CSCE 5310: Methods in Empirical Analysis
Project Increment -1

Project Title: Airbnb Price Predictions

Git - *https://github.com/Saisriteja12/Emperical_analysis_project.git*

Team Members:

Roshan Sah [11574385]
Saisri Teja Pepeti [11555656]
Yamuna Bollepalli [11552426]

Goals and Objectives

Motivation:

- Main motto is to analyze how prices are changing depending on the total reviews and the places where most of the people are booking through Airbnb.

Significance:

- As we said to analyze the prices in the first place, we need to find the relation between certain factors like location(may be busiest) vs demand from customers to the neighborhood group to the number of nights.
- As this is the empirical analysis project, to analyze the data we're using various validation methods and hypothesis test to check the performance and verify how discrete and continuous data is affecting the target variable.

Objectives:

- To conduct statistical analysis on the data set using various stats modules (on New York AirBnB data), including hypothesis testing, tests of the mean (Kruskal Wallis Test, ANOVA - one way an d two way), tests of proportion (z test and chisquared test), and tests of variance (Ftest, Levene test), after ensuring that the three assumptions:
- Normality of the target variable
- Randomness of sampling
- It is assumed that the level of significance is 5% (alpha = 0.05).
- Parametric tests can be run if the assumptions are met; otherwise, non-parametric tests must run.

- We shall be able to determine the dependability and associativity of certain aspects on one another based on the outcomes of the tests that were run. We'll confirm our findings using data visualization approaches.

Related Work:

With the help of the listing's customer evaluations, owner data, and property specs, this study intends to develop a model for estimating the cost of an Airbnb listing. The resulting model can be used by owners and clients to calculate the expected value of an Airbnb listing. On a dataset of Airbnb listings from New York City, linear regression, tree-based models, K-means Clustering, support vector regression (SVR), and neural networks are trained and optimized. The resulting models of this article (Weippl et al. 2021, 1) are then compared in terms of Mean Squared Error, Mean Absolute Error, and R2 score. Customer review features are extracted using sentiment analysis to improve the performance of the chosen predictive models.

Since China will be one of Airbnb's key markets, this essay focuses on the Beijing Airbnb market. For the Beijing Airbnb market, the article has created a pricing prediction model based on machine learning techniques, such as XGBoost and neural networks. The research (IEEE Staff 2021, 2) chooses significant aspects from the analysis of price-related data to build the prediction model. In addition to the precise price prediction model, the research offers advice for hosts on how to raise their rates by including crucial facilities.

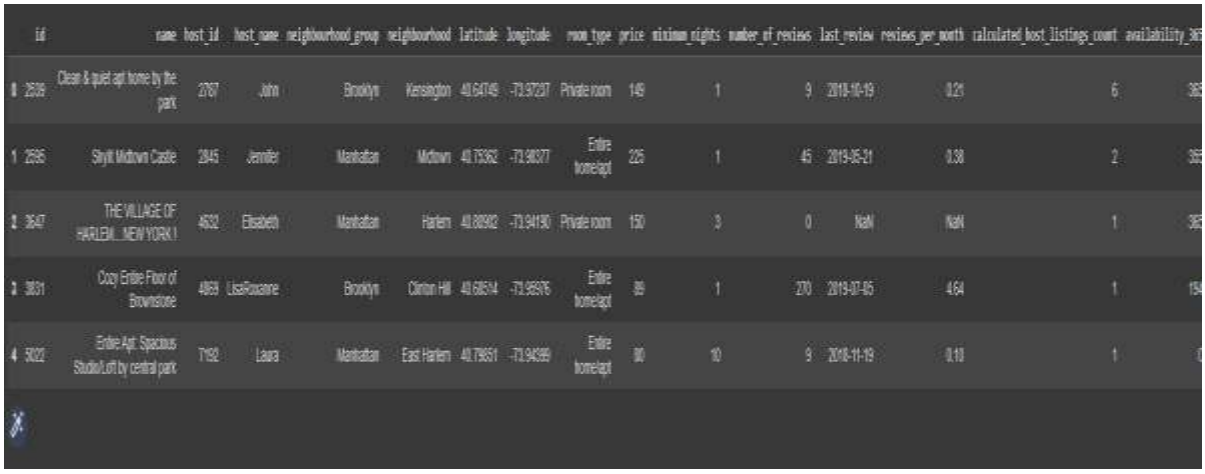
In this paper (IEEE Staff 2016, 3), we first carry out data pre-processing and data cleaning. After that, we do descriptive, prescriptive, and exploratory analysis to learn more about the data's nature. These analyses made it easier to comprehend the crucial factor that must be considered in order to forecast the pricing for our Airbnb listings. Outlier detection was carried out on the dataset and any identified outliers were taken out of it since, even after doing data cleaning on the data set, certain outliers may need to be thoroughly inspected. The models used for price prediction are random forest, logistic regression, and linear regression. The three aforementioned methods were used, and the best model was selected based on the RMSE value.

