New York Airbnb data

Sarah M. Maldonado

Introduction

Airbnb, which is short for Air Bed and Breakfast, is an online marketplace that allows people to rent out their homes to others. It was founded in 2008 and provides users with endless traveling possibilities. The dataset used in this report shows the listing activity in NYC, NY for 2019. The goal of this project is to predict listing prices (low, medium, high) using the factors that are possibly related to the price of the listing. This is crucial as it gives users insight into future listing prices in NYC, NY and allows them to plan their trips accordingly. [1]

Datasets

The dataset used throughout this project is New York Airbnb data. As stated above, it contains all Airbnb listings in NYC, NY for 2019. This dataset contains a variety of different variables that will allow us to predict listing prices. Out of the 16 variables, we found five to be most useful to our research. The other 11 variables were found to be irrelevant to our final prediction. The other five variables include availibility_365, neighbourhood_group, room_type, price, and minimum_nights. The target variable is price and the other four are feature variables. We also ended up adding an extra variable called price_ranges in order to split the features into the low, medium, and high ranges.

Tools Used

To aid our research, we used Orange3 to preprocess and create predictions about the data. This includes processes like imputing the data to remove 0 values in the availibilty_365 variable column. We also used it for data exploration by comparing different variables to each other.

Data Acquisition

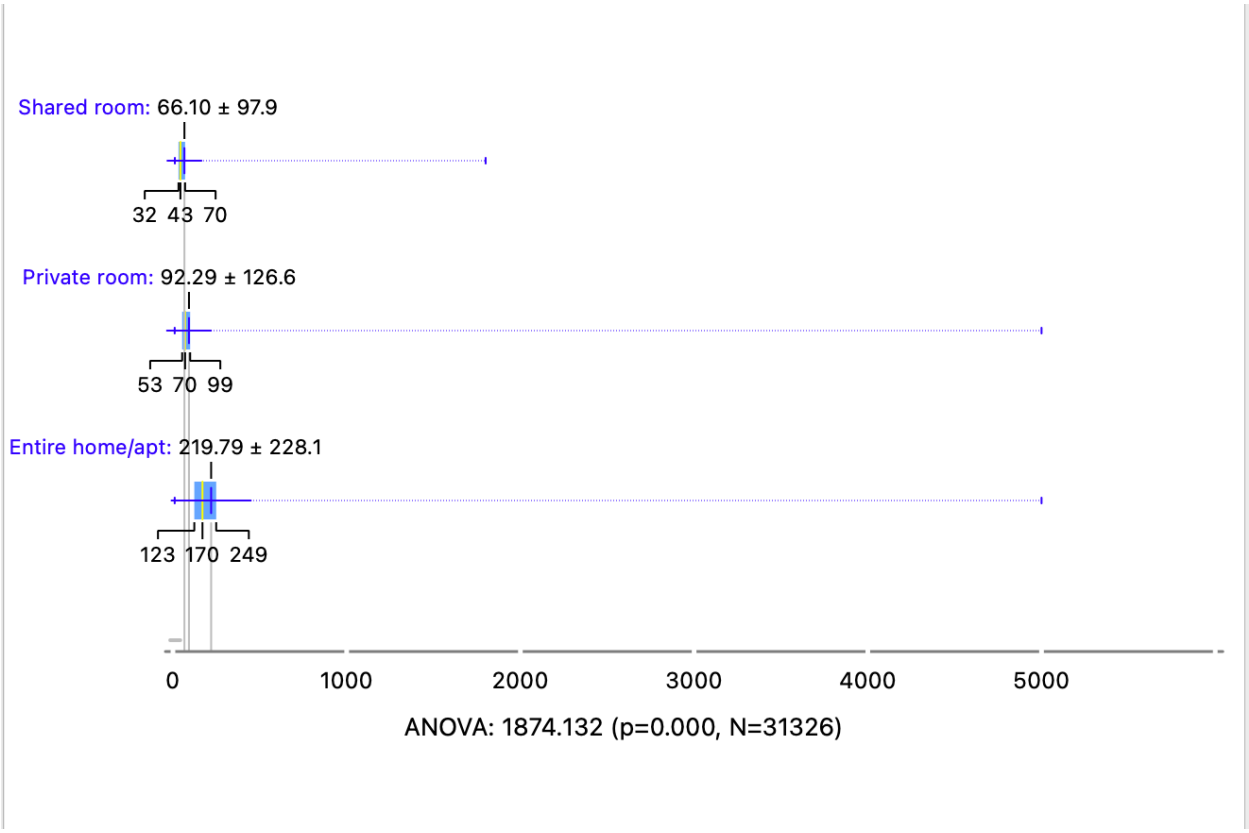
We acquired the original dataset from the New York Airbnb Dataset on the Kaggle website. However, we edited the dataset through data preprocessing to better fit our research as stated above.

