EAI6010 – Applications of Artificial Intelligence

CRN: 71964

Final Project – New York City Airbnb Data Analysis and Prediction

Dishant Rajesh Modi

Karan Kaushal Daiya

[modi.di@northeastern.edu](mailto:modi.di@northeastern.edu)

daiya.k@northeastern.edu

College Professional Studies, Northeastern University.

**INTRODUCTION**

For this project we are taking the Airbnb dataset from New York and using it analyze and build visualizations, lastly we will be using it to train and test our ML models.

We will be using Linear Regression and Gradient Boosting Machine as our ML models. We will also compare accuracy between predicted and actual values

**TASK**

**Explore and Understand data**

We started by using following packages:

Plotly,

Folium,

Wordcloud,

Geopandas,

Since we don’t have those packages installed, we will install it through pip







Next we will import the required packages as follows,

*import numpy as np*

*import pandas as pd*

*import matplotlib.pyplot as plt*

*import seaborn as sns*

*import geopandas as gpd*

*plt.style.use('fivethirtyeight')*

*%matplotlib inline*

*import plotly as plotly*

*import plotly.express as px*

*import plotly.graph\_objects as go*

*from plotly.offline import init\_notebook\_mode, iplot, plot*

*init\_notebook\_mode(connected=True)*

*import folium*

*from folium import plugins*

*import wordcloud*

*from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator*

*from sklearn import preprocessing*

*from sklearn import metrics*

*from sklearn.metrics import r2\_score, mean\_absolute\_error*

*from sklearn.preprocessing import LabelEncoder,OneHotEncoder*

*from sklearn.model\_selection import train\_test\_split*

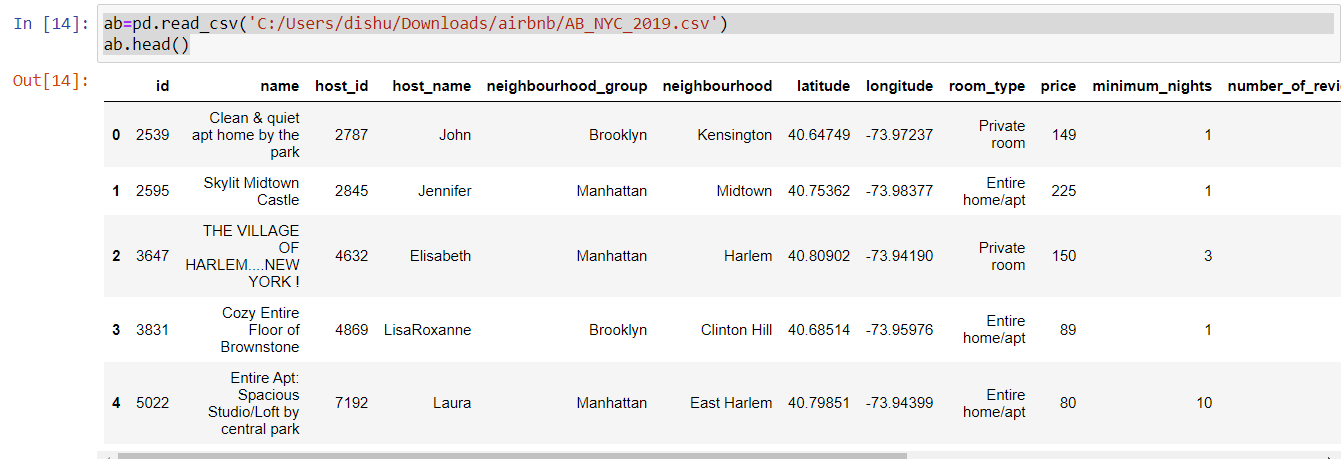
*from sklearn.linear\_model import LinearRegression,LogisticRegression*

