|  |  |  |  |
| --- | --- | --- | --- |
| **Synopsis Report**  **on**  **Airbnb Price Prediction**  **Submitted as partial fulfillment for the award of**  **BACHELOR OF TECHNOLOGY**  **DEGREE**  **Session 2022-23 in**  **CSE-Data Science**  **By:**  **PrakharTiwari**  **200321540040 Somesh Verma**  **2000321540057**  **Priyam Rastogi**  **2000321540042**  **Under the guidance of:**  **Mr. Prabhat Singh Assistant Professor**  **DEPARTMENT OF CSE-DS**  **ABES ENGINEERING COLLEGE, GHAZIABAD** | | | |
|  |  |  | naac A Grade Engineering College |
| **AFFILIATED TO**  **DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW**  **(Formerly UPTU)** | | | |

Student’s Declaration

I / we hereby declare that the work being presented in this report entitled **“** **Airbnb Price Prediction”** is an authentic record of my/ our own work carried out under the supervision of Mr. Prabhat Singh**, Assistant Professor, CSE-DS.** The matter embodied in this report has not been submitted by us for the award of any other degree.

**Date:**

**Signature of student Signature of student Signature of student**

**Name: Prakhar Tiwari) (Name: Somesh Verma) (Name: Priyam Rastogi)**

**(Roll No.2000321540040) (Roll No.2000321540057) (Roll No.2000321540042)**

**Department: CSE-DS Department: CSE-DS Department: CSE-DS**

This is to certify that the above statement made by the candidate(s) is correct to the best of my knowledge.

**Signature of HOD Signature of Supervisor**

**…………………… Mr. Prabhat Singh**

**CSE-DS Assistant Professor**

**Date: CSE-DS**

**Acknowledgement**

We would like to convey our sincere thanks to  **Mr. Prabhat Singh** for giving the motivation, knowledge and support throughout the course of the project. The continuous support helps in a successful completion of project. The knowledge provided is very useful for us.We also like to give a special thanks to the department of Information and Technology for giving us the continuous support and opportunities for fulfilling our project.We would also like to extend our sincere obligation to………………., Head of Department, Information Technology for providing this opportunity to us.

**Signature of student Signature of student Signature of student**

**Name: Prakhar Tiwari) (Name: Somesh Verma) (Name: Priyam Rastogi)**

**(Roll No.2000321540040) (Roll No.2000321540057) (Roll No.2000321540042)**

**Department: CSE-DS Department: CSE-DS Department: CSE-DS**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Contents** | **Page No.** |
|  | Student’s Declaration | i |
|  | Acknowledgement | iii |
|  | List of Figures | v |
|  | List of Tables | v |
|  | Abstract | 1 |
| Chapter 1: | Introduction | 1 |
| Chapter 2: | Related Work/Methodology | 3 |
| 2.1: | Existing Approaches | 3 |
| 2.2: | Comparative Analysis of Existing Works | 4 |
| Chapter 3: | Project Objective | 5 |
| Chapter 4: | Proposed Methodology | 5 |
| Chapter 5: | Design and Implementation | 7 |
| 5.1: | Work Flow Diagram | 5 |
| Chapter 6: | Results and Discussion | 9 |
| Chapter 7: | Conclusion and Future Scope | 11 |
|  | References | 12 |

**List of Figures**

|  |  |
| --- | --- |
| **Figure No and name** | **Page No.** |
| **Fig.1.** Proposed Approach | 4 |
| **Fig.2.** Work Flow Diagram | 7 |
| **Fig.3.** Machine Learning Algorithms Analysis | 9 |

**ABSTRACT**

Figuring out the price of a listed Airbnb rental is an important and difficult task for both the host and the customer. For the former, it can enable them to set a reasonable price without compromising on their profits. For the customer, it helps understand the key drivers for price and also provides them with similarly priced places. This price prediction regression task can also have multiple downstream uses, such as in recommendation of similar rentals based on price. We propose to use geolocation, temporal, visual and natural language features to create a reliable and accurate price prediction algorithm.

## 

## Chapter 1

**Introduction**

Airbnb has hosted over 60 million people in 34,000 cities across the world and is continuing to grow quickly. Airbnb provides a convenient source of income for people who have otherwise vacant space and for guests looking for affordable and convenient housing options. With any service, trying to monitor and understand the underlying pricing dynamics of the Airbnb market is very important both for hosts and guests. As users continue to grow on both the supply and demand side, homeowners may find it hard to properly price their property. Airbnb has recognized this and conducted considerable research into suggesting pricing from a [supply side standpoint](http://spectrum.ieee.org/computing/software/the-secret-of-airbnbs-pricing-algorithm).

