Milestone 5: Final Project Presentation

Airbnb Price Prediction
(New York 2019)

Ankit Jain
Vaishnavi Badame

Problem Statement

Given Airbnb New York data for the year 2019, predict prices of Airbnb listings using factors such as neighbourhood group, latitude, longitude, room type, number of reviews, minimum nights and availability throughout the year.

Motivation

- Airbnb is one of the most widely used application to rent rooms or apartments. It’s price is dependent on many factors like neighbourhood, type of apartment, ratings or reviews, availability throughout the year and so on.
- The motivation was to develop a system that eases customer search experience so that they can have prices of their Airbnb predicted knowing what factors will best suit them.
- This system improves customer experience as they will understand how their needs affect the prices of an Airbnb and also to Airbnb sales as more customers would want to make a booking using their application.

Dataset Description

- Source: https://www.kaggle.com/dgomonov/new-york-city-airbnb-opendata which is in .csv format
- This data set was updated in 2019.
- The data set comprises of 48,895 records of different Airbnb listings around the neighbourhoods of NYC. It has 16 variables, out of which 10 are quantitative and the remaining 6 are qualitative.

Dataset Attributes

- id: Unique id assigned to Airbnb record/listing
- name: Name of the record (Airbnb house/listing)
- host_id: Unique host id
- hostname: Host name
- neighborhood_group: Area of the listing
- neighborhood: Specific area in the neighborhood_group
- latitude: Latitude of the listing
- longitude: Longitude of the listing
- room_type: The type of room provided (either entire house/private room/shared room)
- price: Price of the listing
- minimum_nights: Minimum number of nights the user is required to book the listing
- number_of_reviews: Number of user reviews
- last_review: Date when the last review was given by the user
- reviews_per_month:Number of reviews per month
- calculated_host_listings_count: Number of listings by hosts
- availability_365: Number of days in a year when the listing is open for booking

Implemented Algorithms

- Core Algorithm: Linear Regression

- Candidate Algorithms: Lasso, Ridge, Elastic Net, Best Subset Regression, Stepwise Selection and Random forest.