In order to classify and validate the test scores for greater accuracy against the independent variables to dependent variables, we have referred to pertinent research papers that discuss how we can apply the validation methods on large and well-known fields of data sets that also have more

information as data. by looking at certain case studies, such as analyses of Amazon fine food review data and projections of medical costs and movie box office receipts. After reviewing all these case studies, we discovered that we were missing some data inputs. As a result, we discovered this intriguing case study when we took into account the market's present conditions for hotels and rental services. After that, we considered how to provide specific results for people looking for instant services with location-based analysis.

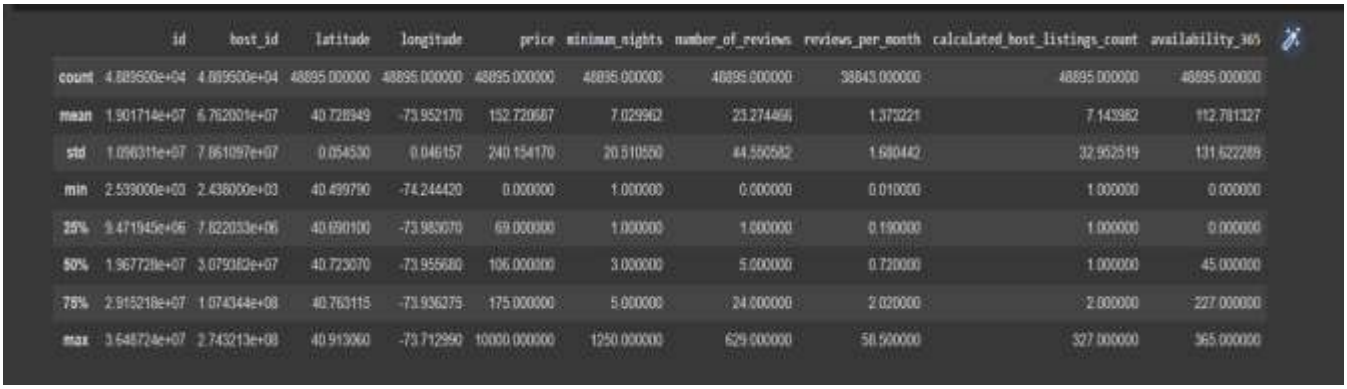
Dataset:

We will take the dataset which contains almost 45k rows with 16 columns for each. It is composed of 3 float types, 7 int types and 6 object types of id, name, host_id, host_name ,neighbourhood_group, neighborhood, latitude, longitude, room_type, price minimum_nights, number_of_reviews, last_review, reviews_per_month calculated_host_listings_count,availability_365.



	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
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0

Figure 1: Sample of dataset



	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889600e+04	4.889600e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

Figure 2: Statistical Description of Dataset

Features:

To bring up the statistical analysis we are definitely working on the important features which are:

1. Variance Check
2. Randomness
3. Normality Check

Along with these we are running the code on:

4. One Way Anova
5. Two Way Anova
6. Shapiro Test we give normality check
7. Kruskal wallis Test
8. Chi-square Test
9. Z-Proportion Test
10. Hypothesis Testing
11. Levene Test
12. F-Test
13. P-Value
14. Performance Metrics
15. Visualizing the analysis
    a. Distance Plot
    b. Pie Chart
    c. Scatter Plot
    d. Bar Graph
    e. Box Plot
    f. Line Plot
    g. Violin Plot

16. Exploratory Data Analysis
17. ML algorithms such as:
    a. Linear Regression
    b. Multiple Linear Regression
    c. Logistic Regression
    d. Naive Bayes
    e. SVM
    f. Decision Tree
    g. KNN

