New York Airbnb Data Analytics and Prediction

Ajitesh Nair
Computer Science and Engineering
PES University
PES2201800681
ajiteshnair@gmail.com

Ishan Padhy
Computer Science and Engineering
PES University
PES2201800158
ishanpadhy21@gmail.com

Nikhil J K
Computer Science and Engineering
PES University
PES2201800303
nikhil.jk02@gmail.com

***Abstract*—This is the final report of our data analytics project titled "New York Airbnb Data Analytics and Prediction" as a part of our Data analytics course UE18CS312.The goal is to analyse and predict the price and other variables in the New york Airbnb data.Also a recommendation system will be built to recommend Airbnb listings according to the user preference.**

Keywords—Data analytics,New york airbnb,Pattern analysis,Prediction,Recommendation

I. INTRODUCTION

This is the final report of our data analytics project titled "New York Airbnb Data Analytics and Prediction" as a part of our Data analytics course UE18CS312.The goal is to analyse and predict the price and other variables in the New york Airbnb data. Various visualisation techniques and other methods are used for exploratory data analysis.The data is searched for patterns.Correlation between the variables are calculated and models are developed so as to accurately predict the target variable. Also a recommendation system will be built to recommend Airbnb listings according to the user preference.

II. LITERATURE SURVEY

Various research papers were analysed to see what kind of analysis and problems were solved by people similar to ours.Based on the similarity to our problem/approach , 3 research papers were analysed and the summaries are given below.

A. Research Paper 1

Name : A socio-economic analysis of Airbnb in New York City [1]

Source:https://www.ceeol.com/search/article-detail?id=578374

Publish Date : July,2017

RELATION: This paper aims to map Airbnb's presence in New York City but, going beyond visual inspection, it analyses the socio-economic factors influencing the spatiality of Airbnb in the American metropolis.

CLAIMS: Before crafting the most efficient and equitable responses, cities need more empirical work to understand the evolution of the short-term rental market and the nature of the externalities associated with it.

Like other research, this study has some limitations. First, is that it is limited to one city only. Second, the study shows only a snapshot of listings. Considering these, hoping this encourages further studies.

TAKEAWAY: Results indicate that Airbnb accommodations and the number of reviews are concentrated in those parts of the city that have a young population, a significant no. of housing units, a high number of points of interest. Noted that, there are significantly more no. of white hosts. And that the ethnic localities have a lower interest compared to that of the white locality. There is no much correlation between the Airbnb price and the selected indicators but price moderately correlates with education, household income and POI supply. Thus, the connection between gentrification and the growing Airbnb offers can be supported.

B. Research Paper 2

Name : Price-Setting Behavior in a Tourism Sharing Economy Accommodation Market: A Hedonic Price Analysis of AirBnB Hosts in the Caribbean [2]

Source : https://mpra.ub.uni-muenchen.de/95475/

Publish Date : August,2019

RELATION: This study investigated the price-setting behavior of hosts in the tourism sharing economy in the Caribbean ,a region that exhibits differences from country to country concerning biodiversity, geography, culture, historical and political background, and economic performance.

CLAIMS: OLS results indicate that 32 of the 36 variables are significant determinants of price-setting behavior. Results from quantile regressions also indicate that these variables do explain price-setting, but these effects vary over the distribution of prices under study. This is evidence of the complexities in the pricing of accommodation in the tourism sharing economy.

TAKEAWAY: This paper sheds some light on the factors behind the substantial pricing heterogeneity observed in AirBnB properties across the Caribbean as well as within the same country.

Understanding these patterns of pricing heterogeneity is necessary to assist policymakers in making informed decisions regarding the sector, in relation to regulation

and other concerns. Findings are important for hosts, as it allows them to better assess the market environment and improve their sales and profits. The study also provides tools for Airbnb, and possibly other P2P platforms in designing tools to help guide hosts in price-setting

C. Research Paper 3

Name :Airbnb Usage Across New York City Neighborhoods: Geographic Patterns and Regulatory Implications [3]
Source:https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3048397
Publish Date : October,2017

RELATION: This paper offers new empirical evidence about actual Airbnb usage patterns and how they vary across neighborhoods in New York City.

