE of 0.147 (defined on ln(price)) on the test set. Another 2019 study(Luo) also confirms that feature selection is an important part of building prediction models. Besides, the author found that continuous features, both the text features and categorical features are helpful for price prediction. McNeil (2020) conducted an interesting study on the impact between variables that are inside of a host and variables outside of the host’s ability to control. The author found that while all the variables have some impact on the price, the variables within the host’s control have a higher impact. It implies that our project could help hosts earn a higher return on the property by modifying the variables which have a high impact on the price.

3. Explore the Data

Before building our machine learning model, exploratory data analysis (EDA) is needed because it can provide valuable insights into the data that will facilitate the establishment of the machine learning model. Tableau and ArgGIS were mainly used in EDA. One of our project goals is to provide advice to tourists who plan to visit San

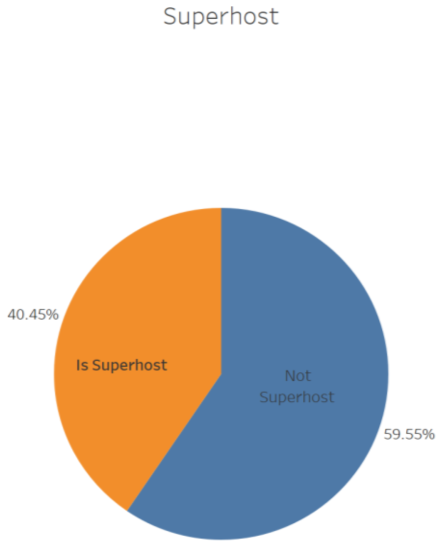
Diego. Therefore, our group tried to cluster the dataset and come out with four main factors (location, host, date, Comment) which most visitors are concerned about. And these three factors comprise three out of four subpages in the Exploratory Data Analysis (EDA) page in our website. Detailed discussion has been mentioned on our website. In the report, we will just empathize some key points.

In terms of “location”, we would like to provide tourists with some insights on which neighborhood they should live in. We have cleaned and sorted out the data to discover the differences between each neighborhood. In this category, we used “listings.csv” dataset and came up with six indicators (“Neighborhood”, “Room Type”, “Availability by Neighborhood”, “Average Daily Price Per Neighborhood”, “Review Score for Location”, “Number of People per Listing”) that help tourists to determine which neighborhood is suitable and meets their needs. For example, if the tourists are price sensitive, they may select the neighborhood with lower average price by looking at our chart called “Average Daily Price Per Neighborhood”.



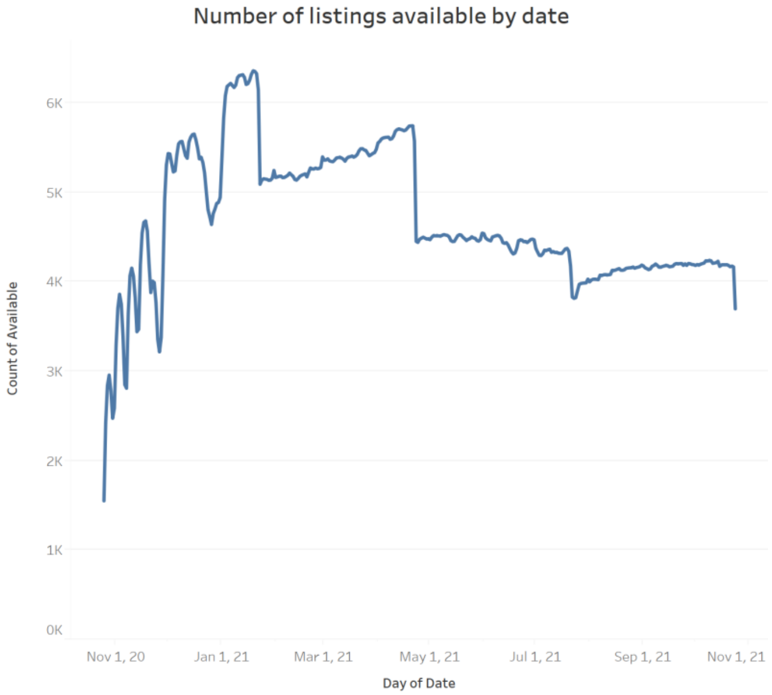
Continuing with “Location”, we generated a chart of “Review Score for Location”. The review scores for location are grouped based on the neighborhood. It is expected the convenience of the location is the key factor affecting the review score. Convenience not only refers to the distance from downtown but also includes other matters. For example, a listing is located outside the downtown but well connected by public transportation, a listing has other facilities (supermarkets, restaurants) nearby, and a listing has free parking. Five neighborhoods (“Yosemite Dr”, “Paseo Ranchoero”, “Lynwood Hills”, “Eastlake Trails”, “Amphitheater and water park”) receive full scores. San Diego has developed public transportation in downtown. It makes car-free and self-guided travel possible. San Diego’s extensive transportation routes allow visitors to easily reach San Diego’s port. By looking at the average score (7.90) with a standard deviation (1.24), we suggest tourists considering the neighborhoods with a review score over 6.66 (7.9 minuses 1.24) because a low review score means the neighborhood is likely inconvenient.