Analysis of Implementation:

Post collection of data set, we planned to preprocess the data before analyzing it to the real scenario validations depending on the data characteristics. By selecting the validation methods during the data selection also at the data preprocessing for the performance of the models. Then picking up the major features to get all the validation parameters which will give the model to predict prices and improve the model. Eventually visualizing all this information will help us understand more about our data in various visualization graphs. Furthermore, to find the relationship between continuous data we will use correlation function. For the efficiency of the specific data taken we have to check the outliers of it. Here, we compare the different characteristics of available information using statistical analysis. The major part of the empirical analysis with the help of Standard deviation, mean, variance, median we analyze our statistics and population. As we advance in our implementation and empirical methods, we will be having t-Test P-value and f and f statistics, in this major element is hypothesis testing. With all these we will process our data to the ML models which are: SVM, KNN, LR, LR, RF, DT. Finally, we will achieve the desired predictions of the selected model with the help of all these empirical analysis methodologies.

Implementation:

- Our models run on statistical analysis so we imported the stats model API which has all the methods and functions related to statistics also along with the necessary libraries like Pandas, Matplotlib, NumPy, Sklearn.
- For finding reports of our data, we have imported classification reports along with that we need some metrics to validate our model like ROC, AUC score.
- To map the correlation between the data, two features of data we plotted confusion matrix

- One of the major techniques we use is f-one-way test and stratified k-fold for our models
- For the classification of ML models, we imported Linear Regression, K -neighbors classifier
- So, after all these further implementations we load our data set and analyze our data information then we have given a check for statistical data analysis to know the mean, standard deviation and also check the missing, null and duplicates values and drop the missing and null values from data to improve the price prediction.
- There are significant outliers in the "price" variable. In order to reduce the impact of outliers and increase the normalcy of the data, we will take the data between the 25th and 75th percentiles.
- After the data preprocess completion, we have identified the highest and lowest affecting features on the target variable.
- For overall picturization we plotted pie chart to know the ratio and relationship among the dataset.
- We check the neighborhood unique values eventually and plotted word cloud.
- So major features of our data set are reviews, bookings and locations which are mainly affecting the Airbnb price all over
- We visualized different types of attributes of data using different plots to know relationship between among these attributes.
- We reject hypothesis for room type Vs price since the p value is almost zero (and consequently less than alpha = 0.05). As a result, there is a difference in the price variation across the various accommodation classifications which is also demonstrated in the figure 3.
- We also see the relationship between price and different categories of room and conclude the relationship between them by hypothesis test.

