Shaoni Mukherjee.docx

by Shaoni Mukherjee

Submission date: 01-Feb-2021 04:56PM (UTC+0000)
Submission ID: 143949699
File name: Shaoni_Mukherjee.docx (3.59M)
Word count: 34276
Character count: 210105

AIRBNB PRICE PREDICTION USING SUPERVISED MACHINE LEARNING

SHAONI MUKHERJEE
LJMUDS00664
M.Sc. DATA SCIENCE

Thesis Report
Liverpool John Moores University – Master of Science in Data Science

FEBRUARY 2021

TABLE OF CONTENTS

ACKNOWLEDGEMENTS v
ABSTRACT vi
LIST OF ABBREVIATIONS viii
LIST OF FIGURES AND PLOTS ix
LIST OF TABLES xi

CHAPTER 1: INTRODUCTION 1
1.1 Background of the Study 2
1.2 Problem Statement 4
1.3 Aim and Objectives 6
1.4 Research Questions 7
1.5 Research Hypothesis 8
1.6 Scope and Limitations of the Study 8
1.7 Significance of the Study 9
1.8 Structure of the Study 9

CHAPTER 2: LITERATURE REVIEW 12
2.1 Introduction 12
2.2 Linear Regression 15
2.3 Gradient Descent 17
2.4 Assumptions of Simple Linear Regression 18
2.5 Multiple Linear Regression 19
2.6 Feature Selection and Feature Engineering in Airbnb Data 20
2.7 Bias Varice Tradeoff 20
2.8 Regularization 22
2.8.1 Ridge 22
2.8.2 LASSO 23
2.9 Bagging and Boosting 23
2.9.1 Bagging 24
2.9.2 Random Forest and Decision Trees 24
2.9.3 Boosting 25
2.11 KNN for Regression 27

2.12 Performance Metrics 28
2.13 Neural Networks in Regression Problems 30
2.14 Hedonic Regression 31
2.15 Literature Review Summary 31
2.16 Discussion 32

CHAPTER 3: RESEARCH METHODOLOGY 34
3.1 Introduction 34
3.2 Research Strategy 35
3.3 Research Approach 35
3.4 Data Selection 36
3.5 Data Collection, Methods, Data Types and Tools 36
3.5.1 Python 37
3.5.2 Seaborn and Matplotlib for visualization 37
3.6 Data Analysis 37
3.7 Data Pre-processing and Transformation 38
3.8 Data Vizualization 39
3.9 Data Mining 40
3.11 Research Design and Methods 40
3.12 Summary 43

CHAPTER 4: ANALYSIS 45
4.1 Introduction 45
4.2 Dataset Description 47
4.3 Data Preparation 52
4.3.1 Data Elimination 52
4.3.2 Data Transformation into Categorical Variables 52
4.3.3 Identification of Missing Values 53
4.3.4 Univariate Analysis 53
4.3.5 Splitting of Original Dataset 60
4.4 Exploratory Data Analysis 60
4.4.1 Correlation Plot 61
4.4.2 Multiple Plots to Understand the Relationship 61
4.4.3 Room Type Distribution 63
4.4.4 Number of Reviews analysis using Bar plots 65

4.5 Natural Language Processing 66
4.6 Summary 69
CHAPTER 5: RESULTS AND DISCUSSIONS 71
5.1 Introduction 71
5.2 Simple Linear Regression 73
5.2.1 Simple Linear Regression using Stats Model 73
5.3 Predictive Modeling using Different Techniques. 76
5.4 Artificial Neural Network 81
5.4.1 ReLu Activation Function 82
5.4.2 Linear Activation Function 83
5.4.3 ADAM Optimizer 86
5.5 KNN for Price Prediction 87
5.6 Comparison with the 2019 Data Set 88
5.7 Discussions. 92
5.8 Summary 94
CHAPTER 6: CONCLUSIONS AND RECCOMENDATIONS 96
6.1 Introduction 96
6.2 Discussions and Conclusions 96
6.3 Contribution to Knowledge 97
6.4 Future Reccomendations. 98

REFERENCES 99
APPENDIX A: RESEARCH PROPOSAL 103
APPENDIX B: RESEARCH CODE LINK 123
APPENDIX C: RESEARCH STRUCTURE 124

ACKNOWLEDGEMENTS

I would like to take a chance to thank everyone who has helped me and supported me to reach this stage of the study and in producing this report.

I would also like to thank my supervisor Ankit Bhatia whose constant guidance and support throughout the project has helped me to successfully complete the report on time.

I would also like to thank Professor Manoj for all his time and advise on each and every detail of report writing.

I cannot forget to thank my parents who encouraged me to pursue Master's degree from Liverpool John Moores University and supported me with all of the financial needs.

I thank my husband who never stopped supporting me with every encouraging words and feelings.

Without all of this help, this report would not have been the way it is and the completion of the project would have not achieved.

ABSTRACT

The study aims to presents the report on the Masters research on "Predicting the price of Airbnb rental properties in NYC using different Machine Learning technique." In this research, we chose one of the best peer to peer property sharing leaders, Airbnb. Since its founding, the company has been growing and has shaped the hospitality industry. In this study we chose New York City for the analysis.

The dataset has helped to generate few important hypothesis which has been answered in the subsequent chapter by further investigation. The project uses the data set of New York City Airbnb 2020 available in insideairbnb.com. The 2019 dataset on Airbnb has been used to make a comparison with the 2020 data.

The collaborative economy where people share goods, resources, etc., through online platform is termed as collaborative economy. This economy has entered the travel and hospitality industry and are serving peer to peer services such as renting their personnel apartments to tourists. One of the successful platform which offers such services in Airbnb. Airbnb was founded in 2008 and provides one of the best and relevant accommodation services. The literature review attempts to provide the background and derive useful information from the previous work.

Determining the price of an Airbnb listing is a challenging task, hence in this paper we aim to develop a ML model which can predict the prices of the listings using the independent variables. Also, to find the determinants of the listing prices in New York. The goal of this paper is to help the hosts of these properties to determine the optimal price for their listings. The paper aims to extract as much useful information from the available data to build a better ML model. Due to the non-linearity in the dataset, there is a need to study ANN as an alternative method. By comparing different ML model's prediction performance with ANN, this study aims to find the best one.

Boosting is a method which improves the model performance. Here, we have also experimented with different boosting algorithm and find out the performance increases. Further, we have also provided importance to feature engineering in this study to build our ML model.

We have used several ML algorithms and methods and compare their performance. The ML models used here for the analysis includes Linear Regression, Random Forest, XGBoost, KNN, and ANN. The algorithms are quite proven to be best for regression analysis.

With Python as the programming language and different libraries provided as the data visualization tool, different methods are to be applied over the data set. The results are validated using different performance measures which are discussed in detail in the later sections.

Currently, the following chapters discusses the problem statement, the hypothesis, we have defined the research question with the aim and the objectives of the study. The literature review section has been discussed with the all the notable works which has been done previously in the topic.

The later study will include the detailed analysis and the results derived and findings from the study.

LIST OF ABBREVIATIONS

ANN….. Artificial Neural Network

D.T…..... Decision Tree

EDA…... Exploratory Data Analysis

I.Q.R….. Interquartile Range

KNN….. K Nearest Neighbour

LASSO…... Least Absolute Shrinkage and Selection Operator

LightGBM.... Light Gradient Boosting Machine

MAE….. Mean Absolute Error

ML…... Machine Learning

MLR….. Multiple Linear Regression

MSE…... Mean Squared Error

NYC……...New York City

NLP…... Natural Language Processing

OLS…... Ordinary Least Square

QR….. Quantile Regression

RMSE…... Root Mean Square Error

ReLU…... Rectified Linear Unit

S.D.…... Standard Deviation

SVM….. Support Vector Machine

TF-IDF…... term frequency–inverse document frequency

V.I.F.....Variance Inflation Factor

XG…...Extreme Gradient

LIST OF FIGURES AND PLOTS

Figure 1.2. Homoscadasticity and Heteroscadasticity 5
Figure 2.2. Linear Regression 15
Figure 2.3. Gradient Descent 17
Figure 2.7. Polynomial vs Straight Line Fit 21
Figure 2.7.2. Model Complexity 21
Figure 2.9.2. Decision Tree Representation 25
Figure 2.11. KNN 27
Figure 3.7. Box Plot 38
Figure 3.11. Correlation Plot 41
Figure 3.11.2. Leaf Wise Tree Growth 42
Figure 4.2.1. Missing Value figure 49
Figure 4.2.2. Target variable distribution 50
Figure 4.2.3. Correlation Plot 51
Figure 4.2.4. Longitude vs Latitude 51
Figure 4.3.4. Univariate Analysis 53
Figure 4.3.4.3. BoxPlot Distribution of Price with outliers 54
Figure 4.3.4.4. BoxPlot Distribution of Price without outliers 55
Figure 4.3.4.5. Q-Q Plot 55
Figure 4.3.4.6. Q-Q Plot 2 56
Figure 4.3.4.7. Neighbourhoods Plot 57
Figure 4.3.4.8. Neighbourhoods group Plot 57
Figure 4.3.4.9. Count Plot of Neighbourhood 58
Figure 4.3.4.11. Box Plot Distributions of Variables 59
Figure 4.3.4.12. Top 20 Most Common words 60
Figure 4.4.2.1. Longitude vs Latitude showing the neighbourhoods 61
Figure 4.4.2.2. Multivariate Plot 62
Figure 4.4.3.1. Longitude vs Latitude showing the room type 64
Figure 4.4.3.2. Price Distribution 64
Figure 4.4.4.1. Number of Reviews per Hosts 65
Figure 4.5.1. Most Frequent Words 66
Figure 4.5.2. Bigrams and Trigrams 66
Figure 4.5.4. Most Frequent Words for Expensive properties 67

Figure 4.5.5. Bigrams and Trigrams for Expensive properties 68
Figure 4.4.7. Word Cloud Representation 68
Figure 5.2.1.1. OLS Results 74
Figure 5.2.1.2. VIF Table 75
Figure 5.2.1.3. VIF Table 2 75
Figure 5.3.1. Real Vs Predicted Values 80
Figure 5.3.2. Feature Importance 79
Figure 5.3.3. Feature Importance Weights 79
Figure 5.4.1. Deep Neural Network 81
Figure 5.4.1.1. ReLu Activation Function 82
Figure 5.4.1.2. Linear Activation Function 83
Figure 5.4.2.1. Predicted Vs Actual Plot 84
Figure 5.4.2.2. Predicted Vs Actual Plot 2 85
Figure 5.5.1. Elbow Curve 87
Figure 5.6.1. Price Description for Price 2019 Dataset 88
Figure 5.6.2. Price Distribution for 2019 and 2020 89
Figure 5.6.3. Room Type Count Plot 90
Figure 5.6.2. Actual Vs Predicted 91
Figure 5.7.1. Confusion Matrix 93

LIST OF TABLES

Table 4.2.1. Variable Description 48
Table 4.2.2. Missing Value 48
Table 4.2.3. Correlation of Variables 50
Table 4.3.2.1. Variable Transformation 53
Table 4.3.4.1 Variables before PowerTransformation 56
Table 4.4.3.1. Room Type Distribution 63
Table 5.2.1. Metric Scores for SLR 73
Table 5.2.1.1. Variables used for SLR 72
Table 5.3.1. Skewness Table 76
Table 5.3.3. Lambda Values 77
Table 5.3.6. Model Performance for Different Models 78
Table 5.4.1. ANN 80
Table 5.4.2.1. ANN 2 84
Table 5.6.2. Model Results 90
Table 5.6.3. Model Results 2 91

CHAPTER 1

INTRODUCTION

Airbnb has helped shaped the hospitality industry since 2008. People travelling around the world are now preferring sharing platforms rather than traditional hotel, holiday homes, motels, etc., Currently the company has 3 million listings in 191 countries. These sharing platforms allows locals to share their property with the tourists at an affordable price. Airbnb does not buy any of the properties listed, but the organization receives benefit from earning a fee on each booking. From a start-up, Airbnb has experienced a rapid growth and is one of the global player.

This report discusses the different methods to be applied to the problem and compare the results. This report also provides the time plan for the study. The literature review section offers a comprehensive background on what has already been achieved in this field. Further, it discusses the different methods which are incorporated till date and the result of the analysis.

This research will try to investigate the variables that has a highest impact on the price variables and will try to find the best ML model to predict the price of the properties by applying different ML techniques.

There are 15 independent variables in the dataset which includes both character and numeric. The variable such as longitude and latitude can be further used in the analysis to find out more about the geographical locations. There are a few variables such as date, name of the properties, host ID which cannot be directly used to build the ML model and needs to be transformed using feature engineering techniques. The pre-processing steps are discussed in the later sections and further will be analysed during the study. The simple ML model to be developed would be the Multivariate Linear Regression Model to assess the relationship between dependent and independent variables. This is because regression analysis helps to describe the changes in each independent variable that effects the change in the dependent variable. Further, the model will be verified through the regression coefficient and the p values associated with each variable. The study of the coefficient represents the average change in the dependent variable if there is a unit change in the independent variable while the other variables are constant.

This research will try to analyse the characteristics of Airbnb listings in New York City that will help predict the price of the rental properties. A set of ML algorithms are selected to answer the research question discussed in section 1.4. The chapter1 to chapter 3 has been entirely dedicated to formulate the base of the research and how to proceed further with the analysis. Thus, the research methodology section discusses the pre-processing steps of the dataset. The details of the dataset are provided in section 3.4.

The detailed research objectives have been discussed in section 1.3. This research aims to extract as much as possible information from the Airbnb dataset provided to achieve the objectives. The chapter 4 discusses the detailed report on implementation and results obtained during the entire study.

Amongst the ML techniques which are to be implemented for model building are: Linear regression; This technique fits a straight line which tries to minimizes the error. KNN; which tries to predict the values of the new point using 'feature similarity.' The algorithm assigns the value based on its resemblance to its nearby points. Random Forest; an ensemble learning technique where the base model is the decision tree. Random Forest is a parallel learning technique where multiple DT are considered and the output is the mean of all the outputs. ANN; A neural network which predicts the output variable as the function of the inputs. Ridge and Lasso; The regularization technique which are be used to reduce model complexity and find the right balance between bias and variance.

The boosting algorithms to boost the model performance and find the best one. This mentioned methods are chosen to build the model. However, there are also few other methods which are not been covered in this study due to time complexity.

Each ML model used are to provide value to the research by analysing the working in depth and by comparing the performance results of each model. The results are be used to test the research hypothesis which are discussed in section 1.5.

Python has been selected as the programming language for the implementation and as the platform for this research study. This is because the tool is very versatile, easy to use and has all the libraries required for the analysis. Coding in python is easy and the language requires fewer code to analyse complex problems. The data analysis in python can be done with ease due to its processing capabilities and availability of powerful packages.

ML techniques, mathematical functions, statistical operations, and visualization plots are some examples can be handled very easily using python. Thus, this language has gained popularity among researchers.

1.1 Background of the Study

Every ten of thousand people, every night decide not to stay in traditional accommodations available such as hotels and guest house but are more inclined towards sharing platforms such as Airbnb. Airbnb allows locals to share their residents to tourists. This has been a major trend shaping the global tourism industry (Guttentag, 2019). Airbnb has been changing the trends in hospitality and tourism service around the world and currently has three million listings in 81,000 cities and 191 countries (Gamboa et al., 2020). This rise of phenomena is notable, as the sharing platforms for accommodation makes up to 7% of the global accommodation supply in 2018 and the rise is expected to be 31% between the year 2013 and 2025 (Bakker et al., 2018). Though sharing platforms were present earlier too but due to rise of internet platforms, mobile phones and change in technology has revolutionized this practise on a large scale (Gutiérrez et al., 2017).

As per New York times Airbnb has been amongst the biggest player in the short-term rental market. Over the years its rapid growth has led to turn residential neighbourhoods into tourist zones despite the recent virus outbreak. Airbnb has grown from a small start-up and has become a global player. Airbnb, by providing the access to rent millions of spaces which includes apartments, villas, castles, treehouses, etc., and has more than 400 million guest arrival recorded in the year 2018. According to Forbes the industry was valued more than world's largest hotel chains such as Hilton, Marriott, Hyatt (Dudás et al., 2020).

Price and low cost are reported as one of the important factors facilitating the growth of this industry globally. As price is one of the most important factors by tourists hence is one of the most well researched topic (Gibbs et al., 2017). As the demand for sharing rental accommodations have increased, pricing has become a significant issue and understanding the Airbnb prices has become crucial for practical and theoretical purpose. Research conducted by Wang and Nicolau in the year 2017 studied the price determinants of Airbnb listings and categorized the factors into different categories. The prices of hotels, hostels and motels are expensive in New York City which makes Airbnb a good option for budget tourists. However, one of

the biggest challenges faced by the Airbnb is the pricing of the listings. Airbnb losses around 46% of their revenue due to inefficient pricing (Gibbs et al., 2018). This loss is one of the consequences of hosts not pricing the property appropriately.

The area selected for this study is New York City. We have chosen this location due to a number of reason. New York is one of the most visited cities in United States hence sharing properties is one of the most common option for tourist who are budget travelling. Further, these rental property prices depend on different factors. Airbnb in the NYC has been extremely active since August 2019 (Atta-Fynn and Zien, 2019).

This research analyses the features of Airbnb properties in New York City that impacts the price. A number of study have been done to investigate the determinants of price however, only a few papers have explored the factors the affects the price of the sharing property especially Airbnb. Hence, the main aim of this research is to find the best ML algorithm to predict the price and highlight the determining factors to predict it.

Multiple ML algorithms have already been used such as hedonic regression, linear regression, decision trees, etc., to conduct this research. Hence there is a scope to further explore these algorithms and find the best one.

The rest of the paper is organized as follows: the next section describes the problem statement, aim and objectives and research related questions along with the scope and limitations of the study. The very next section is the literature review section which reviews the prior studies which is already done in this area. Then, the paper presents the research methodology and describes the data pre-processing and EDA steps. This section provides the research design and the framework. The last section which concludes the interim report is the references section. The next sections are focussed on the implementation and the results obtained from the implementation. Post that we have discussed the limitations and contribution of this study to the literature. We have also discussed the future topics which needs further attention in the last chapter.

1.2 Problem Statement

Pricing an Airbnb listing is still a challenging task for the host as there is a need to consider a number of features and amenities considering the amount of competition in the market. Also, the pricing of the listings is very dynamic and the company is growing at a very fast pace since 2008 (GAMBOA FUENTES, J.E., et al., 2020). Hence there will always be a need to analyse the pricing because of the huge data

which is being generated and the prevailing competition in the market. Airbnb losses 46% of revenues due to inefficient pricing of the listings (Gibbs et al., 2018).

New York being one of the most visited cities in USA and having received the eighth consecutive annual record of approximately 67 million tourists in 2019. Thus, making Airbnb a good option for people who are in budget. This research will help analyse the features which can be a better predictor for the "price".

Also, this paper will help to understand if there are any relation with the property name and the "price" of the property. As we know that words can be a good indicator when it come to the cost of any product. We can use NLP to understand if a particular word has any influence in the "price" variable. To understand the relation, we have used bigrams and trigrams approach to find the most frequently used words for the expensive properties.

Most of the paper has also not captured the non-linearity of the dataset well and has thus resulted in unsatisfying RMSE and MAE values.

We have labelled the data as 1's and 0's for the expensive and non-expensive properties to classify the properties using only the title of the property. We have used LigthGBM classifier to predict the classes.

LightGBM, is a gradient boosting algorithm which can handle large amount of data, uses less memory, computationally less expensive, efficient. The algorithm grows tree leaf wise.

One of the main assumptions of a linear regression is homoscedasticity. The variance should not be increasing or decreasing as the error value changes. The variance should not be following a pattern.

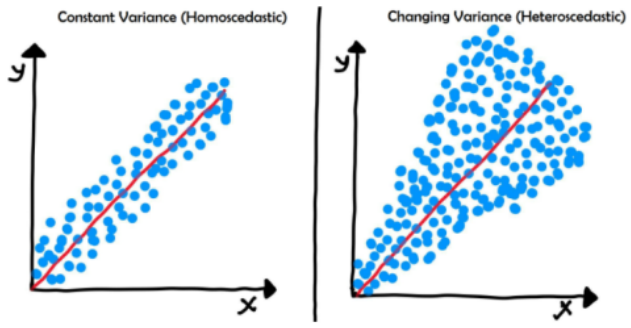


Figure 1.2 (Homoscedasticity and Heteroscedasticity, Upgrad, 2018)

This assumption is often unnoticed and not addressed in the previous work. Heteroscedasticity refers to unequal spread of the error or the residuals over the range of measured values. In OLS regression residuals should not display a pattern which is often cone shaped as displayed in the picture. Heteroscedasticity often results when one or more predictors are not normally distributed or even the target variable is not normally distributed. Here, clearly the target variable "price" is not normally distributed. Also, in order to make any further interpretations the error terms should follow a normal distribution. A very recent method has proven very effective to correct this problem is by using PowerTransformer. The objective of the algorithm is transforming the variable to normal distribution. This module consists of both the box-cox transformation and the Yeo-Johnson transformation.

1.3 Aim and Objectives

The main aim of this research is to develop a "price" predicting model and find the optimal number of features which can be used to create the model. The research aims to use different Machine Learning technique to find the best performing model with optimal number of features. Further, to investigate if there is any relationship between the variable "name of the listings" with the target variable that is "price". The main focus of the study is on feature selection and feature engineering using various ML methods. Also, the analysis aims to find out how a different boosting algorithm works for this dataset and to compare the results with ANN.

Through this paper we hope to achieve a "price" predicting model which can explain the input features well and predict the "price" of the listings accurately. The model is expected to be simple, robust and a generic model which can perform well on unseen data.

Also, in this paper we have focused on using NLP techniques to find the most frequent bigrams and trigrams used to describe the listings and if there exists any correlation with the "price" variable. With the use of NLP, we will aim to understand whether any particular words used by the hosts which indicates the pricing of the properties. The expected results should provide a deeper understanding about the data and the independent variables that influence the target variable "price". Further, to verify the model performance we aim to use MAE, RMSE, MAE, R^2, and adjusted R^2. These performance metrics helps to evaluate the model performance and makes it

easy to present the model to other people or to stake holders. The detailed discussion of these techniques is provided in the literature review section.

In this paper we also to aim to evaluate the model by the Median absolute error. The advantage of using this metric is it is more robust to outliers. Hence large deviations can easily be captured using this metric. We have performed extensive descriptive and exploratory analysis on the data in order to understand the data well and derive useful insights from it.

The main objectives of this research paper are: -

1. This research aims to extract as much feasible information from the Airbnb 2020 data set provided and in order to do that the features or the independent variables are to be examined which are strong predictor of the "price" variable.
2. We will compare the results of different ML algorithm and find the best based on the performance measures as discussed in section 1.3. In order to achieve the best performing model, we will focus on hyperparameter tuning, feature engineering and feature selection.
3. To further analyse the "name" of the listings using NLP techniques to find the correlation between the target variable. In this research our objective is to find the most frequent words used in the name of the properties.

1.4 Research Questions

As a general research question, the Airbnb dataset are to be analysed by using different Machine learning technique, feature engineering, feature selection, data transformation, data visualization to extract as much information as possible. The expected results should give deeper understanding about the target variable and the independent variable that has a better influence on it.

The data set consists of sixteen variables and out of which "price" is the dependent variable. The independent variables are both numerical and string. Some of the research question that needs to be answered in this research are as follows: -

1. If there exist a relationship between the name of the listings and the "price" of the listings?
2. To understand the most frequent words which are used for the properties with high "price"?

3. How well the model performs if we use LightGBM classifier to predict the classes for expensive and non-expensive properties?
4. What the distribution of the error signifies?
5. How the outcome of this research is compared to the outcome of the other's previous works and how the model performs in terms of aforementioned questions?

Some initial questions regarding the data set which will be analysed during the EDA phase are as follows: -

1. How the "price" variable is distributed?
2. Is feature transformation possible using the existing variables?
3. Which Neighbourhood is the most expensive? What areas are more popular than other?
4. What are the top used words for listings names? And if there exists a trend on naming the listings?

Descriptive Analysis is the first analysis performed in quantitative study and the objective includes summarizing the dataset and exploring the behaviour of the data. Using statistical techniques such as measures of central tendency, frequency distribution, measures of dispersion we can understand better and understand the data well for further study. The above questions are to be revised and modified once the data set starts to be analysed.

1.5 Research Hypothesis

The research hypothesis which will be tested in this paper to test the relationships are provided as follows: -

1. The "room type" has an impact on the target variable.
2. The price of the listings depends on the neighbourhood of the property.
3. The variable name of the listings has influence on the dependent variable. There is a trend on naming the expensive property.
4. Less number of reviews are more likely to have low property pricing.

1.6 Scope and Limitations of the Study

This study focusses on building a ML model to accurately predict the price of the listings on Airbnb. The dataset used for the study is NYC 2020 data collected from the Airbnb official site. Further, we have used NYC 2019 dataset to compare the data with the 2020 data. The main aim of the study is to understand the features which have a high impact on the price variable. The model will help the hosts of the properties in determining the optimal rental price. This is one of the main challenges faced by the Airbnb hosts. Once the model is deployed it will allow the property owners to freely find out the price of their properties based on the features such as geographical locations, room type, neighbourhood, etc.,

The study will help us understand each and every variable in detail and there impacts. Further, different ML models are being used to find out the best one and to compare them. In addition, we have also used NLP to process the textual data which will help us analyse the name of the property variable. Further, the study is focussed on NYC which is one of the most famous and busiest location. Also, the duration of the study from data collection to submission of the paper is of nine months.

The techniques which are discussed in details includes Linear regression, Random Forest, Decision Trees, KNN, ANN and certain boosting algorithms such as Gradient descent, XGBoost. However, there are other techniques too which can be used to predict the price but due to time and resources constraints are not included.

The data collected in based on the year 2020 and 2019 and the location considered in this study is New York City. This study does not cover the locations apart from NYC were Airbnb properties are available. The locations which are part of Airbnb properties are not within the scope of this research. The study is also done through understanding the variables which are part of this dataset and has not included variables which might have an impact but is not included in the dataset.

The tools to be used for the study are python as the programming language. Different visualization libraries are used for Data Analysis, Visualization.

The proposed study will help to determine the optimum price for the properties and understand the effects of the variables on the target variable.

1.7 Significance of the Study

The research will be beneficial to the property owners who wants to put up their properties on online platforms for rents. The proposed method can also help us

understand which are the significant variables for pricing. The system can assist the owners to determine the price of their properties however, still decision making will be still as per the owners choice.

Also, the methods used in the research will let us compare the different Machine learning model used by their different performance measures. The different methods to be used have proven to build ML model which are highly accurate. The research will also help us understand the variables well which are provided in the dataset.

This research will also be beneficial to Airbnb industry to understand the main drivers of the property price. By knowing how the prices will change in the future can help provide a competitive edge over the competitors.

This research would be valuable to the future researchers as they can get some useful information that might be needed for their research and some possible questions might be answered using this study. The different methods used here can help understand why certain algorithms are better than the others which will help future researchers choose certain algorithms.

1.8 Structure of the Study

This first three chapters of the report aims to document what has already been done so far in this study and plan for the what has to be done later. This section will cover the designing of the three chapters which are included in the interim report. This report includes the details of the work which has been completed till date. The overview of the contents of each chapter are presented in this report:

Chapter one introduces to the problem and provides an overview of the study and describes the need to build a ML model which will predict the prices of the Airbnb property. Further, the section also describes the aim and objectives of the study. This section also describes the research questions which the study aims to answer in the study. This chapter further discusses the scope and significance of the study.

Chapter two provides the literature review of the of this study which is the previous related works that has already been done in this area. Furthermore, this chapter includes different ML approaches already done in this project and some of them are also to be included in this study. A set of ML models such as KNN, linear regression, random forest, and ANN which are reviewed in the literature review segment. These are the algorithms which have been used extensively used so far in this study for the analysis of the price of the rental properties. Further, the

performance measures are also discussed briefly in this section. The introduction section provides the overview of the techniques which have already been performed. This is one of the most important section in the study as it includes the background of the research and provides a base to the study.

Although this section requires further some additional studies to be included which have not been touched upon yet and are expected to be added as we progress on the research.

Chapter 3 includes the research methodology section which provides a detail of the methods to be included in this study. This section also includes the methodology which has been followed so far. Also, it presents the details of the method which are to be included and has not been done so far. This section includes the research strategy and approach along with the data collection and the tools used. Further, the data pre-processing and transformation techniques are also included.

The section research design discusses the details how the study is designed and the details of the methods to use to build the ML model. This section discusses the new proposed method and does not include the methods which have already been discussed in the literature review section.

The appendices section includes the research plan and the research proposal of the study.