For “Host”, listings on Airbnb websites have the problem that good hosts and bad hosts are intermingled, so knowing how to choose a reassuring good host is very important. In this page, “which host should you choose?” would be our focus. Four indicators (“Super Host”, “becoming an Airbnb Host Since”, “Host Response Rate”, “Host Response Time”) were recognized, and we continue visualizing “listings.csv” dataset. If the tourists want to select a good host in San Diego, they are told on our web page that a good host should have a 90% response rate and probably have a title of “Super Host”. Our website shows to visitors what the criteria a good host should possess, and we also point out the distribution of good hosts in San Diego in our website.



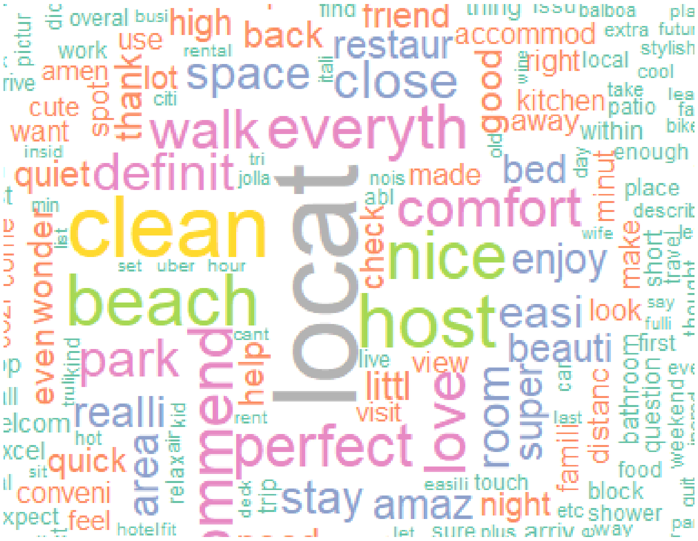
To be a super host, they must maintain an average of 4.8 overall ratings, maintain a 90% response rate or higher, and have served more than 10 groups of guests. Also, the rate of canceling guest appointments is below 1%. as shown on the chart of “Super Host”, in San Diego, 40.45% of hosts are super hosts. According to Scott Shatford (2018), when we look at the worldwide Airbnb lodging service, only 19.4% achieved Super host status in 2017. As you can see, San Diego possesses a higher rate of the super host.

When it comes to “Date”, a new dataset “Calendar” was introduced in this section because only this dataset has a “date” variable which enables us to conduct visualization related to time. The “Calendar” retains 365 records for each listing, which means that for each listing the price and availability by date are specified 365 days ahead. Three indicators (“Number of Listings Available by Date”, “Average Price by Date”, “Average Price by Weekday”) were identified. On this page, our website viewers can tell which date in the coming 12 months has a relatively higher number of listings available to choose, and when they can enjoy a lower price.



In this report, we just picked up a line chart “Number of listings available by date” to illustrate. The chart shows there is a constant increase in accommodations available in the coming three months (November to January). Reasons for this might be that hosts are more actively updating their calendars in this timeframe. Furthermore, the Christmas holiday and student semester break usually fall within this period, which is a peak time of demanding Airbnb lodging service. Overall, for the coming 12 months, the average accommodations available is 4686.28. It has the largest number of listings available in January 2021.

Last but not least, as part of EDA, we have conducted Comment Analysis, which is about the summary of reviews from the past 10 years. We used R to generate a word cloud to summarize the review column from a dataset “reviews.csv.gz”. The review text data requires cleaning works, such as removing common English words, stop words, and punctuation. It turned out the following words are removed during comment analysis: "San Diego", "home", "two", "need", "apartment", "like", "street", "one”, “stay", "house", "will", "great", "place", "felt", "time", "well", "around", "Airbnb", "day", "next", "time", "just", "also”, “diego", "san", "can", "bit". The word cloud tells the website viewers how the previous guests thought of San Diego.



The above word cloud shows some interesting trends. For example, location seems to be the key factor which tourists care a lot about, because the words “locat”, “neighborhood”, and “area” are obviously highlighted in the word cloud. The word “host” was also frequently mentioned, suggesting that hosts play critical roles in tourists’ experience. Finally, San Diego’s attractions “beach” and “park” were repeatedly stated, supporting that visitors were attracted to San Diego by its beach.

4. Database Management and Data Process Procedure

4.1 Data collection

Our data is collected from Airbnb website at http://insideairbnb.com/get-the-data.html.

4.2 Data preprocessing

We used SAS for Data preprocessing. We checked for entries with missing or incorrect values entries and removed those rows from our data. We removed the features that are obviously irrelevant, such as host_picture_url, listing_url scrape_id. Afterward we performed a count for the attribute amenities and verification to test later if the number of accommodates would be a factor that had a strong relationship with price or reviews. A new attribute government identification was also created in SAS with an if statement in the list of identifications included government documentation. Our final data consists of 12150 entries and 27 features:

```
Index(['price', 'number_of_reviews', 'calculated_host_listings_count',
       'availability_365', 'host_since_days', 'host_total_listings_count',
       'accommodates', 'bathrooms', 'bedrooms', 'beds', 'avg_rm',
       'availability_30', 'availability_60', 'availability_90',
       'number_of_reviews_ltm', 'NumberOfVerfication', 'numberofamentities',
       'host_is_superhost_f', 'host_is_superhost_t', 'host_has_profile_pic_f',
       'host_has_profile_pic_t', 'host_identity_verified_f',
       'host_identity_verified_t', 'instant_bookable_f', 'instant_bookable_t',
       'govermentIdentification_no', 'govermentIdentification_yes'],
      dtype='object')
```

