Abstract

Airbnb is a company that offers a home rental platform that allows people to find, list and rent temporary accommodation all around the globe. It has become increasingly famous among people for accommodation as it has brought a latest business model to the hospitality industry. As an online marketplace, its main function is the connection of two third parties such as hosts and guests. It is a challenging task for the house owner to price a rental home and attract customers. Also, customers need to evaluate the price of the rental property based on the listing details and verify if it is worth the deal. In this paper we aim to develop a dependable price prediction system to decide prices of Airbnb listings in various cities across United States using several machine learning techniques to help both the customers and the property owners. For this, we obtained a dataset and performed data cleaning and data pre-processing and then visualized the data to geta better understanding about the type of data. This assisted us in getting the insights about which attributes are most important for predicting the prices. Further, we trained our model using various methods such as Decision Tree Regressor, Random Forest, XGBoost, KNN classifier and KNN regressor. After implementing with several mentioned approaches, the most accurate model was selected based on the RMSE and F1 accuracy score.

Keywords: Machine Learning, Decision Tree, XGBoost, Gaussian Naïve Bayes, KNN Classifier, KNN Regressor, Random Forest, Airbnb.

1. Introduction

The online marketplace is immensely popular with the customers and present a convenient way to compare prices of various rental properties. Airbnb is one such marketplace that has been a critical market driver in addressing a low-cost accommodation problem. It brings forth the lodging experience by connecting people who desire to rent their property with the people looking for accommodation in a particular area. The main concern in this work is the pricing of the houses for the customers as well as the hosts. The pricing in this scenario can be more challenging than the hotels as Airbnb don't have their own pricing system like the hotels do. The main problem is that when it comes to pricing there is no explicit guideline available for Airbnb host, to set a price for their property. On the other hand, customers must pay the offered price with minimal knowledge of the property's true value. Ensuring fair pricing directly affects booking activities, and also matters to the well-being of the e-commerce environment. Thus, studying the reasonable forecast and fair suggestion of prices of Airbnb listings can have huge real-life values and may generalize to other applications as well. Therefore, it is important to address this problem and have a model to predict the pricing for decision making purposes for landlords, consumers as well as stakeholders. Being a student and consumer ourselves, it is important to take into consideration our finances and thus book a property which has reasonable pricing. This project aims to help consumers and landlords estimate the current market value of a house. For this, we will be using different data mining approaches and techniques to analyze the price of the rooms. These models will help in analyzing components of listing prices and estimate the future prices on Airbnb listings with accessible information which will be beneficial in providing accurate pricing strategy to the hosts and other stakeholders along with insight into overall rental accommodation scenario.

2. Problem Statement

Airbnb is an online marketplace that bring forth the lodging experience by connecting people who desire to rent their property with the people looking for accommodation in a particular area. Since its inception, the concept of Airbnb has been focused on helping low-cost lodging residents. The main problem is that when it comes to pricing there is no explicit guideline available for Airbnb host, to set a price for their property. Currently, there is no convenient way for a new Airbnb host to decide the price of his or her listing. New hosts must often rely on the price of neighboring listings when deciding

on the price of their own listing. Nor there is any way for customer to check if they are getting the right price while booking on Airbnb. Customers must pay the offered price with minimal knowledge of the property's true value. Ensuring fair pricing directly affects booking activities, and also matters to the well-being of the e-commerce environment. We build a predictive price modelling tool where host or customer can get an estimate of price for Airbnb based on factors like location of the listing, room type, reviews etc.

3. Literature Review

Analysis of Airbnb Prices using Machine Learning Techniques
Jasleen Dhillon, Nandana Priyanka Eluri, Damanpreet Kaur, Aafreen Chhipa, Ashwin Gadupudi, Rajeswari Cherupulli Eravi, Matin Pirouz.

