**Use Learned-based Model to Predict Airbnb Rent Price**

**Project Reflection Report**

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019. The overall objectives of our project can be put into two categories.

1. **To help customers/renters find a good price**. What we mean by that is, if any customer is looking for a rental place and he/she is finding very different rental rates in that specific neighborhood. Our predictive analysis can provide that customer/renter an insight that he/she should not pay more or less than the price allocated to that place. In that way, customers can save money and have a good idea that which neighborhoods are offering the places at which rates. For example, if a customer looks for a place in Brooklyn and that place is asking for 1200 USD but originally that place is not worth more than 800 USD then our model can help the customer save money.
2. **To help Landlords assess their own Rental rates**. If landlords are asking for too much price as compared to their neighborhood, then our model can help predict what should be the actual rent for that place. For example, if a landlord is asking for 1500 USD whereas the whole neighborhood is asking for no more than 1100 USD then this model can help the landlord to reduce the rent close to 1100 to attract more renters and eventually get more income.

Looking at the objectives, we can clearly see from a business point of view that both parties, customer and renter, are getting benefits from this model whether it be saving money or earning more money respectively.

In Milestone I, we studied different research papers and went over their literature reviews in order to get a better understanding of our topic and gather more knowledge for our project. We explored about different criteria that could be employed in order to get the results. These criterias included linear regression, support vector machine, xgboost and r-square criteria. Moreover, in that Milestone we eliminated a significant amount of columns which were not useful for our purpose in this project. We interpreted the meaning of different columns included in our final dataset and wrote their description and datatypes.

In Milestone II of the project, alongside predictive analysis more of our focus was on the exploratory analysis. We attempted to explore the relationships between different variables and visualized the information in the form of bar charts, boxplot, scatter-plot, histogram and heat map. During visualization, we found some interesting patterns and relationships between the variables. For instance, when plotted the relationship between count (number of customers) and different neighborhoods of New York then it showed that the count for Manhattan and Brooklyn was insanely high than that of Queens, State Island and Bronx. The reason for this behavior is that there are more tourist attraction sites in Manhattan and Brooklyn than the other mentioned towns.

Another interesting observation was to know that a majority of the people like to rent out private rooms or entire home/apartment instead of going for shared rooms. It depicts that a large number of people want to preserve their privacy and do not want to compromise it. One more interesting finding is that even though the rental prices in the areas of Manhattan and Brooklyn are high but still a huge chunk of people tend to go these locations. The other areas such as Bronx, State Island and Queens despite being less expensive remain the comparatively low visited places.

Another interesting and more of an obvious fact that we found is that when we plot the number of customers and price, there is an inverse proportion between these two quantities. As the price goes on increasing the number of customers tend to decrease implying most of the people want do not want to rent out expensive apartments. Another interesting thing we discovered was to know about the size of population in a certain area using the package wordcloud. Bigger the size of a word in the graph, more the number of people in that specific area. Following is the graph we found during our data analysis.

In the Milestone 3, our main focus was on the predictive analysis and find out which of the Machine Learning models was the best option for our data analysis task. We started off by comparing two versions of cleansed dataset (with and without outliers). We applied the OLS regression method on both the versions of our dataset. Based on the resulting R-squared value we found out that the one without outliers was the better version for fitting ML models.

Another reason of selecting the dataset without outliers was that the outliers were representing the wealthy people which were few in number as compared to the people with normal financial condition. After selection of the dataset, we split the dataset in the two sections, training dataset and test dataset with the 80:20 ratio. Then we fit six regression algorithms (linear regression, ridge regression, lasso regression, elastic net regression, XGB and random forest) on our data. When we compared the mean absolute error, root mean squared error and R squared valued for these algorithms, we found out that **XGB** and **Random Forest** outperformed the other algorithms as shown in figure below.