4.3 Measure metrics

a. Measure metrics used on Feature selection

i. Correlation ratio

Correlation ratio is a coefficient of nonlinear association, which indicates the relationship between each variable in a dataset and ranges between 1 and -1. A correlation of 1 shows a perfect positive correlation. One the other hand, if the correlation between two variables is -1, these two variables have perfect negative correlation.

ii. VIF

Variance inflation factor measures if there is a multicollinearity in regression analysis, indicating the increase in the variance of a regression coefficient as a result of collinearity. Variance inflation factor is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone and calculated as: 1/ (1-R2)

Rules for identifying collinearity using VIF technique[1]:

- If all values of VIF are near 1 indicates no collinearity between the predictor variables
- VIF of >1 to 5 indicates moderate collinearity
- VIF of >5 indicates serious collinearity

[1]

iii. Feature importance score

Feature importance score for tree-based models is based on the reduction in the criterion used to select split points, like Gini or entropy.

b. Measure metrics used on Results

The primary metrics testing our result are r-squared score (R^2) and mean squared error (MSE).

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model. R Square value is between 0 to 1 and bigger value indicates a better fit between prediction and actual value.

R Square is a good measure to determine how well the model fits the dependent variables. However, it does not take into consideration the overfitting problem. If your regression model has many independent variables, it may fit very well to the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalize additional independent variables added to the model and adjust the metric to prevent overfitting issues.

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness of fit.

MSE is calculated by the sum of squares of prediction error which is real output minus predicted output and then divided by the number of data points. That is, the average squared difference between the estimated values and the actual value. It gives you a real number to compare against other model results and help you select the best regression model.

4.4 Feature Selection

We used two feature selection methods and compared the results using both. The first feature selection method was used to detect multicollinearity. First, we checked if multicollinearity exists by checking the correlation matrix and split dataset predictors between categorical and continuous. As you can see the figure1, 2, and 3. If the color of the cell is close to yellow, the higher the correlation.

The next step is to check these variables' Variance Inflator Factor (VIF) to see which variable we could drop to eliminate collinearity. VIF values greater than 10 may indicate multicollinearity is influencing the regression results. We will remove unimportant predictors from our models.

The second feature selection method is using Feature importance. It is a filter-based supervised feature selection method.

4.5 Machine learning models

We have built 5 different machine models (Simple linear regression, multiple linear regression model, Polynomial Linear regression model, decision tree regressor, Random Forest regression model, and XGBoost model)

Categorical variables were encoded into dummies variables using scikit-learn built-in function Pipeline(steps=[('onehot',onehot_categorical)])

A dummy variable is a numeric variable that stands for categorical data. We divided data into 75%testing data and 25% training data and used scikit build-in functions to train and test our models. We used python modules panda and NumPy for data manipulation. We used python modules Matplotib and Plotly for graphing and data visualization.

a. Simple linear regression

A linear regression model assumes a linear relationship between dependent and independent variables and takes in only one input variable. Linear regression is easy to implement and interpret. It provides information about relevant variables but is too simplistic for complex dataset.

We found host_total_listings_count is the feature with the highest linear correlation and used it for building the simple linear regression model.

b. Multiple linear regression model

Multiple Linear is similar to simple linear regression but it is able to take in multiple input variables.

c. Polynomial Linear regression model

While a simple and multiple linear regression model assumes a linear relationship between input and output variables, polynomial linear regression model works on nonlinear problems.

d. Decision tree regressor

Decision tree algorithms work by constructing a "tree." A decision tree is a supervised machine learning algorithm that can be used for both classification and regression problems. A decision tree is simply a series of sequential decisions made to reach a specific result. Decision trees over other machine learning algorithms is how easy they make it to visualize data.

e. Random Forest regression model

Random Forest is a tree-based algorithm that randomly creates decision trees. The random forest then combines the results of each random tree to generate the final output. Random forests are commonly reported to have higher accuracy than other learning algorithms.

f. XGBoost model

XGBoost is termed as Extreme Gradient Boosting algorithm works by boosting trees and makes use of a gradient descent algorithm which is the reason that it is called Gradient Boosting. XGBoost uses a continuous score assigned to each leaf which is summed up and provides the final prediction. This allows the algorithm to sequentially grow the trees and learn from previous iterations. The whole idea is to correct the previous mistake done by the model, learn from it and its next step improves the performance. The previous results are rectified, and performance is enhanced.

5. Results

5.1 Study Results

Figure 1, figure 2, and figure 3 is the correlation matrix for continuous & continuous attributes, continuous & categorical attributes, and categorical & categorical attributes respectively, and each cell is the correlation ratio between two variables. As the correlation matrixes show, there is no high correlation (> 0.7) between categorical & categoric, and continuous & categorical variables. However, for the continuous & continuous correlation matrix, we can see that availability_30, 60, and 90 variables, and beds, bedroom, bathroom, and accommodates are highly correlative.