Chapter 4 has been dedicated to the analysis part where we have included the complete data analysis and included each and every step of exploratory data analysis. We have included the figures and plots to describe the attributes and explain the information gathered from the plots. The chapter is useful to understand the relationships with each and every columns and with the target variables. This is useful for model building and provides the base for predictive analysis. The correlation plots are helpful to understand if there exist any multicollinearity in the data. The scatter plots are helpful to understand bivariate relationship. The boxplots are useful to find outliers and find the IQR ranges. Further, the multivariate relation are also discussed in details with each and every step.

Chapter 5 discusses the details of predictive Modeling and model which performed well for this dataset. The results are provided in this section to understand and compare the models. As discussed earlier the performance scores of each model are provided in this section.

Lastly, we have included conclusion and recommendations in chapter 6. Whatever we obtained from the study we have included in this chapter and the future scopes in the study. The chapter also discusses certain areas which have not been covered.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Data Science is one of the hottest fields of research and study because of its strong effect on any domain. It is a blend of Mathematics, Business knowledge, Algorithms, Statistics, Tools, and Machine Learning techniques to derive insights and find patterns from raw data. Today our world generates 2.5 quintillion bytes per person, per day of data. To process this huge amount of data more advancement is being carried out in this field. Data extracting, compiling, processing, analysing, visualizing includes strong knowledge in the area. According to the survey conducted in 2018 by KPMG Global PropTech, concluded that 49% of the participants thought AI, Big Data, and Data Science were technologies that will have the biggest impact on real estate industry on the long run.

Predicting has always been very attractive to human beings and also has significant effects on future. “price” prediction is a regression or a classification, for regression actual “price” is considered however for classification task “price” range is formed. Regardless of being “price” used as a regression or classification the prediction is made using the input features (Zehtab-Salmasi et al., 2020).

In addition, prices for listings often fluctuate according to seasons, weekdays and holidays (Gibbs et al., 2018). The value of position was calculated in the research paper by comparing the general linear model (GLM) and the geographically weighted regression (GWR) model. With a high R-squared value, the GWR model proved effective (Zhang et al., 2017).

Airbnb offers complete independence to the host of the property to fix the price of the property. Few of the challenges which are faced: - are the huge amount of data that is being generated and processed. The requirement of specific skills to process the data and study it. Especially the EDA which involves data cleaning, data processing, extracting correct information from the data. Further, there are a lot of competition in the tourism industry and even a small difference can make a huge impact.

In the case of tourisms sector, as mentioned by Yu and Schwartz in the year 2006 and Torra and Claveria in the year (2015), an intense development in research techniques and a growing interest in AI has been marked over the last decade to increase profitability and sustainability.

Analysis conducted by (Fei, Yue. et al., 2020) used data from California 2020 to get the general pattern of market rental prices to have some fair rental price feedback and later the host can make changes as appropriate. To construct the recommendation framework, ML models such as Ridge Regression, Ridge Regression with K means, Random Forest, SVM, and XGBoost were used to propose "price" based on the information given by the hosts. As anticipated, XGBoost performed well on the training data with the R square equal to .8611 and on the test data with .7304. Random Forest performed best in the training set with R square of .9346.

In the Airbnb data for Copenhagen, study undertaken by (Group et al., 2020) conducted graphical and statistical analysis. The results revealed how the target variable has a major influence on multiple variables such as "room type", "neighborhood", "accommodations", "bathrooms", "bedrooms" and "minstay". In order to recognize the most common terms used in positive feedback and for negative reviews, the same data text analysis was also conducted. In order to find the best performing model with all the important characteristics, five different linear regression models were developed and evaluated. The model calculated the premiums to be smaller than the price of the website. The tourists' remarks on the websites were used by (Ye et al., 2009) to perform opinion analysis on seven destinations: Los Angeles, New York, Las Vegas, Rome, Paris, London, and Venice. NLP was used to translate guest remarks, user interaction, host details, as defined by (Laurent et al., 2015), to gain more details to refine the goal variable.

Recent research proposed by Jorge Enrique Gamboa Fuentes used three models to test the hypothesis that the variable neighbourhood has the most effect on the nightly "price" forecast of an Airbnb rental in NYC, while variables such as restrooms, lodging, room type have less impact on the dependent variable. Linear regression, KNN, and ANN were the ML models used in this study. The results of the Liner Regression model underpinned the presumption that the neighbourhood had the highest price effect. The KNN model has verified that the prediction of the neighbourhood is

homogeneous for all listings. Moreover, the "price" also depends on the neighbourhood. ANN was used to find out the behaviour of other independent variables (Fuentes, Jorge. Et al., 2020).

A research (Kalehbasti et al., 2019) revealed a predictive model using ML, Deep Learning, and NLP techniques. The primary aim of the study was to support both owners/hosts and consumers. A broad variety of approaches such as linear regression, tree-based model, k-means clustering, SVR, and ANN were used to construct the model. In order to test model efficiency, the MAE, MSE, R square score for each model was compared. The model's scores were identical. However, the SVR with the RBF kernel surpassed the R2 score by 69% and the MSE by 0.147 and the best R square score.

Analysis conducted by (Luo et al., 2019) was conducted using a number of regression model methods, including linear regression, nearest neighbour regression, random forest regression, XGBoost, ANN, to estimate the target variable. The baseline model was developed using Linear regression with L2 regularization and K-nearest neighbour (KNN) regression. A completely linked deep neural network was developed using 3 multilayer perceptron's with the ReLU Activation feature. The model performance metrics used were the R2 and MSE ratings. The score achieved by XGBoost and Neural Network ranged from 66% to 67%. It was also observed that the logarithmic transformation of the target variable improved model efficiency. Interaction of continuous variables also showed further improvements in model efficiency. Unigram and Bigram tf-idf have been used to represent the functions. This method was also very recent and beneficial for market estimation and achieved an R2 score of 74%. The date attribute was uncorrelated to the price, so it was later scrapped. The best-performing model was XGBoost and the neural network and received a score of more than 70%.

In an experiment performed by (Lewis et al.,2019), Airbnb estimated a rental price for properties in London using machine learning and deep learning. The best model that outperformed the other model was XGBoost with an accuracy of 73 %, which was much higher than the neural network.

Study performed by (Keating, 2018) to predict the price of Airbnb rentals in Seattle. The key purpose of the analysis was to establish the relationship between the

independent variables and the dependent variables. Here, the basic model was the OLS modeling technique. Later methods, such as nearest k neighbours, random forest and neural networks, were used and model efficiency was evaluated. Here, the most efficient model was a multi-layer neural perceptron net. The distribution of error of each model against the test data revealed that the neural network was more accurate.

Research conducted by (Choudhary et al., 2018) was conducted to analyse the Airbnb listings in the city of San Francisco. The main study, aimed at better understanding attributes such as bedrooms, location, house type, may be helpful in predicting the price accurately. The price can be profitable for both the hosts and the guests. Additional research was carried out to assess the probability of existence of a listing for the guests to consider when making a reservation for the house.

Similar research conducted by (Varma et al., 2018) was based on predicting house prices using ML and the Neural Network. The model was used to analyse the set of parameters chosen by the customers according to their interest. The model was constructed using classical linear regression, forest regression and boosted regression to predict the price. Further Neural Network was designed to increase accuracy. The model helped to establish the strength of the relationship between dependent and independent variables.

Most of the papers used Linear Regression as the base model, section 2.2 describes Linear Regression briefly.

2.2 Linear Regression

The dependent variable is “price” which is numerical and continuous in the Airbnb dataset. The Independent variable includes categorical and numerical variable. Linear regression is one of the best approaches to build the model. Linear regression tries to fit a straight line and model the relationship between two variables. One variable is the explanatory variable while the other is the dependent variable.

The most important concepts behind Linear Regression are first it uses a least square line to fit the data, second the R square calculation, and third to calculate the p value for the R square. The first step is drawing a line through the data next step includes

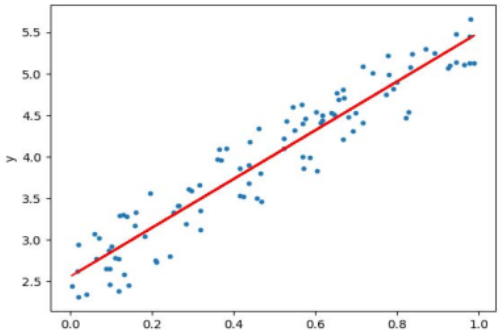


Figure 2.2, Linear Regression, 2018. Animesh Agarwal (Source: Medium)

measuring the distance from the line to the data and squaring each distance and summing up these distances. Distance from the line to the actual data point is called the Residual. The main aim is to find the line which has the least sum of squares to fit the data. (**Eq, 2.2.1**) shows the mathematical equation of the line.

$$y = mx + c \tag{2.2.1}$$

Here, x is the independent variable, y is the dependent variable m is the slope and c is the intercept. The equation shows that whenever the x is zero y equals the intercept c and the slope represents whenever there is a unit change in the independent variable what is the impact on the change in the dependent variable. To find the best fit line a cost function is defined. (**Eq. 2.1.2**) shows the mathematical equation of the cost function.

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{y} - y)^2 \tag{2.2.2}$$

Here, m is the number of data points, y is the actual data points and y hat is the best fit line which is given by the eq 2.1.1. Here, the cost function is also known as

MSE. MSE measures the average squared difference between the observed and the actual value. The goal is to minimize the MSE which further improves the model accuracy.

When performing regression analysis sometimes there is a need to express how confident the model can be to predict the "price" of the property. There is a need to predict a range were the predicted value would lie along with a certain amount of confidence in that range. This is known as quantile regression which is an extension of linear regression, when certain condition of linear regression such as homoscedasticity, normality, linearity is not met. The mathematical equation of quantile regression is provided in (**Eq.2.1.3**)

$$Q_\tau(y_i) = \beta(\tau) + \beta(\tau)x_{i^1} + \cdots + \beta(\tau)x_{in} \qquad (2.1.3)$$

Here, beta coefficients are quantile functions. As mentioned in the paper "Determining Factors in the choice of "prices of tourist rental accommodation," quantile regression results are more stable and robust compared to OLS which has been commonly studied in most of the papers (Moreno-Izquierdo et al., 2020). Further, it has been pointed out by Wang and Nicolau in the year 2017 that OLS limits our understanding of the real situation of the market.

2.3 Gradient Descent

To minimize the MSE an iterative optimization algorithm is applied to find the minimum of the loss function. Gradient descent is plotted as different values of the intercept against sum of squared residuals. It finds the minimum value which takes steps from an initial guess until it reaches the best value. This process is very important in deep learning and Neural Network.

Ordinary linear regression is one the algorithm where gradient descent is used extensively. Let us understand the steps of Gradient Descent.

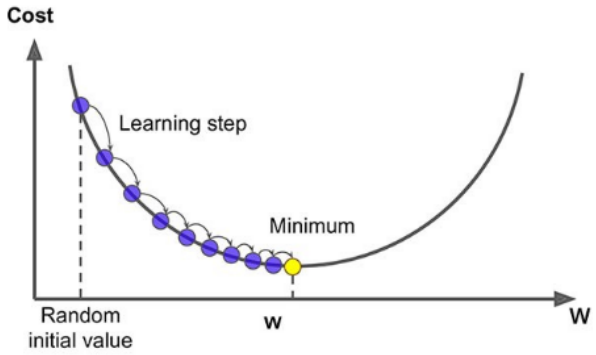


Figure 2.3, Gradient Descent, 2020. Agnes Sauer (Source: morioh)

Steps for gradient descent are as follows: -

1. As per the (**Eq. 2.1.1**) our cost function will be given by (**Eq. 2.3.1**)

$$j(m,c) = \sum_{i=1}^{\eta}\left(y_i - (mx_i + C)\right)^2 \tag{2.3.1}$$

2. Here, n is the number of data points m is the slope, c is the intercept and y is the actual value. Here, our objective is to reduce j(m,c) to reach to global minima. The use of Gradient Descent to minimize the cost function so that our model fits the data in best possible way.

3. Next, step is we need to take the gradient. Here, we can see that there are two parameters in our function m and c. Hence, we need to find partial derivatives with respect to both the parameters. This function needs to be minimized in order to find the optimal value of m and c. Therefore, we need to take the gradient by finding the partial derivative of m and c using chain rule of differentiation. The gradient will be given by: -

$$j(m,\ c) = -2\sum(y_i - (mx_i + b)), -2\sum x_i(y_i - (mx_i + b))$$

4. After solving this we will have a new m and c. Once we have our new m and c, we need to update the equation iteratively till we reach convergence. The gradient tells us the slope of the cost function and the direction to move

to update in order to reach the global minima. The size of our is update is calculated by something called as learning rate. Here, the hyperparameter are the learning rate and number of iterations.

5. However, as the number of features increases model complexity also increases.

2.4 Assumptions of Simple Linear Regression

Here we are making inferences about the population using the sample we are provided with hence an assumption that variables are linearly dependent are not enough to make the results generalizable. Thus, we need certain assumption in order to make certain inferences.

1. There holds a linear relation between the X and y.
   The dependent and the independent variable should display some sort of linear relationship otherwise there won't be any use of fitting the data.
2. The Error terms should be normally distributed with mean at zero.
   In order to make further interpretation about the model the distribution of the error terms are important.
3. The error terms should be independent that is it should not be like a time series data where the next value dependent on the previous value. There should not be any visible pattern for the error terms.
4. Error terms have constant variance.
   The variance should not increase or decrease as the error value changes.

2.5 Multiple Linear Regression

"Multiple" in MLR gives an idea that here there is a relationship between two or more independent variable and one dependent variable. Our data set has one dependent variable ""price"" and fifteen independent variables. The mathematical equation to predict the response variable now becomes (**Eq. 2.5.1)**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + e \qquad (2.5.1)$$

The model instead of fitting a line now fits a hyperplane. The coefficients are calculated using ordinary least square criteria, the assumptions of simple linear regression still hold the same.

However, as more and more variables are added the model tends to overfit the data, there becomes a high chance of multicollinearity, and selecting optimal number of features also becomes an important task. Further, many of the papers have been focussed on feature selection and feature engineering.

In the paper "Analysis and Machine Learning Modeling of New York City Airbnb Data" the major goal of the project was to assess the accuracy of machine learning model to predict the prices of rentals with respect to a set of realistic features (or predictors) (Atta-Fynn and Zien, 2019).

As the number of variables increases the model tends to memorize the training data and also becomes far too complex hence fails to generalise. The model performs poorly for the test data, in this case the model is said to overfit the data. When the training accuracy is higher that shows model has high bias and testing accuracy is lower that shows low variance. Hence most of the papers have focussed on feature selection and feature engineering.

Further, with the increasing number of variables there is a chance of Multicollinearity. Multicollinearity is a phenomenon where the explanatory variables are related to each other. In this case we can simply drop one of the variables.

In the paper "Airbnb "price" prediction using machine learning and sentiment analysis" the experimentation was done with all the features and their findings was the initial experimentation with the baseline linear regression model proved that the abundance of features leads to high variance and weak performance of the model on the validation set compared to the train set (Kalehbasti et al., 2019). The paper showed how to mitigate this problem by using Regularization technique such as Lasso.

2.6 Feature Selection and Feature Engineering

As we understood in the previous section how too many variables can affect the model performance

In one of the recent papers "California Rental "price" Prediction Using Machine Learning Algorithms," data scaling and transformation was done before the model building. Most of the variables in the dataset are in different scales further, the dependent variable "price" was highly skewed. This causes the variables with larger scales have bigger impacts on the prediction of the "price". Standardization of the

numeric variable were performed, the mathematical equation for standardization is given by **(Eq. 2.6.2)**

$$z_i = \frac{v_i - \overline{v_i}}{S} \qquad (2.6.1)$$

Where Zi is the i^{th} variable after rescaling, Vi is the i^{th} variable, $\bar{V}i$ is the mean of the i^{th} variable, and S is the sample standard deviation. For the "price" variable log transformation was performed to remove the skewness (Yue Fei. et al., 2020). Further, in the same paper feature selection was done by checking the correlation among the numeric variables. The variable bedrooms and bathrooms were found to be highly correlated; bedroom was again highly correlated to cleaning fees and guest included. Since the bedroom was highly correlated to the predictor variable hence bedroom was kept for the final model. Further, Random Forest was used to select top 30 features which included bedroom, apartment type, security deposit, etc.,

2.7 Bias Variance Trade off

A simpler model is more generalizable and robust which makes them perform better in the unseen dataset. The model does not undergo significant changes if the data points are changed. The overfitting chances are less when the model is too simple. Overfitting is a phenomenon when a model becomes too specific to the training data and fails to perform well in the test data. On the other hand, the model should not be too simple to underfit the data. This trade-off between the two is known as bias variance trade off.

With a highly unstable model a slight change in the data brings drastic changes to the model, therefore the model is unstable and highly sensitive to changes this is called high variance. In other words, variance refers to the degree of changes in the model with respect to changes in the training data.

Bias shows that how accurate the model is likely to be in the test data. A very simple model is very likely to fail to predict the complex real-world data. A very simple model will have high bias since it is way too simple to be able to learn the complexity as shown in fig. 2.7.

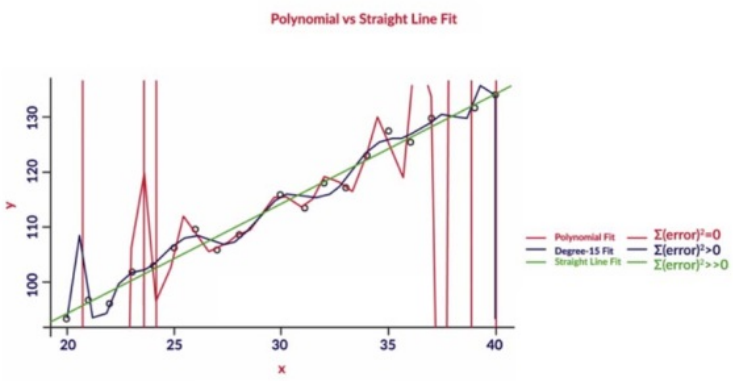


Figure 2.7, Polynomial vs Straight Line, 2018. (Source: Upgrad)

Here, three regression models have been fit to the data a linear, a degree fifteen polynomial and a higher degree polynomial which has fitted the data very well and is passing through all the points. The bias is very high when the model is linear that is too simple and decreases as the polynomial degree increases as shown in fig. 2.8.

2.8 Regularization

To find the right balance between model bias and variance and simplicity, complexity

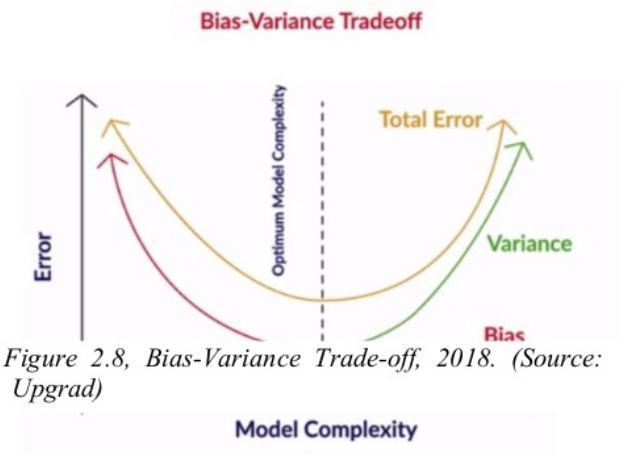


Figure 2.8, Bias-Variance Trade-off, 2018. (Source: Upgrad)

we need technique such as regularization to keep a check on the model. Regularization

is a process of simplifying model to achieve the correct balance in the model. This technique discourages the model to become too complex even though the model is able to explain the training data very well. In regularised regression, the objective function has two parts: the error term and the regularised term. The two most common regularised regression are: Ridge and Lasso. Both of these methods are used to balance the trade-off.

2.8.1 Ridge Regularization

In ridge regression along with the error term we add 'sum of the squares of the coefficients' whereas in lasso we add a 'sum of the absolute value of the coefficients.'

Research conducted by Yue Fei implemented regression models using ridge regression, ridge regression with k means, and XGBoost (Fei, Yue 2020.) Ridge regression also known as L2 regularization prevents the model to overfit the data by penalizing the model. In Ridge regression we not only minimize the sum of squared residuals but along with that we add regularization term which is lambda multiplied by the slope square. Mathematical equation provided in (**Eq. 2.8.1.1)**

$$\text{Cost function} = RSS + \lambda \sum_{j=1}^{n} \beta_j \qquad (2.8.1.1)$$

Here lambda is called the tuning parameter which take the value from 0 to and to any positive number. When lambda equals zero the model follows a Least square model. By adding the regularised term, we introduce a small amount of bias into the model in order to reduce high the variance that is high error in the test data. As the lambda value increases the variance decreases however the bias increases. Beta is the coefficient or the slope of the features. When the line is very steep that is when the slope is higher the penalty term or the regularized term increases. Here, the beta values tend to reach very close to zero but is never equal to zero.

2.8.2 Lasso Regularization

As we have noticed in the ridge regression, in a similar way lasso also performs however there is a small drawback when it comes to ridge regularization. The penalty term will shrink the coefficient towards zero but will not make any of the coefficient equal to zero. So, we need again a selection process to select the features which have an impact on the predictor variable. This problem is addressed by Lasso regularization. Lasso also known as Least absolute shrinkage and selection operator is also used for feature selection as the technique shrinks the coefficient of few variables to zero. Also, Lasso-based feature selection technique reduces the variance (Kalehbasti et al., 2019). The mathematical equation for lasso is provided by (**Eq. 2.8.2.1)**

$$\text{Cost function} = RSS + \lambda \sum_{j=1}^{n} |\beta_j|$$

Here we provide the absolute value of the coefficient of the slope. This technique is also called L1 regularization technique.

In the paper "Melbourne "price" Prediction model" lasso was performed in order to prevent overfitting. To accommodate potential clusters Lasso was run on three parts of the data they were: listings with "price" under $250, between $250 and $500 and above $500. Further 5-fold cross validation was used to tune the hyperparameters and to select the best features (Cai et al., n.d.).

2.9 Bagging and Boosting

Bagging and Boosting are two types of ensemble Modeling technique. Ensemble learning is a technique where multiple Machine Learning model are trained to solve the same problem and are combined together to generate best result. Most of the time it happens that model has low bias and high variance; ensemble method can be used to convert this high variance to low by combining multiple models. This combining of multiple weak learners to construct a strong learner that achieves better performance is called ensemble learning.

2.9.1 Bagging

Bagging also known as Bootstrap Aggregation and one of the most used algorithms to solve the Airbnb "price" prediction model was Random Forest. We create multiple base learners and for each model a sample of data set is provided. Again, for the next model resampling is done and another sample of data is provided to the second model. Hence, this process is also called as row sampling with replacement (Few records in the sample may get repeated.) The model gets trained on the sampled data provided and output is generated from each model. So, when test data is provided to all the model and output is generated, we use a voting classifier in the case of classification problem to arrive at a conclusion. A voting classifier takes the majority vote to predict the class of the target variable. Here, the row sampling with replacement is called Bootstrapping and last stage where we are generating the result using voting classifier to combine the result is known as aggregation. The diversity is model ensures individual model making prediction independent to each other. This also ensures the problem of overfitting hence resulting in lower variance in the test data.

2.9.2 Random Forest and Decision Tree

One of the most important technique used as a bagging technique is called as Random Forest. As mentioned in section 2.9.1 we train multiple base learners and use majority voting classifier to arrive at a conclusion. In Random Forest these base models are the decision trees. Further, we use the same technique row sampling with replacement to sample the data along with that we also consider feature sampling with replacement. This process is repeated and samples are provided with replacement to each decision tree to get trained on the data. All the results from multiple decision trees are combined to get the final results.

Whenever we construct a to its complete depth Decision Tree it gives us low bias and high variance which leads to overfitting. However, in Random Forest we are combining multiple Decision Tree which makes this high variance to low variance as we are providing the different samples to each and every model to get trained. This makes the model expert when new test data is provided.

A Random Forest classifier was used to classify the important feature for the "price" prediction. Here, the novel idea was to classify each listing whether it will be easy or difficult to predict. A Random Forest Classifier was trained and important features

were listed accordingly which included “Month,” “Number of Reviews,” “Cleaning Fee,” “Latitude,” “Longitude,” etc., (Choudhary et al., 2018).

Further, tree-based models such Random Forest regression, Gradient Boosting regression, and Extreme Gradient Boosting explained the “price” variation in the training data set and even for test dataset the scores were quite well. Here, the “room-type” variable has given highest importance (Atta-Fynn and Zien, 2019).

It was noticed that Random Forest obtained was the best performing model on the training data. The model was able to explain 93% of the training data (R square was 0.93.) Also, the RMSE and MAE was lowest for Random Forest. (Fei Yue, 2020).

Random Forests are collection of multiple decision trees. Random Forest is an ensemble method that takes into account multiple decision trees by performing bootstrap aggregation which helps to decorrelate the trees (Hastie et al., 2009). Decision trees mimics the human decision-making process as it is not focussed to find a linear relationship between the independent and the target variables rather decision tree can model highly nonlinear data. The factors leading to a particular decision can easily be explained using decision trees. (Fan et. al., 2006) used the decision tree approach to examine the relationship between house price and housing characteristics. This was done using the case study of Singapore resale public housing market.

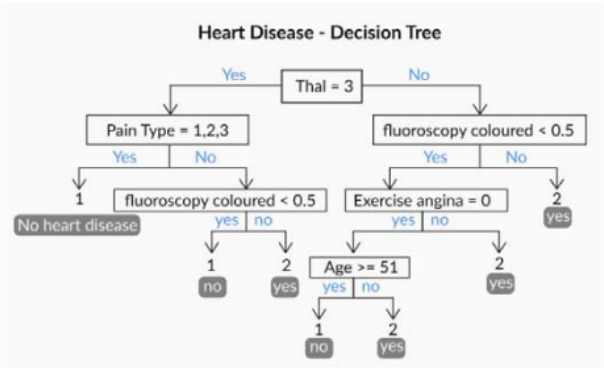


Figure 2.9, Decision Tree Representation, 2018. (Source: Upgrad)

(Figure represents a Heart Disease Decision Tree which is a binary classification problem to find out whether the patient has heart disease or not.)

2.9.3 Boosting

Boosting is an ensemble learning technique where models are trained sequentially and unlike bagging the models are dependent on each other. The models here are weak learners which together when combined becomes a strong learner. For the Airbnb "price" prediction problem almost all the papers have experimented with this technique and have achieved best results. The most common boosting techniques includes.

1. AdaBoost
2. Gradient Boosting
3. XGBoost
4. LightGBM

Boosting reduces both the bias and variance and hence creates a generalized model. XGBoost is one of the algorithms which is used extensively as it works well for large datasets and is not computationally expensive. XGBoost uses weak learners sequentially and the error is reduced each time by the next model. The loss function used by the model for regression is provided in **(Eq. 2.9.3.1).**

$$\sum_{i=1}^{n} l_i(y_i, \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.9.3.1}$$

XGBoost was developed by Taiqi Chen, to minimize the regularized objective function provided in equation **(Eq. 2.9.3.2).**

$$obj = \sum_{i=1}^{n} l_i\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_t) \tag{2.9.3.2}$$

Where, (f_t) means prediction coming from the t^{th} tree. The first term controls the loss function and the second term $\sum_{i=1}^{t} \Omega(f_t)$ penalizes the complexity of the model in order to avoid overfitting. The XGBoost objective function constitutes of the loss function of the overall model, all predictions and sum of the regularized term for all the predictors. Once objective function is defined the very next thing to do is optimizing the parameters. When the regularized parameter is set to zero the objective is back to the traditional Gradient Boosting tree (Chen and Guestrin, 2016)

$$obj = \sum_{i=1}^{n} l_i\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_t) \quad (2.9.3.2)$$

$$= \sum_{i=1}^{n} l_i\left(y_i, \hat{y}_i^{(t-1)}\right) + \left(f_t(x_i)\right) + \Omega(f_t) \quad (2.9.3.3)$$

After taking the taylor expansion and re-formulating the objective, the final objective becomes **(Eq. 2.9.3.4).**

$$= -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (2.9.3.4)$$

Were, gamma is the regularization term (Yue, 2020).

Another, approach on Gradient Boosting was done on the paper "Customized Regression Model for Airbnb Dynamic "price" Prediction" to predict the booking probability. The idea was to train separate GBM model for each market with an adaptive training data sampling (Ye et al., 2018).

Gradient Boosting combines decision trees in a sequential manner where each tree tries to correct the mistakes done by the previous tree (Hastie et. al., 2009).

2.11 KNN for Regression

The KNN algorithm uses feature similarity to predict the values of new data point. The new values is the average of the nearby points (Luo et al., 2019).

