NEW YORK AIRBNB PRICE AND AVAILABILITY PREDICTION

Sai Teja Dampanaboina
B.Sc. Computer Science
Otto von Guericke Universität
Magdeburg, Germany
saitejadampanaboyana@gmail.com

***Abstract*—A dataset is utilized in this project to predict Airbnb prices and availability. The datasets are imported, cleaned, pre-processed, and then load into machine learning models after performing certain visualization techniques to comprehend the insights. Airbnb pricing prediction using the current dataset in New York City is a regression issue. In conclusion, predictions were made using machine learning models such as the Univariate Analysis, Random Forest Regressor, XG Boost, and MinMax Scaler transforms.**

***Index Terms*—Airbnb price, Machine learning models, classification, Regression, Data mining. Data exploration, price prediction, availability prediction.**

I. Introduction

Despite the difficulties and dangers, the world is currently facing, the travel and tourism industry significantly contributes to the global economy. According to the World Travel and Tourism Council, travel and tourism have a $2.3 trillion direct economic impact on the world and an indirect economic impact that is nearly three times as great. The sector's contribution to international wealth creation, the advancement of culture globally, and the creation of jobs all add to its relevance (World, 2017).

Policymakers and practitioners in the public and commercial sectors have long shown interest in the hospitality sector, which is a component of the travel and tourist business. Planning for tourism is a common component of mainstream public policy objectives in many industrialised economies. It takes a methodical and scientific strategy to support growth in this economic sector, and planning is a complicated process because the tourist industry is interconnected with many other businesses (Dredg & Jamal, 2015). Managers are curious about the elements affecting the industry in regard to their revenue management systems. They keep a tight eye on pricing fluctuations in order to foresee any movement that can affect their company. They deal with perishable goods, therefore this is very crucial. Numerous goods and services in this industry must be consumed immediately and cannot be stored (Legohérel, Poutier, & Fyall, 2014). This study was conducted utilising the most cutting-edge forecasting technologies now available to look into the hospitality industry, in the context of the rising literature on hotel pricing prediction.

Some business professionals have thought about gradually replacing conventional management techniques with revenue management practises to boost performance (Rodrguez-Algeciras & Talon-Ballestero, 2017). To further enhance performance, this method might be used to gradually include complex analytical procedures employing AI algorithms. The objective of revenue management is to match the right product with the right consumer at the right time and place. In recent years, classification- and regression-oriented AI models namely support vector machines and random forests , as well as neural networks, have proven their worth, winning forecasting competitions all over the world.

There are several internal and external elements that affect costs in this complicated sector. To forecast using machine learning techniques

People who prepare data must take these elements into account and collect all of this data to train the model accurately.

Among the internal factors, there are historical prices,the location, the number and size of rooms,amenities and services provided ,demand fluctuations based on seasonality and holidays, and more. External factors include pricing strategies of competitors,economic changes,events such as sports competitions or festivals,exchange rates and political situations, etc.

It's crucial to know which elements should be considered and which ones may be disregarded at the same time.

Nowadays, the house rental market is expanding quickly, necessitating the adoption of machine learning techniques for pricing prediction on Airbnb in order to improve the lodging industry. With the help of Airbnb, anyone may advertise, search for, and book homes for its customers all around the world. Guests are free to remain on a host's property for whatever long they like. They differ from the typical hotel room in that visitors can experience and learn about the local way of life while staying in these accommodations .The listings on Airbnb provide a wide range of options, all on one platform, from shared beds to opulent homes. Customers can make recommendations through a peer-review process after making reservations. These reviews frequently dictate how much owners may book or make. The project's major goal is to create classification and regression models with the fewest characteristics and most accurate data feasible. In order to get insights from a dataset, it was first imported, cleaned, and pre-processed, and then some visualizations were carried out. All datasets were then fitted into machine learning models. The

best fit is determined when using models like Random Forest Regressor and boosting regression to estimate Airbnb prices and availability using regression analysis.

A. NYC Airbnb Open Data

The New York Airbnb Open data Datasets comprises one CSV file with 16 columns and 48.8k in order to increase the number of trip options. We must make a regression prediction about the cost of an Airbnb. Website: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

B. Target Variable

The desired variable is price and availability. This issue involves regression. We can forecast the price based on the other variables. Additionally, several KDD features for data cleansing are incorporated.

C. objectives

To fulfil the aim of the research, following objectives must be satisfied. 1. Identifying a dataset that will be used for the study. 2. Identify the supervised machine learning models that are capable of predicting price and availability values.