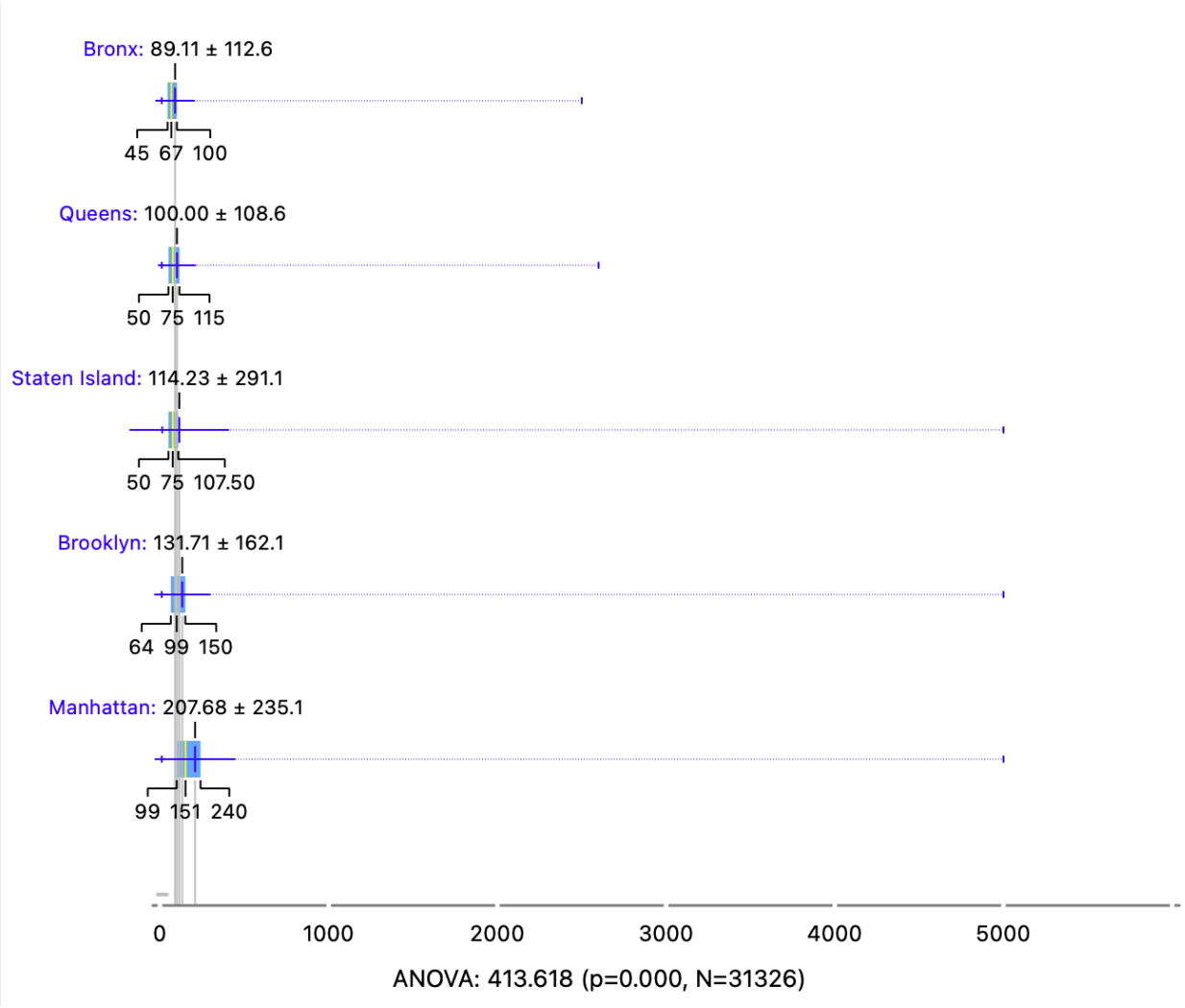
Data Analysis and Results

Basic Characteristics of the Dataset

Characteristics	Number of Instances	N or Mean	Median	Percentage or Standard Deviation
Listing Price				