*from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor*

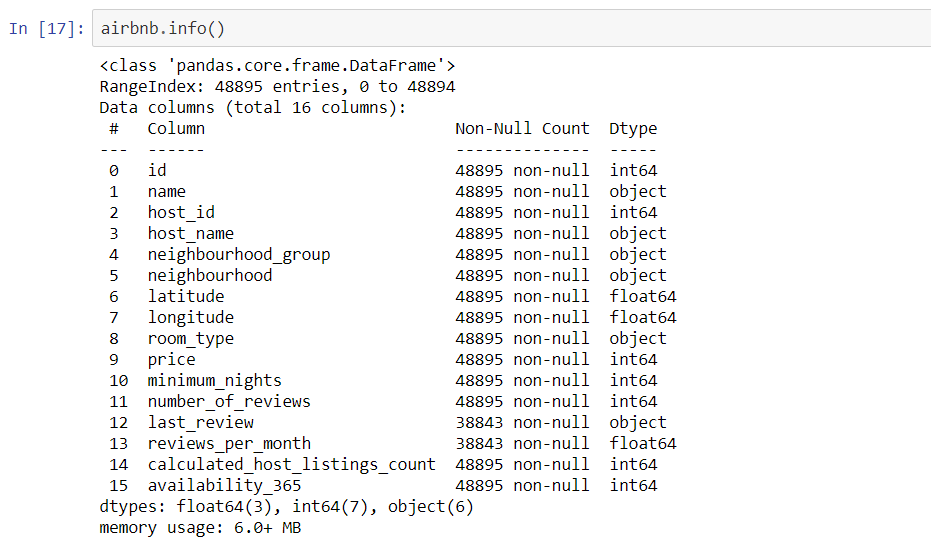
Next we will import our dataset using pandas dataframe.

*ab=pd.read\_csv('C:/Users/dishu/Downloads/airbnb/AB\_NYC\_2019.csv')*

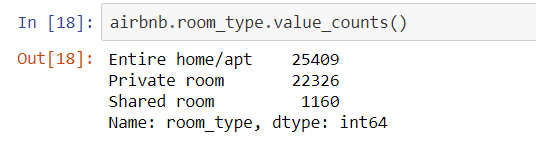
*ab.head()*



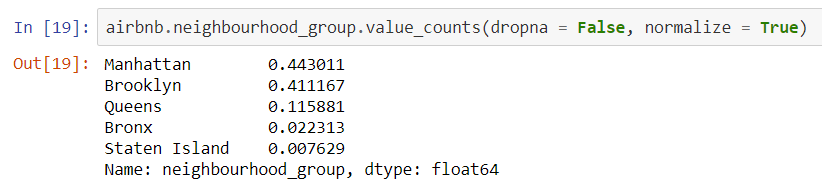
We can now do preliminary data analysis and EDA using .describe(), .info() etc.



Lets see how many different types of rooms are there.



WE can also see how many percentage of accommodations are allocated to each locality.