We seek to analyze over 27,000 listings in the NYC area in order to better understand how the use of listing attributes such as bedrooms, location, ratings, and more can be used to accurately predict the optimal listing price both for the host and guest. Holiday and seasonality is another useful component that can attract more customers and drive higher prices, but it is unclear how much of a premium one should pay per holiday. With better price suggestion estimates, Airbnb home providers can reach an equilibrium price that optimizes profit and affordability. The objective of this project is to build a model that predicts the optimal price of a property taking into account listing features and seasonality. The end goal is so users can understand what features of an AirBnB listing are most important as well as how prices should be fluctuating based on seasonality

**Problem Statement :-**

1. 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.
2. Airbnb losses 46% of revenues due to inefficient pricing of the listings. Further there is a need the study also helped to understand if there was any post COVID-19 effects in pricing.
3. Most of the paper has also not captured the non-linearity of the dataset well and has thus resulted in unsatisfying RMSE and MAE valuesTo study different machine learning algorithms for sentiment analysis and to find out the best one that fits it.

## Chapter 2

## Related Work

The related work associated with our project is given below:

### Existing Approaches

* + Based on the large number of public datasets provided by Airbnb, Airbnb price prediction with machine learning technique becomes a popular study. Yuanhang Luo, Xuanyu, Zhou, and Yulian Zhou [1] developed an Airbnb price prediction model using Random Forest, XGBoost, and Neural Network. They eliminated several features, such as host\_id and customer\_name to reduce noise and keep features like country\_code and number of bedrooms to build the model. Textual data, such as house description and neighberhood\_review, is also considered as textual features. After performing extensive feature engineering and extraction on the New York and Paris dataset, XGBoost and Neural Network perform better than other models. R-Squared and Median Squared Error (MSE) are used to examine the result. The R-Squared error for XGBoost is 0.722, and 0.769 for Neural Networks.
  + Research conducted by (Fei, Yue. et al., 2020) used California 2020 data to get the general trend of rental price on the marketplace to provide some reasonable suggestions about rental price and later the host can make adjustments as per their need.
  + Research which was conducted by (Group et al., 2020) performed graphical and statistical analysis in the 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.
  + As described by (Laurent et al., 2015) NLP was used to convert guest comments, interaction between users, host descriptions to obtain more information in optimising the target variable.
  + Research conducted by (Luo et al., 2019) was done using a variety of regression model approaches which included, Linear Regression, k nearest neighbour regression, random forest regression, XGBoost, ANN to predict the target variable.
  + Research which was conducted by (Keating, 2018) to predict the pricing of Airbnb rentals in Seattle. The main aim of the study was to find out the relationships between the independent variables and the dependent variables.19[1].

### Comparative Analysis of Existing Works

* + - We try to perform extensive feature engineering, which can aid our models to predict prices accurately, along with a variety of tuned models.
    - We believe that customer reviews are also an important factor that is helpful to build the price predictor. In this project, we analyze and classify customer reviews for each listing as a feature to improve the prediction accuracy

## Chapter 3

**Project Objective**

* The main aim of the research was to predict the price of Airbnb house based on multiple factors like locality, reviews per month, availability, room type etc.
* . The find the attributes which has a significant impact on pricing.
* To find if there exist a relationship between the name of the property and the price of the properties.
* To use different ML model and find the best one.This project will analyze different Machine Learning Algorithms and finds the one with best accuracy.

## 

## Chapter 4

## Proposed Methodology

The proposed methodology related to our project is given below:

Step 1**: Data extraction :**

The dataset is taken from Kaggle, it is named as Airbnb New York data set.

Step 2:  **Data preprocessing :**

In this part, first we need to understand the attributes and behavior of data like, how many columns are there, what are the data types and how much data is missing? We treated missing values, then filled mean values in place of null in some columns and simply remove the whole row in some column. Also remove insignificant columns from the dataset. It involves the following steps:

* Removal of null values.
* Removal of links, gifs, emoji, images and special characters.
* Removal of missing data.
* Removal of non-English words.

Step 3:  **Exploratory data analysis :**

It is an approach to analyze the dataset and summarize the data and find patterns in it. We had also analyzed categorical and numeric features by univariate and bivariate analysis.

Step 4: F**eature engineering :**

It is a process of using domain knowledge to extract feature from raw data. These features are used to improve the performance of machine learning model. In this step converted categorical features like neighbourhood group, root\_type and neighbourhood into numerical features because these are not processed in machine learning. Also, I had to perform log transformation on Price attribute because it is left skewed.

Giving the step 3 output in different machine learning algorithms and analyze it to find the algorithm with best accuracy.

Step 5: **Model building :**

in this step, I had applied Linear Regression, Ridge and Lasso Regression and Decision Tree on pre processed data. I found Decision Tree for predictive model with the help of R-squared score because its accuracy is the higher than all other models.

Step 6th : **Model deployment :**

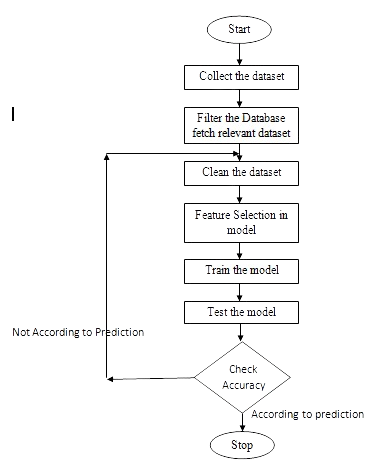
first dumped the most accurate model into pickel file. We created a Flask app which takes input as Neighbourhood group, room type availability and show predicted result on screen.