CLAIM: The problem and data used in this paper is somewhat similar to the ones chosen by us.The focus is mainly on Airbnb usage patterns and how they vary across neighborhoods in New York City.It is found that Airbnb listings have become more geographically dispersed, although centrality remains an important predictor of listing location.They have used data from various sources like airbnb,zillow and trip advisor to extract unique insights about the relationship between short-term rentals, long-term rentals, hotels, and the characteristics of the neighborhoods in which these types of lodging are located.

TAKEAWAY: This paper describes how the airbnb listings have varied from 2011 to 2016.It lso describes how the pricing of the listings varies across regions.Overall the paper gives us some insights on what kind of patterns can be searched for and also provided insights which we would not have been able to provide due to data source limitations.

III. PROPOSED PROBLEM AND APPROACH

Since our dataset contains several missing values,preprocessing has to be done.Missing values will either be removed or replaced with the column mean based on how important the attribute is. Also with respect to preprocessing the datatype of certain attributes like last_review has to be changed to make processing easier.

Our main goal is to analyse and find interesting patterns between the variables in our dataset.Visualisation is an important aspect of finding patterns. Hence several visualisation techniques like bar graph , scatter plot , correlogram etc will be plotted to gain insights

We then plan to predict certain variables such as price by using predictive models. Several models will be explored and models with the best accuracy will be selected.

We also plan to make some kind of recommendation system which would recommend Airbnb listings based on a particular keyword given by the user.

IV. EXPLORATORY DATA ANALYSIS

A. Dataset

Name:New York City Airbnb Open Data

Source:Kaggle(https://www.kaggle.com/dgomonov/new_york-city-airbnb-open-data)

Number of Columns:16

Number of Samples:48895

Number of quantitative variables:10

Number of qualitative variables:6

Attributes:

id, name,

host_id, host_name,

neighbourhood-group,

neighbourhood,

latitude, longitude,

room_type, price,

minimum_nights,

number_of_reviews,

last_review,reviews_per_month,

calculated_host_listings_count,

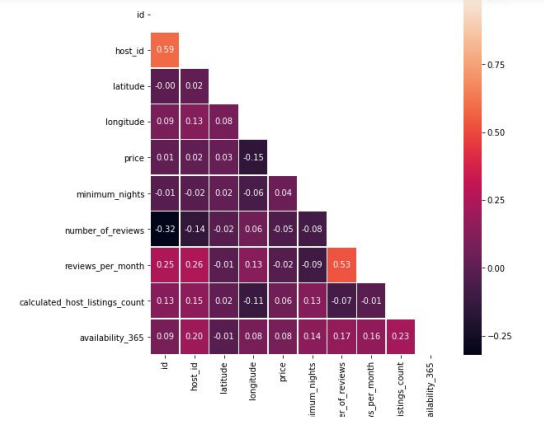
availability_365.

B. Exploratory Data Analysis Summary

-Data type of one column (Last_review) had to be changed from object to date-time.

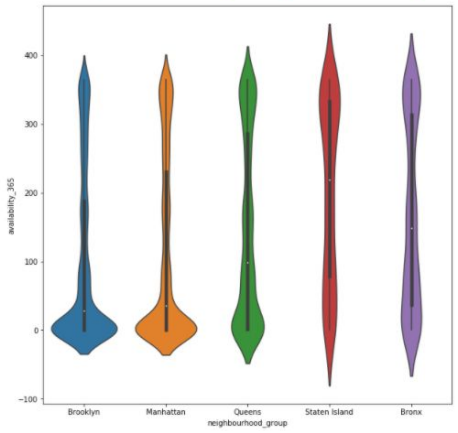
-Few columns like name,host name,last review had many missing values and they were not of any importance for analysis,hence they were deleted.