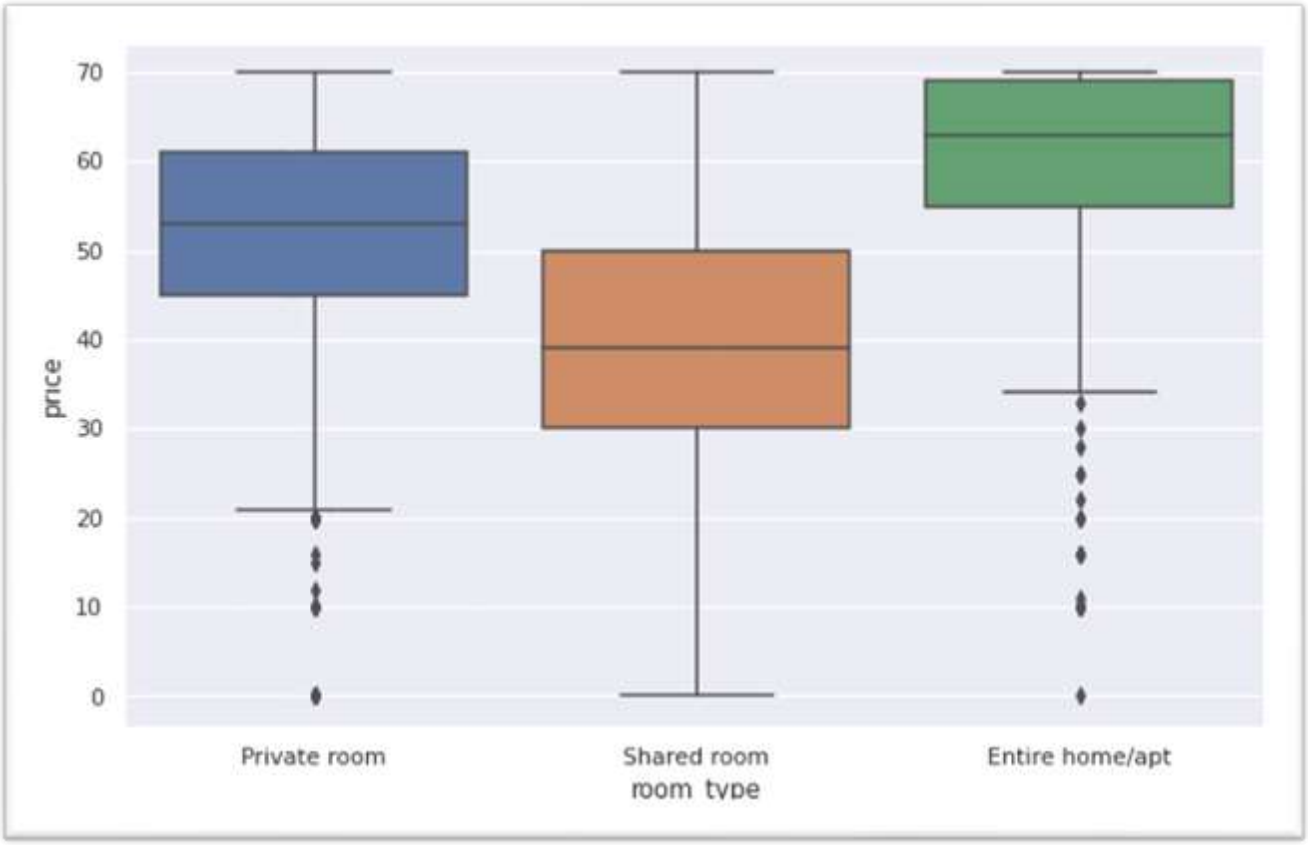


Figure 3: Room type Vs Price

- The categorical variable of "neighborhood group" has more than two categories, so, we use the One Way ANOVA test to know link between price and neighborhood group, and we get to know that the price is based on the neighborhood group that the house is offered in.
- We have performed chi-squared test for room types and neighborhood group and conclude that we reject null hypothesis. Since, we can say that there is relationship between the neighborhood group and room type and also shown as stacked bar plot as shown in figure 4.