## Chapter 5

**Design and Implementation**

The design and implementation of our project is as follows:

### Work Flow Diagram



**Fig.1.** Work Flow Diagram

The dataset has been extracted from Kaggle .Python library Numpy is used for the numerical computation and pandas is used for the data manipulation. Natural Language Toolkit is used for the preprocessing of the dataset.

Matplotlib is used for the graphical representation of results.

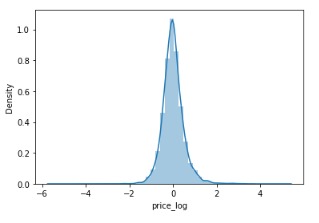
Technology used here are:

1. Python , Flask
2. Numpy, Pandas , Matplotlib, Sklearn, Seaborn
3. Folium
4. Pickle
5. Algorithms – Linear Regression , Ridge regression , lasso regression , Decision Tree

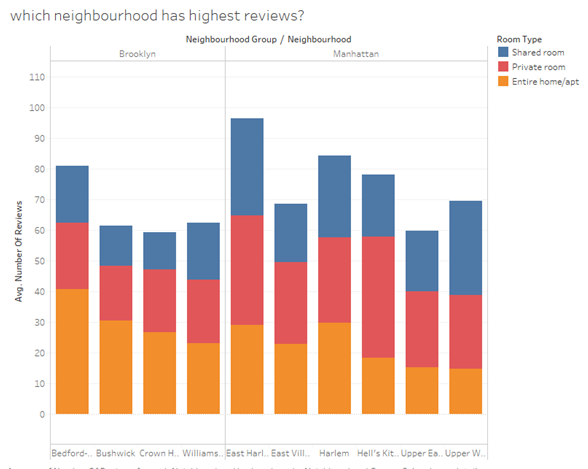
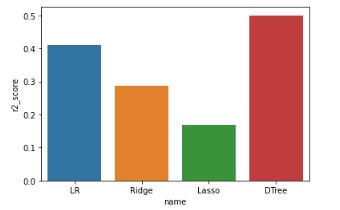
## Chapter 6

## Results and Discussion

The result we got from analyzing the data is given below in Fig.3.



**Fig.2.**Machine Learning Algorithms Analysis



**Fig.3. Analysis between Neighbourhood and Availability**

## Chapter 7

**Conclusion and Future Scope**

* Through this research we are able to understand the 2019 data and the price distribution of the rental properties.
* Avg. Price of Airbnb house in Manhatten is 31% which is maximum among all the neighbourhood group. Bronex is the cheapest as its contribution in pi chart is 14%. Most of the persons prefers Entire Room/ Apartment and very less people prefers Shared Room to stay.
* Average price of Entire Room/Apartment is 160 which is twice of the price of Private Room.
* Further the average price in both the cases is below 200 $ which shows that in an average the properties are quite affordable.
* We can also add some state and location specific features to improve the model.
* For future works, we expect the price prediction model to be improved using a larger dataset with balanced customer reviews since the mean squared error in the Neural Network model is still decreasing at the end of the training phase. Additionally, public Airbnb datasets contain more positive reviews than negative reviews. A well balanced dataset should be helpful to build a more accurate price prediction model. Other than customer reviews, historical prices might be another key factor to evaluate the price. Customers expect a lower price if the price is constantly decreasing.
* We would like to know if the model performs the same using datasets for different cities.

## References

1. Luo, Y. et al. “Predicting Airbnb Listing Price Across Different Cities.” (2019).
2. Tang, E.. “Neighborhood and Price Prediction for San Francisco Airbnb Listings.” (2015).
3. Nguyen, QuangTrung. “QuangTrungNguyen/Airbnb-Pricing-Prediction.” GitHub, github.com/QuangTrungNguyen/Airbnb-pricing-prediction.
4. Cmdline. “Linear Regression Using Matrix Multiplication in Python Using NumPy.” Python and R Tips, 18 Mar. 2020, cmdlinetips.com/2020/03/linearregression-using-matrix-multiplication-in-python-using-numpy/.
5. Breiman, L. (June 1997). "Arcing The Edge". Technical Report 486. Statistics Department, University of California, Berkeley.
6. Mason, L.; Baxter, J.; Bartlett, P. L.; Frean, Marcus (1999). "Boosting Algorithms as Gradient Descent" (PDF). In S.A. Solla and T.K. Leen and K. Müller (ed.). Advances in Neural Information Processing Systems 12. MIT Press. pp. 512–518.
7. Friedman, Jerome. (2000). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 29. 10.1214/aos/1013203451.
8. Kennedy, Peter (2003). A Guide to Econometrics (Fifth ed.). Cambridge: The MIT Press. pp. 205–206. ISBN 0-262-61183-X.
9. An Introduction to Support Vector Regression, towardsdatascience.com/anintroduction-to-support-vector-regression-svr-a3ebc 1672c2.