**CIS 8005 Group Project Team 15**

**Satyaraj Reddy Nandi:** [**snandi1@student.gsu.edu**](mailto:snandi1@student.gsu.edu)

**Aashrith Sangani:** [**asangani1@student.gsu.edu**](mailto:asangani1@student.gsu.edu)

**Neha Samhitha Pinjala:npinjala1@student.gsu.edu**

**Project Summary:**

**Problem statement:**

**The main influence for this project came in as we often find ourselves immersed in the how a giant like Airbnb predicts the prices for their hosts:**

* What can we learn about different hosts and areas?
* What can we learn from the predictions? (ex: locations, prices, reviews, etc.)
* Which hosts are the busiest and why?
* Is there any noticeable difference of traffic among different areas and what could be the reason for it?

**This project aims to solve the problems using ML techniques to predict the prices for listings in the NYC area along with performing EDA in the due course.**

**Data Source(s):**

* The dataset has been sourced from **kaggle.com**, from which any personal identifying data such as name,host\_name have been removed.
* This public dataset is part of Airbnb, and the original source can be found on this website This dataset describes the **listing activity and metrics in NYC, NY for 2019.** This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.
* **Link to the dataset:** [**http://insideairbnb.com**](http://insideairbnb.com)

**A snippet of how the data looks like:**

**A screenshot of a computer

Description automatically generated with medium confidence**

* **Contains pricing information[including categorical variables] such as**:
  + Host\_ID
  + Neighborhood group along with the neighborhood name
  + Latitude & longitude of the places
  + Room type listed by the host
  + The number of reviews along with reviews\_per\_month
  + Availability of the place in a cycle of 365 days
  + Prices of the individual listings

**We’ll be using the dataset to model and predict the below:**

* *Performing EDA on the dataset [Data Cleaning, Data Visualization & Word Cloud]:*
  + Predict the price of the listings by employing **Random Forest Regression & XG Boost algorithms** after proper data cleaning, label encoding, normalization & feature engineering.

**Platform used: Jupyter Notebook**

**Data Preparation:**

**Data Cleaning:**

* Dropping duplicates values(if any) in the dataset: duplicated() & drop\_duplicates
* Removing the NaN values from the dataset: isnull().sum() & dropna()
* Filling NaN values in the dataset with appropriate values: fillna()
  + The NaN values in the respective column, reviews\_per\_month were replaced by 0 as it was logical to do so.
* Dropping unnecessary data : drop()
  + Accordingly dropped certain columns from the dataframe as they were irrelevant for the models & included personal data of the hosts as well.

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Data Preprocessing:**

* + **Label Encoder: LabelEncoder()**
    - We encoded the categorical variables required to execute the correlation b/w the features.

Graphical user interface, application

Description automatically generated

* + - **Scaling: StandardScaler()**
      * Scaling was also performed on the data before building the model as the values for a few features were exorbitantly high.

Text

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

**Removing all the unnecessary data**

Chart

Description automatically generated

**Checking for correlation among factors and key contributors to price.**

Graphical user interface, text, application, email

Description automatically generated

**Checking for Multicollinearity in the data**

**Data Visualization:**

Graphical user interface

Description automatically generated

**Correlation between the variables**

**Chart, histogram

Description automatically generated**

**A histogram for prices less than $1000**

Chart

Description automatically generated

**Let’s see the spread of the popularity of the neighborhood groups.**

Chart, scatter chart

Description automatically generated

**Let’s use the scatterplot to depict the availability of the rooms.**

Chart, bar chart

Description automatically generated

**Let's plot the most popular neighborhood area in terms of the Airbnb hostings there.**

Chart, bar chart

Description automatically generated

**Let’s use a catplot to depict the spread of room\_types among different neighbourhoods split by the neighbourhood groups.**

Chart, bar chart, histogram

Description automatically generated

**Let’s see how the words play a role in listing the houses on Airbnb.**

A picture containing text, newspaper

Description automatically generated

**A word-cloud analysis on the neighborhood feature.**

**Data Partition:**

**For the modeling, we split the dataset into train and test using train\_test\_split().**

**Graphical user interface, text, application

Description automatically generated**

**Before we started modelling, we performed label encoding along with feature processing on the data in order to make it ready for modelling: Removing the price skewness.**

**Text

Description automatically generated**

**Normalizing the data using Standard Scaler Technique:**

**Text

Description automatically generated with low confidence**

**Splitting the data into different variable types along with label encoding:**

**Graphical user interface, text, application, chat or text message

Description automatically generated**

***We have split the data into 80 : 20***

**A custom function for printing out the MAE, MSE & R2E for the following models:**

**Text

Description automatically generated**

**Modelling:**

***XG Boost***

**Code Snippet for the XG Boost model depicting the MAE, MSE & R2E values:**

**Graphical user interface, text, application, email

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**The inputs were accordingly encoded in the label-encoding stage. The output here is the price. A plot showing the model performance depicting measured and predicted values.**

***Random Forest Regression***

**Code Snippet for the Random Forest Regression model depicting the MAE, MSE & R2E values:**

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

**A plot showing the model performance depicting measured and predicted values.**

**Evaluation**

* The code for the project runs from understanding how the data is given.
* Next phase is data cleaning: which involves checking for NaN values, dropping unnecessary data, removing the NaN values from the dataset: isnull().sum() & dropna(), filling NaN values in the dataset with appropriate values: fillna()
* After proper data cleaning, it’s time for data preparation which mostly involves *Data Processing*, as explained in the preceding pages of the document.
* Now, it’s time to visualize the data. We published major findings on how the dataset gave us immense results. It helped us understanding the demographics of the hosts / neighborhoods / correlation between the different variables.
* And the final step in the whole evaluation of the project was modelling. Here, we uniquely wrote a function that takes the model names in its constructor and prints out the metrics useful for comparison. We gently employed XG Boost & Random Forest Regression to test the predictions.

**Recommendations**

***The accuracy can be much more enhanced with the following recommendations:***

* The data sample size is relatively small in this case, especially with a perspective of applying this in real world.
* The number of features is also limited in this dataset. With an addition of 5 or more valid/logical features, the predictions can be improved much further.
* The data is kind of skewed with regards to price for a few listings with certain cases, going upwards of >9,000 where as the average price is 153. Even after normalizing the price, there are still some outliers affecting the accuracy of the models.

**Future Work**

* Introduce more models and compare their accuracy metrics with regard to the price.
* With the housing boom at an all time high since the pandemic, there’s massive scope for further analysis with petabytes of data being pumped in every month. More data will make the analysis more interesting.
* And lastly, with an assumption that travel will rise in the future, companies like Airbnb / Booking.com will look ahead to meet the amped up demand post pandemic. Vast data with n number of features will prove to be beneficial for building robust prediction models.