1. The distance between the training data and each point is calculated.

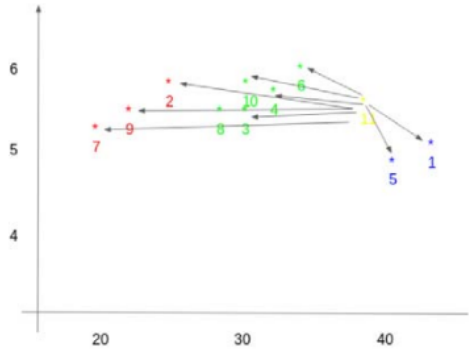


Figure 2.11, KNN, 2018. Aishwarya Singh (Source: Analytics Vidya)

2. The data point that are closest to the training data are selected. The number of data point selected is the value of k.
3. The average is taken as the final prediction for the training data.

There are many methods to calculate the distance between the new point and the training data. The most common methods of regression problems are Euclidean and Manhattan distance.

1. Euclidean distance is calculated as the squared root of the sum of the difference between the data point squared.
2. Manhattan distance is the sum of the absolute difference between the difference between the data point.

2.12 Performance Metrics

Best performance measure for regression analysis are R squared and root mean square or RMSE.

Regression model are the prediction when we have a continuous target variable and we are trying to predict with the help of several correlated independent variable. The various performance metrics to evaluate the results are as follows: -

1. Mean Squared Error (MSE)
2. Root Mean Squared Error (RMSE)
3. Mean Absolute Error (MAE)
4. R^2 or Coefficient of Determination
5. Adjusted R^2

MSE is the most preferred metrics to determine the performance of the model. However, in most of the papers MSE along with R^2 is used to test the performance.

MSE is the average of the squared difference between the actual value and predicted value. Since the term is squared it can penalize the model even the error is small. MSE is a differentiable metric and hence is preferred as it can be optimized for better performance.

$$\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2 \tag{2.11.1}$$

RMSE is the squared root of the averaged squared difference between the predicted and the actual value. This poses a high penalty to large errors as the metrics is averaged and then taken the root.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2}$$

MAE is the absolute difference between the predicted value and the actual value. The metrics is more robust to outliers

$$\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y}) \tag{2.11.3}$$

R^2 is one of the most used metrics for evaluating the predicted "price". This metric helps to compare the current model with the baseline model and lets us decide the model performance.

$$1 - \frac{RSS}{Tss} \quad (2.11.4)$$

Here, RSS is the residual sum of squares, TSS is the total sum of squares.

Adjusted R^2 is an improvement to R^2 value hence in most of the cases Adjusted R^2 is preferred. The R square value increases every time we add a new variable to the model. The variable added might not have any impact on the overall model performance. However, adjusted R square increases only when the variable has a significant impact on the dependent variable.

$$1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right] \quad (2.11.5)$$

Where, n is the number of observations, k is the number of independent variables.

2.13 Neural Network

Neural Network is a ML technique which also can be used for regression problem and has provided good results when used for the pricing problem. The Artificial Neural Network's (ANN) tries to replicate the human brain, generalise the response and provide result. Usually, the model is unstable at the beginning but after a number of iterations it adjusts itself. A neural network with more than one hidden layer is called a Deep Learning Network. The output layer contains of one node if it is a regression problem.

ANN consists of a replica of human brain containing parallel processing neural elements which are quite similar to living beings' brain. This network can store large amount of experimental information to be used for appropriate model prediction (Gonzalez-Fernandez et al., 2019).

Neural Network's predictive capacity increases with an increasing number of layers and neurons in it, although two to three layers are sufficient to solve majority of practical tasks of classification, regression, and prediction (Hornik et al., 1989).

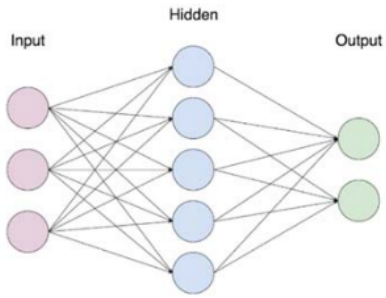


Figure 2.13, ANN, 2020. Chun Hei Michael Chan (Source: Towards Data Science)

The study by (Moreno-Izquierdo et al., 2019) developed an ANN, a network was created using three layers with sigmoid activation functions used in the neurons. The regularization algorithm was done in two iterations to eliminate variables with extremely small coefficients. The remaining variables were significant for constructing the model. The paper also focussed on comparing "price" model with ANN model. More than 20 variables which included both elements of property and others from the economic, social, and tourist environment of the destinations considered. A methodological comparison was made and it was found that ANN's are more robust model.

As compared to hedonic pricing model, ANN can be useful to describe the non-linear relationships between the variables and therefore provides better predictive capacity than multivariate analysis (Wilson et al., 2002). A similar study was conducted by (Limsombunchai et al., 2004) where effect of factors such as size of the property, the number of bedrooms, location, was again analysed using hedonic regression and neural networks.

2.14 Hedonic Regressions

Hedonic regression is the implementation of regression model to estimate the influence of different factors which has an impact on the "price" or demand of any good. The dependent variable is the "price", and the independent variable are the attributes or factors. Hedonic regression is used mainly in real estate pricing. A study on hedonic regression was conducted and summarized the relationship

between the individual and the government effect. The study examined the determinants of house prices in Turkey. Both ANN and Hedonic regression Modeling approaches was implemented. The results from Hedonic model revealed that pool, number of

rooms, water system, house size, type of house, locational characteristic and type of house were the most significant factors. Due to the non-linearity in the hedonic function, ANN was used to predict the prices. It was found that ANN was a better alternative for predicting prices in Turkey (Selim, 2009). The paper by introduced house-hold level data into the hedonic models to measure the heterogeneity of prices regarding household type, age, income, education, tenure status of the buyers (Kestens et al., 2006). Hedonic Pricing has been widely analysed to determine the qualities that are best to consider the demand and supply (Moreno-Izquierdo et al., 2020).

2.15 Literature Review Summary

There has been a significant amount of research done on optimising the pricing of rental property. Research conducted by (Gibbs et al., 2018) claims that almost all hosts fails to maximize the profit due to the poor "price" listings. Most of papers reflects the CRISP-DM methodology framework and has been successfully implemented in the process. A growing number of hosts and the listings has given rise to a reasonable and competitive market for rental "price" that helps to maximize the benefits for both the hosts and guests (Yue Fei, 2020). The resent research has been conducted to look into the variables that has a significant effect on the pricing. To understand the pricing of Airbnb listings, the study of the progression of Airbnb pricing's history is also important. In 2012, a "price" recommendation tool based on simple characteristics was offered. Even though the tools can be used to recommend the "price" of the property however, the ultimate decision was left to the owners (Mcneil, 2020). One of the recent researches was done to test the hypothesis that the prices of the listings in Airbnb New York City was more influenced by the physical locations rather than other features and characteristics of the property. Number of Machine Learning model has been implemented to analyse the different characteristics that can help predict better listing prices. The algorithm such as KNN, Linear Regression, XGBoost, Random Forest, Ridge Regression, K means, SVM has been proved to be very effective. XGBoost has outperform most of the techniques, however in XGBoost the model is usually a black box model, interpretation of the model is quite difficult.

Data cleaning, visualization and exploration plays an important role which helps in a better understanding of the data and hence builds the foundation of the project. The Paper "Determining factors in the choice of prices of tourist rental accommodation"

by (Moreno-Izquierdo et al., 2020) provided new evidence using quantile regression approach. QR helps to better interpret the results and the variables used for model building. Also, it was found that higher priced rental properties have stricter cancellation policies mostly due to their negotiating power in the market.

A great majority of work has been done using classical method of OLS estimation which helped to analyse the relationship between the variables and conclude the results. Previous work has looked at the specific variables which has a significant effect on the pricing.

One of the papers was focussed on grouping the variables into two groups on the basis of host's ability to control that variable. The results showed that variable which were in control of the host's ability had more impact on "price" versus the variables which were not in control of the hosts (Mcneil, 2020).

2.16 Discussion

A thorough review of the previous work revealed that tree-based algorithm and Neural Network performed very well when compared to OLS methods. While many papers have confirmed the features which directly have a significant impact on the predictor variable, there is always a chance that the impact may change as pricing a product is very dynamic field. Further, as we are aware of Covid19 situation, it has a tremendous impact on global economy. Airbnb a company which belongs to the hospitality domain has also been impacted heavily and the listing prices are also affected due to the situation. In the current literature review it was noticed that most of the papers contributes to identifying the important variables for the "price" with the use of Machine Learning model such OLS, Random Forest, Decision Trees, etc., In the recent paper "California Rental "price" Prediction 2020" XGBoost gave the most satisfying result. XGBoost was chosen for two reason as it gave the highest R^2 and small RMSE and MAE value for both training and test data (E. Ezalia et al.,2020). LightGBM is also a tree based boosting algorithm however it differs from other boosting techniques. While other tree algorithms grow horizontally LightGBM grows trees vertically leaf wise. This process helps in reducing more loss. Further, the features contained in the dataset are to be analysed and studied to extract knowledge regarding the target variable. Since the size of the data is increasing day by day the traditional boosting algorithm is becoming difficult to provide faster results. LightGBM, as the name suggests is light due to its high speed and the capacity to

handle large amount of data. The algorithm is easy to implement and provides better accuracy. However, parameter tuning is complicated compared to others. This paper aims to implement this algorithm as a classifier and check the results in predicting the classes of the properties.

Further, the feature "listing name" has not been analysed much in the previous work. As the words used in the name of the listings can convey some insights regarding the property and the "price". The paper "Determining factors in the choice of price of tourist rental accommodation" it was observed that an expensive property has a detailed description provided to the users through interactive photos (Moreno-Izquierdo et al., 2020). Also, further experimentation with neural network can be done to check how the model is performing.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Research is a systematic and logical search for new and useful information regarding a particular topic (Rajasekar et. al. 2006). Research helps to investigate and find solutions to a problem through objectives and systematic analysis. A research is done with the help of study, carrying out experiments, comparison and correct reasoning.

This chapter discusses the research methodology which has to be followed in the analysis for the data set (i. e., the Airbnb 2020 New York data set). In this part the outline of the research strategy, methods, the research approach, methods of data collection, research process, type of analysis, research limitations are explained in detail. Here, we have also justified the proposed methods to better understand why certain methods are considered for the analysis. The EDA part has not been included in this section.

Airbnb listing data was collected from Inside Airbnb website which is a data collection website and provides clean and updated data around the world every month. For the New York area, a total of 48,572 Airbnb listings were collected with fifteen variables and one target variable "price."

The main type of analysis to be applied are the different ML techniques and compare the results of each techniques. These methods are linear regression, regression using ridge and lasso, artificial neural network, random forest, decision trees, KNN, and applying different boosting algorithms such as gradient boosting, extreme gradient boosting, and Light GBM classifier. These methods are roughly explained in the literature review section and will be thoroughly discussed in this section.

As the results for the different ML algorithms, a comparison between the results is needed to reach a conclusion. One method of comparison is to compare their performance metrics such as MAE, RMSE, R squared and Adjusted R squared values.

The initial pre-processing steps includes exploratory data analysis, missing value treatment, outlier treatment, feature engineering and selection. Most of the time of any ML project is dedicated to this step and the rest is to building the model. Section

3.2.2 describes the pre-processing which are included in this research in detail and the methodology for preparing the data before applying the any ML algorithm.

Further, in most of the papers the variable "name" has been omitted for analysis however, in this research we also aim to analyse this variable and find out if there are any patterns for naming the listings which are high priced. The words are almost always are good indicator of the cost hence we have tried to find the most frequent words, bigrams, trigrams, wordings for the expensive properties, and if there exists a correlation with the price variable. With the use NLP to understand if the words indicate the pricing variable.

Section 3.4 includes the tools we have used for our analysis. We have a number of libraries in python which are to be used for data visualization, model building, boosting etc.,

3.2 Research Strategy

The research with respect to the dissertation is an applied one, but not new. Numerous papers of previous research exist regarding the price prediction for Airbnb properties, not only for New York in specific, but also for other tourist destinations. However, the methods to be used in this research is novel and has not been conducted before. As such, the proposed research will be a new form but on existing research subject.

3.3 Research Approach

The research approach to be followed for the purpose of this research are described briefly in this section which includes data pre-processing, data visualization, statistical analysis, data transformation. In this research our main aim is to come up with a price prediction model by applying different ML techniques and also to understand the variables which have an influence on the target variable. We have tried to understand well what are the effects of different independent variables towards the dependent variable. To gain more understanding of the data we have used descriptive statistics which will form the base of the model building. Descriptive analyses are the first data manipulations to be performed for quantitative study which helps in summarizing the data set and explore the behaviour of the variables with respect to the target column. This can be achieved using statistical technique such as frequency distribution, measures of dispersion, and central tendency.

We are carrying a quantitative approach to our problem that enables us to determine the existing relationship between the price variable and the features in the data set. Further, in quantitative study it is very important to understand what type of data is used for the analyses. The type of data involved for the study is discussed in the next section.

3.4 Data Selection

The data has been collected from the third-party website www.insideairbnb.com which includes the public data of a number of locations where Airbnb has properties. The dataset has a rich set of variables which are a good predictor of the price variable. A specific listing is provided in each row that is available for renting on Airbnb in NYC. That means each row is a specific data point.

Since the data belongs to US hence the dependent variable price is in US dollar as the currency in our dataset. The data belongs to New York, as it is one of the busiest place for tourists. Hence the process of determining the price of the listings will be interesting and helpful to understand the problem better.

The data set contains both textual and numeric data hence a number of ML algorithms can be applied such as NLP to extract as much information from the data. Further we have obtained the data in csv format hence loading the data to the tool can be done conveniently.

3.5 Data Collection Methods, Data Types and Tools

For the purpose of this research, we have used a secondary data for our analysis. We have considered Airbnb 2020 dataset of New York city (http://insideairbnb.com), which is a third-party provider. Airbnb does not release any data however, this group named inside Airbnb has extracted data for the major cities on the website. This is an independent website for data collection, the website provides us clean and updated data of different tourist locations. The data was available in csv format hence making it easier to upload using pandas library in python.

The data is divided into two main types: qualitative and quantitative.

Qualitative variables are categorical variables which cannot be measured numerically such as name of the hosts, name of the listings, room type.

Quantitative variables in the other hand are the variables which can be measured such as minimum nights, latitude, longitude, number of reviews, etc., Here, the target variable price is continuous quantitative variable.

The variables such as name of the property cannot be used directly for model building. Further, the variable "neighbourhood group" and "neighbourhood" contains names of the neighbourhood which is a categorical variable. Similarly, "room type" is also a categorical variable. Few variables which contains "id" and "host name" are insignificant and irrelevant and will not provide much information hence can be dropped for analysis. The main advantage of secondary data collection is that it is readily available and does not require much effort in data collection hence it is time saving. Also, since the data is available in Airbnb official website hence making it cost effective too. The data has also been used previous researches thus making it easier to carry out further research. The data is in structured format. The data set has very rich information regarding the properties for deep exploration.

As we will be using python for our analysis, we will need pandas library to load the data appropriately in a pandas data frame format for analysis. Pandas is an open-source high performance data analysis tool for python language. Data frames allows store and manipulate data which are in rows and column format.

3.5.1 Python

As discussed earlier, we have chosen python 3.9.0 which has been recently released on 5th October 2020 (python.org). Python is an open source, high level, object-oriented language. There are a number of open-source libraries which can be used for mathematical functions, data science and analytics, data visualization. Since the language is very easy to use and simple syntax makes it very popular among researchers. Python's available deep learning frameworks with python API's and the scientific packages have made the language extremely productive and versatile.

For applications such as NLP and text analysis developers choose Python over any other language because of the large libraries that can help in solve complex problem easily and build a string application. Some of the very common libraries are Numpy, Pandas, Scipy, Matplotlib, Seaborn, NLTK, Scikit Learn, etc.,

3.5.2 Seaborn and Matplotlib for Visualization

Seaborn and Matplotlib are two of the most powerful libraries for data visualization. These two libraries are most widely used in python for business insights and data analysis. It helps creating interactive graphs and plots in the form of bar charts, scatter plots, pie plots, violin plots, etc., in order to gain data understanding. Compared to Matplotlib, Seaborn uses fewer syntax and has a number of default themes.

The variables available in the dataset can be easily visualized using this two libraries.

3.6 Data Analysis

Data analysis is the process of cleaning transforming, inspecting and modelling the data with an aim to derive useful information and conclusions for decision making. In today's world it plays one of the main role in industries to help business running more effectively. In this study we will consider divide the process into three parts: Descriptive Analysis, Diagnostic Analysis and Predictive Analysis.

3.7 Data Pre-processing and Transformation

The data set contains object, integer and float data types which should be treated differently to feed to the model. There are missing values present in the data represented by NaN which requires cleaning and handling. One way is to delete the missing values however there is a chance we will lose information in this case we will try to impute the missing values. One of the best practices considered is imputing the missing values by mean, median or mode. Further, the columns such as room type, neighbourhood group and neighbourhood are categorical columns which can be later treated by one hot encoding and label encoding.

The first step for our analysis work is to understand the variables of the dataset. The use of statistical techniques such as frequency distribution tables, histograms and bar plots to understand the phenomena under the study. A frequency distribution table helps to represent the data in tabular format to display the number of observations within a given interval or range. For the variables longitude and latitude, scatterplot can be created using heatmaps to find out which area are most popular.

One of the most important pre-processing steps involves treatment of the outliers. Outliers are the data points which does not fall inside the overall distribution of the

data. Different techniques such as Box plots, scatter plots, IQE, or Z score can be used to find the outlier and treat them.

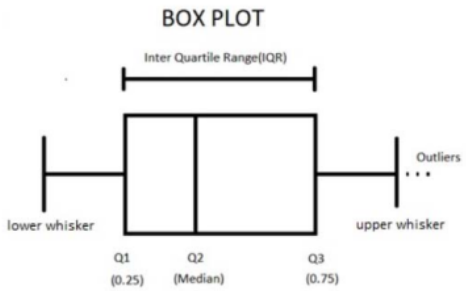


Figure 3.7, Boxplot, 2019. Sangita Yemulwar (Source: Analytics Vidya)

As shown in the figure the data point which falls outside 1.5 times the IQR range above the 3rd quantile and below 1st quantile are considered as outlier. Further, by using the z score outlier can be detected, any point greater than three S.D. away are considered as outlier. One of the best treatments is to completely delete the outlier. Transforming the variables can also help in removing the outliers. Using Natural log transformation reduces the variation caused by the outliers. Binning can also be used for variable transformation.

Variable transformation and creation are techniques of feature engineering which is one of the vital process in data exploration. Since the variables are all in different scales, we would need to change the scale or standardize the variable.

Variable creation is a process to generate new variables from the existing one. As we notice in the dataset there is date variable which is the last review date. We can generate day, month, week information that may have a better relationship with the target variable price. ‘

Further, for the variable listing name there is a need to remove the punctuation, special characters if any exists, or digits to further process this variable. Also, there is a need to remove the stop words such as articles, conjunction, etc., Once the variables are in correct format it can be used for the analysis.

Data exploration and transformation takes the maximum amount of time and effort to build a good model. The main aim of EDA is to provide deeper understanding of the data and also provide useful insights for model building. EDA helps to model provide accurate prediction and also can be used for interpreting the result.

3.8 Data Visualization

Exploring the data by visualizing the values of the features is also one of the important part in data analysis. By visualizing the data using bar plots, scatter plots, histogram, etc., we can find the distribution and the trend of the data.

To find the frequency distribution of variables we have used histograms which can be used to divide the dataset into uniform and non-uniform classes. For example, we can create a histogram for the target variable to find the region which has a larger concentration.

Bar graphs are also similar to histograms with rectangular bars which can be both horizontal or vertical. However, the bar graphs are mainly used for comparisons and the size of the bar depends on the frequency. Bar graphs are mainly to be constructed for the categorical variables such as the room type and neighbourhood to understand how these categories are distributed. The neighbourhood variable has more than 200 categories hence it was a challenge to handle this variable. Further, having these many categories may also lead to curse of dimensionality. In order to resolve this issue, we have found the top 10 most frequent neighbourhood and use them for our analysis. Similarly, for the variable neighbourhood group we have used bar and polar graphs to find out how the different categories are distributed.

We have information on longitude and latitude which have been used to understand the data on a geographic scale. One way to view this information is using scatter plot to find the which area is the busiest. Also, we can use heatmap to visualize all the listings of NYC. Further, to analyse the price variable based on the location, a colour coded map can be very helpful. Also, we need to do the outlier treatment before the analysis.

The goal is to explore the behaviour of the attributes and derive insights from the data.

3.9 Data Mining

A process to extract useful information from raw data is known as data mining. Data mining is a process of diagnostic analysis which helps us understand why something happened and what happened. This information can be further used to explore the unknown patterns, relationship, and the anomalies which are present in the data. Data mining provides us the ability to find new insights and unknowns from the dataset. Here, our data consists of 16 variables out of which one is the target variable price.

The data set can be used to predict the future price of the listings using the features. Data mining is the subset of business analytics to find future trends. While data mining and ML are both about learning from the data however, they are both different technique as ML goes beyond the trend to predict the future outcomes. Further, data mining is more of a manual process as we have discussed in the previous section to visualize the data, transforming the data. ML is an automatic process where certain algorithms are fed and results are generated. In this research we have used both the techniques to reach to the conclusion.

3.11 Research Design and Methods

The study is designed as to understand the data first and then to proceed with the plan to answer the research questions discussed through different methods.

To understand the data, we have discussed about exploratory data analysis in the previous section. As per the first research question we will need to understand if there exists a relationship between the name of the listings and the price of the variable. NLP is a technique used to process the textual data and is a subfield of linguistics. In our case we have carried out this technique to analyse the textual data like the names, descriptions. To understand this, we have pre-processed the name column. We have found the top 20 used words for the listings to understand if there exist a trend used by the hosts to describe their property. These frequent words with the price variable are used to find if there exists any correlation using heatmap. We have found the most frequent bigrams and trigrams which are used to describe the NYC apartments. The TF-IDF technique to process the name of the listings. TF-IDF stands for term frequency inverse document frequency is a popular technique in NLP to process the human language. Cleaning the text document is a vital process for any text pre-processing. Once the text is cleaned it can be converted to numerical value to be represented by a matrix or word vectors known as word embedding. Bigrams are the two consecutive words which appears in the document. Similarly, trigrams are the three consecutive words. We have used this approach to find top sequences of the words used in the title of Airbnb NYC flats. This analysis will help find the words which are used for expensive properties. A word cloud to be created for visual representation of the words used to describe the property, where the size of the word increases if it has been used more often.

In this dataset we have a number of independent variables hence before model building one of the aspects, we need to look is multicollinearity. Correlation with different pairs of independent variables can be checked using correlation heatmaps. Heatmaps are graphical representation where values are represented by colours. As the value depends on the intensity of the colour. We have used seaborn library in python to create heatmap for the data. Here, we can also find out which variables has an impact on the target variable.

The figure 3.11 represents a heatmap.

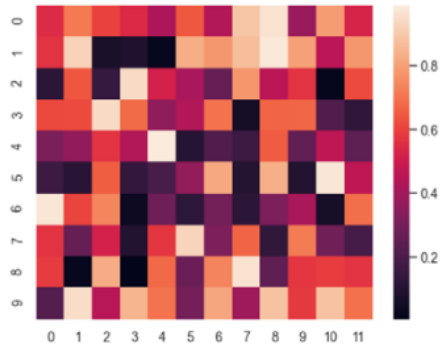


Figure 3.11, Correlation Plot, 2019. (Source: Upgrad)

Using the columns longitude, latitude and neighbourhood groups, spatial analysis can be carried out to analyse the geography of the NY city. The most expensive neighbourhoods can be identified using this plot.

Predictive modelling is the last process in the study which includes model building, model comparisons and hyperparameter tuning. Predictive analysis summarizes the data to make predictions for the future data.

Further, as discussed in the problem statement section, we have used the Light GBM technique to classify the properties using only the title of the properties. Boosting is a technique used in ML which converts a weak model to a strong one. It creates an ensemble model where base learners are decision trees. Boosting leverages the fact that a series of model is built which specifically targets the incorrectly predicted data points. If a series of model is built the error keeps reducing, hence we have an ensemble having high accuracy.

Gradient boosting was developed specifically to solve regression problems and XGBoost was developed on top of Gradient boosting involving shallow decision trees. Most of the papers have used Gradient boosting and XGBoost algorithms.

Light GBM is a boosting algorithm which grows tree vertically that is leaf wise while others algorithm grows trees level wise. This leaf wise algorithm can therefore reduce more loss.

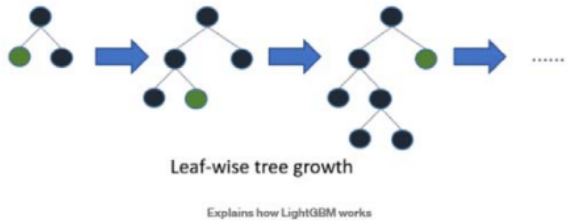


Figure 3.11.2, LightGBM, 2017. Pushkar Mandot (Source: Medium)

Light GBM is called light due to its high processing speed and ability to handle large volume of data. Also, the algorithm takes in lower memory compared to other boosting framework to run. Further, the implementation is very easy which makes it one of the most popular algorithm.

Further, the model performance will be checked using the R squared, Adjusted R squared, MSE, MAE, RMSE. These techniques are discussed in the literature review section.

We will also create ANN to predict the prices of the listings. The input features will be combined to produce a high-level predictors. We have created a four-layer NN with densely connected layers and ReLU activation function for the hidden layer and a linear activation function for the output layer. The ReLU activation function also known as rectified linear activation function is the default activation function when multilayer neural network is developed. ReLU activation function has a simple formula **(Eq. 3.9.1)**

$$\max(y, 0) \qquad (3.9.1)$$

Here, y is the output. The output of this linear function is either positive or zero. It is the default activation function as it achieves better performance and the model becomes easier to train. Neural network are implemented using Scikit learn and Keras library. However, experimentation with the hyperparameter and the architecture of the neural network is required once we start building the model.

Most of the papers have shown that neural network has been more robust and provide better results in terms of performance measures.

3.12 Summary

The Airbnb 2020 dataset is very rich and contains a variety of columns that can be used to extract information and for deep exploration. First step is the data pre-processing and transformation step to get a clean data which forms the base of the model building. EDA is the most important step in model building if not done correctly can result in incorrect model. Action such as missing value treatment, outlier removal dropping unnecessary columns are included in the initial steps.

There is a need to analyse the data using techniques of data visualization to analyse the columns available in our dataset. Plots such as violin plot, pie plot, bar charts, scatter plots, box plots are very useful for qualitative and quantitative analysis. Density and distribution of the target variable can also provide information regarding the dataset. Using the latitude and longitude columns we can create the geographical heatmap by creating colour coding for the price variable.

To perform the descriptive analysis, we will be applying statistical technique such as measures of dispersion, central tendency to further understand the data well. This process will help to confirm the assumptions and hypothesis build.

Also, we will take the help NLP techniques as discussed to process the textual data and find out if there exists any correlation with the price variable. Further, by creating word cloud we can find the most used words in the dataset for example name of the listings or reviews.

Also, using matrix correlation plots such as heatmaps we will find if there exists multicollinearity between the independent variables. These steps are included in the diagnostic analysis.

Next step includes feature engineering and feature selection. We have dropped features using different techniques such as vif, recursive feature elimination, or by using p values to find out the features which are contributing most to predict the target variable. With an increasing number of feature there is a chance of curse of dimensionality. Due to this model becomes less generalizable and does not perform well to unseen dataset. Often algorithms such ANN and Linear regression fails due to curse of dimensionality.

We have features such as date of last review which can be further transformed to extract the days, months and weeks information. Again, we have few categorical variables which cannot be directly fed to the ML model hence we would need to transform this variables using encoding techniques.

Last step will be predictive analysis which will include model building and model evaluation in order to predict the price of the listings. We will be using boosting techniques such as Gradient boosting and extreme gradient boosting to boost the model performance and compare the results.

We will create ANN with ReLU activation and three hidden layers to create a model using deep learning.

We will consider the R squared, Adjusted R squared, MSE, MAE, and RMSE values to compare the model performance and find the best model.

CHAPTER 4

ANALYSIS

4.1 Introduction

The purpose of this chapter is to summarize the techniques and the detailed analysis which has been carried out in the entire study.

The data collection has been done from the site Inside Airbnb site which contains a rich collection of Airbnb listings and the details of the properties. The site sources its data directly from Airbnb site. The dataset was scraped for the month Jan-Dec 2020 and 2019 data for New York City with 570365 listings and 16 columns. Then the data was compiled into one large dataset. Once compiled the duplicates were removed. These steps were done using pandas and globe module.