II. Related Work

A. Price Prediction

It has been decided to organize hotel room rates in order to encourage visitors to stay in hotels. Depending on the host-guest connection, specific Airbnb assets are donated to a specific cause. In order to determine if there is a difference in costs between those before and after the founding of the Airbnb firm, it is necessary to understand the variables and the impact of hotel room rates. individuals transferring from Airbnb to hotels is one of the main causes, but in order to better understand the shift, we need to look at the kinds of individuals who prefer Airbnb to hotels. It's also crucial to note whether they switched from hotels to Airbnb permanently, whether they wish to book accommodations through both platforms, and if they haven't changed at all. The Airbnb host has to be informed of what is expected of the home in contrast to hotels. Airbnb pricing is influenced by the listing's many aspects, the interaction between various attributes, and how they vary while displaying the price, therefore it's crucial to consider how they vary while reflecting the price. As a consequence, the host was able to select the features that would aid the organization while still keeping an eye on the price. In this study, the natural language processing approach is used to assess over 48896 hotels listed on Airbnb.com in New York City and construct a prediction model. There were several models utilized, including the generalized additive model, a deep neural network, linear regression, and random forest. The XGBoost, Bagging (combination of the five models) was the outcome of the final forecast. The XGBoost, Random Forest, and Airbnb bagging data provided the greatest results.

B. Availability prediction

Airbnb's availability forecasts in New York City can be influenced by several factors, including B. Seasonality, Events, and Demand. In general, New York City has a high demand for short-term rentals, especially during peak seasons such as the summer and vacation periods. Therefore, it is imperative that hosts accurately forecast and adjust prices based on market demand in order to maximize rental income. There are several data-driven approaches and models that can be used to predict Airbnb availability in New York City. These models can consider a variety of variables such as past booking dates, seasonality, days of the week, weather, location, and property characteristics such as size, amenities, and ratings. Machine learning algorithms such as regression, decision trees, and neural networks can be used to build predictive models that estimate the likelihood that a property will be booked on a given date or time period. Several companies and research groups are actively working on Airbnb's availability prediction models using various data sources and methodologies. For example, Inside Airbnb provides an open-source dataset of his Airbnb listings and reviews in New York City that can be used to build predictive models. And some research groups, such as The Data Incubator, have published studies on demand and prices for his Airbnb in New York City using machine learning models.

III. Methodology and Pre-processing

A. Dataset

The dataset consists of estimated prices for Airbnb, information about the hosts, geographic availability, some important parameters to do prediction, and lastly, we can draw some inferences from this dataset. This is a regression issue, and the dataset is one CSV file with 48k rows.

B. Methodology for Prediction

First, the data set from the Kaggle Repository was chosen and cleaned. The KDD Methodology is applied. The first step is to determine whether the dataset has any missing values; if so, the missing values must be removed. Check to see whether there are any Null values in the dataset, and if there are, use isnull() to select and delete them. Both pre-processing and visualization have been completed. various insights are gained from the visualization, such as how the heap map and the folium map show the relationships between latitude and longitude and the variable, respectively, and how various plots help us understand the data sets better.

C. Data Pre-processing and Transformation

In order to limit the amount of data for price prediction, several columns in the data—such as latitude, longitude, neighboring group, and neighbourhood—have been eliminated. For predicting availability, id, name, host id, and host name columns have been eliminated to limit the amount of data for processing.

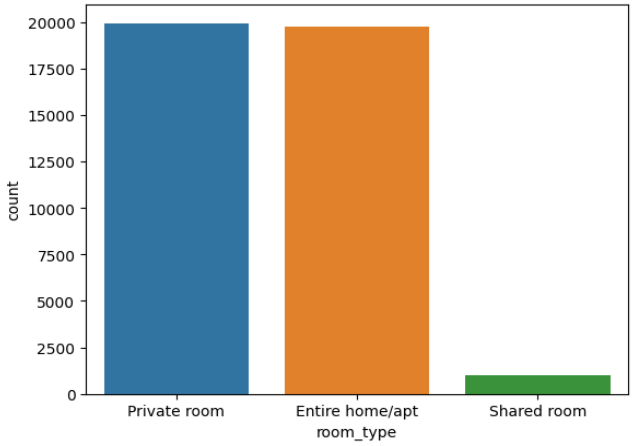

Fig. 1.