This paper aims to make predictions about the Airbnb listing prices from various cities across US, linking the prices with the locations and then analyzing minimum and maximum number of bookings per month. For this a descriptive, exploratory and prescriptive analysis is performed for understanding the data which helped to gain important insights that needs to be considered for predicting prices. Further, in this paper they have used models such as linear and logistic regression and random forest. Out of which the best results were obtained on Random Forest model. Thus this paper will be successful in giving the host accurate information as per the requirements such as location, ratings, prices and further it can be improved by using reviews and summary attributes.

Machine Learning Predictions of New York Airbnb Prices
Ang Zhu, Rong Li, Zehao Xie

In this paper, the insights to latest business model is provided. Nearly a sample of 48 896 listings in New York City are considered and analyzed for making a price prediction model with natural language processing and machine learning techniques. For predicting prices, they have used methods such as linear regression, deep neural network, generalized additive model, Random Forest and XGBoost. From these methods the best were Random Forest and XGBoost. They have also performed K fold validation as the RMSE was underestimated. A five-fold cross validation was carried out on each of the different methods to see the results. The deep neural network uses a back propagation algorithm for training model. But this will also result in decrease in degrees of freedom, resulting in over-fitting. Also it is noted that having combination of different models brings the benefits of multiple models into one, while decreasing the dominating power of one specific model.

Learning-based Airbnb Price Prediction Model
Siqi Yang

The purpose of this work is to do the analysis of the features related to the price and the research considers important ones for developing prediction model. Along with this, the research also gives suggestions for hosts about how to increase their price by making some changes. Further, they carried feature selection to obtain most significant features and develop an accurate price prediction model. In this paper, the machine learning techniques include tree-based model XGBoost and Neural Network with feature importance selection. And, the XGBoost model performs best in this research.

Airbnb Price Prediction Using Machine Learning and Sentiment Analysis
Pouya Rezazadeh Kalehbasti , Liubov Nikolenko, Hoormazd

This paper has come up with the best-performing model for predicting the Airbnb prices based on a limited set of features including property specifications, owner information, and customer reviews on the listings. They have developed a reliable price prediction model using machine learning, deep learning, and natural language processing techniques to aid both the property owners and the customers with price evaluation given minimal available information about the property. They have

used models such as support-vector regression (SVR), K-means Clustering (KMC), and neural networks (NNs) for creating the prediction model.

Melbourne Airbnb Price Prediction.
Tiancheng Cai, Kevin Han, Han Wu

They proposed a model for predicting the Airbnb prices in Melbourne by using various regression models and made comparisons between different methods. They have used traditional ML methods such as linear regression, ridge regression, support vector regression, random forest regression, gradient boosting and neural networks to output the predicted prices of listings. In Neural Network, they have considered continuous and categorical features, only description data, by using all text data and by using all the features. They plan to consider feature selection and carry out two-step modeling, which would divide training sets into K groups based on price range and build separate models for each group.

4. Methods and Techniques

We implemented the following models in our project:

- Decision Tree
- Random Forest
- XGBoost
- Gaussian Naïve Bayes
- KNN Classifier
- KNN Regressor

a) Decision Tree:

A Decision Tree is a tree-structured plan of a set of attributes to test in order to predict the output. Decision Trees are a type of Supervised Machine Learning, where the data is continuously split according to a certain parameter. We imported the Decision Tree Regressor for this. To get the best possible results from our model we need to decide the perfect combination of hyper-parameters. Selection of right combination of parameters or features plays very important role in getting better accuracy. It involves training the Model with different values for a set of parameters.
We tried different set of parameters and got best result by using
'calculated_host_listings_count' and 'room_type'.

b) Random Forest:

Random Forest is improved from bagged decision tree, and uses modified tree learning algorithm with feature bagging. The model output an average of all k trees' prediction. Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression

c) KNN Regressor:

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations. We evaluated our model using KNN Regressor to check if it is the optimum method in our case.

d) KNN Classifier:

K-Nearest Neighbors (KNN) is one of the easiest algorithms used in Machine Learning for regression and classification problem. It uses data and classify new data points based on similarity measures. It uses majority vote to its neigbors for classification. To achieve better accuracy, we implemented our model using KNN classifiers and we used the minkowski distance for that purpose. The reason for selecting this distance metric was that it is intended for real valued vector spaces, so in this case the distances can be represented as a vector with length which is not negative.

e) XGBoost:

We also tried using XGBoost since it is an implementation of gradient boosted decision trees and is better for obtaining good performance and speed. We used XGBoost Classifier by importing XGBClassifier from xgboost.