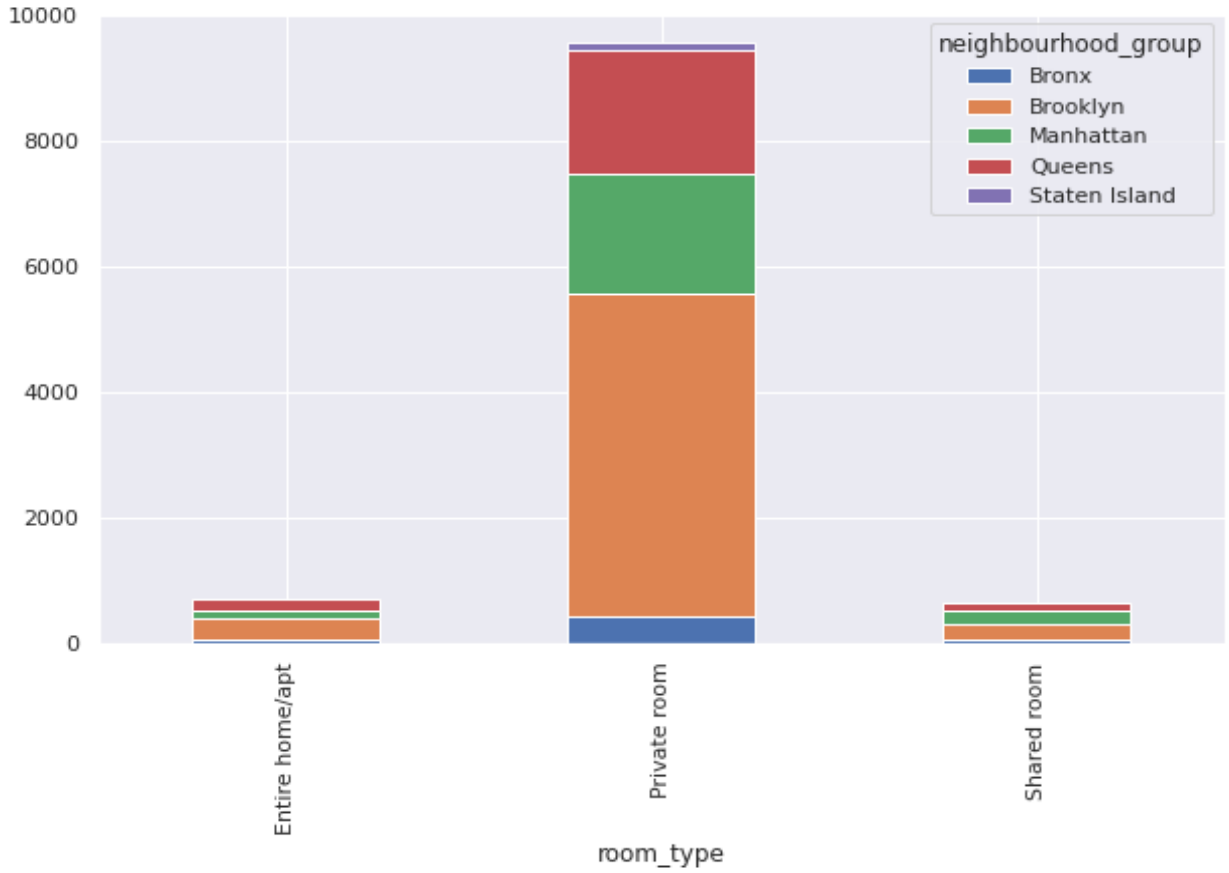


Figure 4: Room type Vs. Neighborhood Group.

- We have performed chi-squared test for neighborhood and neighborhood group and conclude that we reject null hypothesis. Hence, there is relationship between neighborhood and neighborhood group.
- For this type of data, we check the variance between the bookings of the different room types
- Here testing for variance is called liven test
- We performed T- Test on neighborhood group and availability_365 and get the P-value 0.663 which is above the 0.05. Since, we fail to reject the Hypothesis and hence, we accept the null hypothesis and conclude that there is not any relationship between them.
- Using Distpot we plotted the graph for number of available days in the whole area and here we measured the density of the data

- After all the statistical analysis on dataset, we did the Encoding for the categorical dataset.
- We did the encoding for the neighborhood group into ng_Brooklyn, ng_Manhattan, ng_Queens, ng_Staten Island, rt_Private room and rt_Shared room which will understand the machine in the period of training.
- We implement Machine learning models which are:

  Linear Regression - Here using this we validated our statistical model in this we find mean square error and root mean square error, below are the parameters we got.

  RMSE is: 0.8213420228839912

RMSE is: 0.8213420228839912

OLS Regression Results

Dep. Variable:	y	R-squared:	0.325
Model:	OLS	Adj. R-squared:	0.325
Method:	Least Squares	F-statistic:	5256.
Date:	Thu, 10 Nov 2022	Prob (F-statistic):	0.00
Time:	21:04:45	Log-Likelihood:	-13320.
No. Observations:	10899	AIC:	2.664e+04
Df Residuals:	10897	BIC:	2.666e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.802e-17	0.008	-7.37e-15	1.000	-0.015	0.015
x1	0.5704	0.008	72.500	0.000	0.555	0.586

Omnibus:	5379.693	Durbin-Watson:	1.198
Prob(Omnibus):	0.000	Jarque-Bera (JB):	38332.273
Skew:	2.278	Prob(JB):	0.00
Kurtosis:	10.978	Cond. No.	1.00

Figure 5: Linear Regression Results

Multiple Linear Regression - Here we got the same but for various independent variables which will affect the dependent outcome.