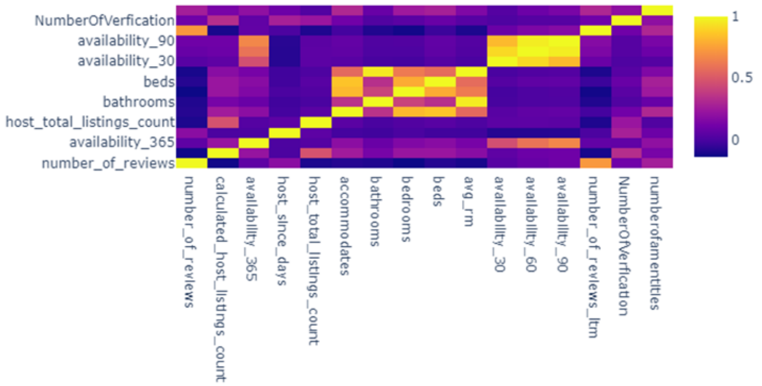


Figure1

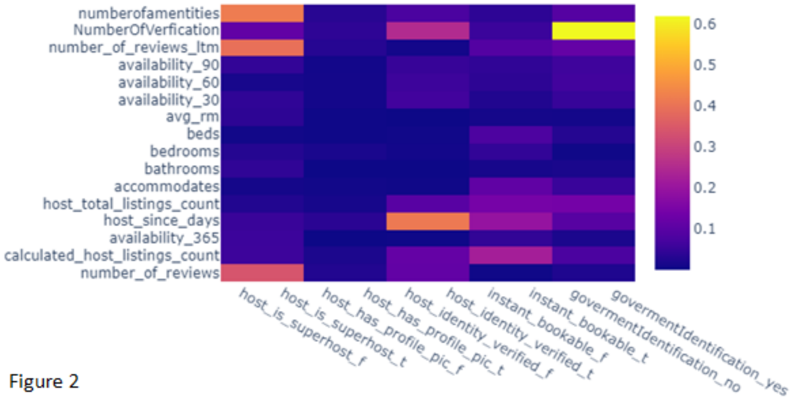


Figure 2

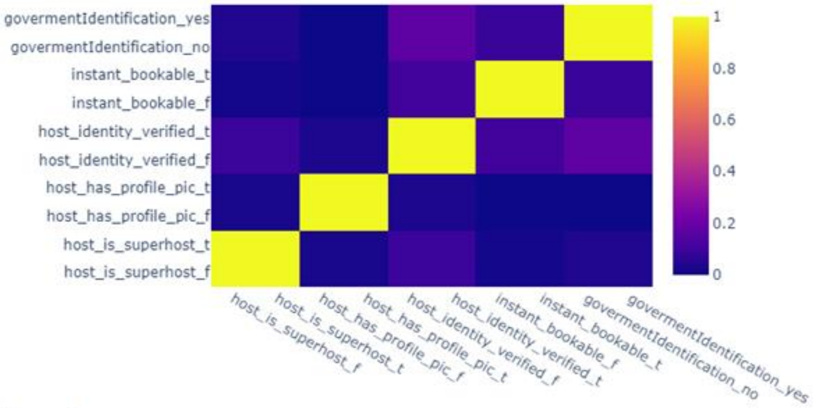
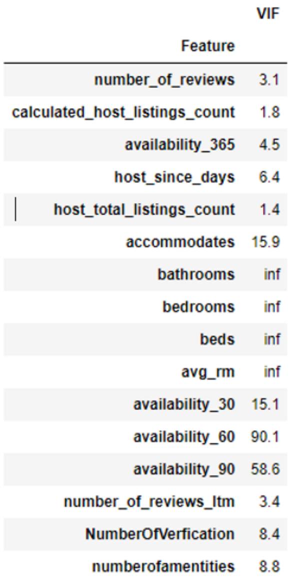


Figure 3

The following shows that attributes or features we would drop based on VIF.



Feature	VIF
number_of_reviews	3.1
calculated_host_listings_count	1.8
availability_365	4.5
host_since_days	6.4
host_total_listings_count	1.4
accommodates	15.9
bathrooms	inf
bedrooms	inf
beds	inf
avg_rm	inf
availability_30	15.1
availability_60	90.1
availability_90	58.6
number_of_reviews_ltm	3.4
NumberOfVerfication	8.4
numberofamentities	8.8

The attributes we dropped.