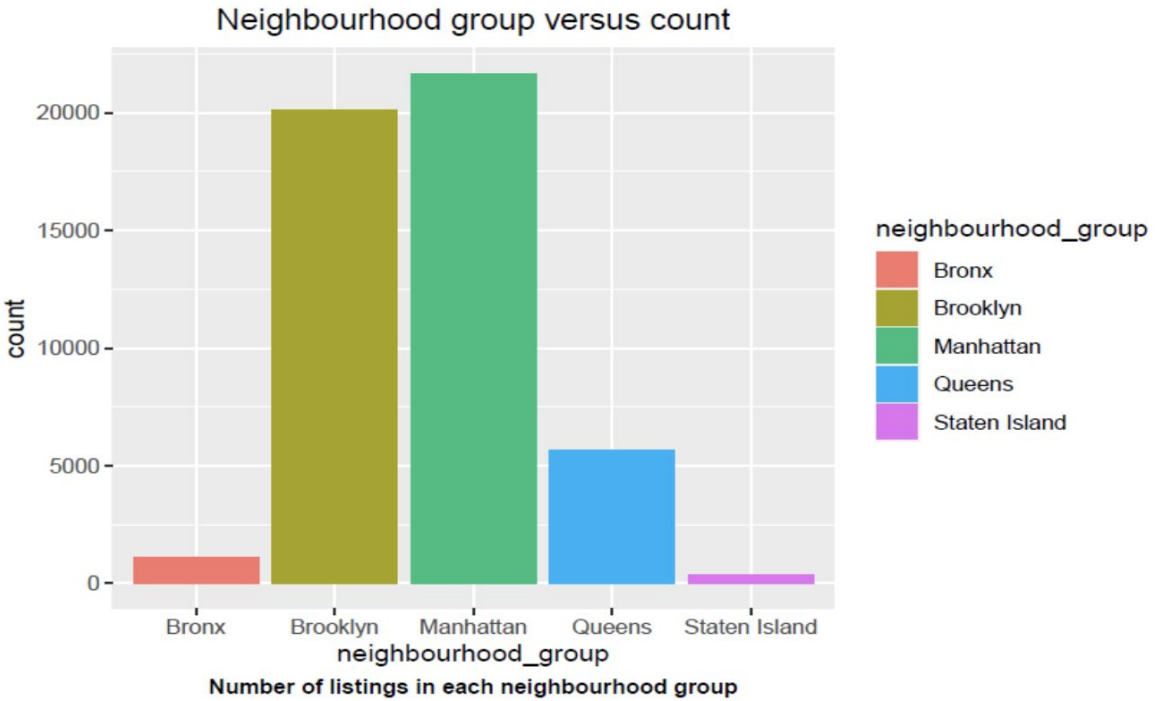
Assumptions

1. Since there are a total of 221 unique neighbourhoods for the given 5 neighbourhood groups, we have restricted analysis on this neighbourhoods.
2. We as a user usually don’t consider about who is the host of the house (given we do not have any host rating) or what is the name of the listing to make booking and as a result. Hence attributes not taken into consideration were: id, name, host_id and hostname

Data Pre-processing and Cleaning

- The data, initially, resides in a single csv file. We segregated the data into training and test data
- The ratio of training to test data is 70:30 respectively as a model needs to be trained and then tested.
- Columns like ‘reviews_per_month’ and ‘last_review’ have null values(na) in the dataset. We have substituted these ‘na’ values as ‘0’ in order to maintain consistency in the dataset.
- We eliminate the rows from the dataset where the price is 0. This is not possible and are incorrect records.

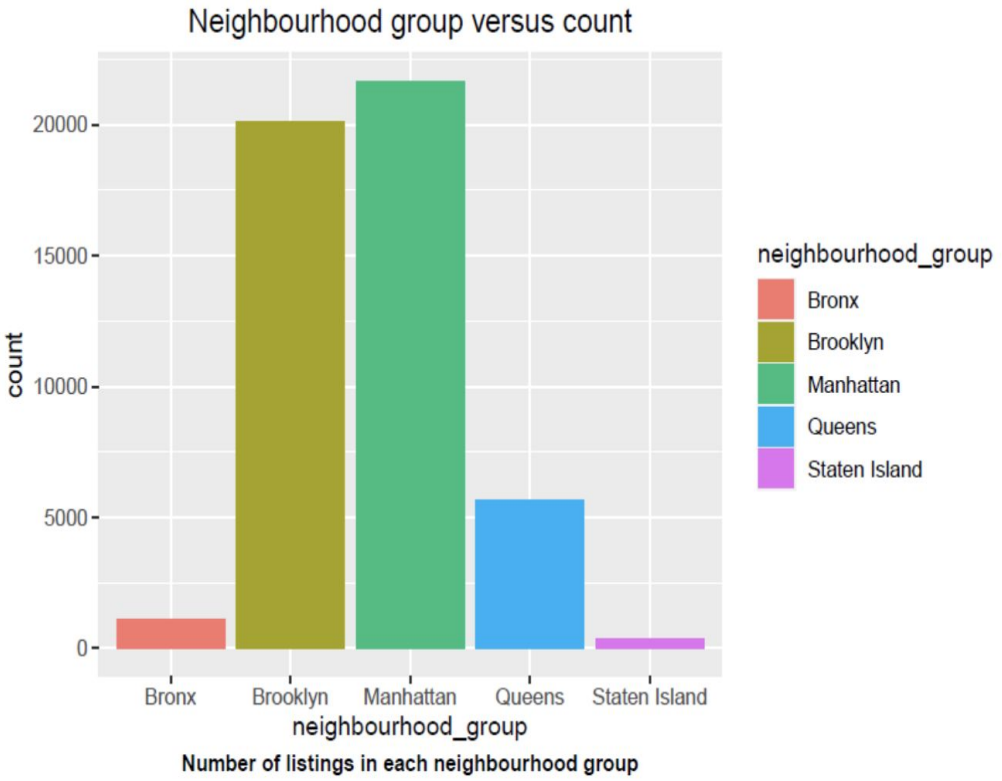
Data Visualization



Data Visualization

(Variable :Neighbourhood_group)

Manhattan and Brooklyn being the most visited places during holidays, it makes sense that these two places have the most number of Airbnb listings.