5. Discussion and Results

- **Datasets**

A few existing sources that were referred are mentioned below:

The Dataset used for this project is available on Kaggle. The dataset consists of 17 columns and approximately 226030 rows. The dataset includes NaNs, and data is of mixed types

Dataset Name: “AB_US_2020.csv"

Few important variables:

- Neighbourhood_group: neighbourhood of the listing in United States
- latitude: latitude coordinates
- Id: unique listing id
- name: name of listing
- host_id: unique host id
- city: name of city
- neighbourhood: neighbourhood listing
- host_name: name of host
- longitude: longitude coordinates
- room_type: listing space type
- price: price in dollars
- minimum_nights: minimum number of nights
- number_of_ reviews: number of reviews of the listing
- reviews_per_month: number of reviews per month of the listing
- calculated_host_listings_count: number of listings per host
- availability_365: number of days per year that listing is available

Variables that are irrelevant to the analysis, such as “id”, “neighbourhood group”, “last review”, “city”, “room_type”, “neighbourhood”, “price_range”, “host_name”, “name”, were excluded. Additional

adjustments were applied to variables such as “price_range” and “room_type”. We label encoded price_range and did one hot encoding on room_type.

We created an additional column to categorised prices based on ranges into 4 categories.

The table below lists the final variables in the dataset:

- neighbourhood_group: neighbourhood of the listing in United States
- latitude: latitude coordinates
- longitude: longitude coordinates
- room_type: listing space type
- Price_range_category: prices categorised into 4 categories such as economic, low-mid,high mid
- minimum_nights: minimum number of nights
- number_of_reviews: number of reviews of the listing
- reviews_per_month: number of reviews per month of the listing
- calculated_host_listings count: number of listings per host
- availability_365: number of days per year that listing is available

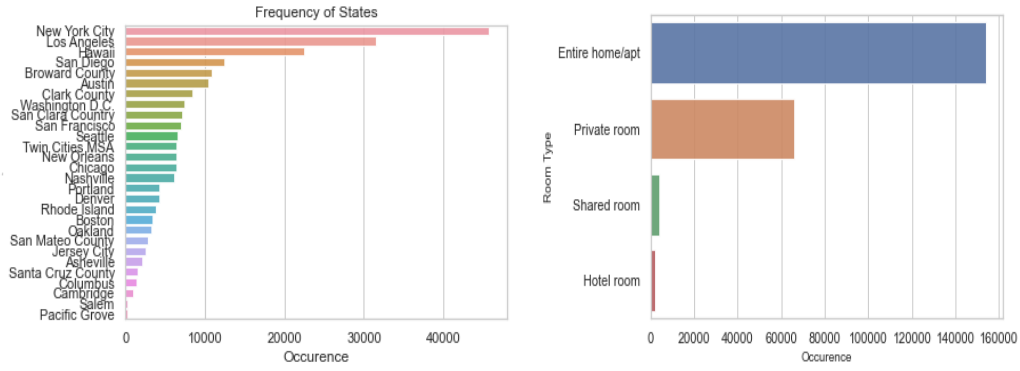
- **Data Visualization**

To get a better insight into where the listings are located, the number of listings in various cities are plotted in the figure below. In the given dataset, New York City has most number of listings followed by Los Angeles and Hawaii.