-Reviews per month column had a lot of missing rows but is important for analysis,hence missing values are replaced with the mean of that column.

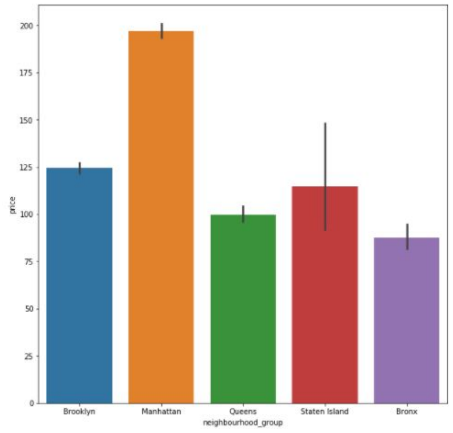


-Correlation heatmap showed a high correlation between number of reviews and reviews per month.

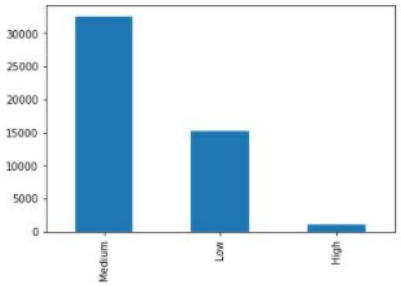
-Relationship between the neighbourhood_group and the availability_365 was explored using violin plot. It was seen that Staten Island has the highest mean availability value around 220-250 days compared to others



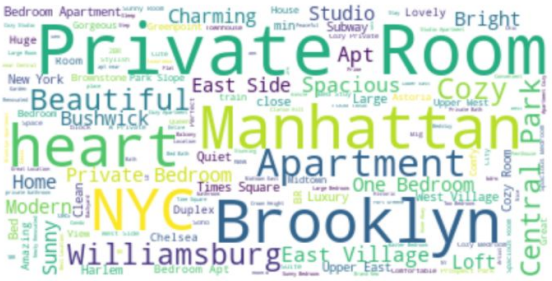
-Bar plot was plotted between neighbourhood group and price and Manhattan was seen to have the most expensive prices.



-The listings were divided based on price and bar graph was plotted. Most of the rooms fall in the Medium (75-500 $) category followed by Low and High range rooms.Very few people prefer to live in high priced rooms



Our word cloud shows the words that are more often used in the Name of this listings.We can see most of the words are related to the description,place and the experience in the room.



-Low cost rooms or in range 0-50 $ have more reviews.This shows us that people who pay more for the rooms generally dont write reviews.It is observed people write reviews more if they are not happy with their experience.In case of costly rooms there is high possibility that the customers are happy.So they dont write much reviews.

V. MODELS

A. LGBM CLASSIFIERS

There are 16 names fields as NaN.Hence they are replaced with empty string.Then punctuation, digits and special characters are removed.Then stop words removed . Now lets define a price above 300 as expensive and below 300 as cheap.

	clean_name	final_name
0	clean quiet apt home by the park	clean quiet apt home park
1	skylit midtown castle	skylit midtown castle
2	the village of harlem new york	village harlem new york
3	cozy entire floor of brownstone	cozy entire floor brownstone
4	entire apt spacious studio loft by central park	entire apt spacious studio loft central park

We divide dataset into 80% training and 20% as test dataset

We use TF-IDF metric which we have learnt in class to represent the weights of each feature.

LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms, used for ranking, classification and many other machine learning tasks.

LGBM classifier is used which is similar to decision trees.

	precision	recall	f1-score	support
0	0.97	0.82	0.89	9108
1	0.22	0.68	0.33	671
accuracy			0.81	9779
macro avg	0.59	0.75	0.61	9779
weighted avg	0.92	0.81	0.85	9779

Accuracy: 0.809
Recall: 0.680

The scores aren't perfect but taking into account that the predicting model is built solely on textual descriptions of a listing, it seems like the words in Airbnb titles actually do matter!