Brooklyn	12245	131.71	99	162.1
Bronx	913	89.11	67	112.6
Manhattan	13541	207.68	151	235.1
Queens	4296	100.0	75	108.6
Staten Island	331	114.23	75	291.1
Room type price				
Shared room	861	66.10	43	97.9
Private room	13953	92.29	70	126.6
Entire home/apt	16512	219.79	170	228.1





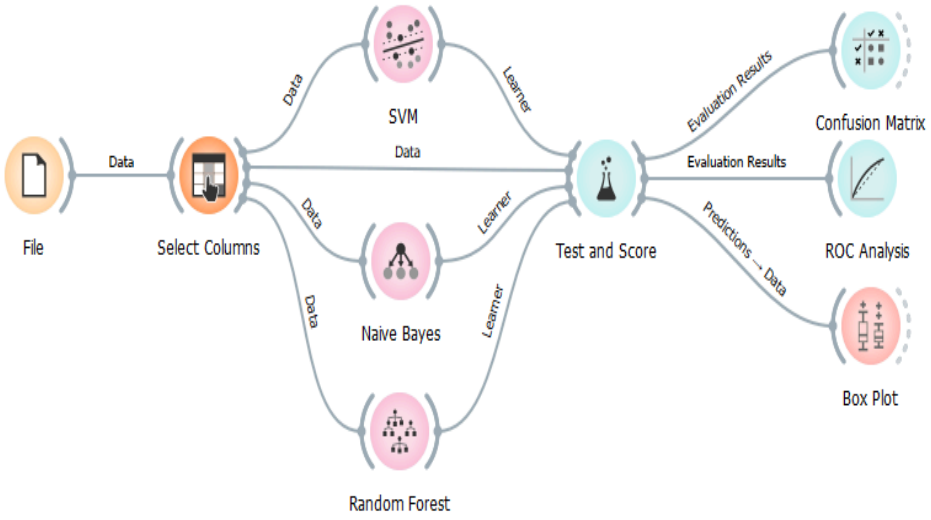
Above are box plot representations of the data listed in the table.

Data preprocessing

We used Orange 3 to complete the data preprocessing. The first step to this was removing any values that did not make sense to the dataset. This included any features where the price was equal to zero or the availability was above 365 days. We also removed the outliers which left us with all prices below $5000. Before completing the preprocessing we had 48,895 records and after completing it we were left with 31,326 records. Another step we took was removing variables that we did not see fit to our problem statement. Out of the original 16 records we were left with five plus the extra variable we added, price_ranges.

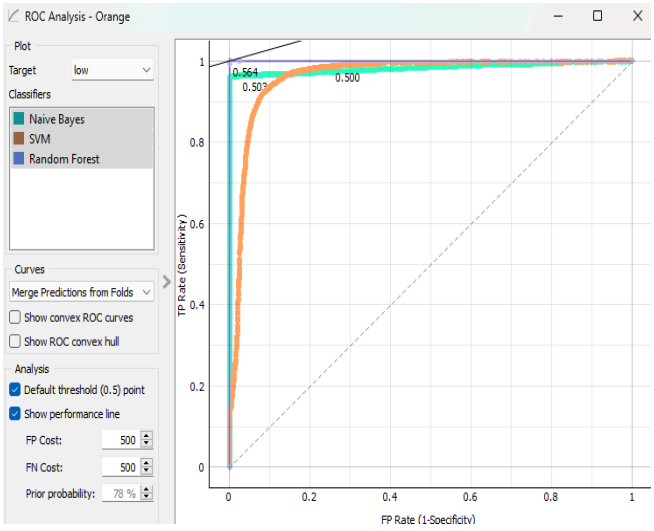
Data Modeling and Evaluation

For this part of our research, we used the SVM, Naive Bayes, and Random Forest machine learning algorithms to test our target variable, price. Using these three models we were able to cross-validate in order to reveal how efficient each model is at correctly predicting the listing prices low, medium, and high.