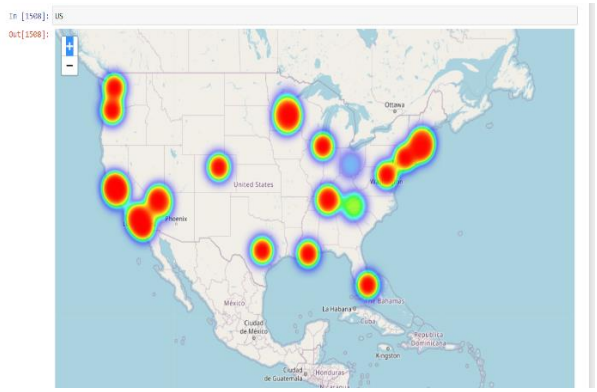
Airbnb offers three types of listings:

- Entire home/apartment
- Private Room
- Shared Room

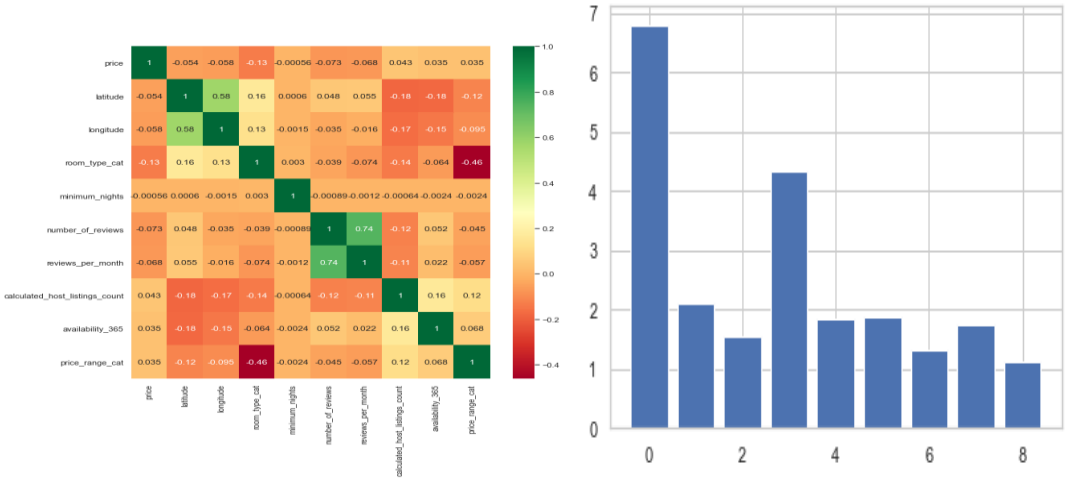
Entire home/apartment is the most popular type of listing followed by Private room and then Shared room.



Further, we plotted our dataset on the U.S. map to get clear idea of our dataset, which can be seen as follows:



We then plotted a heatmap to represent our data in form of colors, where darker shades indicate higher values than the lighter shades. Also, we performed feature importance to find out which features are more score and therefore are important.

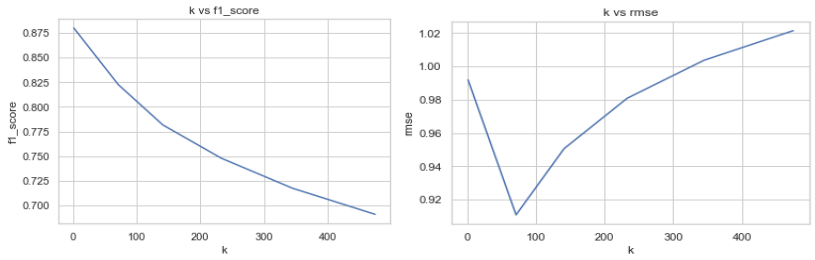


- **Evaluation Metrics**

  Since this is a regression task as we have to predict the price of listing we can use various evaluation metrics such as Root Mean Squared Error, F1 accuracy score, Mean Absolute Error, etc. RMSE penalizes errors to a greater magnitude than MAE, therefore that was used in this project. We checked the 'rsquare' score of KNN Regressor, Decision Tree Regressor, Random Forest Regressor and F1 score of KNN Classifier model on train and test data sets to see how our model performed when pricing the price. We used the Minkowski distance metric as it is for real valued vector spaces so in this the distances can be represented as a vector with length which is non negative.