The initial steps which includes data gathering, selecting, and transforming the data to answer the problem stated in the chapter 1 Introduction section. Initial pre-processing included removal of the duplicates from the dataset. Once the data was uploaded as a data frame it was noticed that around 1,26,638 data points were duplicates and were removed from the data. Further, there were missing values present in the data as well. Few unnecessary column which would have caused "Curse of Dimensionality" were removed from the data set such as "last review date," "id." These columns would have not contributed to study hence were removed. The ML model which were included were Linear Regression, Random Forest, XGBoost, Light GBM, KNN, ANN to predict the price of the listings. Further, NLP was used to find if there exist a relationship with the name of the properties and pricing.

The initial model which was created using Linear Regression, had low performance however feature engineering and feature selection helped to achieve a good model score.

Artificial Neural network gave a decent performance and also didn't show overfitting. ANN was initially trained using 4 layer which included 13 neurons in the first layer, 13 neurons in the second hidden layer, 6 neurons in the third hidden layer, and 1 neuron in the output layer.

First three layer included 'relu' as the activation function and last contained 'linear' activation function. The ANN model was trained using keras and TensorFlow in the

backend. Further, the ANN model performance increased due to hyperparameter optimization.

All the ML model were evaluated using the performance metrics such as MSE, RMSE, MAE, R^2, and Adj. R^2. Further the error terms were plotted to find the distribution.

The model performed better when feature engineering was done correctly in the data set. There was high skewness noticed in few of the columns such as the target variable price, host listings counts, minimum nights, etc., These variable were treated after the first model was built to increase the model performance.

In this section we will perform the exploratory data analysis which will help to find what lies behind the data. The univariate, bivariate and multivariate analysis are performed to find out relationship between the attributes and the patterns. Few questions which are answered in this section are as follows: -

1. Which is the most expensive neighbourhood group in New York City?
2. Which is the most expensive room types among the 5?
3. How the price variable is distributed?
4. If there exists multicollinearity between variables.
5. If there exists any difference between 2019 and 2020 data set.

The data was available for 5 neighbourhood groups: “Manhattan,” “Brooklyn,” “Queens,” “Staten Island,” “Bronx.” Out of these the Manhattan had the most expensive properties followed by Brooklyn. The hotel rooms in Manhattan were most expensive. Also, Manhattan has the greatest number of properties and Staten Island had the least number. However, it was seen that the Queens had the greatest number of neighbourhood among the 5 neighbourhood groups.

The variable “number of reviews” was plotted against the price variable and it was noticed that number was high for less expensive properties.

The variable “host_id” was not dropped initially and was later dropped when ML model was executed. The host id was grouped and plotted against number of reviews and it was noticed that top 10 hosts had more than 14926 number of reviews.

The target column “price” had 144 properties which had a price as 0 and were dropped out in data cleaning step as it not possible for a property to be absolutely free. Also, few properties were priced as 25000$ which seemed very abnormal. Further investigating showed that only 3 rows were priced as 25000$ and all the three were in Manhattan which seemed to skew the prices. Rest of the outliers were treated using

"Tukey Test." Even after removing the outlier the price variable did not showed normal distribution.

The price variable was highly skewed and 99% of values were below 800$. The mean of the price was around 154$ with a standard deviation of 378. The price variable was the most treated variable in the entire dataset. None of the variable were seen to be highly correlated with the price variable.

Hence, multicollinearity was not present in the dataset. The price variable was highly away from the gaussian distribution. After removing the outliers, the variable was transformed using boxcox transformation.

The categorical variables were transformed using label encoder and one hot encoding both of the techniques are considered best to transform any object variable to fed to the ML model.

Since the target variable was a continuous hence the first algorithm which was applied was regression analysis, followed by other algorithms. To test the model accuracy the data was split to train test split. A 5-fold cross validation was applied to increase check the model performance.

The best score was provided by XGboost model which even reduced overfitting. The random forest had also shown a good score however the model was sensitive to the test data. Later ANN was built in order to make a comparison.

4.2 Dataset Description

The Airbnb data contains different attributes and the price variable to define each listings. Each row contained the name of the property, host name, neighbourhood details, and the review details.

The data was a collection of both integer and object type variable which was treated differently in the entire analysis. The shaped of the data before the analysis was (570365, 16). The table 4.2.1 provides the detail of the attributes provided.

Variable Name	Description
id	Listings ID
Name	Name of the listing
Host_id	Host id
Host_name	Name of the host
Neighbourhood_group	Location
Neighbourhood	Area
Latitude	Coordinates
Longitude	Coordinates
Room_type	Space Type
Price	Price in dollars
Minimum_nights	Minimum number of nights
Number_of_reviews	Reviews number
Last_review	Last review date
Reviews_per_month	Reviews per month
Calculated_Host_listings_count	Host listings count
Avaliability_365	Availability

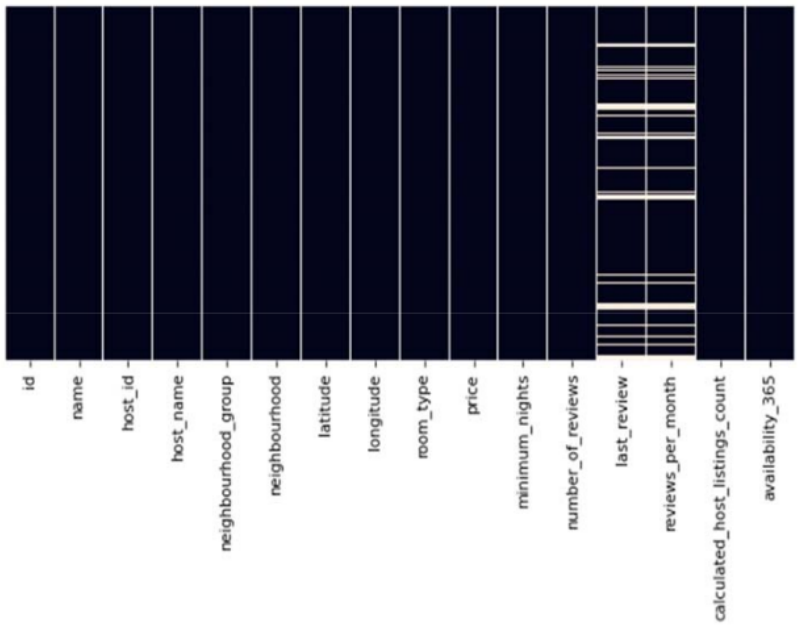
The following equation will represent the linear regression model to fit the data.

$$y = mx1 + mx2 + mx3 + mx4 + \ldots . + mx12 + c \qquad (4.2.1.1)$$

However, we removed the duplicates and used the remaining data for the analysis. The data shape after removal was (443727, 16). Columns such as "name," "host name," "last review," and "reviews per month" had a few number of missing values. Table 4.2.2 and Fig. 4.2.1 shows the missing value counts in different variables.

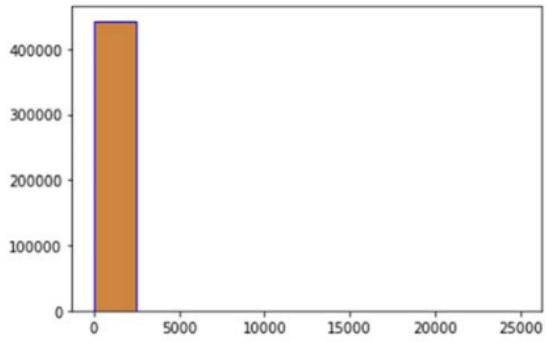
Variables	Missing Value
id	0
Name	73
Host_id	0
Host_name	271
Neighbourhood_group	0
Neighbourhood	0
Latitude	0
Longitude	0
Room_type	0

Price	0
Minimum_nights	0
Number_of_reviews	0
Last_review	68197
Reviews_per_month	68198
Calculated_Host_listings_count	0
Avaliability_365	0



The columns “last review,” and “reviews per month” had almost similar number of missing values clearly because reviews per month can be calculated using “last review.” The treatment of these missing values are discussed in the section 4.3 Data Preparation.

Most of the numeric columns were skewed to the left including the price variable. The maximum price was 25000$ which mostly skewed the price variable. The mean and standard deviation was 154.87$ and 378.26. Fig 4.1.2 shows the bar plot of the target variable price.



90% of data points were below the 250$ and 99% of value were below 800$. On further investigation it was found that only three properties belonging to Manhattan had a price of 25000$. The variable showed positive skewness with a value of 22.21 and kurtosis 626.67.

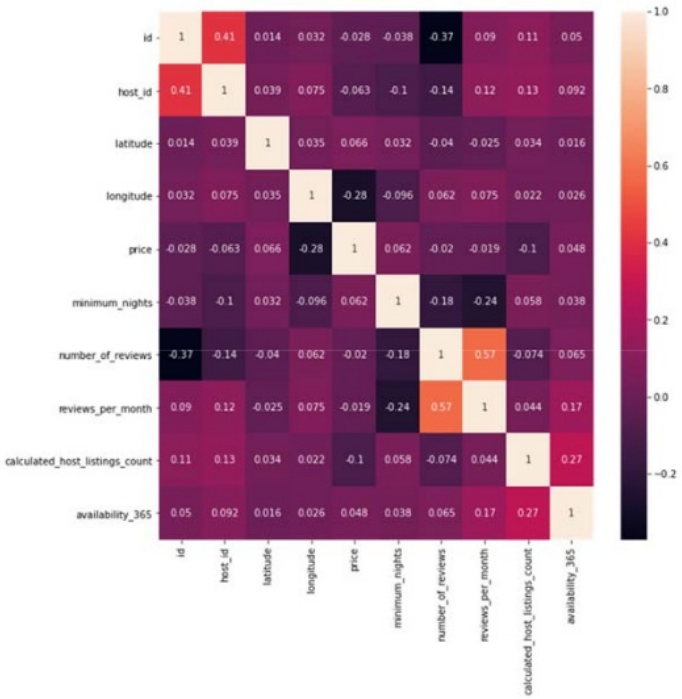
The variable minimum nights had a maximum value 1250 which was a very abnormal value. A minimum number of reviews was 0 and maximum was 748 with a mean value of 29.8. The column "reviews per month" had a minimum value of 0 and maximum value 66.36. The "calculated hosts listing count" had a minimum value 0f 1 and maximum of 307 which was again an anomaly. The variable "availability 365," had a minimum value of 0 and maximum of 365.

The variables present in the dataset did not show multicollinearity as shown in the Fig 4.2.3.

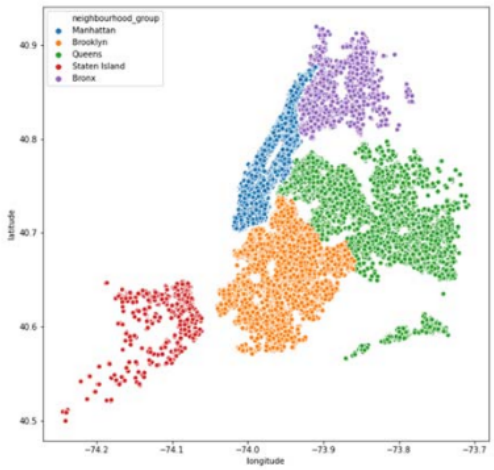
Many authors suggested that multicollinearity may be a serious issue in regression problems (Yang et al., 2016) hence to check this issue we first created the correlation plot using the method known as "Kendall." The method measures the correlation as a distribution free test of independence and measures the strength of relation between two variables. The table 4.2.3 shows the correlation of a variable with the target variable.

Variables	Correlation
id	0.02736
Host_id	0.047812
Latitude	0.027291
Longitude	-0.098794
Price	1
Minimum_nights	0.00094

Number_of_reviews	-0.038269
Reviews_per_month	-0.013936
Calculated_Host_listings_count	0.011229
Avaliability_365	0.055674



After the filtering of the data a total of 443727 records were remained. The scatter plot of different neighbourhood groups using the longitude and latitude value is plotted in the Fig 4.2.4. The figure shows the geographic distribution of the listings.



4.3 Data Preparation

One of the most important and crucial stage during the entire analysis was the data preparation stage. The first model performance was not well enough as the data was not prepared for modelling. The most time was consumed during this stage. The reason is that each dataset from January 2020 to December 2020 was a rich collection of data points and highly specific to the study. The step involved exploiting the data well to uncover the underlying trends and exposing the unknowns. Hence this step was done very carefully and a lot of trial and error was also performed to understand each and every column well enough. The detailed steps are discussed in the sub sections.

4.3.1 Data Elimination

Initially the data contained 16 attributes as listed in the data description section both object, integer and float type variable. The variables multicollinearity was checked however there was no correlation present among variables hence vif was not tested further. The name of the listings were initially used to find the more about the price variables. However, during the model building this variable was removed. Further variables such as “id” and “last review” were removed. The variables “host name” and “host id” were used to find the most popular hosts and the count the number of reviews they have. In all we have removed 4 attributes from the dataset.

The duplicates were also removed from the data before the pre-processing steps.

The outliers from each columns were analysed and removed so that our ML model is not bias.

4.3.2 Data Transformation into Categorical Variables

Columns such as neighbourhood group, neighbourhood, and room type were used in the predictive model. These columns cannot be directly used to model the data as the model requires the data to be in numeric format. Initially we used one hot encoding technique to create the dummy variables. One-hot encoding is a technique were the integer is encoded as binary elements 0 and 1, this binary variable is called a dummy variable. Further, we dropped the neighbourhood as the information it was providing was similar to neighbourhood group. Hence, to reduce the dimension this variable was dropped. However, this transformation did not result in a good model.

For the next model, we used label encoder to encode the variable "neighbourhood" and "neighbourhood_ group." For the variable "room type" we used one hot encoding. The table 4.3.2.1 shows the categorical variable transformation.

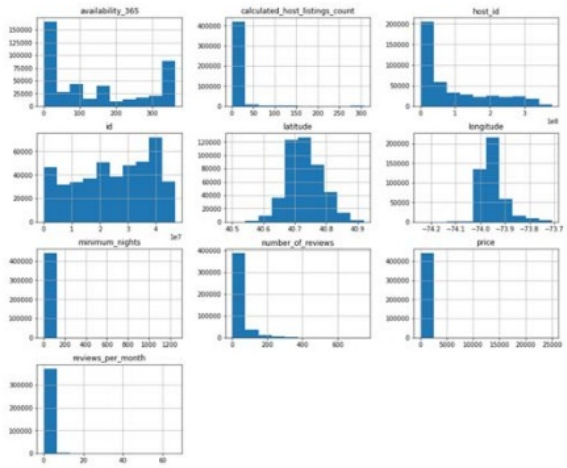
Neighbourhood_group	neighbourhood	room_type_Entire home/apt	room_type_Hotel room	room_type_Private room	room_type_Shared room
2	129	1	0	0	0
1	41	1	0	0	0
2	139	1	0	0	0
1	13	0	0	1	0

4.3.3 Identification of Missing Values

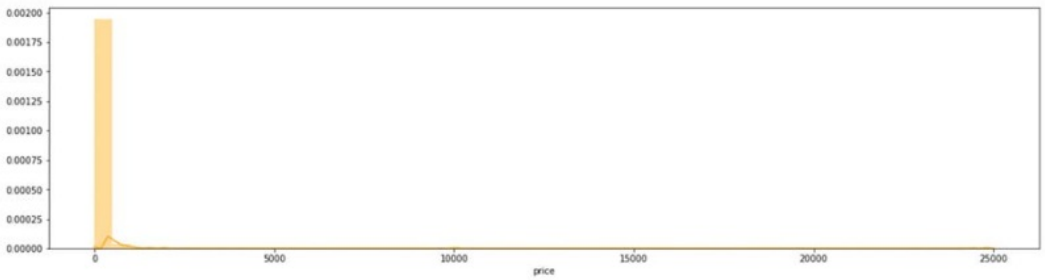
Missing values needs to identified and handled carefully to avoid its effect in the analysis. In this data set there was a very less amount of missing data present. The name columns missing data was filled with an empty string. The host name variable was removed from the column hence missing value imputation would not have made sense here. The id column was also removed from the dataset as it would not have made sense to use the variable to build the model. The last review and reviews per month had almost similar number of missing data. As it was evident from the fact the both the variable are correlated. The variable "last review" was removed and we have imputed "0" for the reviews per month variable.

4.3.4 Univariate Analysis

One of the simplest way to visualize and understand data is by using univariate analysis. Here, we analyse one variable at a time and find out more about the distribution of the data. This analysis does not deal with causes and effects of a relationship. It is mainly used to take in data, summarize the data and find pattern in the data. The fig. 4.3.4.1 shows univariate analysis for the numerical variables.



As observed that most of the variables are skewed towards right and is quite away from normal distribution including the price variable.

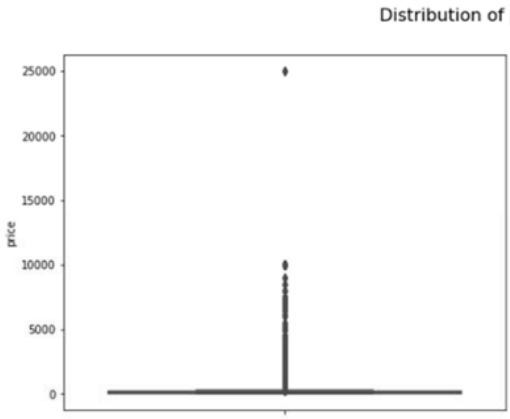


skewness: 22.218865828748672

kurtosis: 626.6701528278874

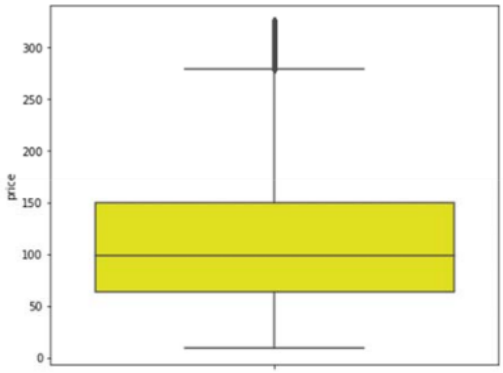
The fig. 4.3.4.2 shows the price variable which has a long tail and is positively skewed before any transformation. Few of the anomalies which was in the target variables were removed before any outlier test, they were as follows:

1. Removal of data points which priced 25000$
2. Removal of data points which priced 0$



	price
count	443727.0
mean	154.87
std	378.26
min	0.0
25%	65.0
50%	100.0
75%	169.0
max	25000.0

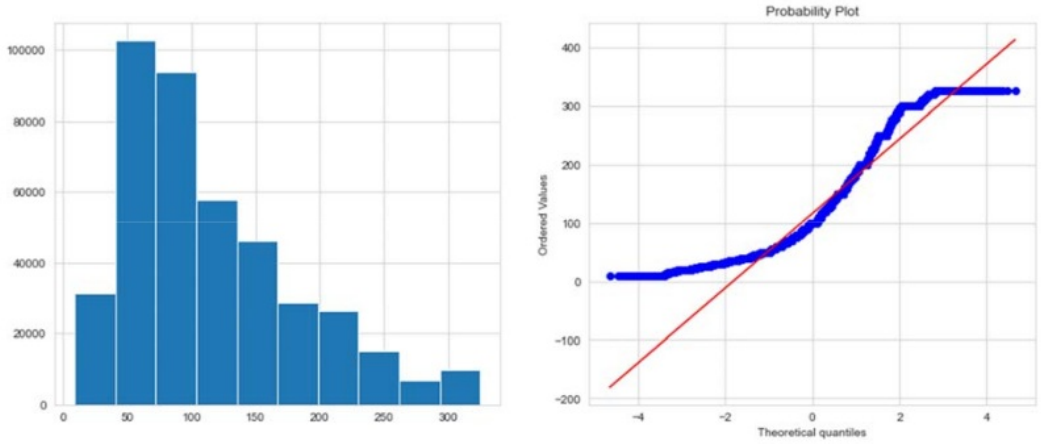
The fig. 4.3.4.3 shows the price using boxplot before removing the anomalies. The mean price is below 200$ which shows the properties are quite affordable. However, not surprised to find out the difference between the minimum and maximum value is high. But minimum price being 0 is not something normal.



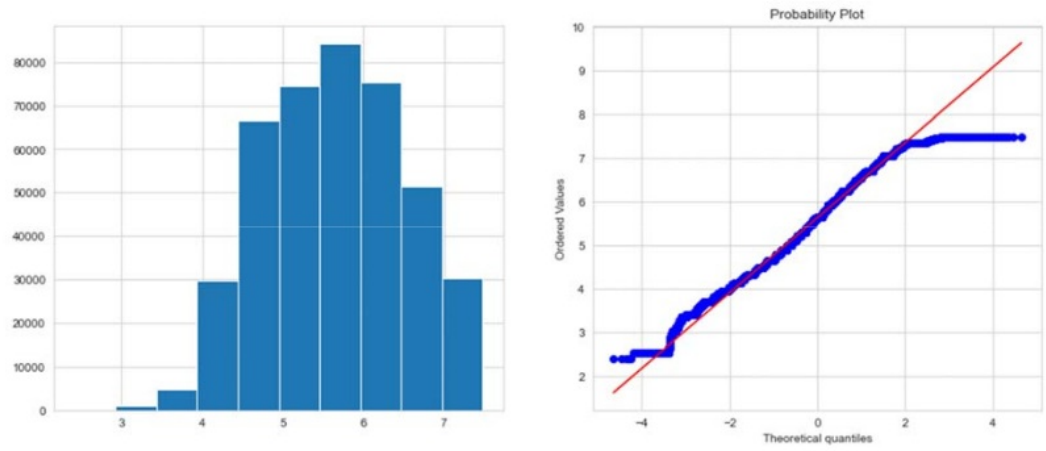
	price
count	418301.0
mean	115.7
std	66.66
min	9.0
25%	64.0
50%	99.0
75%	150.0
max	325.0

The fig. 4.3.4.4 shows after removing the anomalies from the price variable.

Further, we used Q-Q plot find out if the variable follow normal distribution after the transformation. The fig 4.3.4.5 shows the plot.



To transform the target variable, we used box-cox transformation from the library scipy stats which helped to transform the variable. The fig 4.3.4.6 shows the price variable after the transformation.

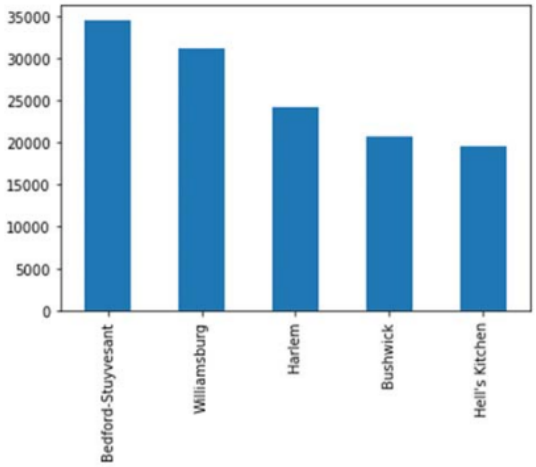


The remaining numerical variables were similarly transformed using box cox transformation and yeo-johnson. Power transformation is a method which transforms the variable to make the data look more gaussian like. This is useful because as we are using regression to model our data. To satisfy the regression assumption where heteroscedasticity is desired. Currently, power transformer supports both box cox and

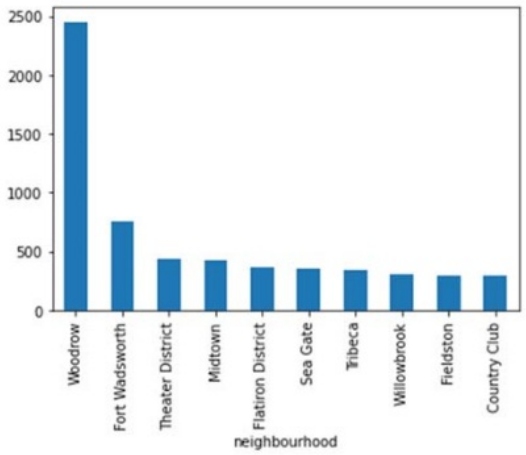
yeo-johnson transformation. The table 4.3.4.1 provides the details for skewness and minimum values before any transformation on the variables.

minimum_nights Min value: 1	Skewness: 14.94
calculated_host_listings_count Min value: 1	Skewness: 06.70
number_of_reviews Min value: 0	Skewness: 03.41
reviews_per_month Min value: 0	Skewness: 03.98
availability_365 Min value: 0	Skewness: 00.51

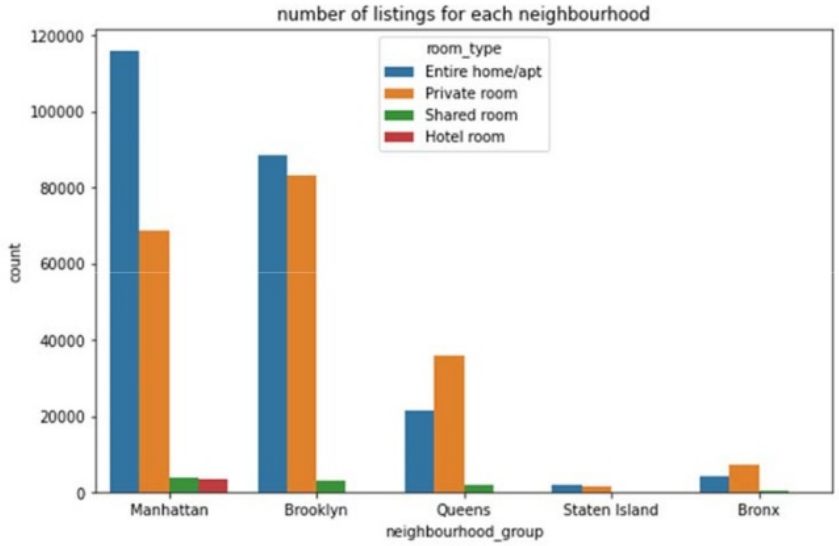
The neighbourhoods such as Bedford-Stuyvesant, Williamsburg, Harlem, Bushwick, Hell's Kitchen are the top 5 areas with most listings available. The plot 4.3.4.7 shows the distribution.



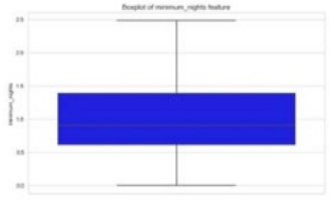
Further, the top 5 most expensive neighbourhoods were Woodrow, Fort Wadsworth, Theater District, Midtown, Flatiron District. The plot shows the expensive neighbourhoods in NYC.



The count plot of the neighbourhood group showed Manhattan has the greatest number of listings with the maximum number of listings being as entire home/apt as the room type. Manhattan followed by Brooklyn has the largest count. fig. 4.3.4.8.



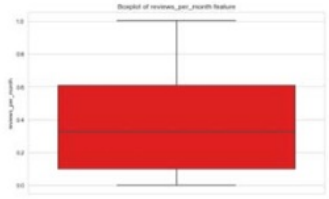
	price
count	418301.0
mean	5.64
std	0.87
min	2.42
25%	4.99
50%	5.63
75%	6.25
max	7.48



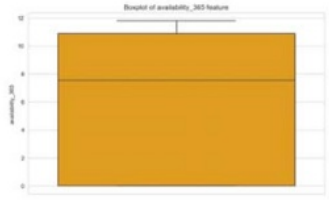
	minimum_nights
count	418301.0
mean	0.91
std	0.67
min	0.0
25%	0.61
50%	0.9
75%	1.38
max	2.48

	calculated_host_listings_count
count	418301.0
mean	0.29
std	0.37
min	0.0
25%	0.0
50%	0.0
75%	0.67
max	1.0

	number_of_reviews
count	418301.0
mean	2.1
std	1.38
min	0.0
25%	1.06
50%	2.15
75%	3.21
max	5.49



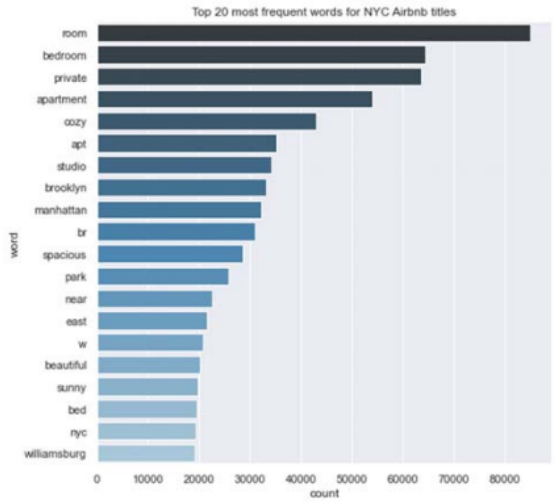
	reviews_per_month
count	418301.0
mean	0.36
std	0.28
min	0.0
25%	0.1
50%	0.33
75%	0.61
max	1.01



	availability_365
count	418301.0
mean	6.26
std	4.73
min	0.0
25%	0.0
50%	7.53
75%	10.89
max	11.78

The boxplot distributions of the variables after the transformation are shown in the fig. 4.3.4.9.