Data Visualization

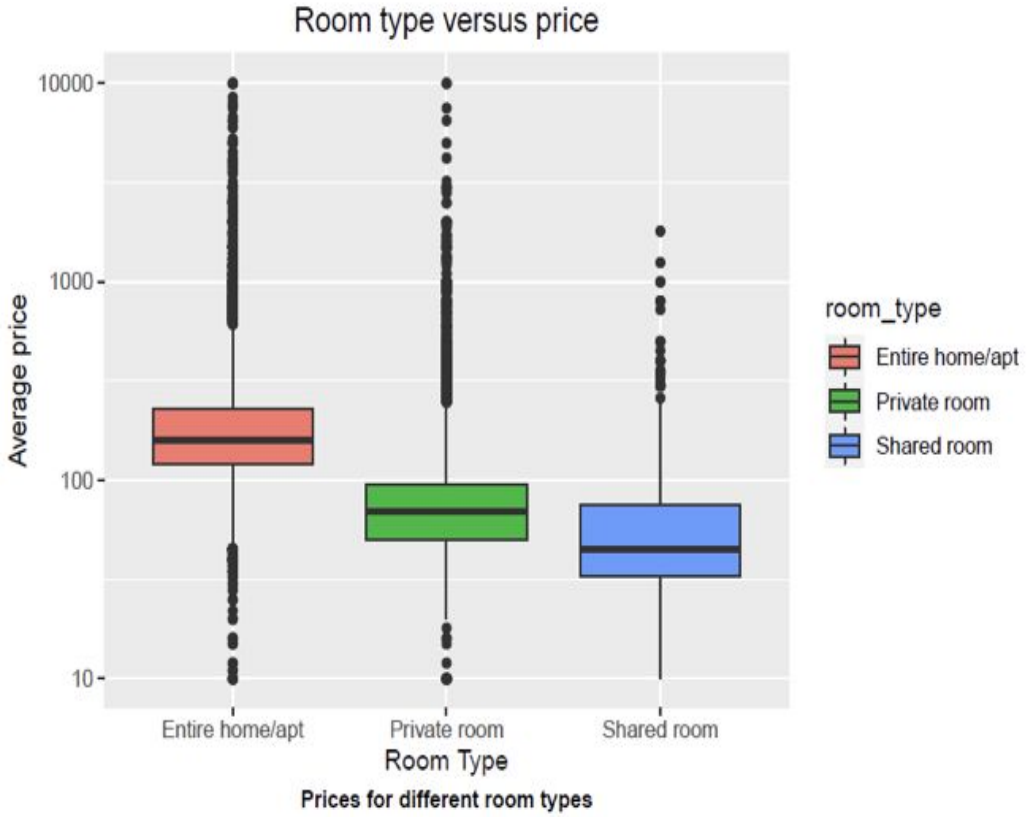
(Variables :Neighbourhood_group, price)

Manhattan and Brooklyn have the maximum average prices which it should as they are primary locations where people visit during holidays and the fact that they are expensive neighbourhoods in general.

Data Visualization

(Variable :Room_type, price)

An an entire apartment is expensive than a private room or a shared room which is true.



Data Visualization (Correlation Matrix)

	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
price	1.00000000	0.04279933	-0.04795423	0.05747169	0.08182883
minimum_nights	0.04279933	1.00000000	-0.08011607	0.12795963	0.14430306
number_of_reviews	-0.04795423	-0.08011607	1.00000000	-0.07237606	0.17202758
calculated_host_listings_count	0.05747169	0.12795963	-0.07237606	1.00000000	0.22570137
availability_365	0.08182883	0.14430306	0.17202758	0.22570137	1.00000000

As we can see, the values for price against all the parameters are low and thus, no single component alone could be used to predict the price. But the combination or a subset of them could be helpful.