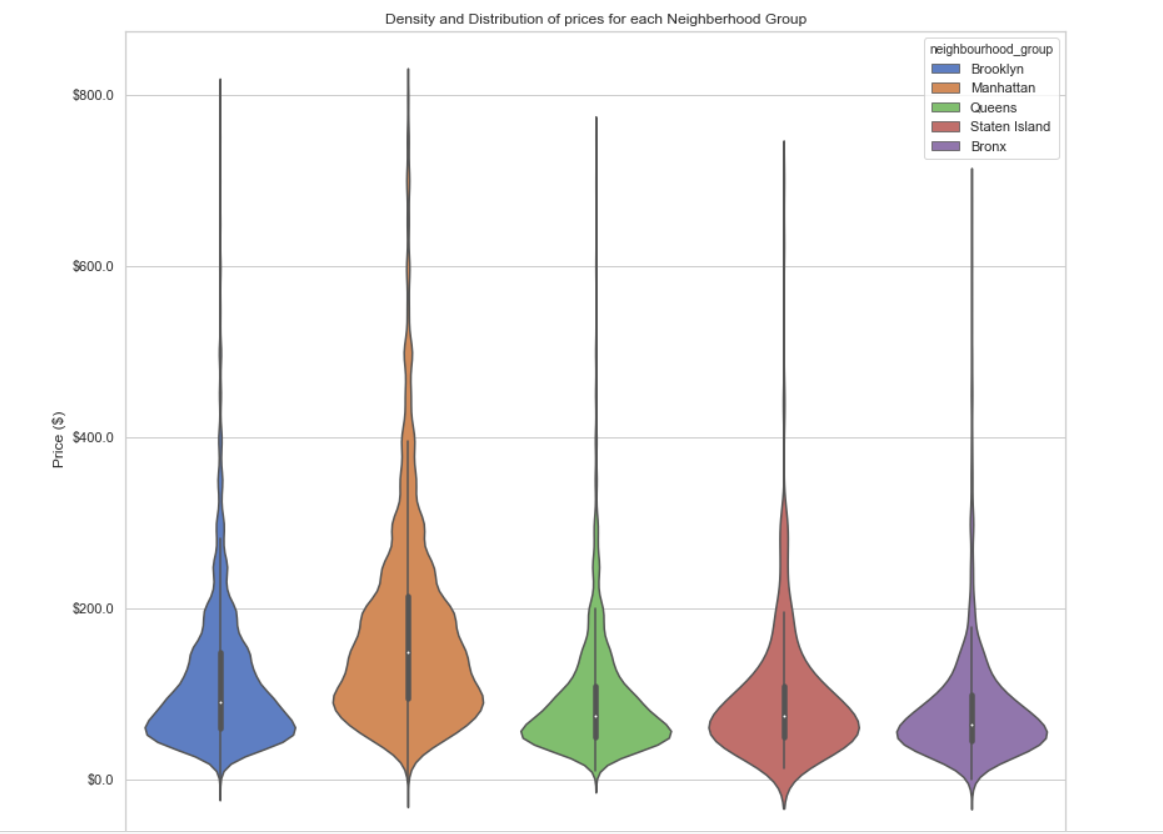


We also found out most rented neighborhoods



We can see highest is Manhattan and it is most rented in Airbnb.

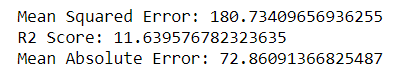
We also found density distribution of prices classified by neighbourhods.



We can now train our model and test it using the data that we have.

Firstly, we perform linear rregression.

Using lm.predict we can check for the prediction and compare it.

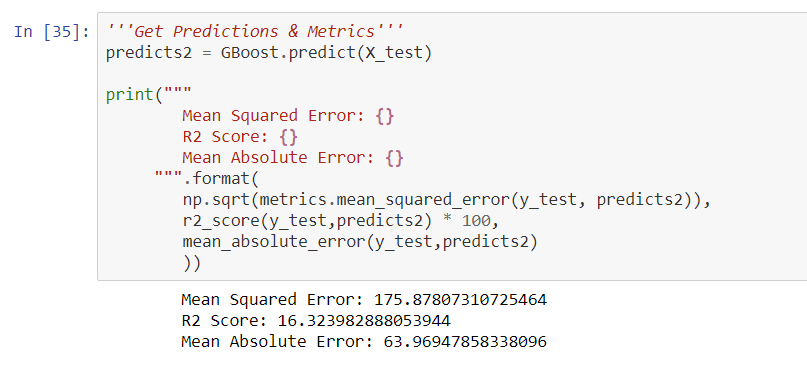


For linear regression, we got r2 score of 11 which is not a goof fit.

Lets compare the actual vs predicted through bar graph



Next we created GBM model and checked for its accuracy.



Comparison for GBM model



**CONCLUSION**

This is experiment gives us a nice understanding of how ML models works and how they collectively constitute an AI system.

With more complex datasets, we can achieve more complex results.