We further analysed the name of the listings separately to find out if there exist any relationship with the price variable. We used NLP techniques to find out the top most common words used for the titles to describe the properties. It was found that most common words were the words used to describe the room type to describe the apartments. The fig 4.3.4.11 shows the top 20 most frequent words.



4.3.5 Splitting of original dataset

One of the first thing we do when we start to build the ML model is to decide upon how to utilize the entire dataset. The technique used to split the data into group referred as training and test data. The training set is used to train the data and create the model. This is the reason regression problems are referred to as supervised ML technique. The remaining data or the test data is used to test the model performance.

We have used the "train_test_split" object from the module sklearn to split the dataset into 70-30, were 70% data was used as training data and 30 % is used as test data. For the Next iteration we have used 80-20 split.

4.4 Exploratory Data Analysis (Bivariate Analysis)

Bivariate analysis involves two variables to determine the relationship between them. Similar to univariate analysis, bivariate analysis can be both descriptive or inferential.

The analysis is a special case of multivariate analysis where multiple relations are examined. In this analysis we have used different plots to find out the relationship between the variables such as boxplots, scatterplots, bar plots, and violin plot.

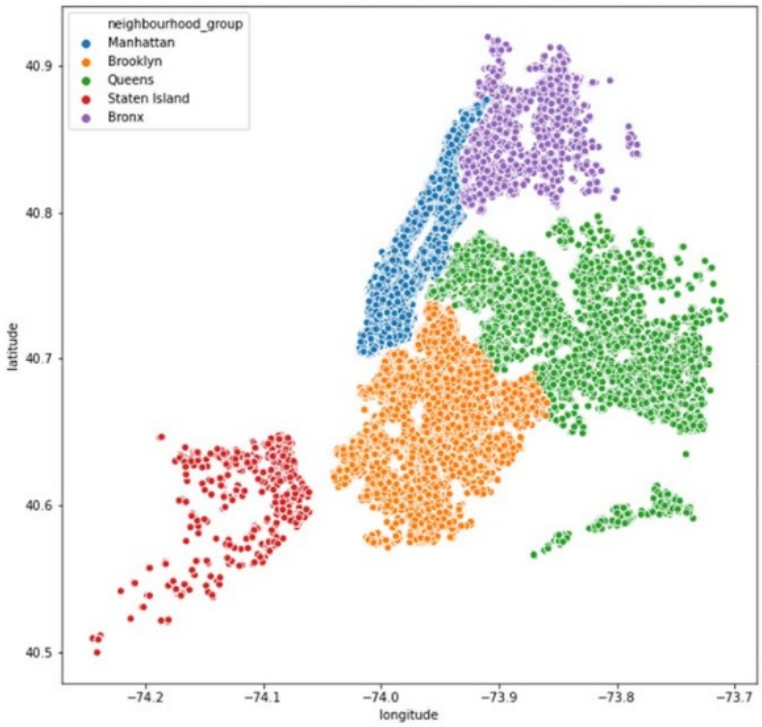
4.4.1 Correlation Plot

The correlation plot helps to understand if there exist any multicollinearity between the variables apart from the target variables. However, it was seen that there was no collinearity was present. The plot 4.2.3 and Fig. 4.2.3 shows the correlation plot of the variables.

4.4.2 Multiple Plots to Understand the Relationship

The scatter plot are useful to understand the relationship between two variables which can be plotted in both the axis.

To make use of the latitude and longitude variables we have used the data to plot a scatterplot. We used the spatial information to understand the distribution of the different neighbourhood groups in the New York City. Fig 4.4.2.1 shows the plot latitude vs longitude.



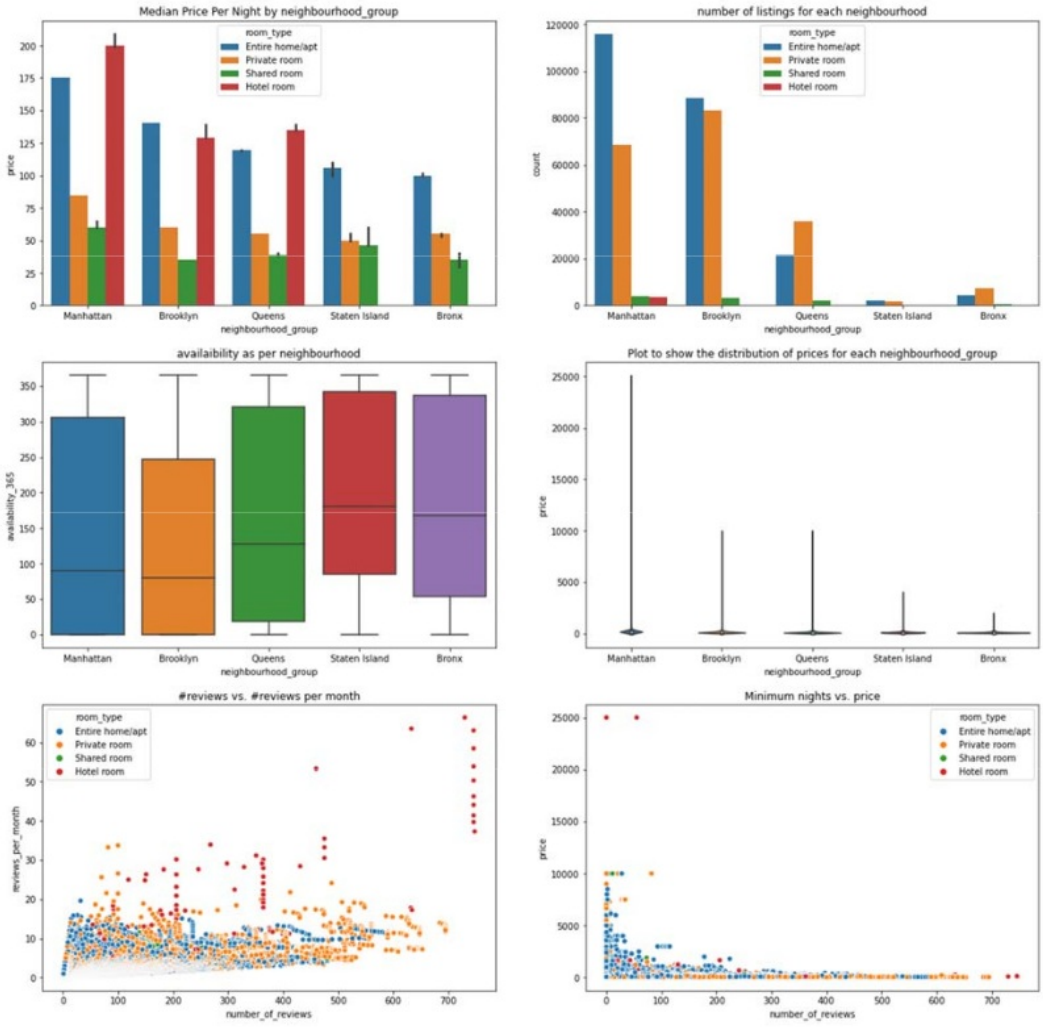


Fig 4.4.2.2 shows different plots for the variables.

1. The plot shows the relative pricing amongst each neighbourhood group and the pricing among each room type. We can see that Manhattan is the most expensive area and hotel room are the most expensive in all the neighbourhoods.
2. Plot 2 shows the count of the listings per neighbourhood Manhattan, Brooklyn, Queens are the places where most listings happen.

3. Plot 3 shows availability of the listings round the year. Brooklyn has the lowest availability and the plot shows that Manhattan and Brooklyn are the busiest.
4. Plot 4 shows the price distribution however the plot is not very much interpretable due to the presence of outliers however we have analysed this variable separately.
5. Plot 5 shows the relationship between number of reviews and reviews per month. The plot shows a slight correlation between them.
6. Plot 6 shows the relationship between number of reviews and the target variable price. It is noticed that high price variable has low number of reviews and less expensive properties have high number of review. Further, the room type for these properties are private rooms.

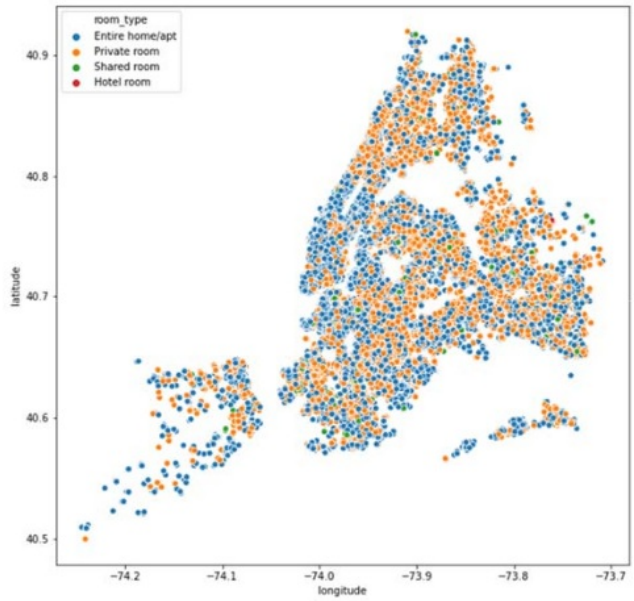
4.4.3 Room Type Distribution

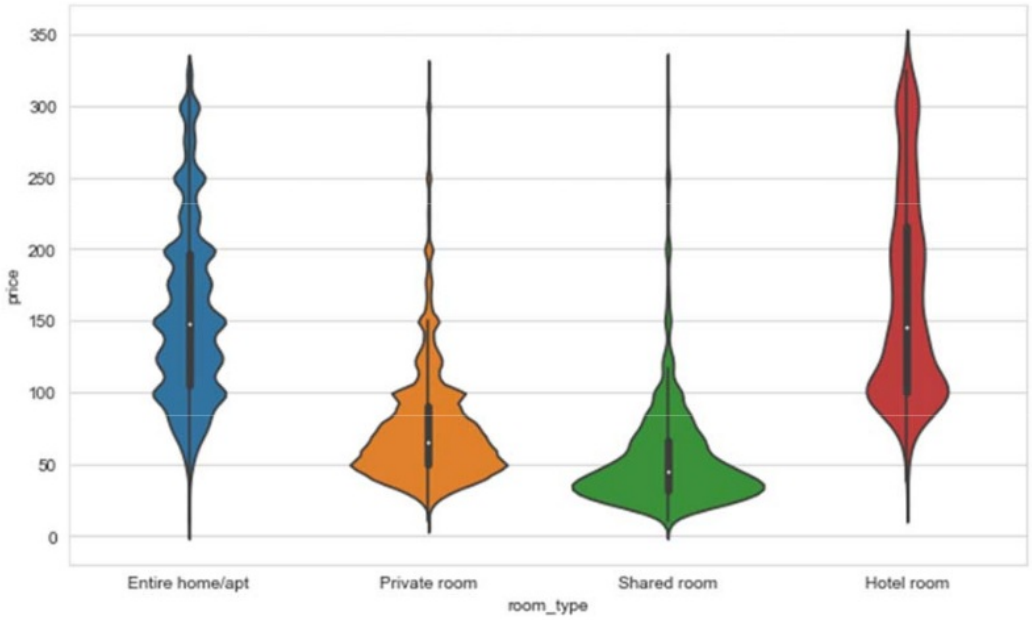
The Airbnb properties are distributed among 4 room types which are entire home/apt, hotel room, private room, shared room. The price of the entire home/apt and hotel rooms are expensive compared to private rooms and shared rooms. The average price of the room types are provided in the table 4.4.3.1.

room type	
Entire home/apt	154.852985
Hotel room	166.125076
Private room	75.349172
Shared room	54.662529
Name: price	

Further, we plotted latitude and the longitude variables to find the distribution of the room types. It was also found that the count of entire home/apt and private room are higher compared to other types.

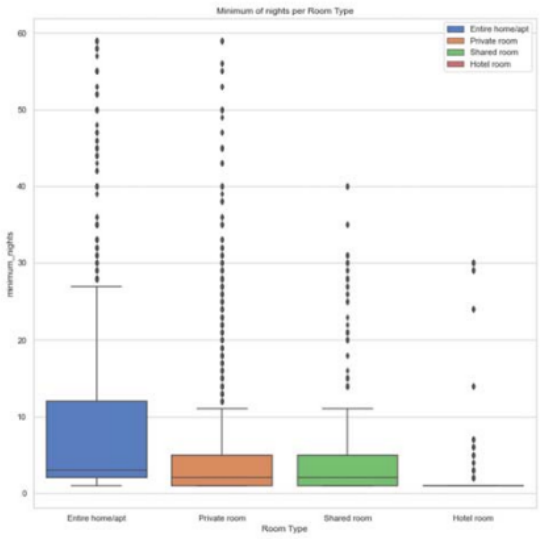
Fig 4.4.3.1 shows the distribution room type and Fig 4.4.3.2 shows the price distribution among the room types.





The price of the hotel room is higher compared to other room type. Also, the plot shows presence of outliers in the data.

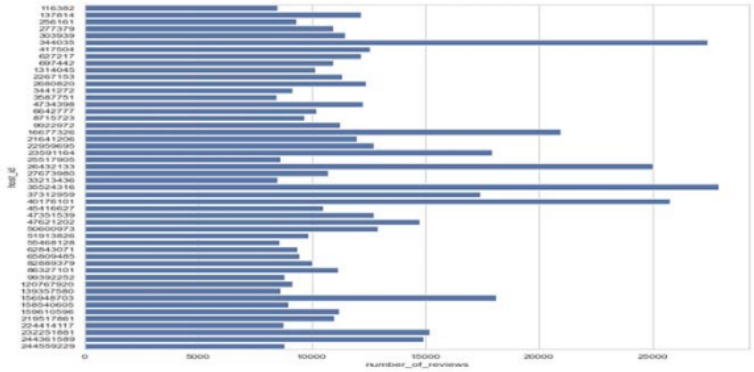
Further, the variable was examined with minimum number of nights variable which was plotted as a boxplot. Fig 4.4.3.3 shows the boxplot.



The plot shows that the entire home/apt has the highest number of night bookings. The private rooms and shared room types have almost the same numb er of minimum nights booking. The plot shows the present of outliers in the data.

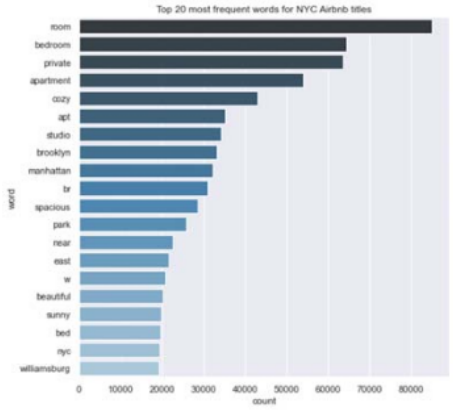
4.4.4 Number of Reviews analysis using Bar plots

Number of reviews column has been analysed using the host id to find out the greatest number of reviews received by the host. It was found that few of the hosts received more than 20,000 reviews. The fig. 4.4.4.1 shows the number of reviews per hosts.



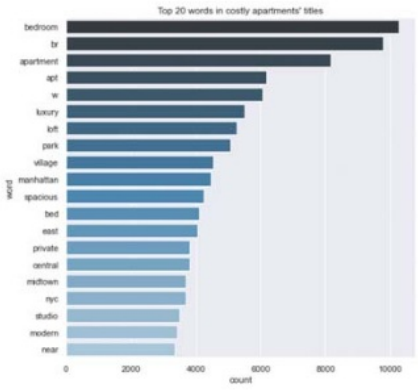
4.5 Natural Language Processing

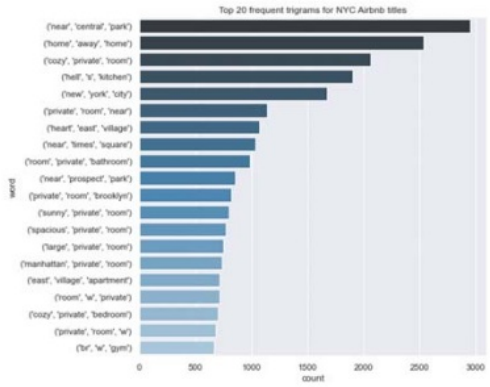
To analyse the name of the listings we used certain techniques of NLP to derive some information regarding this column. The initial steps included removal of the punctuation marks, certain digits and special characters. Further, we removed the stopwords and converted the data to tokens. The token were used to find the frequency of the terms used to name the properties. Also, these token are used to build the N-grams of words. The frequency of each term was used to analyse the most common words. The plot is shown in fig 4.5.1. It was observed that most



frequent words were the room type of the properties.

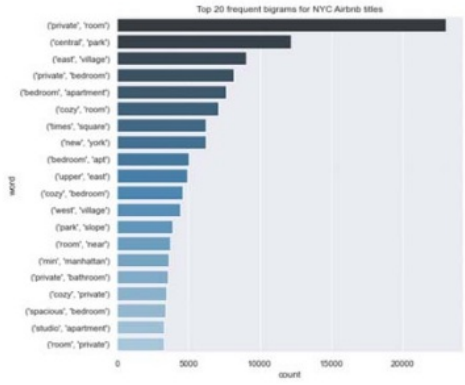
We further found the bigrams and trigrams which was most used most commonly. The use of N-grams technique are the traditional statistical techniques of text analysis. N-grams are used to find the sequence of the words. The top 20 frequently used bigrams and trigrams are plotted in the fig 4.5.2 and 4.5.3.



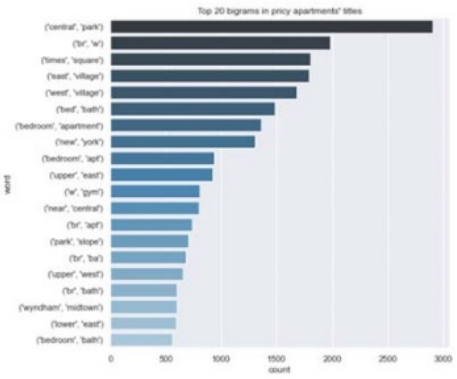


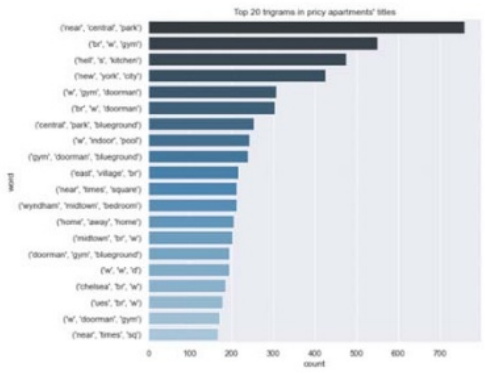
We then analysed the price variable again to find out the most common words used for the expensive places. We used the properties which was above the 200$. The mean value of the price was 153$ hence we took 200$ which was considered expensive apartments.

We found the top 20 words used to describe the expensive apartments are shown in the fig 4.5.4



The most frequent words were bedroom, apartment, luxury, Manhattan, etc., Also, we analysed the top 20 bigrams and trigrams used to describe the expensive apartments. The fig 4.5.5 and fig 4.5.6 shows the plots.





With the name variable we constructed a word cloud. Word clouds are graphical representation of words, were the frequency of the words decides the size of the word in the word cloud. Words like Private bedroom, Central Park, East Village, New York came out to be the most frequent word's.

Fig 4.5.7 shows the word cloud representation of the name of the properties.

4.6 Summary

EDA is a very essential first step to be conducted before any ML model (Rahayu, 2019). Preliminary data munging has already been conducted by the site Inside Airbnb to provide the users to conduct any research and analysis and present their findings. The data has been scraped for the year 2019 and 2020 Jan to Dec. Further the data was then compiled to get a master dataset which to be used for the analysis.

We have thoroughly examined the entire 2020 dataset along with all the attributes to derive maximum information needed to build the ML model. The data visualization has been done broadly on all the important columns to find the relationship between variables.

After the exploratory data analysis, we have a brief understanding of the data set. Also, we notice there are some number of drawbacks in our data. For example, the presence of outliers, almost all the variables were skewed and needed transformations, the variables were all present in different units, the presence of missing values. The skewness of the price variable was removed using tukey test, further the variable was transformed using box cox transformation to fit the normal distribution. The dataset has been treated to ensure data quality, removing unwanted columns that are not important for the research, the duplicate values were removed, the dependent and the independent variables had zero missing data after the processing.

The outliers were also removed from the dataset before the beginning of the analysis. The columns such as last review date had a number of missing values and removed from the analysis. The independent variables which will provide value to the research are kept and are used to predict the pricing of the listings.

Categorical variables were encoded to be fed to the ML model. We used label encoding and one hot encoding techniques to convert these variables.

It was found in the EDA that Manhattan is the most expensive neighbourhood group in New York City in 2020. The anomaly value of price which was 25000$ were belonging to Manhattan Hotel room. In NYC 99% of values were lying below 800$ price.

The entire home/apt and hotel room were the most expensive room types.

A similar approach was conducted for 2019 data to analyse if there are any effects of Covid-19 on the pricing of the properties. The mean price for 2020 data was found to be 154$ and for the 2019 data it was 150$. Though there was not much difference in the average pricing. However, we found the price for the year 2020, the 100-

percentile value was 25000$ and for 2019, the 100-percentile value was 1000$. Manhattan and Brooklyn were the most expensive neighbourhood groups in both the years. It was also noticed that the mean price of Manhattan was less in the year 2020 than the previous year. However not much difference in average price was noticed for rest of the neighbourhoods. Compared to 2020 the 2019 data set showed less variation in the price variable.

We have a number of plots in this section to provide the details regarding the attributes present. In the next section we will discuss the ML model which was implemented on both the dataset and the results obtained from them. We will also compare the results of each of the model and find out the best ML model to predict the pricing.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Introduction

In this chapter we will go through the details about the application of the Machine Learning algorithms to find the rental prices of the Airbnb listings. We will use the techniques to build a recommendation system which will help the hosts to find the pricing for their properties and they can make adjustments to the predicted prices. The Implemented regression models are Simple Linear regression, Random Forest regressor, XGBoost regressor, and ANN with K-fold cross validation. Their performances are evaluated by $R^{2,}$ Mean Squared Error, Mean Absolute Error, and RMSE. The detailed implementation are discussed in this section.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (5.1.1)$$

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}} \quad (5.1.2)$$

$$MAE = \frac{1}{n}\Sigma|y_i - \hat{y}_i| \quad (5.1.3)$$

$$MSE = \frac{1}{n}\Sigma(y_i - \hat{y}_i)^2 \quad (5.1.4)$$

The first algorithm which was implemented was the linear regression which did not provide a good score. Further, we did the feature engineering again and implemented the algorithms.

We used OLS model from library stats model in the third iteration for Linear regression. The model performance increased after feature engineering.

Out of all the Random Forest proved to be the best one. However, there was a slight overfitting issue.

The paper investigates the relationship between the dependent variable and the independent variables (room type, availability, neighbourhood, minimum nights, etc.,)

Finally we implemented ANN to understand how ANN predicts the price of the listings. Further, to compare the results with other algorithms. We further implemented KNN and Light GBM to understand how the model works. Although or analysis was based on 2020 dataset, however we scrapped the 2019 data to study if there are any new trends or pattern. There was no notable difference found in 2019 dataset however there was a slight variation in the price variable.

Out of all the ML model Random Forest and XGBoost performed well compared to other ML algorithms.

Neighbourhood Group
Neighbourhood
Latitude
Longitude
Minimum Nights
Number of Reviews
Reviews Per Month
Calculated Host Listings
Availability 365
Room Type Entire Home/Apt
Room Type Hotel Room
Room Type Private Room
Room Type Shared Room

Table 5.2.1.1

5.2 Simple Linear Regression

The first we created was a simple linear regression model which was built after we standardised the numerical features. The metric scores which we received are provided in the table 5.2.1.

Mean Squared Error	0.7454286170108727
R2 Score	44.43361769412577
Mean Absolute Error	0.5543928947975881

The R square was very low and below 50 which showed that the independent variables was able to explain only 44% of the variation of the dependent variable. Higher R square is always better.

The object type variable were encoded using one hot encoding. The numerical variables were transformed using standard scaler from sklearn. Standard scaler standardizes the features with mean equal's zero and unit variance. The formula used for standardizing the variables is provided in the equation 5.2.1.1

$$z = (x - \mu) / s \qquad (5.2.1.1)$$

The coefficient of the variables were found to be 0.04850079, -0.20713002, -0.06433568, -0.00690147, -0.02277242, 0.0048468, 0.10820096, 0.08055158, 0.2385045, 0, 0, 0, -1.14927357, 0.

5.2.1 Simple Linear Regression using Stats Model

We used Simple Linear regression model from statsmodel to build a simple linear regression model. The model was created after feature transformation which is discussed in section 5.3. The columns used for the analysis are provided in table 5.2.1.1.

In order to use the stats, we need to manually add a constant which is equal to 1. The OLS regression model results are provided in table 5.2.1.2.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.497
Model:	OLS	Adj. R-squared:	0.496
Method:	Least Squares	F-statistic:	2.750e+04
Date:	Thu, 21 Jan 2021	Prob (F-statistic):	0.00
Time:	12:28:27	Log-Likelihood:	-3.1177e+05
No. Observations:	334640	AIC:	6.236e+05
Df Residuals:	334627	BIC:	6.237e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-245.7543	1.508	-162.969	0.000	-248.710	-242.799
neighbourhood_group	0.0387	0.001	27.076	0.000	0.036	0.041
neighbourhood	0.0006	1.59e-05	34.825	0.000	0.001	0.001
latitude	1.0313	0.020	51.842	0.000	0.992	1.070
longitude	-3.6598	0.022	-165.211	0.000	-3.703	-3.616
minimum_nights	-0.1426	0.002	-78.794	0.000	-0.146	-0.139
number_of_reviews	0.0145	0.001	10.135	0.000	0.012	0.017
reviews_per_month	-0.2050	0.007	-27.351	0.000	-0.220	-0.190
calculated_host_listings_count	-0.1036	0.003	-31.884	0.000	-0.110	-0.097
availability_365	0.0163	0.000	65.783	0.000	0.016	0.017
room_type_Entire home/apt	-60.7348	0.377	-161.145	0.000	-61.474	-59.996
room_type_Hotel room	-60.8673	0.378	-161.110	0.000	-61.608	-60.127
room_type_Private room	-61.8039	0.377	-164.085	0.000	-62.542	-61.066
room_type_Shared room	-62.3483	0.377	-165.456	0.000	-63.087	-61.610

Omnibus:	4542.286	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6828.771
Skew:	0.152	Prob(JB):	0.00
Kurtosis:	3.631	Cond. No.	2.63e+17

Here, as we can see the p-values associated with all the variables are significant. The R-squared and Adjusted R-squared have slightly increased in this model. Next, we checked if there exist any multicollinearity between variables using vif.

Few of the features vif values were noted to be infinity as provided in table 5.2.1.2. An infinite vif indicates that the variable can be expressed as the combination of other variables. The Eq. 5.2.1.1 provides the vif equation. The table 5.2.1.2 shows the VIF of different variables.

$$VIF = \frac{1}{1 - R_i^2} \qquad (5.2.1.1)$$

Features	VIF
room_type_Entire home/apt	inf
room_type_Hotel room	inf
room_type_Private room	inf
room_type_Shared room	inf
reviews_per_month	3.82
number_of_reviews	3.45
minimum_nights	1.31
calculated_host_listings_count	1.26
availability_365	1.22
latitude	1.11
longitude	1.10
neighbourhood_group	1.09
neighbourhood	1.09
const	0.00

In the next iteration we dropped the variable room type entire home. The reason to drop the variable are: -

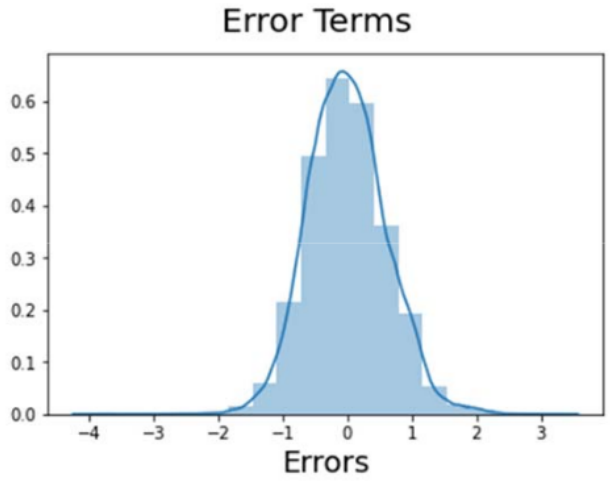
1. The variable room type was transformed using one hot encoding. However, we did not drop one column which caused the curse of dimensionality.
2. The vif value was infinity.