B. LINEAR REGRESSION MODEL

First convert the categorical features into numeric by using encoding.Prices are not normally distributed as well as there is a lot of noise.Hence instead of considering price, we consider log(price) for all further models.

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

Then a linear regression model was built ,trained and tested using the training and testing dataset that is 80% and 20% of the dataset.

The following are the output:

Mean Squared Error: 0.23993364054603417

R2 Score: 41.24192706830667

Mean Absolute Error: 0.17677653373327856

C. DECISION TREE MODEL

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

A decision tree model was built next with maximum depth set as three. The model was built using the sklearn library.The model was trained and tested and the following output was achieved:

Median absolute deviation (MAD):0.07953946647160076

Mean Squared Error: 0.22889689464980478

R2 Score: 46.52324351141029

Mean Absolute Error: 0.16583322887806484

D. RIDGE REGRESSION MODEL

Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

Next a ridge model was built using the same sklearn library. The parameters alpha was set as 0.01 and normalisation was set as true. The model was then trained and tested and the following output was achieved.

Mean Squared Error: 0.23992898296923088

R2 Score: 41.24420826222875

Mean Absolute Error: 0.1767871985511428

E. LASSO REGRESSION MODEL

Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

Next a Lasso model was constructed using the skearn library.The parameter alpha was set as 0.001.The model was trained and tested and following output was achieved:

Mean Squared Error: 0.23990803282018353

R2 Score: 41.254468705442136

Mean Absolute Error: 0.1767949649280997

We can see that the Decision Tree model has the highest R^2 value and least error. Hence it is the best model.

VI RECOMMENDER SYSTEM

```
In [9]: from fuzzywuzzy import process

        def airbnb_finder(title):
            all_titles = mydata['final_name'].tolist()
            closest_match = process.extractOne(title,all_titles)
            return closest_match[0]
        title = airbnb_finder('village')
        title

Out[9]: 'village harlem new york'
```

We have implemented a recommender system which suggests an airbnb listing using the keywords the user has provided. It returns the most relevant airbnb listing.The fuzzywuzzy library was used to extract the airbnb listing most similar to the keyword provided by the user.

VII .CONCLUSION

We were able to successfully analyse the dataset and gain several useful insights.We also developed several models specifically Lgbm classifier,linear regression, decision tree, ridge regression, lasso regression etc out of which decision tree seemed to be the most accurate.Then we also built a basic recommendation system which takes a keyword or a set of keywords from the user and returns the most similar airbnb listing based on name.

The workload was divided equally among the three team members with each member contributing to preprocessing,visualisations,model building and recommender system.

V. REFERENCES

[1] Dudás Gábor, Vida György, Kovalcsik Tamás, Boros Lajos,A socio-economic analysis of Airbnb in New York City (July,2017) Available at :https://www.ceeol.com/search/article-detail?id=578374

[2] Lorde, Troy and Jacob, Jadon and Weekes, Quinn (2018): Price-Setting Behavior in a Tourism Sharing Economy Accommodation Market: A Hedonic Price Analysis of AirBnB Hosts in the Caribbean. Published in: Tourism Management Perspectives , Vol. 30, (2019): pp. 251-261.Available at : https://mpra.ub.uni-muenchen.de/95475/

[3] Coles, Peter A. and Egesdal, Michael and Ellen, Ingrid Gould and Li, Xiaodi and Sundararajan, Arun, Airbnb Usage Across New York City Neighborhoods: Geographic Patterns and Regulatory Implications (October 12, 2017). Forthcoming, Cambridge Handbook on the Law of the Sharing Economy, Available at SSRN: https://ssrn.com/abstract=3048397 or http://dx.doi.org/10.2139/ssrn.3048397