- **Experimental Results**

For the results in this section, we used the default settings for all machine learning approaches imported from the sklearn package. We evaluated the model using Decision Tree Classifier at first but did not get expected results as the accuracy achieved was very low. So, we thought of using Decision Tree Regressor. For this we need to select proper combination of features from the available features. We calculated RMSE score and got result as 1.19 after using features calculated_host_listings_count and room_type while price_range was our target variable. To further check if we achieve better results, we evaluated our model using Random Forest Regressor. We calculated the RMSE score and obtained result as 1.12 which was slightly better than Decision Tree Regressor but still not as expected. We got bad result for Random Forest Classifier. Next, we tried using KNN Regressor and Classifier. For which we needed two important components that are distance metric and value of K nearest neighbors. For selecting value of k we decided to calculate the square root of total number of rows of our dataset as generally square root is the optimum value, which gave us values as 475. So, we plotted multiple values ranging from 0 to 475 at different intervals and the square root was not the most optimal value in our case. And the distance selected was Minkowski distance. KNN regressor gave us RMSE score of 1.16 while the KNN Classifier gave us F1 accuracy score of 0.83. Which were the best results compared to other models.



Further we implemented using XGBoost Classifier but it did not give us the expected results as the RMSE score was not good. But we achieved good result for XGBoost Regressor (RMSE = 1.10). We also tried using Gaussian NB but did not get expected results.

.

6. Conclusion

We have used the United States dataset for making a model to predict the price of each property or house in different regions in U.S. based on the location, different type of reviews, availability, city and previous prices, etc. For this we explored the dataset using data visualization to gain insights from it. Also, we removed the irrelevant data from the dataset to achieve better accuracy. We divided the dataset into Train and Test split as 75/25. We also tried to find out which Machine Learning algorithm works best for this dataset in predicting the prices with better accuracy. In our research, we have found out that Random Forest model works best with this kind of data. Next, we have performed extensive feature extraction and engineering, and experimented with various machine learning approaches such as XGBoost, Random Forest, Decision Tree, KNN Regressor, KNN Classifier in predicting Airbnb listing price. We showed that KNN Classifier gave us the best results compared to others, while we got good results for KNN regression. To conclude, we can say that the best top features for correctly determining the price of Airbnb listings are “calculates host listings and room type.

- **Directions for Future Work**

Since our data is not balanced, we plan to handle it effectively to reduce the location based bias. Also, we plan to find the popularity of a listing based on available features. We also plan to estimate the popularity of the property based on the given features. Further, based on different

factors we will try to recommend a title to the host for his/her listing.

References

1. Analysis of Airbnb Prices using Machine Learning Techniques
Jasleen Dhillon, Nandana Priyanka Eluri, Damanpreet Kaur, Aafreen Chhipa, Ashwin Gadupudi, Rajeswari Cherupulli Eravi, Matin Pirouz.
https://ieeexplore.ieee.org/document/9376144

2. Machine Learning Predictions of New York Airbnb Prices
Ang Zhu, Rong Li, Zehao Xie
https://ieeexplore.ieee.org/document/9253078

3. Learning-based Airbnb Price Prediction Model
Siqi Yang
https://ieeexplore.ieee.org/document/9406836

4. Airbnb Price Prediction Using Machine Learning and Sentiment Analysis
Pouya Rezazadeh Kalehbasti , Liubov Nikolenko, Hoormazd
https://www.researchgate.net/publication/334783073_Airbnb_Price_Prediction_Using_Machine_Learning_and_Sentiment_Analysis

5. Melbourne Airbnb Price Prediction.
Tiancheng Cai, Kevin Han, Han Wu
http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26586189.pdf

6. Predicting Airbnb Listing Price Across Different Cities
http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647491.pdf

7. Machine Learning Prediction of New York Airbnb Prices.
Zhu, A., Li, R., & Xie, Z. https://ieeexplore.ieee.org/document/9253078

8. Boston Airbnb Price Prediction
https://medium.com/geekculture/boston-airbnb-price-prediction-67f65aa54768