['accommodates',

'bathrooms',

'bedrooms',

'beds',

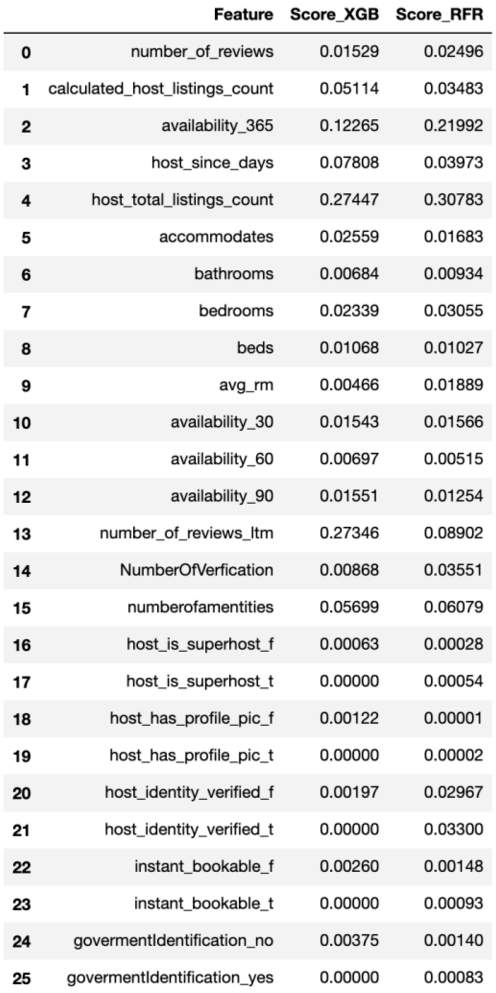
'avg_rm',

'availability_30',

'availability_60',

'availability_90']

For feature importance we drop the attributes whose score is less than 0.001.



	Feature	Score_XGB	Score_RFR
0	number_of_reviews	0.01529	0.02496
1	calculated_host_listings_count	0.05114	0.03483
2	availability_365	0.12265	0.21992
3	host_since_days	0.07808	0.03973
4	host_total_listings_count	0.27447	0.30783
5	accommodates	0.02559	0.01683
6	bathrooms	0.00684	0.00934
7	bedrooms	0.02339	0.03055
8	beds	0.01068	0.01027
9	avg_rm	0.00466	0.01889
10	availability_30	0.01543	0.01566
11	availability_60	0.00697	0.00515
12	availability_90	0.01551	0.01254
13	number_of_reviews_ltm	0.27346	0.08902
14	NumberOfVerfication	0.00868	0.03551
15	numberofamentities	0.05699	0.06079
16	host_is_superhost_f	0.00063	0.00028
17	host_is_superhost_t	0.00000	0.00054
18	host_has_profile_pic_f	0.00122	0.00001
19	host_has_profile_pic_t	0.00000	0.00002
20	host_identity_verified_f	0.00197	0.02967
21	host_identity_verified_t	0.00000	0.03300
22	instant_bookable_f	0.00260	0.00148
23	instant_bookable_t	0.00000	0.00093
24	govermentIdentification_no	0.00375	0.00140
25	govermentIdentification_yes	0.00000	0.00083

Summary of MSE and R2 for each model:

- Linear regression Mean squared error: 1068161.82, R^2 score: 0.04,
- Multiple linear regression model has Mean squared error: 12862.88 and R^2 score: 0.51
- Polynomial Linear regression model has Mean squared error: 22752.51 and R^2 score: 0.13
- Decision tree regressor has Mean squared error: 22752.51 and R^2 score: 0.13

Graph below shows MSE before and after dropping features for Multiple linear regression, Random Forest Regression, and XGboost based on the VIF and XGBoost feature importance score.