After dropping the column, the R-squared and Adj R-squared value did not change however the vif value came down as shown in the table 5.2.1.3.

room_type_Entire home/apt	11.75
room_type_Private room	11.46
reviews_per_month	3.82
number_of_reviews	3.45
room_type_Hotel room	1.36
minimum_nights	1.31
calculated_host_listings_count	1.26
availability_365	1.22
latitude	1.11
longitude	1.10
neighbourhood_group	1.09
neighbourhood	1.09

The error terms were found to be normally distributed with mean equal to zero satisfying the assumption of SLR.

One repercussion of the error term not being normally distributed is that the p-values obtained during the hypothesis test which determines the significance of the coefficients becomes unreliable. The fig 5.2.1.1 shows the plot for the error term.



5.3 Predictive Modeling using Different Techniques

In the next iteration we analysed the features variables again and transformed the variables to find the best model. After removing the outliers using tukey's text, we analysed the numerical variables using Q-Q plot. The variables were found to be quite away from the normal distribution including the price variable. The table 5.3.1 and 5.3.2 provides the skewness and the Min value of the numerical columns after removing the outliers for 2020 and 2019 dataset.

price	Skewness: 00.98	Min value: 9
minimum_nights	Skewness: 14.94	Min value: 1
calculated_host_listings_count	Skewness: 06.70	Min value:1
number_of_reviews	Skewness: 03.41	Min value: 0
reviews_per_month	Skewness: 03.98	Min value: 0.0
availability_365	Skewness: 00.51	Min value: 0
price	Skewness: 00.92	Min value: 10
minimum_nights	Skewness: 20.60	Min value: 1
calculated_host_listings_count	Skewness: 08.15	Min value: 1
number_of_reviews	Skewness: 03.28	Min value: 0
reviews_per_month	Skewness: 02.79	Min value: 0.0

availability_365	Skewness: 00.54	Min value: 0

We used two methods yeo-johnson and box-cox transformation from Power Transformer to transform the variables. The variables which were transformed are price, minimum-nights, calculated host listings count, number of reviews, reviews per month, availability 365 days.

Since some of the features have got values near to 0, which causes error in boxcox transformation hence we calculated lambda coefficient manually and then used box cox transformation.

The lambda values of the variables are listed in the table 5.3.3 and 5.3.4 for 2020 and 2019 data.

“price lambda":	0.08544784980940612
"minimum_nights lambda":	-0.37577910280344384
calculated_host_listings_count:	-0.9952732586928884
"number_of_reviews lambda: [-0.05825883]" #Transformed by calculating lambda manually	
"reviews_per_month lambda: [-0.97882508]" #Transformed by calculating lambda manually	
"availability_365 lambda: [0.21229905]" #Transformed by calculating lambda manually	
"price lambda":	0.0895143049332677
"minimum_nights lambda":	-0.47759482815426113
“calculated_host_listings_count”:	-1.0586907796163991
"number_of_reviews lambda: [-0.04402662]" #Transformed by calculating lambda manually	

The categorical variables were transformed to encoded variables. We used label encoder to convert the neighbourhood and neighbourhood group variable and used one hot encoding to transform the room type variable. The data was divided into training and test data, the dimensions of the data are provided in the table 5.3.5.

Dimensions of the training feature matrix: (334640, 13)

Dimensions of the training target vector: (334640,)

Dimensions of the test feature matrix: (83661, 13)

Dimensions of the test target vector: (83661,)

We used Linear regression, Random Forest regressor, XGBoost regressor to predict the model performance. The model performance are listed in the table 5.3.6.

Model	CV error	CV std	RMSE train	RMSE test	R2 train	R2 test
Linear Regression	0.377	0.002	0.614300	0.615500	0.496500	0.494300

lin_reg. coef [3.86908352e-02, 5.54531449e-04, 1.03125049e+00, -3.65981248e+00, -1.42559144e-01, 1.44840496e-02, -2.04957747e-01, -1.03550026e-01, 1.63161932e-02, 7.03752830e-01, 5.71236498e-01, -3.65316045e-01, -9.09673284e-01]

Mean Squared Error:	0.6154675388616478
R2 Score:	49.427318859394944
Mean Absolute Error:	0.4848213702730525

Model	CV error	CV std	RMSE train	RMSE test	R2 train	R2 test
Random Forest Regressor	0.104	0.001	0.118100	0.306100	0.981400	0.874900

Mean Squared Error:	0.3060823196302204
R2 Score:	87.49216446293678
Mean Absolute Error:	0.18446876106083634

Model	CV error	CV std	RMSE train	RMSE test	R2 train	R2 test
XGBRegressor	0.174	0.001	0.364700	0.414600	0.822500	0.770500

Mean Squared Error:	0.414627056083954
R2 Score:	77.04797805386922
Mean Absolute Error:	0.3089370856440916

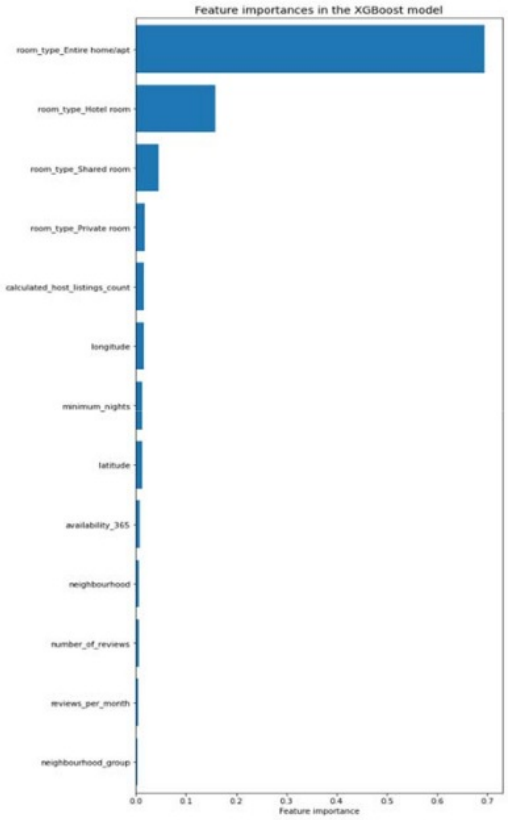
The Random Forest and XGBoost regressor showed good model performance.

The random forest model showed overfitting.

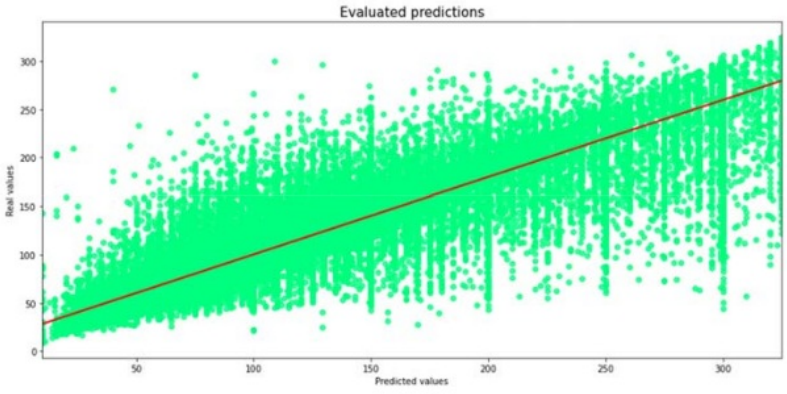
The scores were high for both the training and test data in case of random forest. However, for training set the R-square was 98%. The fig. 5.3.1 shows the plot between the real values and predicted values.

The XGBoost model is more generalised model as the results in the training set are not overfitting the model.

XGBoost calculates the feature importance on the predictive model. These importance scores are available in feature_importances_ variable for a trained model. The fig 5.3.2 and the table 5.3.3 shows the feature importance to predict the price of the properties. From the plot we can say that room type has a significant impact in predicting the price.



	weight
neighbourhood_group	0.003500
reviews_per_month	0.004769
number_of_reviews	0.005965
neighbourhood	0.006568
availability_365	0.007185
latitude	0.011800
minimum_nights	0.011849
longitude	0.015254
calculated_host_listings_count	0.015752
room_type_Private room	0.018099
room_type_Shared room	0.045329
room_type_Hotel room	0.158402
room_type_Entire home/apt	0.695528



Next, we have built an Artificial Neural Network to compare the model performance.

5.4 Artificial Neural Network

The initial artificial neural network was built using 4 layer, the first layer included 13 neurons, second layer included 13 neurons, third layer included 6 neurons and the last layer included 1 neurons.

The table 5.4.1 shows the ANN

Model: "sequential_2"

Layer (type)	Output Shape	Param #
dense_3 (Dense)	(None, 13)	182
dense_4 (Dense)	(None, 13)	182
dense_5 (Dense)	(None, 6)	84
dense_6 (Dense)	(None, 1)	7

Total params: 455

Trainable params: 455

Non-trainable params: 0

None

We have used 'ReLU' activation function in all the three layers and in the last layer we used 'linear' activation function. Activation function are mathematical equation which is used to determine the output of a neural network. This activation function is used to activate a neuron or not, based on each neurons input. The fig. 5.4.1 shows a deep ANN.

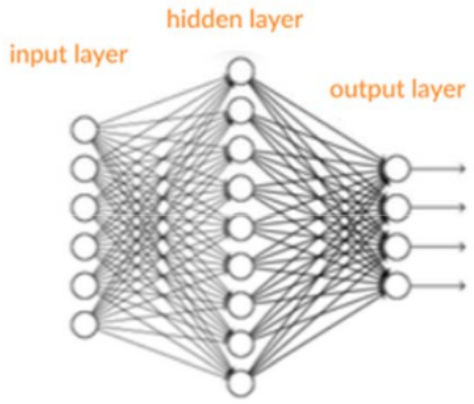


Figure 5.4.1 (Deep Neural Network, missinglink.ai)

5.4.1 ReLU Activation Function

The initial layer or the first layer consists of the features variable in this case we have used all the 13 features. Numeric data points are fed into the model and each neuron is multiplied by the weights which provides the output of the neuron. The weights are updated using a technique known as back propagation. The activation function decides whether a neuron is to be activated or not. The input is fed to the input layer, the neurons present in the layer performs the transformations using the weights and biases. As provided in the Eq. 5.4.1.1

$$x = (w^{*}input) + b \qquad (5.4.1.1)$$

Next, an activation function is applied on the above result.

$$y = Activation(\Sigma(weight * input) + bias) \qquad (5.4.1.2)$$

This output is moved to the next layer and the same process is repetitive. This forward movement is known as forward propagation.

The next question arises if it is necessary to provide an activation function. The neural network is just a simple linear regression model without an activation function. There

are a number of activation function however we have discussed few of them in this paper.

The activation functions are of two types: -

1. Linear Activation Function
2. Non-Linear Activation Function

Linear activation function are straight line hence does not have any boundaries.

The Non-Linear activation function are the most used as it adapts the variation in the data. A few examples of non-linear activation function are sigmoid, tanh, ReLu, etc.,

We have used ReLU (Rectified Linear Unit) function, the equation 5.5.1.3 represents the function.

$$f(x) = max(0, x) \qquad (5.4.1.3)$$

The neurons gets activated only if the output is more than 0. The plot 5.4.1 helps to understand the function.

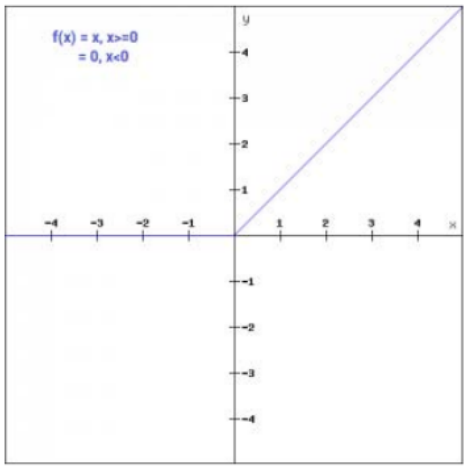


Figure 5.4.1.1 ReLU function, Analytics Vidya, 2020

5.4.2 Linear Activation Function

Since the output was a continuous value and the problem was a regression hence, we have used linear activation function in the last layer. The function is defined in the Eq. 5.4.2.1. The plot 5.4.2.1 shows the function.

$$f(x) = ax \qquad (5.4.2.1)$$

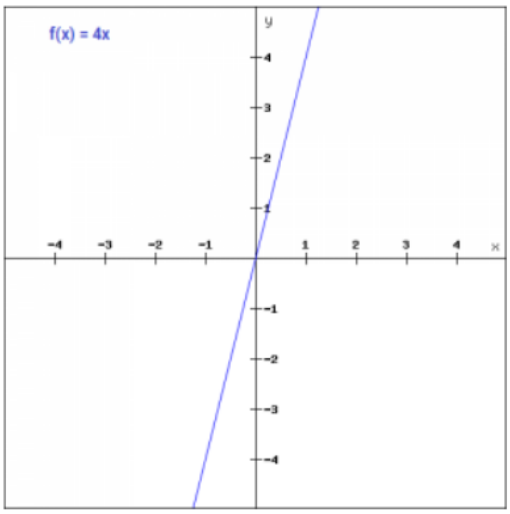


Figure 5.4.2.1 Linear function, Analytics Vidya, 2020

We have fitted the model using 100 epochs with a batch size of 256 and 0.1 validation split.

The model performance using this neural network structure are as follows: -

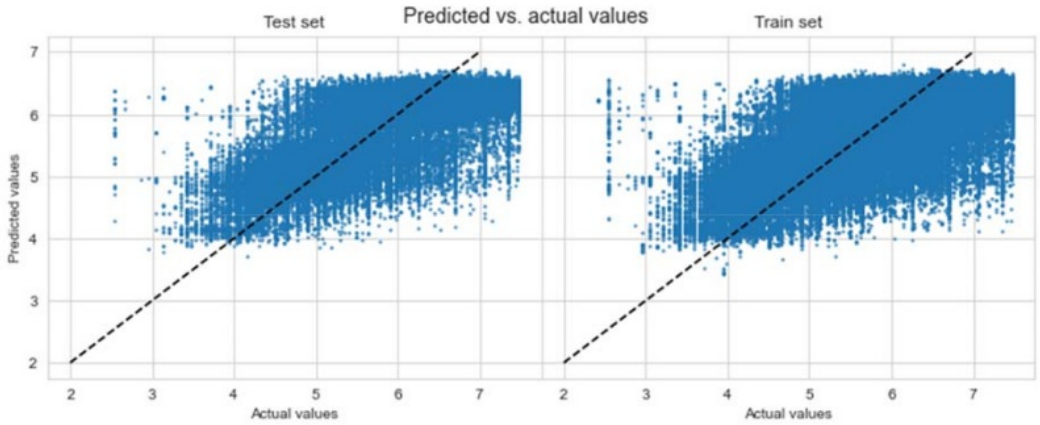
Training MSE: 0.3502

Validation MSE: 0.3514

Training r2: 0.5327

Validation r2: 0.5308

The fig. 5.4.2.1 shows the predicted vs actual plot for training and the test set.



In the next iteration, we used four-layer ANN with L1 regularization and a greater number of epochs. As shown in the table 5.4.2.1.

Model: "sequential_3"

Layer (type)	Output Shape	Param #
dense_7 (Dense)	(None, 20)	280
dense_8 (Dense)	(None, 13)	273
dense_9 (Dense)	(None, 6)	84
dense_10 (Dense)	(None, 3)	21
dense_11 (Dense)	(None, 1)	4

Total params: 662

Trainable params: 662

Non-trainable params: 0

None

L1 regularization was used to prevent overfitting as NN have a tendency to overfit if there are a greater number of hidden layers. Regularization is a technique to control overfitting and provides the model to become complex by penalizing the model. In L1 regularization we add a regularization term. Due to this term the weight matrices decreases hence making the model simpler by penalizing the absolute value of the weights. The Eq. 5.4.2.2 describes the equation.

$$Cost\ function = Loss + \boldsymbol{\lambda/2m} * (\boldsymbol{\Sigma}|\boldsymbol{w}|) \qquad (5.4.2.2)$$

The regularization allows to compress the model as it reduces few weights to zero. In keras, we can use any of the regularization technique.

We ran 150 epochs with batch size of 256 and validation split of 0.1. However, our model scores decreased.

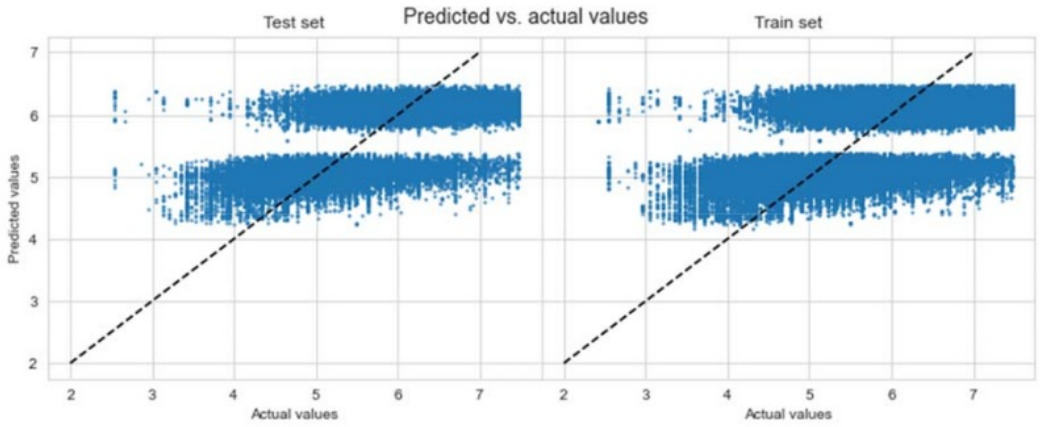
Training MSE: 0.4082

Validation MSE: 0.4097

Training r2: 0.4553

Validation r2: 0.453

The fig. 5.4.2.2 shows the predicted vs actual plot for training and test set.



5.4.3 ADAM Optimizer

To create our ANN, we have considered ADAM optimizer. The reason to consider this as our optimizer are listed down.

1. The optimizer is quite straightforward to implement
2. Requires less memory
3. It is computationally less expensive
4. Quite well suited for large scale data
5. Hyper parameter requires less tuning

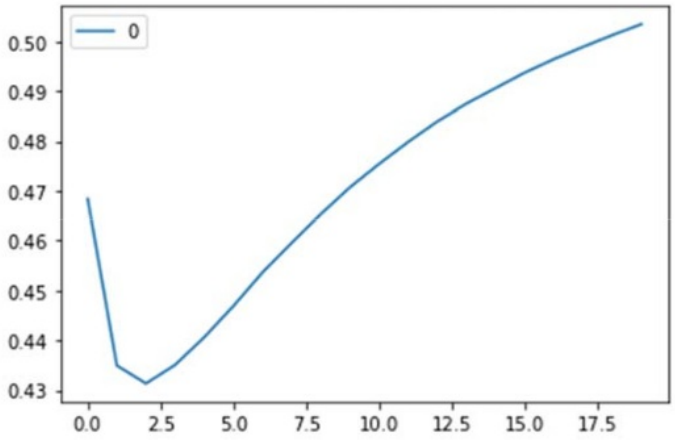
Adam optimizer is an adaptive learning rate optimization that has been designed to train deep artificial neural network. Optimizers are used to update the weights during the back propagation iteratively. Adam optimizer is a combination of RMS prop and Stochastic gradient descent with momentum. Adam optimizer is used instead of SGD. In SGD a single learning rate for all the weight is maintained further the learning rate does changes during training.

Whereas Adam is a combination of Adaptive gradient algorithm and Root mean square propagation.

"Adam is an algorithm for first order gradient based optimization of stochastic objective function, based on adaptive estimates of lower-order moments" (Kingma and Ba, 2015).

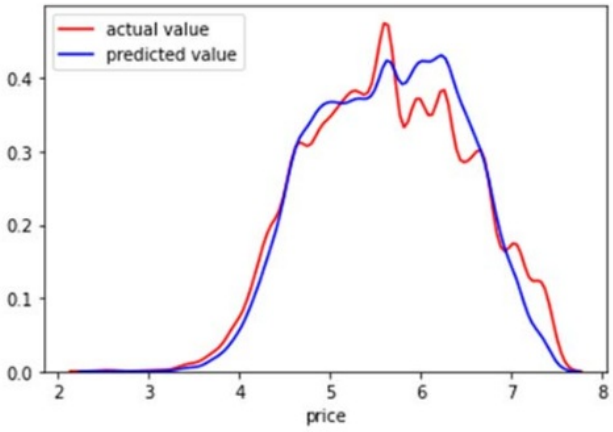
5.5 KNN for Price Prediction

The last model which we created was using KNN, the algorithm is one of the simplest algorithm among the supervised learning techniques. The algorithm is easy to implement and to understand however, the algorithm becomes slow as the data pints increases. It assumes that similar things exists in closer proximity. "Birds of same feather flock together" more or less the same principle is applied here. The similarity is calculated using Euclidean distance, Manhattan distance, Hamming distance, etc., However the most common is Euclidean distance. We used the test data to find the best value for k. The plot 5.5.1 shows the different RMSE values plotted against different k values.



The grid search results showed the best k value to be 2. However, the model performed worst by applying this value.

In the next iteration we removed the longitude and the latitude value the model performance significantly improved after removing these two columns. The best k value was chosen using elbow curve and grid search which came out to be 3. The plot 5.5.2 shows the predicted and test value.



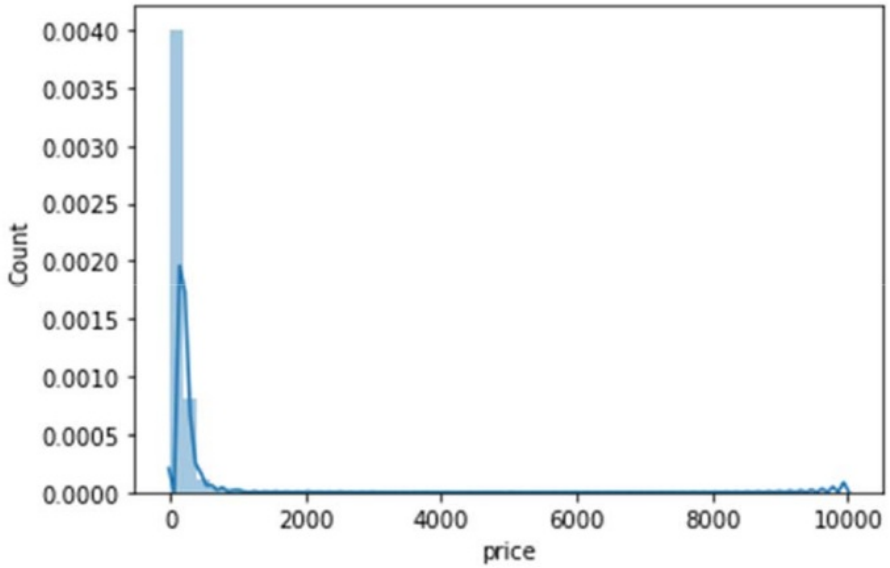
The r-square was found to be around 71% and rmse value was .465. The model was able to perform better than linear regression and was also did not show overfitting.

5.6 Comparison with the 2019 Data Set

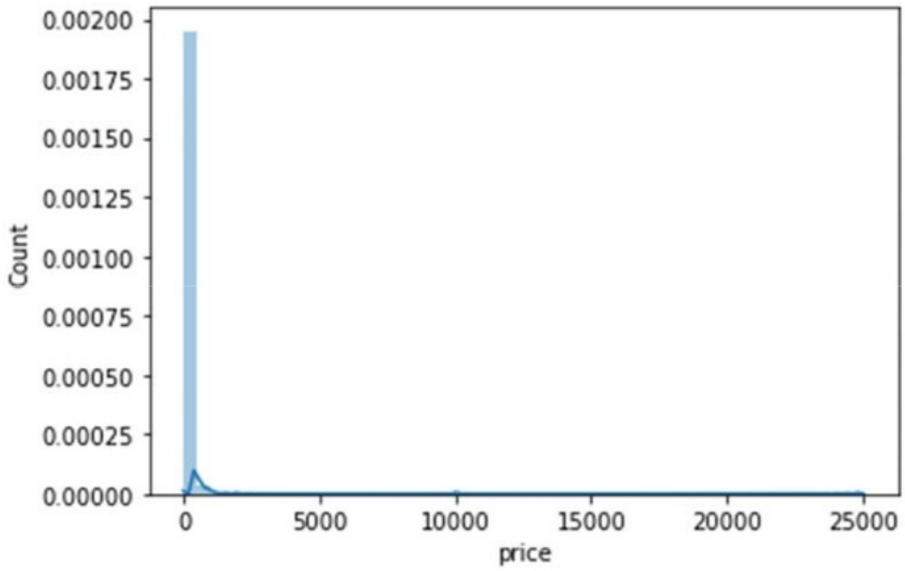
We scraped the 2019 data set from Jan to Dec in order to compare both the data set and find out if there has been any noticeable effect on prices and other factors post covid-19. Although the study was completely based on 2020 dataset however, we moved a step forward to find out about the previous year data. The descriptive statistics for the price variable is provided in the table 5.6.1.

	price
count	444746.000000
mean	150.622830
std	236.747523
min	0.000000
25%	69.000000
50%	108.000000
75%	175.000000
max	10000.000000

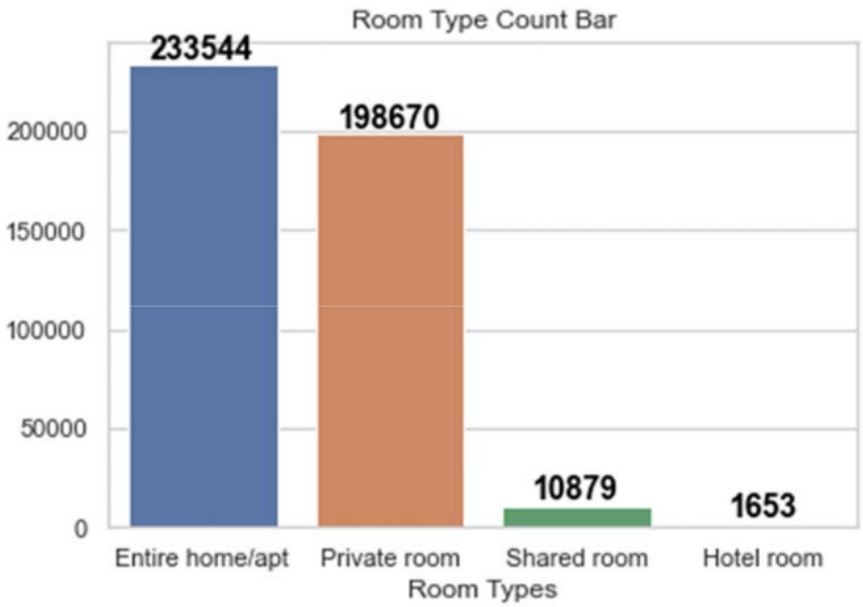
The maximum value in this case was 1000$ however for 2020 data the maximum value was 25000$. In 2020, 99% of value was below 800$ and in the case of 2019, the value was below 750$. The distribution of price variable for 2019 and 2020 data set are shown in the plot 5.6.1 and 5.6.2.



For the year, 2019, the price variable showed a slight variation compared to 2020 dataset.



Manhattan and Brooklyn were most expensive and busiest neighbourhood groups in New York City for both the years. Also, the count of the Entire home/apt was more compared to other room types.



The fig.5.6.1 shows the count plot for the room types in NYC.

Also, the hotel room and entire home/apt were the expensive room types compared to other room types.

We could not find any new trends in the 2019 data set apart from the outlier value. For both the year; 2019 and 2020 the data set followed quite similar trend.

The ML model which was implemented in this data set was Linear Regression, Random Forest, XGBRegressor, and ANN.