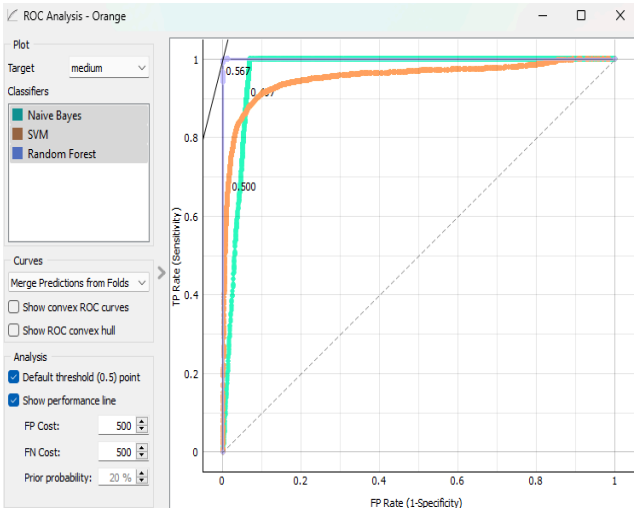
Model	AUC	CA	F1	Precision	Recall	MCC
Naive Bayes	**0.971**	**0.939**	**0.929**	**0.923**	**0.939**	**0.837**
SVM	**0.961**	**0.925**	**0.927**	**0.933**	**0.925**	**0.804**
Random Forest	**1.00**	**0.999**	**0.999**	**0.999**	**0.999**	**0.998**

As you can see from the results above, we were able to determine that random forest is the most efficient at predicting the listing prices. It outperforms Naive Bayes and SVM across all metrics, indicating that it is the most effective model for predicting listing prices. The near-perfect accuracy, precision, recall, and AUC scores demonstrate its excellence in distinguishing and accurately predicting the different price range categories.

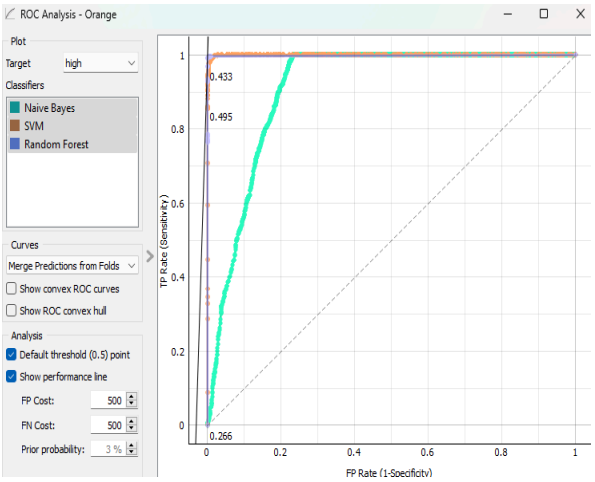


Low

Medium



High



The above graphs show the ROC Curves for the low, medium, and high variables. The ROC curve is meant to measure the accuracy of the model and because the area is close to 0.5 the machine learning algorithms prove to be very accurate.

Discussion

From our research, we were able to predict the prices between low, medium, and high and we were able to determine the precision, recall, F1 score, sensitivity, and specificity to determine the accuracy of the data. We grouped the final dataset after cleaning the data into low ($0-$200), medium ($201-$500), and high ($501-up) to find the average listing price to be $158.77. However, during this process, we did encounter a few challenges that we had to overcome. The first challenge was determining which variables played a role in the listing price. Many of the variables that were in the original dataset were found to be useless to the actual

problem statement. To overcome this we discussed which variables we would use and eliminated the rest of them. The factors we found useful for predicting the listing prices between low, medium, and high were room_type, neighbourhood_group, price, minimum_nights, availibility_365, and our added variable, price_range. Another challenge we faced was deciding which machine learning models to use. We chose Support Vector Machine, Naive Bayes, and Random Forest for predicting Airbnb listing prices in NYC due to their diverse modeling approaches, which encompass linear, probabilistic, and ensemble-based methodologies, respectively. These algorithms were selected to ensure comprehensive coverage of potential patterns and relationships within the dataset, allowing for robust predictions across different price categories. Through this, we were able to discover that random forest was the most accurate model for our data prediction with an accuracy of 0.999. The most valuable takeaways from our project include our successful prediction of listing prices across low, medium, and high categories, as well as the thorough evaluation of precision, recall, F1 score, sensitivity, and specificity to gauge the accuracy of our predictions.

Timeline for Completion

Data Cleaning	February 18, 2024
Midterm Report	March 5, 2024
Midterm Powerpoint Presentation	March 7, 2024
Final Research complete	April 18, 2024
Final Powerpoint Presentation	April 21, 2024
Final Report	April 28, 2024

Team Workload and Roles

Presentations and midterm report were split between all three team members.

References

https://q4launch.com/blog/how-airbnb-works-for-hosts/#:~:text=Airbnb%2C%20as%20in%20%E2%80%9CAir%20Bed%20and%20Breakfast%2C%E2%80%9D,rooms%2C%20or%20the%20entire%20property%20for%20themselves. [1]