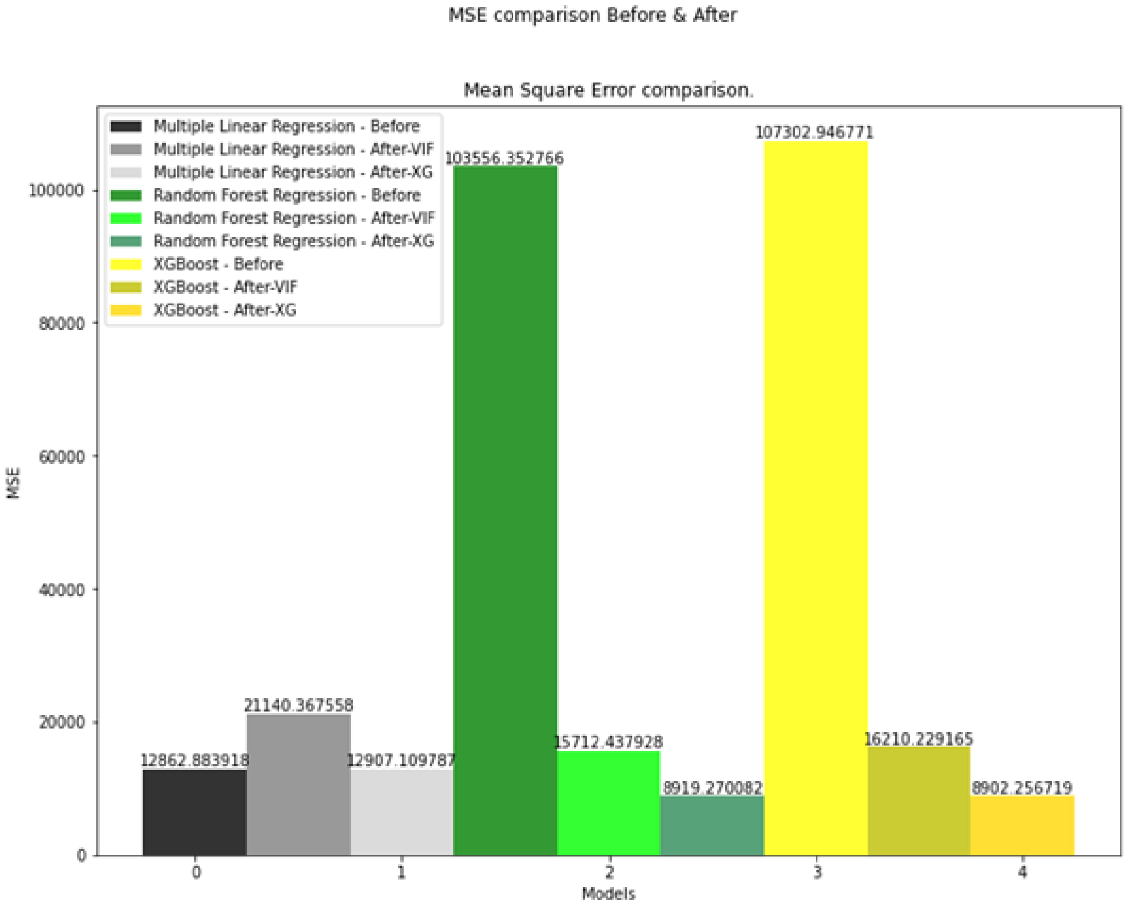


Figure 8

Feature selection using both methods had improved the results. Using the VIF feature selection method is better than feature importance in multiple linear regression models. Feature selection using importance methods produce better results in random forest regression and XGBoost models. The model with the lowest MSE is XGBoost after dropping the features with low importance scores, which has Mean squared error: 8713.62 and R^2 score: 0.67.

Due to limited time and processing power, there are other features not included and evaluated in this project. Comment analysis could be incorporated to use customers' reviews as attributes. Time series could also be incorporated to take time into account in the building machine learning models.

5.2 Website

We used Google Site to build this website, it is a great resource as it provides simple templates and allows you to upload visuals with ease. The landing page will be where visitors can see our project's overview, goals, and tools we used. We also embedded our video in the starting page and viewers that want more concrete visualizable and analytics of our study can use the bar on top to access different analysis or our data files in the database tab.

On our website, there are three main contents (Exploratory Data Analysis, Comment Analysis, Machine Learning). In the page of “Exploratory Data Analysis”, four sub-pages (Location, Host, Date, other) are included. To give a better understanding of location, we also created maps in ArcGIS that show with our other charts and graphs. In the page of “Comment Analysis”, we generated a word cloud using R to see which words are frequently used when visitors left their comment for the apartment they lived in. In the page of “Machine Learning”, we used Python to test which model is better to predict the price.

Apart from the main pages, we also embedded our dataset into our website, allowing our websites visitors to get access to them in case they would like to conduct further research based on our results. Last page is “About Us’ which has contact information about each team member. Such that, website visitors who have questions about our research could contact us.

Website URL

https://sites.google.com/d/1sFimR8DhepB6tZYiXUtrBbfUkP_UEJvj/p/10uwCnbaNP-3jiAbAabdWlA_MlJIAJdnm/edit

6. Discussion & Conclusion

Our goal for machine learning is to find out which model is better to predict Airbnb rental price. The datasets being studied were able to provide information in terms of room availability across neighborhood and date, and price prediction. When gathering all the results, some interesting and informative findings are available.

More works can be done for machine learning modeling especially in feature engineering. Data plays a crucial role in machine learning modeling. Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy. A feature is an attribute that is useful or meaningful to your problem. With good features, the result could address the underlying problem and a representation of all the data you have available. In a given dataset, there are some features that will be more important than others in terms of model accuracy by calculating feature importance score.

In the Airbnb dataset we used to train our machine learning models, we did not use some attributes such as date, locations, and response rate which could be good features to improve model accuracy. Due to time constraints, we did not dig into these attributes to see if these features could be aggregated or combined to create new features. We could create a new numerical feature called Hour_of_Day for the hour that might help a regression model.

There are many tools that we can use to determine if the attribute is good at predicting the target variable. We list a few tools such as correlation ratio, VIF, MSE and feature importance score. For this project, there are three machine learning models (XGBoost, Random forest, and Decision tree) that have feature importance function, but we only use the feature importance scores that XGBoost generated to filter the unimportant features out. There are other tools (t-value for each attribute) available for feature selection to improve model accuracy. We also can combine the scores generated by these tools to have different feature combinations and test which combination has the highest model accuracy.

We also can use correlation ratio between target and attribute to measure if any attributes are there that have high positive relationship with the target attribute which we can keep them to improve the model’s performance, on the other hand, we can drop the negative correlation ratio, making sure every attribute in the dataset has improved the accuracy of the models.

Based on the machine learning modeling results, there are a couple of suggestions for the hosts. From the feature importance bar chart generated from the XGBoost algorithm, there are some features that could help hosts increase their rental fees, and the top 3 features are accommodates, bedrooms, host_total_listing_count.

The top 3 features of the feature importance of the decision tree regressor are host_total_listing_count, number_of_verfitication, and availability_365. For the top 3 features of random forest regressor, these features are host_total_listing_count, availability_365, and number_of_review. Overall, host_total_listing_count and host_total_listing_count are top 2 important features that hosts should pay attention to.