The results obtained are as provided in the table 5.6.2.: -

Model	CV error	CV std	RMSE train	RMSE test	R2 train	R2 test
LinearRegression	0.354	0.002	0.594600	0.595800	0.543900	0.541100

Mean Squared Error:	0.5958223186543604
R2 Score:	54.1133305237846
Mean Absolute Error:	0.4661708639382209

Model	CV error	CV std	RMSE train	RMSE test	R2 train	R2 test
RandomForestRegressor	0.070	0.001	0.090800	0.241400	0.989400	0.924700

Mean Squared Error:	0.24142804628776
R2 Score:	92.46 595096411687
Mean Absolute Error:	0.1349288223755594

Model	CV error	CV std	RMSE train	RMSE test	R2 train	R2 test
XGBRegressor	0.138	0.001	0.328800	0.368200	0.860600	0.824700

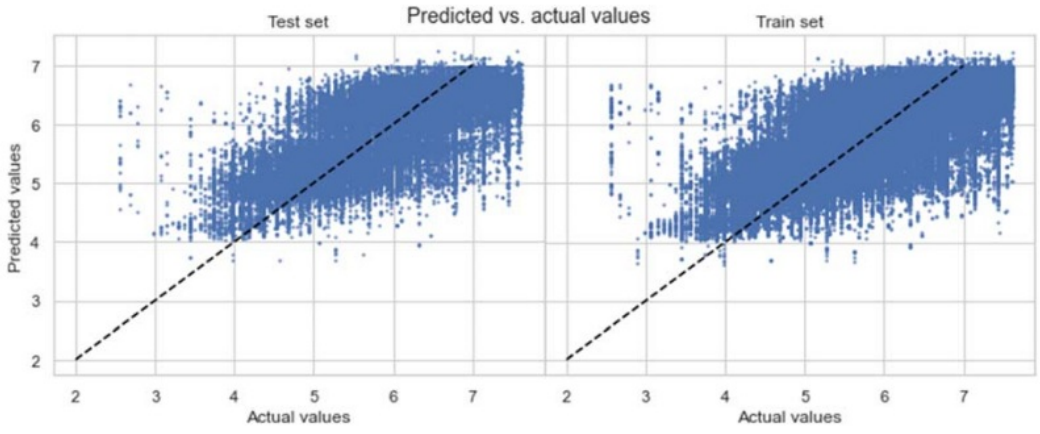
Mean Squared Error:	0.3682162703877934
R2 Score:	82.47496265182751
Mean Absolute Error:	0.2715425653847894

A 4 layered ANN was built using Adam optimizer and ReLU activation function in the first three layer and linear activation function in the last layer. A batch size of 256 and 100 epochs was run using 0.1 validation split. The results are obtained are provided in the table 5.6.3.: -

Training MSE: 0.3306
Validation MSE: 0.3317

Training r2: 0.5735
Validation r2: 0.5712

The fig 5.6.2 shows the predicted vs actual plot for training and the test set.



As we noticed that the best results were obtained using Random forest and XGBoost. However Random forest showed overfitting.

5.7 Discussions

We analysed the data well to understand each and every column independently to find out more about the features of Airbnb properties. The method for this research consisted of different ML models, the first of which was Linear regression to predict the price of the property using the independent variables, to then compare the results between different algorithms.

The implementation of first model where not much feature engineering was done and in the second iteration it showed the importance of feature transformation. Among the different ML algorithms Random Forest and XGBoost performed well on both the training and the test data.

Further we performed NLP on the name column to understand if there exist any relationship with the price variable. We used the NLP techniques to find the most common words used to describe the apartments using bigrams and trigrams approach. The most common approaches to describe the apartments are usually by describing the room types and the neighbourhood group.

We further implemented KNN to find how the algorithm works for this dataset. The first iteration did not score well however the second iteration we got scores. The second iteration was done after removing the longitude and the latitude columns. This showed the two columns were insignificant. The k was chosen to be 3 and metrics used was r squared and root mean square error. The r square was found to be around 71% and rmse value was 0.465.

As discussed in the Problem Statement section 1.2 the use of PowerTransformer has been proven very effective when the data is skewed. We have used both the approaches yeo-johnson and box-cox transformation to transform the variables as discussed in section 5.3. The variable transformation drastically helped the model performance to increase. The variables were more gaussian like after the transformation.

1. If there exist a relationship between the name of the listings and the “price” of the listings?

By looking at the plots in section 4.5 we can easily identify some kind of pattern. Titles of apartments with high prices includes: -

Luxury

Manhattan,

Bedroom or br

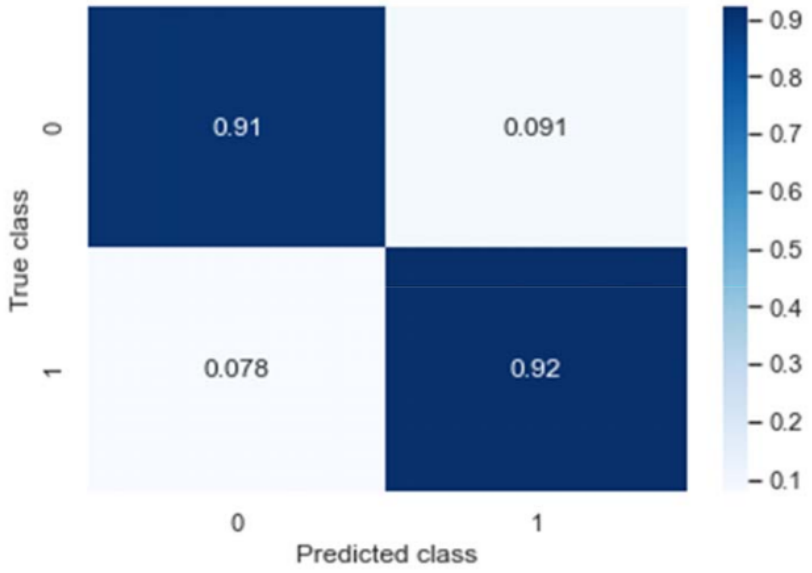
Apartment or apt

Loft

Park

The trend is normal as luxury apartments with loft or park are expensive also Manhattan neighbourhood is the priciest among the other neighbourhood. Further we created a ML model where price above 800 dollars where considered expensive and were labelled 1 and below 800 were considered low and were labelled 0. To balance the classes, we used random oversampling and used Light GBM to predict the classes based on titles.

The fig. 5.7.1 shows the confusion matrix built to predict the classes.



Although we could not derive much from the model as only one feature which is the name of the apartments were used. However, looking at the bigrams and trigrams plots we can say there is a trend in naming the properties.

2. What are the most frequent words used for the properties with high “price”?

The most frequent words used for high priced apartments are: -

Bedroom

Apartment

Luxury

Loft

Park

Manhattan

Spacious

Private

3. How well the model performs if we use LightGBM for boosting the model performance?

The LightGBM classifier was used to predict the classes of the apartments using titles of the apartments. The model performance was not worst however we notice slight overfitting. Further, we did not use this algorithm to predict the prices of the apartments.

We did not include Light GBM boosting algorithm to predict the prices of the apartments due to time constraints we were unable to tune the algorithm and perform further research on this algorithm.

In the first iteration we did not receive a good score however we have implemented 5 ML algorithms out of which Random Forest and XGBoost outperformed. Further, KNN model was a generalised model and also was able to explain 70% of variation.

Further we were able to include the algorithm to predict the classes. The expensive properties was labelled as 1, also the properties were considered expensive which were above 800$. The reason to consider 800$ as expensive because 99% of value were below 800$. The algorithm performed well to classify the prices. However, the model did overfit the data and was not able to explain anything further.

5.8 Summary

The method for this research consisted of a set of different ML algorithms, the first of which was linear regression to predict the price of the dependent variable price.

Next different ML algorithms were built to compare the performance and predictions between algorithms.

The results of the models are discussed and well elaborated in the results section. We were able to attain a good score for the XGBoost and KNN model. We also found the importance of feature engineering and feature transformation.

The study helped us answer the question discussed in the problem statement section and also helped use understand the data well. The results from linear regression outputs an r squared .494300, meaning that the set of independent variables analysed to build the model only explains 49% of the price. The results from OLS model was used to support the hypothesis discussed.

The XGBoost algorithm was used to understand the feature importance to predict the price of the properties. It was found that room type has a significant impact to predict the price, hence supporting the hypothesis. The algorithm displayed a sorted list of features with the weights associated with it to understand the significance of each variable.

The results of all the model are compared with the ANN outputs. The ANN model scores was better when we used a simple neural network with 4 layers and adam optimizers. As we increased the complexity the model performance kept decreasing.

Further, we analysed the name of the properties and if there exists a relationship with the price. We were able to find patters while naming the expensive properties. The bigrams and trigrams approach was quite useful to find the most frequently used words.

Though we were unable to implement LightGBM to predict the price of the properties but we were able to implement the algorithm to classify the expensive and non-expensive properties based on the title. We were able to attain a good accuracy in predicting the classes.

Due to time complexity, we were unable to implement LightGBM but we applied 5 ML algorithm in the data set and all the algorithms were able to add knowledge to study.

As we are aware of the Covid-19 that has a tremendous effects in almost all the business sectors, and the global economy. Airbnb a company which revolves around tourism industry has also been effected by the virus. While the comparison between

the 2019 and 2020 data set did not show any significant difference. However, we were able to find some difference in the average prices of the properties.

In this paper we focussed mainly on the 2020 dataset and were able to analyse each and every column individually and built the ML model.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1. Introduction

As Airbnb is becoming popular amongst tourists more people are considering renting apartments rather paying for hotels and other staying options. There is a question if there are any suggestions for hosts and guests to set up rental properties and to rent a property. Because of this motivation we started this project. We included dataset from NYC, the data cleaning step involved a great deal of data formatting and transformation. Also, for this project exploratory data analysis and data visualization played a very important role.

Due to feature transformation and feature understanding we have a better understanding of features. During the ML algorithms implementation, we used different techniques to increase the model performance and evaluate them. As we expected XGBoost gives the most satisfying result. Also, the model built using KNN gave satisfying result. This study also helped to find the most significant features to predict the prices. Certainly, this is one of the situation were ML model performs better than ANN. The model could possibly be improved with hyperparameter tuning. However even in the best performing model, the model was able to explain 82% of variation in price.

This section describes the findings of the study in details and provides a conclusion of the study. Also, the future works that can be done to improve the study. While we have tried to implement the most important ML algorithms and compare their performance. However due to time complexity we were unable to research more on ANN and LightGBM.

6.2 Discussions and Conclusions

This Airbnb dataset for the year 2020 was used for our analysis however we scrapped the 2019 data to compare and find if there exists any notable different patterns. The dataset is rich and has a variety of columns for the analysis. The columns allowed us to dig deeper and explore the data well. We have used a systematic approach to compete the study which included several steps.

First, we have pre-processed the data to clean and refine such as missing values treatment, removal of less important columns, dropping of duplicates, standardizing the

attributes. Next, we explored the columns using the basic plots and visualization such as bar plots, charts, scatter plots, etc., These plots helped to analyse the room types, price distribution, expensive neighbourhoods, distribution of neighbourhood. Also hosts takes good advantages of the Airbnb platforms to provide their properties for rents. Further we analysed which are the busiest neighbourhoods and the areas which are most popular. Also, we made use of the longitude and latitude columns to create a geographical map with color coding to understand the data well. Lastly, we used predictive analytics to compare different model performance. We have used the most optimized and the newest algorithms where we got positive results.

The results from the OLS model supported the hypothesis that the features have a significant impact to predict the price. The feature importance from the XGBoost model helped to obtain the most important features. The KNN model was a generalised model and the results obtained were quite satisfactory.

The title to describe the properties were analysed using NLP techniques and we were able to obtain the trend used to name the expensive properties. The word cloud obtained helped to find the most frequently used words.

Nevertheless, the study contributes to the existing literature on the determinants of price of sharing economy. In practical use, the analysis may offer potential suggestions to the stakeholders of traditional accommodation industry. The study well help the decision maker committee to analyse and evaluate the market situation and strategize to improve their services. The market study can be done using the visualizations provided in the paper.

Further, the present paper can provide the hosts with an insights to strategize how to set up the prices of the property which can be both affordable to the guests too. The paper may also help Airbnb employees to design tools and provide offers and tips that can attract more customers.

6.3 Contribution to Knowledge

The research attempts to contribute to knowledge by understanding the data well and applying different ML algorithms to predict the prices of the properties.

In this study we have investigated whether and how accommodation attributes are related to price in touristic region like NYC. In line with the previous researches, the findings confirms that the property attributes has a significant influence in the prices.

Through this research we are able to understand the 2019 and 2020 data and the price distribution of the rental properties. Though there is no significant difference uncovered between the two dataset however we found that the difference between the minimum price and the maximum price is much higher in 2020 data. Further the average price in both the cases is below 200 $ which shows that in an average the properties are quite affordable.

Our results also illustrated that the attributes have similar effects with the previous findings. The features significance obtained in XGBoost model showed room type has the most impact in predicting the price. This result is in line with study of attributes related to size which has a positive correlation with price that is the price increases with the size of the rentals (Cai et al., 2019; Chen & Xie, 2017; Gibbs et al., 2018; Kakar et al., 2016; Wang & Nicolau, 2017).

One of the assumptions which is often unnoticed in the previous works has been looked upon. The residuals from the OLS model were found to be normally distributed. The variables were transformed using the power transformation which proved very effective.

The study helped to understand the most used titles used for describing the properties of the Airbnb NYC. We were able to analyse the pattern used to name the expensive properties. Further the use of word cloud helped to find the most frequent words.

We also experimented with ANN to find out how the algorithms works for regression problems. It was found that a simple neural network is more generalized. We were also able to understand the each and every attributes using univariate, bivariate and multivariate analysis.

6.4 Future Recommendations

We have included all the important ML algorithms and have discussed each and every step-in detail. We acknowledge that like any piece of research, this study has certain limitations as well which needs to be analysed and highlighted. First, the ANN model requires hyperparameter tuning and find the best results. Further, the dataset we have considered is only for a specific time period that is for the year 2019 and 2020. Also, we have not looked upon the seasonality factor in this analysis. Further, we have explored only New York City hence variation between different cities are not captured and explored here.

Future works can also explore and experiment with ANN to find the best model.

In conclusion, the present research does provide relevant insights, however the research underlines the need for further research. The study has been able to answer the question stated in chapter 1. Overall, we discovered a very good number of interesting relationship and insights from the data and each and every step has been explained clearly.

REFERENCES

Anon (2020) *Airbnb Listings in New York City : Price Prediction and Analysis by Jorge Enrique Gamboa Fuentes A Capstone Project Submitted to the Faculty of Utica College May 2020 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data S.*

Anon (n.d.) *Air bnb.pdf*.

Atta-Fynn, R. and Zien, C., (2019) Analysis and Machine Learning Modeling of New York City Airbnb Data. *NYC Data Science Academy*, [online] pp.1–17. Available at: https://nycdatascience.com/blog/student-works/analysis-and-machine-learning-modeling-of-new-york-city-airbnb-data/.

Cai, T., Han, K. and Wu, H., (n.d.) Melbourne Airbnb Price Prediction. [online] pp.1–6. Available at: https://drive.google.com/open?id=1D32jVpSfvEYCDCoVt6FYxS98KVFhp3Fm.

Chen, T. and Guestrin, C., (2016) XGBoost : A Scalable Tree Boosting System. pp.785–794.

Choudhary, P., Jain, A. and Baijal, R., (2018) Unravelling Airbnb Predicting Price for New Listing. [online] Available at: http://arxiv.org/abs/1805.12101.

Dudás, G., Kovalcsik, T., Vida, G., Boros, L. and Nagy, G., (2020) Price determinants of airbnb listing prices in lake balaton touristic region, Hungary. *European Journal of Tourism Research*, 24February.

Gibbs, C., Guttentag, D., Gretzel, U., Yao, L. and Morton, J., (2018) Use of dynamic pricing strategies by Airbnb hosts. *International Journal of Contemporary Hospitality Management*, 301, pp.2–20.

Gonzalez-Fernandez, I., Iglesias-Otero, M.A., Esteki, M., Moldes, O.A., Mejuto, J.C. and Simal-Gandara, J., (2019) A critical review on the use of artificial neural networks in olive oil production, characterization and authentication. *Critical Reviews in Food Science and Nutrition*, [online] 5912, pp.1913–1926. Available at: https://doi.org/10.1080/10408398.2018.1433628.

Group, G., Stelmer, C., Tang, C. and Tanghøj, N., (2020) A study of Airbnb prices

in Copenhagen 2 . A general overview of Airbnb in. June 2016, pp.1–12.

Guttentag, D., (2019) Progress on Airbnb: a literature review. *Journal of Hospitality and Tourism Technology*, 103, pp.233–263.

Hornik, K., Stinchcombe, M. and White, H., (1989) Kornick et. al. *Neural Networks*, 2, pp.359–366.

Kalehbasti, P.R., Nikolenko, L. and Rezaei, H., (2019) Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. [online] Available at: http://arxiv.org/abs/1907.12665.

Kingma, D.P. and Ba, J.L., (2015) Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp.1–15.

Li, Y., Pan, Q., Yang, T. and Guo, L., (2016) Reasonable price recommendation on Airbnb using Multi-Scale clustering. *Chinese Control Conference, CCC*, 2016-Augus, pp.7038–7041.

Luo, Y., Zhou, X. and Zhou, Y., (2019) Predicting Airbnb Listing Price Across Different Cities. pp.1–6.

Mcneil, B., (2020) Price Prediction in the Sharing Economy : A Case Study with Airbnb data by. *University of New Hampshire Scholars' Repository*. [online] Available at: https://scholars.unh.edu/cgi/viewcontent.cgi?article=1511&context=honors.

Moreno-Izquierdo, L., Egorova, G., Peretó-Rovira, A. and Más-Ferrando, A., (2019) Exploring the use of artificial intelligence in price maximisation in the tourism sector: Its application in the case of airbnb in the valencian community. *Investigaciones Regionales*, 201942, pp.113–128.

Moreno-Izquierdo, L., Rubia-Serrano, A., Perles-Ribes, J.F., Ramón-Rodríguez, A.B. and Such-Devesa, M.J., (2020) Determining factors in the choice of prices of tourist rental accommodation. New evidence using the quantile regression approach. *Tourism Management Perspectives*, [online] 33November 2019, p.100632. Available at: https://doi.org/10.1016/j.tmp.2019.100632.

Selim, H., (2009) Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, [online] 362 PART 2,

pp.2843–2852. Available at: http://dx.doi.org/10.1016/j.eswa.2008.01.044.

Varma, A., Sarma, A., Doshi, S. and Nair, R., (2018) House Price Prediction Using Machine Learning and Neural Networks. *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, pp.1936–1939.

World Health Organization., World Bank., Ruiz-Ibán, M.A., Seijas, R., Sallent, A., Ares, O., Marín-Peña, O., Muriel, A., Cuéllar, R., Mobasheri, A., Batt, M., Quintana, J.M., Escobar, A., Arostegui, I., Bilbao, A., Azkarate, J., Goenaga, J.I., Arenaza, J.C., Murphy, L.B., Helmick, C.G., Schwartz, T.A., Renner, J.B., Tudor, G., Koch, G.G., Dragomir, A.D., Kalsbeek, W.D., Luta, G., Jordan, J.M., Martín-Fernández, J., Gray-Laymón, P., Molina-Siguero, A., Martínez-Martín, J., García-Maroto, R., García-Sánchez, I., García-Pérez, L., Ramos-García, V., Castro-Casas, O., Bilbao, A., Kloppenburg, M., Berenbaum, F., Kohring, J.M., Pelt, C.E., Anderson, M.B., Peters, C.L., Gililland, J.M., Macías-Hernández, S.I., Zepeda-Borbón, E.R., Lara-Vázquez, B.I., Cuevas-Quintero, N.M., Morones-Alba, J.D., Cruz-Medina, E., Nava-Bringas, T.I., Miranda-Duarte, A., Joseph, D.S., Walton-Paxton, E., Ostendorf, M., van Stel, H.F., Buskens, E., Schrijvers, A.J.P., Marting, L.N., Verbout, A.J., Dhert, W.J.A., Seijas, R., Ares, Ó., Sallent, A., Gómez-Valero, S., García-Pérez, F., Flórez-García, M.T., Miangolarra-Page, J.C., Maillot, C., Harman, C., Al-Zibari, M., Sarsam, K., Rivière, C., Herdman, M., Badia, X., Berra, S., Cnudde, P., Rees, H.W., Greene, M.E., Learmonth, I.D., Young, C., Rorabeck, C., Zhang, W., Moskowitz, R.W., Nuki, G., Abramson, S., Altman, R.D., Arden, N., Bierma-Zeinstra, S., Brandt, K.D., Croft, P., Doherty, M., Dougados, M., Hochberg, M., Hunter, D.J., Kwoh, K., Lohmander, L.S., Tugwell, P., Ramadani, R.F., Erastus Mosha, Ramadani, R.F., Crawford, R.W., Murray, D.W., Peláez-Ballestas, I., Sanin, L.H., Moreno-Montoya, J., Alvarez-Nemegyei, J., Burgos-Vargas, R., Garza-Elizondo, M., Rodríguez-Amado, J., Goycochea-Robles, M.V., Madariaga, M., Zamudio, J., Santana, N. and Cardiel, M.H., (2020) No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title. *Osteoarthritis and Cartilage*, [online] 282, pp.1–43. Available at: http://journals.sagepub.com/doi/10.1177/1120700020921110%0Ahttps://doi.org/10.1016/j.reuma.2018.06.001%0Ahttps://doi.org/10.1016/j.arth.2018.03.044%0Ahttps://reader.elsevier.com/reader/sd/pii/S1063458420300078?token=C039B8B13922A2079230DC9AF11A333E295

FCD8.

Ye, P., Wu, C.H., Qian, J., Zhou, Y., Chen, J., De Mars, S., Yang, F. and Zhang, L., (2018) Customized regression model for Airbnb dynamic pricing. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.932–940.

Zehtab-Salmasi, A., Feizi-Derakhshi, A.-R., Nikzad-Khasmakhi, N., Asgari-Chenaghlu, M. and Nabipour, S., (2020) Multimodal price prediction. [online] July. Available at: http://arxiv.org/abs/2007.05056.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* pp. 785-794.

APPENDIX A

RESEARCH PROPOSAL

SEPTEMBER 2020

Table of Contents

Abstract		3
1.	Introduction	4
2.	Background	4
3.	Problem Statement	5
4.	Research Questions	7
5.	Aim and Objectives	7
6.	Significance of the Study	8
7.	Scope of the Study	8
8.	Research Methodology	9
9.	Required Resources	11
10.	Research Plan	13

Airbnb Price Prediction Using Machine Learning

Abstract

Airbnb is an American vacation rental company that offers arrangement for homestays. The company acts as a broker between the owner and the tenant. The company has been one of the most popular sites among travellers. Airbnb was founded in the year 2008 and in June 2012 announced its 10,000,000th night booked approximately doubling the business.

Pricing of a property is an important factor to maintain the supply and demand. Hence, a reasonable pricing is a key point for both the host and the guest.

Machine Learning is the Scientific study of algorithms and build statistical model that can be used by the computer to efficiently perform some specific task without any explicit information. These tasks are mainly based on patterns and inferences.
Machine learning is not only focussed on building model but also to derive insights from the data. The data to be used should have some relationship and to further analyze these relationships and draw conclusion based on it.

The main focus of the research is to build a machine learning model which can predict the pricing of the rental property on Airbnb. Here, the main aim of research is solely not to find the best machine learning model to predict the pricing but also to study the different features, apply deep learning techniques and different tree-based ensemble boosting algorithms like Light GBM, XGBoost to prepare the model.

The study will include both Descriptive and Predictive analysis.

Descriptive analysis is a study which includes statistical analysis such as data collection, interpretation, measuring the central tendency. This process is considered as the base for model creation and transforms the raw data into something meaningful.

Predictive analysis is analysing the current data and historical data to predict the probability of future outcomes in this case the price of the rental property. This includes data mining and model training using ML.

In a number of research paper tree-based ensemble model are created to increase the model performance such as XGBoost, Random Forest, Gradient Boosting, etc., Ensembles model mitigate the problem of overfitting the data.

XGBoost has mostly been used in almost all data science competition to increase the model performance. However, when given lots of data the algorithm slows down. Recently **Light GBM** has been one of the most widely used algorithm. It is fast, high performing boosting framework also based on **Decision Tree**.

Light GBM short for Light Gradient Boosted Machine is an open source distributed Gradient boosting algorithm for ML which was originally discovered by Microsoft.

The algorithm splits the tree leaf wise unlike other learning algorithm which splits the tree depth wise. The leaf wise split reduces more loss compared to others hence resulting in better accuracy in less time.

1. Introduction

Data Science is a blend of Analytics, Computer Science, and Business domain knowledge to solve business problems. The main aim of any data science project is to make accurate prediction and solve any business problem with the help of available data. Data science involves less of Mathematics and Coding and relies more on data and building new systems to process the data.

The field is divided into three parts: -

- Supervised
- Unsupervised
- Reinforcement

Supervised Learning model is a type of algorithm that needs a labelled data to generate predictions on a new data. Few of the types include Regression and Classification problem.

Unsupervised learning in the other hand works on unlabelled dataset and the algorithm tries to make use of the features and underlying patterns to reach to a conclusion. Cases such as Clustering, Anomaly detection, Autoencoders are few types.

Reinforcement learning is less supervised ML algorithm and depends on an agent which is a software that is being trained on. This agent learns to take actions to maximise the output.

Any data science project lifecycle includes proper business understanding, data collection, proper understanding of the data, data pre-processing and data preparation, modelling, model evaluation, and model deployment. However, the process does not end here, once the model is deployed it needs to be updated regularly and if needed reiterate the whole process. The model can

only be helpful if it can provide insights to the end users. Any ML model is successful after it is deployed in a platform.

2. Background

This paper aims to predict the price of the rental property with the help of Machine learning technique and the right amount of data. Here, the target variable is Price of the Property which makes it a supervised ML model. Supervised ML model learns from the labelled data to predict on any new given data. Predicting the price correctly is important and difficult at the same time as price too high can lead to a smaller number of customer and too low will lead to a loss for the owner. A number of research's are conducted in this field which includes **Hedonic Regression** model. Hedonic Regression model estimates the influence of various factors on price of any product. This technique is mainly used in real estate pricing problems.

In most of the research's **Artificial neural network** performed better with **Relu** activation function. **ANN** are based on collection of many connected hidden layers and neurons. It is one of the deep learning techniques. The input is passed through multiple hidden layers with an activation function and output is generated. The training of the neural network is usually done by determining the difference between output and the target output.

To predict the price of any product which is the dependent variable, the input variables or the feature variables also known as the independent variables plays a major role. Often the features are correlated to each other which results in lowering the predictive power of the model. **Principal component analysis** is a technique which helps to mitigate this problem. It creates new features using the features provided. Further, PCA benefits by reducing the dimension of large datasets thus helping in improving model performance. Further, to visualize a complex data is not possible with the help of 2-D plots. Hence, to find the relationship between variables is difficult. PCA helps in capturing these relationships by creating combinations from original features.

Many research works included tree-based ensemble model to create the model.

3. Problem Statement

Airbnb is a US based online property rental marketplace. Airbnb offers lodging, homestays, or tourism arrangements. The company offers complete independence to host the price of the property to the owners.

Pricing of a rental property is still a challenge mainly in big cities. Few of the challenges are the huge amount of data which is being generated and processing the data requires specific skills. Especially the EDA which

involves data cleaning, data processing, extracting correct information from the data. Further, there are a lot of competition and even a small difference in price can make a huge impact.

Related Work

I. Study conducted by (Kalehbasti et al., 2019) proposed a predictive model using ML, Deep Learning, and NLP technique. The main focus of the research was to benefit both the property owners and the customers. A range of methods such linear regression, tree-based model, SVR, k means clustering and ANN were used for creating the model. To evaluate the performance MAE, MSE, R square score were compared for each model. The model had similar scores however, SVR with RBF kernel outperformed with R2 score of 69% and MSE of 0.147 and highest R square score.

II. Study conducted by (Keating, 2018) to predict the pricing of Airbnb rentals in Seattle. The main focus of this study was to find out the relationships between the independent variables and the dependent variables. Here, the baseline model was OLS modelling technique. Later techniques such as k nearest neighbours, random forest and neural network was used and model performance was evaluated. Here, the best performing model was multi-layer perceptron neural net. The spread of error of each model against the test data showed neural network to be more accurate.

III. Study conducted by (Atta-Fynn and Zien, 2019) was done to analyze the New York City data using Machine Learning. 2018-2019 Airbnb data was used for the study. The paper was focused on to learn the patterns from the data. Four model were trained: Lasso (Linear regression model with L1 regularization) and three tree-based model were built: Random Forest Regression, Gradient Boosting Regression, and Extreme Gradient Boosting. Grid search CV was used for hyperparameter tuning. RMSE score was used as the to evaluate the performance of each model. The tree-based model performed better generating the accuracy in the range of 76% to 77%. It was found that the most frequent feature was the “Wi-Fi internet” service. Further, 80% of the listings had an average nightly price of 180$.

IV. Study by (Luo et al., 2019) used a variety of regression approaches such as Linear Regression, k nearest neighbour regression, random forest regression, XGBoost, ANN to predict the target variable. The baseline model was build using Linear regression with L2 regularization and K-nearest neighbour (KNN) regression. Neural Network was built using 3 fully connected multilayer perceptron with Relu Activation function. The model evaluation metrics were R2 score and MSE. XGBoost and Neural Network achieved the score ranged between 66% to 67%. Also, it was noted that logarithmic transformation of the target variable increased the model performance. Also, interaction terms of the continuous variables increased the model performance

further. Unigram and Bigram tf-idf were used to represent the features. This approach was also very new and helpful for the price prediction and obtained a R2 score of 74%. Date feature was uncorrelated to the Price hence was dropped later. The best performing model was XGBoost and neural network and obtained a score of more than 70%.