Algorithm Testing

Following Algorithms were tested on the data-set:

- Linear Regression
- Best Subset Regression
- Stepwise Selection
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Random Forest

Fine Tuning Core Algorithm

Data Cleaning: We have excluded the outliers in the dataset as outliers can greatly affect the slope of the regression line. An outlier may represent bad or incorrect data.

Handling Multicollinearity: The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model. It also reduces the precision of the estimate coefficients, which weakens the statistical power of your regression model.

Fine Tuning Core Algorithm

Polynomial Transformations: We have transformed the variables using different polynomial functions so as to get a smaller p-value and thus tuning the algorithm by reducing values of MAE, MSE and RMSE altogether.

Performance Comparison of Algorithms

Regression_Method	MAE	MSE	RMSE	MAPE
Linear(without fine tuning)	80.65974	9862.187	99.30854	0.8414769
Linear(with fine tuning)	30.63416	1623.901	40.29765	0.2714912
Ridge	32.29558	1770.007	42.07145	0.2897226
Lasso	32.17972	1769.301	42.06306	0.2875063
Elastic Net	32.18080	1769.273	42.06273	0.2875306

Performance Analysis and Conclusion

Thus by taking a look at the different values in the above table we can understand that :

- The worst performing algorithm among all the above algorithm is the linear regression with the highest MSE.
- While the same algorithm when fine tuned gives us the best values for MAE. MSE, MAPE and RMSE(refer row 2) as compared to the other regression algorithms indicating it to be a better fit.

Performance Analysis and Conclusion

- Among the other candidate algorithms i.e Lasso Regression, Ridge Regression and Elastic Regression algorithm the performance has a very little variance. The performance of the above algorithms can be arranged as:
- Linear(with fine tuning) > Elastic Net > Lasso > Ridge > Linear(without fine tuning).

Related Work

There is a lot of work done around Airbnb price prediction not only in R but also in python. In general the following are things to note:

- Some of the related work did not have any fine tuning done and only focused on getting rid of prices where it was zero. But in our work we have removed outliers, taken care of multicollinearity which is a big factor when dealing with many factors such as ours.
- Another case where our algorithm was a little better was when we included interaction terms and polynomial terms which had a good effect on fine tuning.
- Some work done was also better than our where implementation of text mining on basis of reviews was done, inclusion of seasonal and every day data was included to give a more accurate and precise price output.

Issues and Alternatives

- Our work does not consider analysis around neighborhood as there were around 200 unique ones. We could have set some standard or logic where we reduce them to say around 40-50 and then have a better knowledge on the listing in a specific neighborhood within a neighborhood group.
- We do not have any owner review which is a limitation. The alternative was to use a different dataset that would have all the attributes we worked upon and also the owner review as this can be a good filtering criteria.

Potential Extensions or Future Work

- Include a wider geographic area like more states or entire US and major cities around the world.
- There are scenarios where the guest pays by the day and does not prefer booking a set number of days. In this case, the prices vary everyday. So, in addition to predicting base prices, a model could be created to calculate daily rates using data on seasonality and occupancy.

Potential Extensions or Future Work

- Include Airbnb listing owner rating so that we can filter a room/apartment by high owner ratings.
- We could focus on the neighbourhood to price correlation more deeply. Upscale neighbourhoods will have higher priced listings, and finding the qualities of that encourage people to pay higher prices may be an important application for real estate planning.

Key Contributions

- Best Subset Regression, Step wise selection and Lasso Regression were implemented by Ankit.
- Ridge, Elastic Net Regression and Random Forest were implemented by Vaishnavi Badame.
- Data Visualization, core algorithm testing and fine tuning was a combined effort of the two of us.