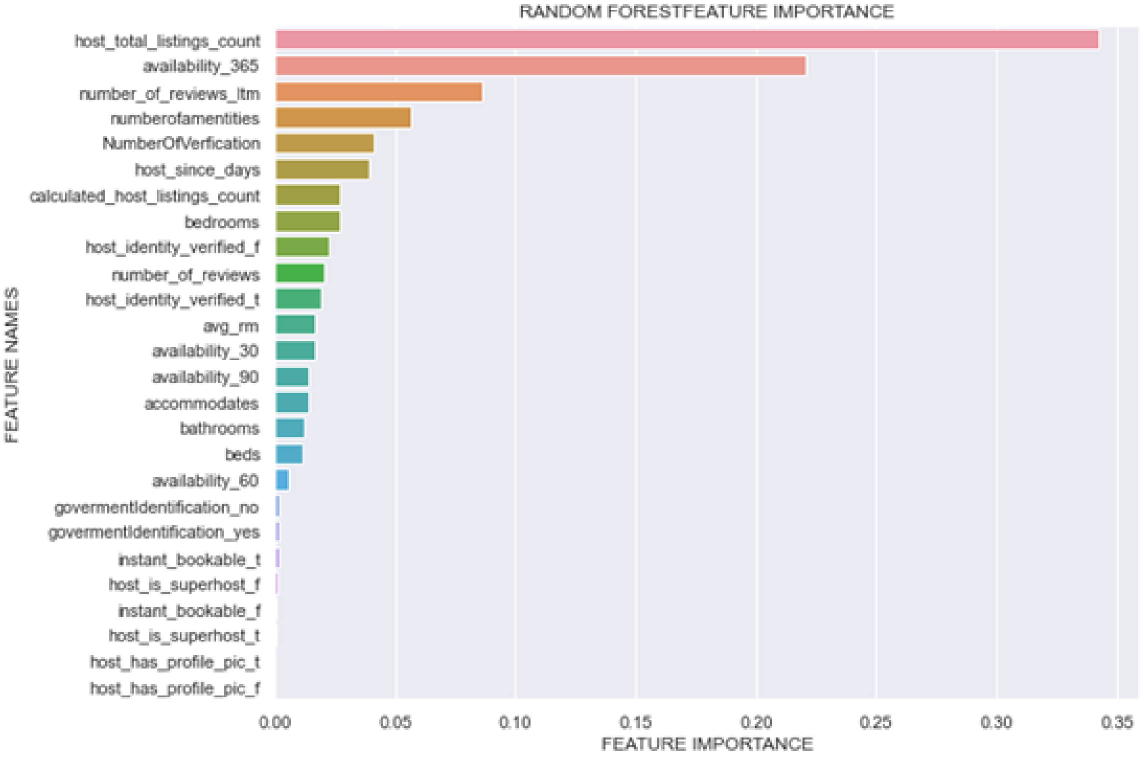


Figure 5

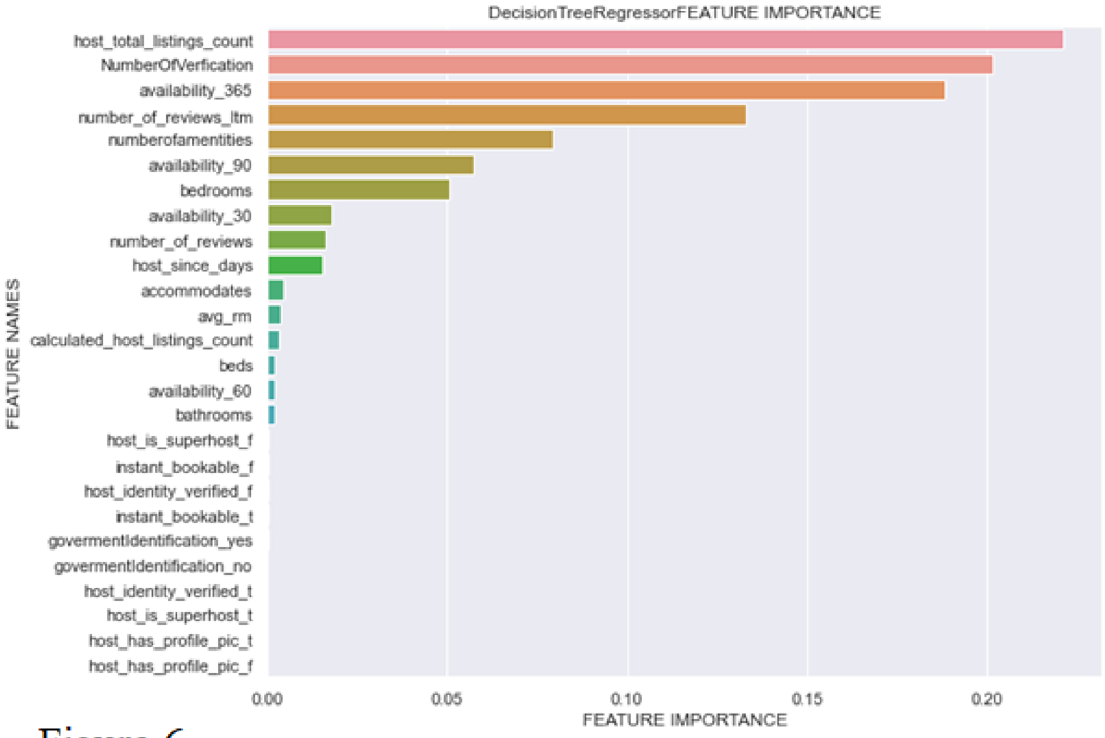


Figure 6

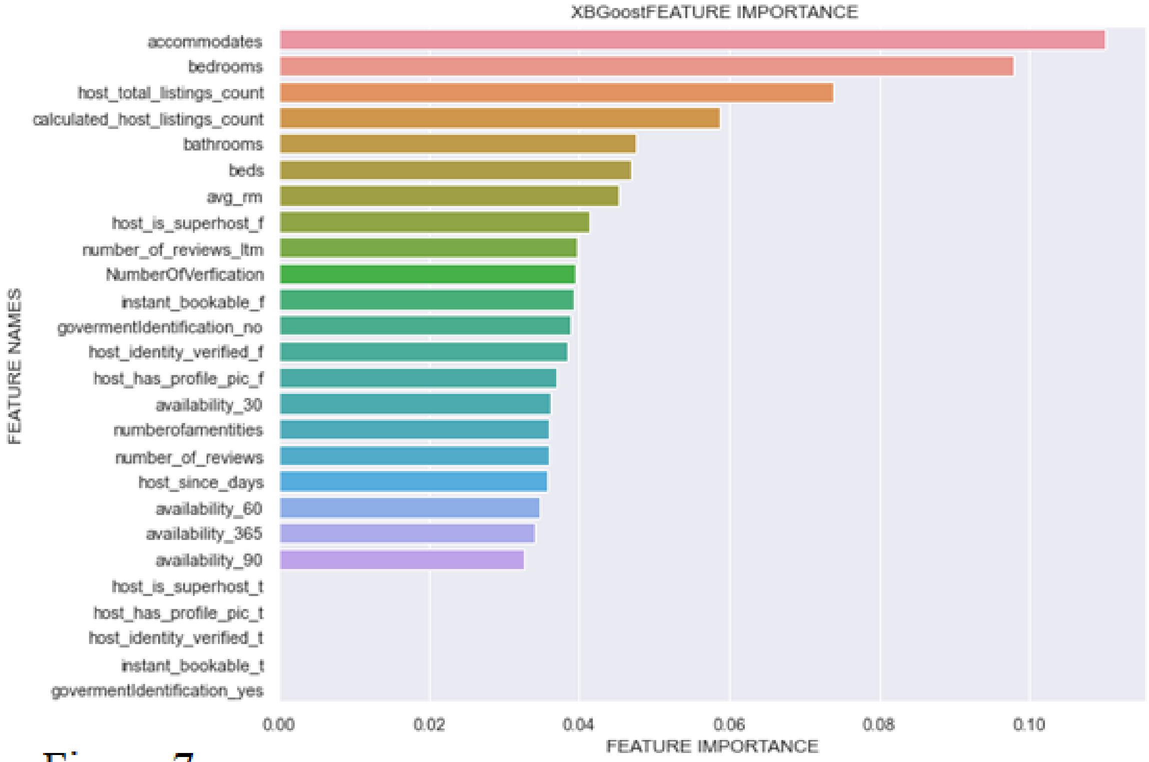

XBGoostFEATURE IMPORTANCE
FEATURE NAMES
accommodates
bedrooms
host_total_listings_count
calculated_host_listings_count
bathrooms
beds
avg_rm
host_is_superhost_f
number_of_reviews_ltm
NumberOfVerfication
instant_bookable_f
govermentIdentification_no
host_identity_verified_f
host_has_profile_pic_f
availability_30
numberofamentities
number_of_reviews
host_since_days
availability_60
availability_365
availability_90
host_is_superhost_t
host_has_profile_pic_t
host_identity_verified_t
instant_bookable_t
govermentIdentification_yes
0.00
0.02
0.04
0.06
0.08
0.10
FEATURE IMPORTANCE

Figure 7

7. References

Kalehbasti, P. R., Nikolenko, L., & Rezaei, H. (2019). Airbnb price prediction using machine learning and sentiment analysis. *arXiv preprint arXiv:1907.12665*.

Luo, Y., Zhou, X., & Zhou, Y. (2019). Predicting Airbnb Listing Price Across Different Cities.

McNeil, B. (2020). Price Prediction in the Sharing Economy: A Case Study with Airbnb data.

Scott Shatford (2018). What is Airbnb's Superhost Status Really Worth? Retrieved from https://www.airdna.co/blog/airbnb_superhost_status#:~:text=Why%20Hosts%20Aren't%20Qualifying,achieved%20Superhost%20status%20in%202017.

Discover Feature Engineering, How to Engineer Features and How to Get Good at It https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

http://insideairbnb.com/get-the-data.html

https://www.kaggle.com/erikbruin/airbnb-the-amsterdam-story-with-interactive-maps

https://www.kaggle.com/duygut/airbnb-nyc-price-prediction

https://www.kaggle.com/chirag9073/airbnb-analysis-visualization-and-prediction/comments#647941

https://www.kaggle.com/kostyabahshetsyan/boston-airbnb-visualization

https://www.kaggle.com/yogi045/how-to-become-top-earner-in-airbnb

https://www.kaggle.com/mpanfil/nyc-airbnb-data-science-ml-project

https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b

https://towardsdatascience.com/which-evaluation-metric-should-you-use-in-machine-learning-regression-problems-20cdaef258e

https://machinelearningmastery.com/calculate-feature-importance-with-python/