V. Research conducted by (Wang, Dan & Nicolau, Juan. 2017) Airbnb dataset which included 33 cities and identified top 25 price determinants. The paper was based on Ordinary Least Square and Quantile regression.

VI. Research conducted by (Li et al., 2016) was based on price recommendation system using Multi-Scale Affinity Propagation (MSAP). This paper used to build a house price prediction model. The paper also used linear regression model with normal noise (LRNN) to understand the increase of the renting rate. The paper concluded that MSAP
can be used to cluster the house capably and aggregate the house into different price-zone, which further helps in providing the reasonable price uniquely.

VII. Similar research conducted by (Varma et al., 2018) was based on predicting house price using ML and Neural Network. The model was used to analyse the set of parameters selected by the customers as per their interest. The model was built using classical linear regression, forest regression and boosted regression for predicting the price. Further Neural Network was built to increase the accuracy. The model helped establish the relationship strength between dependent and independent variable.

VIII. Research conducted by (Choudhary et al., 2018) was conducted to analyse the Airbnb listings in the city of San Francisco. The study aimed to better understand attributes such as bedrooms, location, house type can be helpful to accurately predict the price. The price can be both profitable to the hosts as well as to the guest. Additional analysis was performed to establish the likelihood of a listing's availability for the guests to consider while making a booking the property.

IX. Research conducted by (Group et al., 2020) performed graphical and statistical analysis of Airbnb data for Copenhagen. The findings showed how different variables such as "room type", "neighbourhood", "accommodations", "bathrooms", "bedrooms" and "minstay" has a significant effect on the target variable. Further, text analysis was performed to understand the most frequent words used in positive reviews and for negative reviews. Five different linear regression model was tested and the best performing model had all the significant features. The model predicted lower than the website prices.

X. In an experiment conducted by (Lewis et al.,2019) predicted the Airbnb rental price for properties in London by using machine learning and deep

learning. The best model which outperformed was XGBoost with an accuracy of 73%, which was even better than the neural network.

Research Gap

The literature review suggests that a number of research and study has been conducted. Most of the paper suggests customer preferences, model which predicted the target variable accurately, the important features, correlation among the variables, etc., With a change in the data there arise a need to study and analyse the data again. There is also a need to better understand the factors which have brought a paradigm shift in pricing of the rental property. This study will help in major decision making for both the hosts and the guest.

The scopes of most of the existing research paper are limited to creating tree-based ensembles which splits the data depth wise or level wise however LightGBM splits the tree leaf wise. LightGBM is fast, distributed, and high-performance boosting framework. This leaf wise split reduces more loss than other algorithms. Further, in most of the papers ANN has not proved to be a better performing model hence there is a need to further investigate in this area.

Moreover, identifying the most important features to predict the price of the rental property can be helpful in strengthening the market. Also, the proposed method can be extended to similar price prediction problem.

4. Research Question

Few of the research question which will be addressed are as follows: -

1. What is the most significant variable in predicting the price of the rental property?
2. What are the most frequent words used by the property owners to describe their places? Do these words have a significant impact on the predictor variable?
3. Are there any patterns to identify the expensive apartments?
4. In which area the most expensive properties are located?
5. Which Machine Learning algorithm gives the better accuracy and results?

5. Aim and Objectives

The goal of the research is to make a model which accurately predicts the price of the rental property in New York. Further, predicting the price of any rental property is a challenging task. Here, the main focus will be feature engineering and feature selection. The independent variables or features are extremely important for a mathematical model. Extensive

descriptive and exploratory analysis of the data to be performed in order to understand how each feature behave individually.

Any model fails when there are a greater number of features. It is seen in most case the accuracy increases when we add features but after a certain point this accuracy starts decreasing as the model has to learn the insignificant features too. The abundance of number of features leads to high variance that is high error in the test data.

Objectives:

The primary objective of the research paper is to use different Machine Learning algorithm to assess the Price of the property. Based on the identified research gap the overall objective is to explore the data well enough and draw conclusion based on it.

1. To examine the specific features which are strong predictor for the Price.
2. To compare the performance of different models by evaluating the model parameters such as precision, recall, F-measure, and the accuracy.
3. To further analyze the features which are in text format.

6. Significance of the Study

The study of Airbnb price prediction using Machine learning will be helpful to find more about the variables or the main drivers which are directly affecting the pricing. ML provides a very unique way of combining both fundamental and technical analysis for price prediction. The study will involve both descriptive and predictive analysis. By combining all these more accuracy in prediction can be achieved. Further, how much the techniques such as ANN, Random forest and other boosting algorithm are helpful in regression problems.

Comparing the different models will help to gain more understanding about the working of the algorithms in predicting the price. Price prediction is not only confined to real estate domain but also the study can be extended to other domain as well. Further, the data used here is not only numerical or categorical but also text-based data.

Text data is usually unstructured data that cannot be directly fed to any ML algorithm. NLP is such a field which will help to process these data. As known the text-based data reveal lot of information if used correctly. Applying natural language processing techniques such unigram, bigram, or ngram approach can reveal lot of information about the data. The goal of NLP here is to help understand what are most frequent words used and if it pays any importance to predicting the target value. The study of these texts

can help gain more insights into customers willingness to pay for the property and can thus be helpful in pricing strategies.

7. Scope of the Study

This study will focus on developing a price prediction model using machine learning techniques.

Data used here is the recent 2020 Airbnb data and the study limits its coverage to this year's available data. Further, due to the missing reviews section in the recent data. Guest review sentiment analysis will not be included in the research.

The analysis and prediction will be based on the new data and no previous year data is used in the study.

The geographical location considered here is the New York City data. The other locations which are part of Airbnb are not within the scope of the study. The research aims to find out the most influential features to predict the pricing and how the text-based data such as the property description data can be useful to understand the target variable. The study will not cover other factors which are responsible for pricing which are not part of the data.

The programming language used here will be Python 3.7 and will not cover any other programming languages such R, SAS, Java, etc.,

8. Research Methodology

In this research New York City dataset is chosen to understand the distribution of price of a rental property and to understand the importance of each and every feature in determining the price variable. A number of research paper were referenced to understand and gain in-depth knowledge of the topic. Here the research problem is to come up with a Machine Learning model which will help in predicting the price of the property. The data used for this purpose was collected from official Airbnb site. The data is a secondary data which contains both quantitative and qualitative data.

This research will be divided into descriptive and exploratory analysis of the data. The entire analysis will follow a simple and direct structure. Simple statistical techniques will be implemented such as frequency distribution table, measures of central tendencies, histograms, distribution of the data, boxplots to find the outliers. These steps will serve as a guide for further steps.

The following methodology will be implemented for the research are:

I. Data Pre-processing: - Data Pre-processing is one of the most important step and takes most of the time in any data science project. It is one of the most important step because this creates the base for the model and if not done correctly can highly impact the predictive model and the results obtain will be meaningless. This includes initial data filtering, cleaning, treating the outliers and missing values, removing the invalid entries.

II. Data Exploration: - Data exploration involves creating interacting features from the existing variables this will help in reducing the dimension of the data, visualizing the data to find patterns, trends and anomalies if present. Data visualization further helps to find the distribution of the features with the dependent variables and with other independent variables. A better data understanding can be done by using simple charts and bars. The data contains different data types which includes numerical, categorical, and text data and involves different data treatments.

III. Data Analysis: - Analysing the final data which can now be used to build the model. Generating descriptive statistics and predictive analysis will be part of this step. Summarizing the central tendency and dispersion of the data will be included here.

IV. Modelling: - Linear regression will be used to create the baseline model. Later, Decision tree, Random forest regression will be created. NLP techniques will be implemented to process the text data and analyse the text-based feature's correlation with the predictor. Further, experiment will be performed using Artificial Neural Network with hyperparameter tuning. Boosting algorithm such XGBoost, Light GBM which is one if the very recent tree based boosting algorithm can be performed to check the model performance on the training data.

V. Model Evaluation: -Different model evaluation technique such as R Square also known as coefficient of determination, Mean Squared Error (MSE), Adjusted R square, Mean-Absolute-Error (MAE), Root-Mean-Squared-Error (RMSE) will be used to compare the different model performance. Since the dependent variable here is continuous variable the mentioned techniques are among the best to find out the model accuracy.

VI. Model Deployment: -Once the model is created it will be deployed with the help of the platform's available to provide insights to the end user.

Evaluation Criteria

The performance of the model will be validated by **MSE** or Mean squared error. This method is one of the most preferred metrics for any regression model. It is the average of squared difference between the target value and

the predicted outcome by the model. As the value is squared it penalizes even if there is a small error.

Another metric will be **RMSE** or root mean squared error. This is also one of the most widely used metric and is the square root of the average squared difference between the predicted value and average value.

MAE or mean absolute error is the absolute difference between the target value and the predicted value by the model divided by the total number of data points. Though, it does not penalize the model like MSE but is more robust to the outliers.

R square also known as the coefficient of determination will be another technique to evaluate the model performance. R square can be stated as explained variance divided by the total variance. R square is always between 0 to 100%. 0% indicates the model has not explained none of the variation and 100% means model explained all the variation of the response data around the mean. R square is given by 1-(ss_R)/(ss_T) ss_R is sum of squared regression error and ss_T is sum of squared total error.

Adjusted R square is an improvement of R square. However, r square increases every time any new feature is added to the model this leads to curse of dimensionality. To mitigate this problem adjusted r square is used. Adjusted r square increases only when a predictor has a significant impact to the model.

Chapterization

The entire research work will be divided into four chapters and they are as follows: -

I. **Introduction** which will convey the entire background of the work, the problem statement, aim, significance and the importance of the research, hypotheses and the objectives.

II. **Literature Review** will describe the previous work in the same or similar field, concepts and definitions followed by the methods used previously. Further, a brief description of the results will follow.

III. **Research Methodology** which will be covering the methods used in attain the goal, research design, data collection, data sampling, analysis and findings.

IV. **Experimental Results** will be the results obtained in the experiment and analysis.

V. **Discussion and Findings** will cover the findings, results obtained, and the conclusion of the results followed by future works which can be done.

VI. **Bibliography and References.**

Dataset Used

The dataset used is the most recent one and has been taken from the official Airbnb site (http://insideairbnb.com/get-the-data.html). The data describes the listing activity and metrics in New York City for the year 2020. The file contains all the required information to find out more about the hosts, geographical availability, and important metrics to make predictions and draw conclusions.

The dataset has sixteen features out of which "Price" is the dependent feature and forty-eight thousand five hundred and eighty-nine rows. There are 15 features which provides a very rich amount of information for data insights and exploration. The data used here has some interesting features which can be helpful for data visualisation and descriptive statistics. Descriptive study is the quantitative study to summarize the dataset.

Though there are few missing values in the dataset. This will require some amount of time to explore and later cleaning of the data.
These includes numerical, categorical, and text-based feature which makes the data very interesting for the study. Further, there are features such as "Latitude," and "Longitude" which can be further used in feature engineering. The "host_name" column might not be a significant factor for the model and which can be removed in the EDA part. The column "last_review" is a date column. There are few missing values in this column as well. One reason might be there are no reviews for the property which makes the value missing.

9. Required Resources

Jupyter Notebook
Pycharm
Anaconda
Tableau for visualisation
QlikView
MS Office

Python 3.7.7

Cloud Services

Amazon Web Services (AWS) or
Microsoft Azure
Google Colab

Hardware requirements

16 gigs of RAM
Processor anything above i5

Libraries required for model creation

Lightgbm
Scikit Learn
NumPy
Pandas
Matplotlotlib
Seaborn
TensorFlow
Keras
Imblearn
SciPy
Counter
XGBoost

Books Required

I. An Introduction to Statistical Learning
II. Elements of Statistical Learning
III. OpenIntro Statistics
IV. Python for Everybody
V. Python for Data Analysis
VI. All of Statistics: A Concise Course in Statistical Inference
VII. Core Python Programming

Model Deployment

Heroku
Docker
Google Cloud Platform
Digital Ocean

10. Research Plan

Airbnb Price Prediction using ML and SA	
ACTIVITY	PLAN DURATION
	21 Weeks
Literature search	2
Literature review	1
Design Machine Learning Model	2
Investigate & Evaluate ANNs	2
Design ANN	2
Develop & test ANN	2
Get iAirbnb irecent imarket idata	
Train ANN	2
Use ML model	2
Review statistical tests	2
Analyse & Evaluate	2
Complete report	1
Deployment	1

References

Anon (2020) *Airbnb Listings in New York City : Price Prediction and Analysis by Jorge Enrique Gamboa Fuentes A Capstone Project Submitted to the Faculty of Utica College May 2020 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data S.*

Anon (n.d.) *Air bnb.pdf*.

Atta-Fynn, R. and Zien, C., (2019) Analysis and Machine Learning Modeling of New York City Airbnb Data. *NYC Data Science Academy*, [online] pp.1–17. Available at: https://nycdatascience.com/blog/student-works/analysis-and-machine-learning-modeling-of-new-york-city-airbnb-data/.

Cai, T., Han, K. and Wu, H., (n.d.) Melbourne Airbnb Price Prediction. [online] pp.1–6. Available at: https://drive.google.com/open?id=1D32jVpSfvEYCDCoVt6FYxS98KVFhp3Fm.

Chen, T. and Guestrin, C., (2016) XGBoost : A Scalable Tree Boosting System. pp.785–794.

Choudhary, P., Jain, A. and Baijal, R., (2018) Unravelling Airbnb Predicting Price for New Listing. [online] Available at: http://arxiv.org/abs/1805.12101.

Dudás, G., Kovalcsik, T., Vida, G., Boros, L. and Nagy, G., (2020) Price determinants of airbnb listing prices in lake balaton touristic region, Hungary. *European Journal of Tourism Research*, 24February.

Gibbs, C., Guttentag, D., Gretzel, U., Yao, L. and Morton, J., (2018) Use of dynamic pricing strategies by Airbnb hosts. *International Journal of Contemporary Hospitality Management*, 301, pp.2–20.

Gonzalez-Fernandez, I., Iglesias-Otero, M.A., Esteki, M., Moldes, O.A., Mejuto, J.C. and Simal-Gandara, J., (2019) A critical review on the use of artificial neural networks in olive oil production, characterization and authentication. *Critical Reviews in Food Science and Nutrition*, [online] 5912, pp.1913–1926. Available at: https://doi.org/10.1080/10408398.2018.1433628.

Group, G., Stelmer, C., Tang, C. and Tanghøj, N., (2020) A study of Airbnb prices in Copenhagen 2 . A general overview of Airbnb in. June 2016, pp.1–12.

Guttentag, D., (2019) Progress on Airbnb: a literature review. *Journal of Hospitality and Tourism Technology*, 103, pp.233–263.

Hornik, K., Stinchcombe, M. and White, H., (1989) Kornick et. al. *Neural Networks*, 2, pp.359–366.

Kalehbasti, P.R., Nikolenko, L. and Rezaei, H., (2019) Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. [online] Available at: http://arxiv.org/abs/1907.12665.

Kingma, D.P. and Ba, J.L., (2015) Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track*

Proceedings, pp.1–15.

Li, Y., Pan, Q., Yang, T. and Guo, L., (2016) Reasonable price recommendation on Airbnb using Multi-Scale clustering. *Chinese Control Conference, CCC*, 2016-Augus, pp.7038–7041.

Luo, Y., Zhou, X. and Zhou, Y., (2019) Predicting Airbnb Listing Price Across Different Cities. pp.1–6.

Mcneil, B., (2020) Price Prediction in the Sharing Economy : A Case Study with Airbnb data by. *University of New Hampshire Scholars' Repository*. [online] Available at: https://scholars.unh.edu/cgi/viewcontent.cgi?article=1511&context=honors.

Moreno-Izquierdo, L., Egorova, G., Peretó-Rovira, A. and Más-Ferrando, A., (2019) Exploring the use of artificial intelligence in price maximisation in the tourism sector: Its application in the case of airbnb in the valencian community. *Investigaciones Regionales*, 201942, pp.113–128.

Moreno-Izquierdo, L., Rubia-Serrano, A., Perles-Ribes, J.F., Ramón-Rodríguez, A.B. and Such-Devesa, M.J., (2020) Determining factors in the choice of prices of tourist rental accommodation. New evidence using the quantile regression approach. *Tourism Management Perspectives*, [online] 33November 2019, p.100632. Available at: https://doi.org/10.1016/j.tmp.2019.100632.

Selim, H., (2009) Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, [online] 362 PART 2, pp.2843–2852. Available at: http://dx.doi.org/10.1016/j.eswa.2008.01.044.

Varma, A., Sarma, A., Doshi, S. and Nair, R., (2018) House Price Prediction Using Machine Learning and Neural Networks. *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, pp.1936–1939.

World Health Organization., World Bank., Ruiz-Ibán, M.A., Seijas, R., Sallent, A., Ares, O., Marín-Peña, O., Muriel, A., Cuéllar, R., Mobasheri, A., Batt, M., Quintana, J.M., Escobar, A., Arostegui, I., Bilbao, A., Azkarate, J., Goenaga, J.I., Arenaza, J.C., Murphy, L.B., Helmick, C.G., Schwartz, T.A., Renner, J.B., Tudor, G., Koch, G.G., Dragomir, A.D., Kalsbeek, W.D., Luta, G., Jordan, J.M., Martín-Fernández, J., Gray-Laymón, P., Molina-Siguero, A., Martínez-Martín, J., García-Maroto, R., García-Sánchez, I., García-Pérez, L., Ramos-García, V., Castro-Casas, O., Bilbao, A., Kloppenburg, M., Berenbaum, F., Kohring, J.M., Pelt, C.E., Anderson, M.B., Peters, C.L., Gililland, J.M., Macías-Hernández, S.I., Zepeda-Borbón, E.R., Lara-Vázquez, B.I., Cuevas-Quintero, N.M., Morones-Alba, J.D., Cruz-Medina, E., Nava-Bringas, T.I., Miranda-Duarte, A., Joseph, D.S., Walton-Paxton, E., Ostendorf, M., van Stel, H.F., Buskens, E., Schrijvers, A.J.P., Marting, L.N., Verbout, A.J., Dhert, W.J.A., Seijas, R., Ares, Ó., Sallent, A., Gómez-Valero, S., García-Pérez, F., Flórez-García, M.T., Miangolarra-Page, J.C., Maillot, C., Harman, C., Al-Zibari, M., Sarsam, K., Rivière, C., Herdman, M., Badia, X., Berra, S., Cnudde, P., Rees, H.W., Greene, M.E., Learmonth, I.D., Young, C., Rorabeck, C., Zhang, W., Moskowitz, R.W., Nuki, G., Abramson, S., Altman, R.D., Arden, N., Bierma-Zeinstra, S., Brandt, K.D., Croft, P., Doherty, M., Dougados, M., Hochberg, M., Hunter, D.J., Kwoh, K., Lohmander, L.S., Tugwell, P., Ramadani, R.F., Erastus Mosha, Ramadani, R.F., Crawford, R.W., Murray, D.W., Peláez-Ballestas, I., Sanin, L.H., Moreno-Montoya, J., Alvarez-Nemegyei, J., Burgos-Vargas, R., Garza-Elizondo, M., Rodríguez-Amado, J.,

Goycochea-Robles, M.V., Madariaga, M., Zamudio, J., Santana, N. and Cardiel, M.H., (2020) No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title. *Osteoarthritis and Cartilage*, [online] 282, pp.1–43. Available at: http://journals.sagepub.com/doi/10.1177/1120700020921110%0Ahttps://doi.org/10.1016/j.reuma.2018.06.001%0Ahttps://doi.org/10.1016/j.arth.2018.03.044%0Ahttps://reader.elsevier.com/reader/sd/pii/S1063458420300078?token=C039B8B13922A2079230DC9AF11A333E295FCD8.

Ye, P., Wu, C.H., Qian, J., Zhou, Y., Chen, J., De Mars, S., Yang, F. and Zhang, L., (2018) Customized regression model for Airbnb dynamic pricing. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.932–940.

Zehtab-Salmasi, A., Feizi-Derakhshi, A.-R., Nikzad-Khasmakhi, N., Asgari-Chenaghlu, M. and Nabipour, S., (2020) Multimodal price prediction. [online] July. Available at: http://arxiv.org/abs/2007.05056.

Mishra, D., (2019) Regression: An Explanation of Regression Metrics and What Can Go WrongAnon (2020) *Airbnb Listings in New York City : Price Prediction and Analysis by Jorge Enrique Gamboa Fuentes A Capstone Project Submitted to the Faculty of Utica College May 2020 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data S.*

Anon (n.d.) *Air bnb.pdf*.

Atta-Fynn, R. and Zien, C., (2019) Analysis and Machine Learning Modeling of New York City Airbnb Data. *NYC Data Science Academy*, [online] pp.1–17. Available at: https://nycdatascience.com/blog/student-works/analysis-and-machine-learning-modeling-of-new-york-city-airbnb-data/.

Cai, T., Han, K. and Wu, H., (n.d.) Melbourne Airbnb Price Prediction. [online] pp.1–6. Available at: https://drive.google.com/open?id=1D32jVpSfvEYCDCoVt6FYxS98KVFhp3Fm.

Chen, T. and Guestrin, C., (2016) XGBoost : A Scalable Tree Boosting System. pp.785–794.

Choudhary, P., Jain, A. and Baijal, R., (2018) Unravelling Airbnb Predicting Price for New Listing. [online] Available at: http://arxiv.org/abs/1805.12101.

Dudás, G., Kovalcsik, T., Vida, G., Boros, L. and Nagy, G., (2020) Price determinants of airbnb listing prices in lake balaton touristic region, Hungary. *European Journal of Tourism Research*, 24February.

Gibbs, C., Guttentag, D., Gretzel, U., Yao, L. and Morton, J., (2018) Use of dynamic pricing strategies by Airbnb hosts. *International Journal of Contemporary Hospitality Management*, 301, pp.2–20.

Gonzalez-Fernandez, I., Iglesias-Otero, M.A., Esteki, M., Moldes, O.A., Mejuto, J.C. and Simal-Gandara, J., (2019) A critical review on the use of artificial neural networks in olive oil production, characterization and authentication. *Critical Reviews in Food Science and Nutrition*, [online] 5912, pp.1913–1926. Available at: https://doi.org/10.1080/10408398.2018.1433628.

Group, G., Stelmer, C., Tang, C. and Tanghøj, N., (2020) A study of Airbnb prices in Copenhagen 2 . A general overview of Airbnb in. June 2016, pp.1–12.

Guttentag, D., (2019) Progress on Airbnb: a literature review. *Journal of Hospitality and Tourism Technology*, 103, pp.233–263.

Hornik, K., Stinchcombe, M. and White, H., (1989) Kornick et. al. *Neural Networks*, 2, pp.359–366.

Kalehbasti, P.R., Nikolenko, L. and Rezaei, H., (2019) Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. [online] Available at: http://arxiv.org/abs/1907.12665.

Kingma, D.P. and Ba, J.L., (2015) Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp.1–15.

Li, Y., Pan, Q., Yang, T. and Guo, L., (2016) Reasonable price recommendation on Airbnb using Multi-Scale clustering. *Chinese Control Conference, CCC*, 2016-Augus, pp.7038–7041.

Luo, Y., Zhou, X. and Zhou, Y., (2019) Predicting Airbnb Listing Price Across Different Cities. pp.1–6.

Mcneil, B., (2020) Price Prediction in the Sharing Economy : A Case Study with Airbnb data by. *University of New Hampshire Scholars' Repository*. [online] Available at: https://scholars.unh.edu/cgi/viewcontent.cgi?article=1511&context=honors.

Moreno-Izquierdo, L., Egorova, G., Peretó-Rovira, A. and Más-Ferrando, A., (2019) Exploring the use of artificial intelligence in price maximisation in the tourism sector: Its application in the case of airbnb in the valencian community. *Investigaciones Regionales*, 201942, pp.113–128.

Moreno-Izquierdo, L., Rubia-Serrano, A., Perles-Ribes, J.F., Ramón-Rodríguez, A.B. and Such-Devesa, M.J., (2020) Determining factors in the choice of prices of tourist rental accommodation. New evidence using the quantile regression approach. *Tourism Management Perspectives*, [online] 33November 2019, p.100632. Available at: https://doi.org/10.1016/j.tmp.2019.100632.

Selim, H., (2009) Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, [online] 362 PART 2, pp.2843–2852. Available at: http://dx.doi.org/10.1016/j.eswa.2008.01.044.

Varma, A., Sarma, A., Doshi, S. and Nair, R., (2018) House Price Prediction Using Machine Learning and Neural Networks. *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, pp.1936–1939.

World Health Organization., World Bank., Ruiz-Ibán, M.A., Seijas, R., Sallent, A., Ares, O., Marín-Peña, O., Muriel, A., Cuéllar, R., Mobasheri, A., Batt, M., Quintana, J.M., Escobar, A., Arostegui, I., Bilbao, A., Azkarate, J., Goenaga, J.I., Arenaza, J.C., Murphy, L.B., Helmick, C.G., Schwartz, T.A., Renner, J.B., Tudor, G., Koch, G.G., Dragomir, A.D., Kalsbeek, W.D., Luta, G., Jordan, J.M., Martín-Fernández, J., Gray-Laymón, P., Molina-Siguero, A., Martínez-Martín, J., García-Maroto, R., García-Sánchez, I., García-Pérez, L., Ramos-García, V., Castro-Casas, O., Bilbao, A.,

Kloppenburg, M., Berenbaum, F., Kohring, J.M., Pelt, C.E., Anderson, M.B., Peters, C.L., Gililland, J.M., Macías-Hernández, S.I., Zepeda-Borbón, E.R., Lara-Vázquez, B.I., Cuevas-Quintero, N.M., Morones-Alba, J.D., Cruz-Medina, E., Nava-Bringas, T.I., Miranda-Duarte, A., Joseph, D.S., Walton-Paxton, E., Ostendorf, M., van Stel, H.F., Buskens, E., Schrijvers, A.J.P., Marting, L.N., Verbout, A.J., Dhert, W.J.A., Seijas, R., Ares, Ó., Sallent, A., Gómez-Valero, S., García-Pérez, F., Flórez-García, M.T., Miangolarra-Page, J.C., Maillot, C., Harman, C., Al-Zibari, M., Sarsam, K., Rivière, C., Herdman, M., Badia, X., Berra, S., Cnudde, P., Rees, H.W., Greene, M.E., Learmonth, I.D., Young, C., Rorabeck, C., Zhang, W., Moskowitz, R.W., Nuki, G., Abramson, S., Altman, R.D., Arden, N., Bierma-Zeinstra, S., Brandt, K.D., Croft, P., Doherty, M., Dougados, M., Hochberg, M., Hunter, D.J., Kwoh, K., Lohmander, L.S., Tugwell, P., Ramadani, R.F., Erastus Mosha, Ramadani, R.F., Crawford, R.W., Murray, D.W., Peláez-Ballestas, I., Sanin, L.H., Moreno-Montoya, J., Alvarez-Nemegyei, J., Burgos-Vargas, R., Garza-Elizondo, M., Rodríguez-Amado, J., Goycochea-Robles, M.V., Madariaga, M., Zamudio, J., Santana, N. and Cardiel, M.H., (2020) No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title. *Osteoarthritis and Cartilage*, [online] 282, pp.1–43. Available at: http://journals.sagepub.com/doi/10.1177/1120700020921110%0Ahttps://doi.org/10.1016/j.reuma.2018.06.001%0Ahttps://doi.org/10.1016/j.arth.2018.03.044%0Ahttps://reader.elsevier.com/reader/sd/pii/S1063458420300078?token=C039B8B13922A2079230DC9AF11A333E295FCD8.

Ye, P., Wu, C.H., Qian, J., Zhou, Y., Chen, J., De Mars, S., Yang, F. and Zhang, L., (2018) Customized regression model for Airbnb dynamic pricing. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.932–940.

Zehtab-Salmasi, A., Feizi-Derakhshi, A.-R., Nikzad-Khasmakhi, N., Asgari-Chenaghlu, M. and Nabipour, S., (2020) Multimodal price prediction. [online] July. Available at: http://arxiv.org/abs/2007.05056.

Mishra, D., (2019) Regression: An Explanation of Regression Metrics and What Can Go Wrong

APPENDIX B

GitHub Link for the code used in the entire study: -

https://github.com/ShaoniMukherjee/Airbnb-Price-Prediction

APPENDIX C


Research Topic Selection
17 Aug 2020
Research Proposal Submission
28 Sept 2020
Research Interim Report Submission
30 Nov 2020
Final Thesis Submission
01 Feb 2021
Video Presentation Submission
04 Feb 2020

Shaoni Mukherjee.docx

ORIGINALITY REPORT

1%	1%	1%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	hdf.ncsa.uiuc.edu Internet Source	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		