R-Square Value 0.13935718762733007

mean_absolute_error: 0.7604323398842785

mean_squared_error: 0.8615707446682286

root_mean_squared_error: 0.9282083519707354

and the RMSE score (across experiments): 0.8198990775340705

```
print("R squared scores:\n", scores1)
print("Average R squared score (across experiments):",scores1.mean())


print("RMSE scores:\n", scores3)
print("Average RMSE score (across experiments):",scores3.mean())
```

```
R squared scores:
 [-2.68983001  0.35656977  0.52437256  0.77114788  0.73031733  0.63570778
  0.46359052  0.20194511  0.00656353 -0.72625699]
Average R squared score (across experiments): 0.027412748488921234
RMSE scores:
 [1.21413378 0.5732473  0.48369005 0.42590084 0.50297081 0.58300332
 0.74908638 0.91220924 1.16354122 1.59120783]
Average RMSE score (across experiments): 0.8198990775340705
```

Figure 6: Multilinear Regression result

Preliminary Results:

As we know that Airbnb is the biggest online marketplace for the various price comparisons of each property, so here they will have the large data sets from the various companies so the properties against each location for the bookings is very difficult to predict. Hence, we have the best models to predict the prices or various parameters using empirical analysis of statistical methods which could bring up the desired results. So, at the analysis stage of project selection about Airbnb we came up with a thought to implement various ML models for the dataset to predict the prices as per the locations so that we can bring up the variation to the current existing models which will help Airbnb to provide the better services. As proposed, we are implementing all the statistical analysis of the data along with the performance metrics and visualized well comparatively. For the ML models we are able to predict the prices in the measurement of the RMSC value which is a bit above for the baseline models we got so we are aiming to bring up the best RMSC in the upcoming models in the coming increment which will lead to the better preliminary results.

Project Management:

Work completed:

We have completed the data analysis and validation of our data using statistical methods and implemented the comparison between the independent variables and dependent variables. Calculated the means, variances, deviation between each type of data which are affecting the outcome of the target variable. used the T-test, p-value and hypothesis testing values to evaluate the performance of models with the Airbnb Dataset. For implementation of statistical methods, we've also used linear and Multi linear regression.

Responsibility(Task, Person):

EDA, Statistical Analysis and Report Writing: Roshan Sah
Model Implementation, Model evaluation and Report Writing: Saisri Teja Pepeti
Data collection, Preprocessing and Cleaning, Report Writing: Yamuna Bollepalli

Contributions (members/percentage):

Roshan Shah - 33% Saisri

Teja Pepeti - 33%

Yamuna Bollepalli-34%

Work to be completed:

For the further increment we are going to implement a few more ML models to compare the statistical method values of the different models to select the best one to predict our Airbnb data. Also, to get a better RMSE score we will use further ML models to bring up the best price prediction. Along with this we are going to implement which is when the traffic data is more how the booking impacts this works on a conditional basis. After finding the relation between the data we use the different classification models like KNN, Decision Trees, SVM, Logistic Regression, Naive Bayes to improve the model's performance by reducing the variance and deviation and get the best accuracy score. For these models we also find AUC score and plot a confusion matrix

Responsibilities(Task,Person):

Model Implementation, Model evaluation and Report Writing: Roshan Sah
Model Implementation, Model evaluation and Report Writing: Saisri Teja Pepeti
Model Implementation, Model evaluation and Report Writing: Yamuna Bollepalli

References

[1] Weippl, Edgar, A. M. Tjoa, Peter Kieseberg, and Andreas Holzinger, eds. 2021. Machine Learning and Knowledge Extraction

[2] S. Yang, "Learning-based Airbnb Price Prediction Model," 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), 2021, pp. 283-288, doi: 10.1109/ECIT52743.2021.00068.

[3] J. Dhillon et al., "Analysis of Airbnb Prices using Machine Learning Techniques," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 0297-0303, doi: 10.1109/CCWC51732.2021.9376144.

[4] https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1979&context=etd_projects

[5] Predicting Movies' Box Office Result - A Large Scale Study Across Hollywood and Bollywood

[6] https://snap.stanford.edu/data/web-FineFoods.html

[7]https://journalofethics.ama-assn.org/article/challenge-understanding-health-care-costs-and-charges/2015-11

[8] https://assets-global.website-files.com/5ca95f7a3be192f65a7b4e4f/617afbf92fc4b3a49b9aa0a4_how-much-is-an-image-worth-research-paper.pdf

[9] https://medium.com/@daazene/airbnb-seattle-dataset-basic-exploratory-data-analysis-b2615b2a44ef

Recording Link : https://youtu.be/hcxo4-STNew