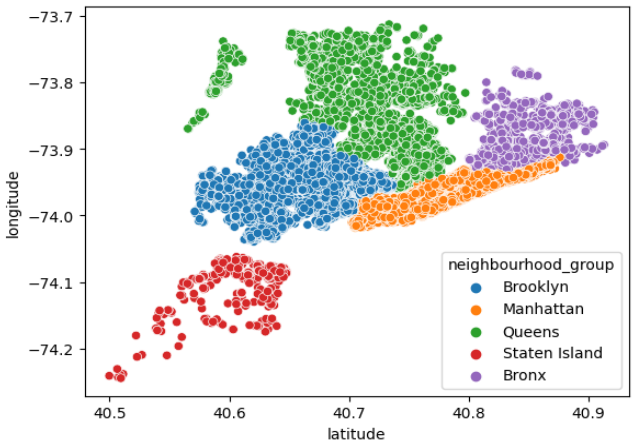

Fig. 2. A map of 5 various neighborhoods groups.

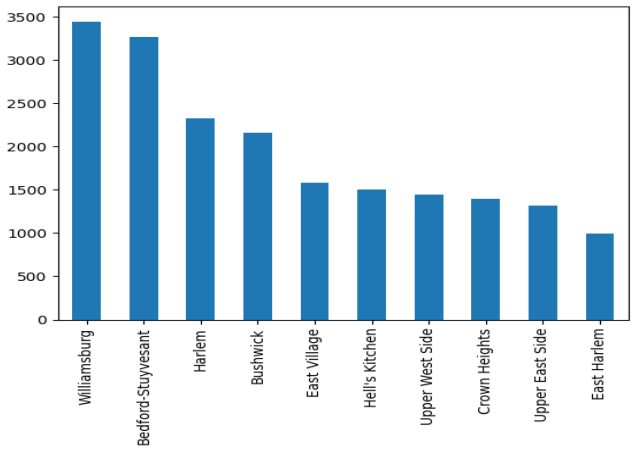

Fig. 3.

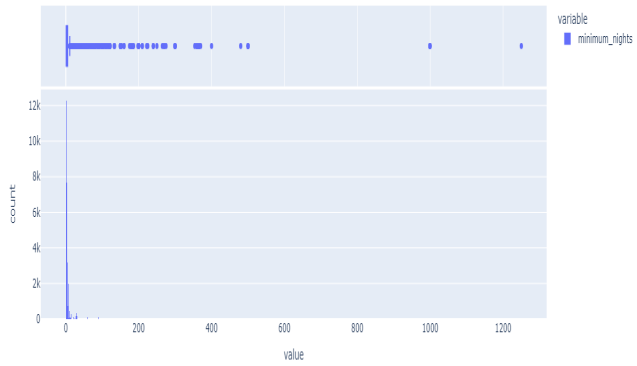

Fig. 4. histogram minimun nights

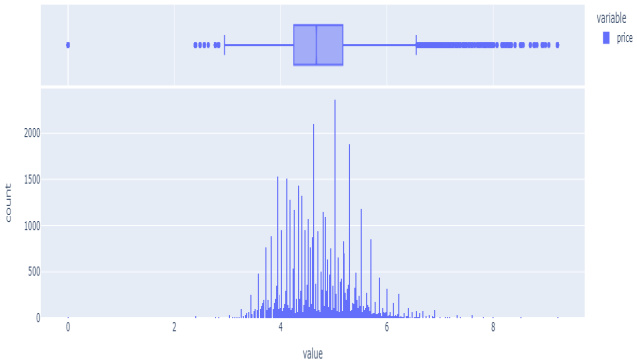

Fig. 5. histogram for price

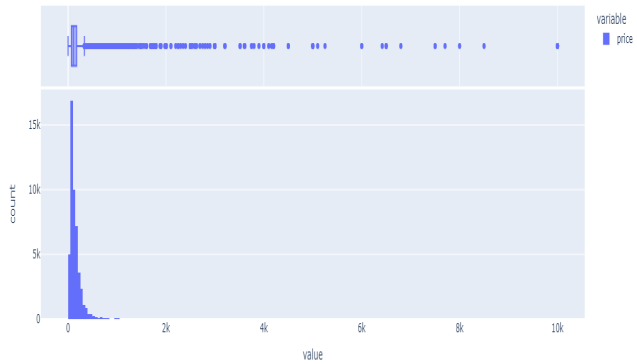

Fig. 6. histogram for price

D. Data Mining

Based on random sampling, the data set for a classification issue is split into a training (20%) and testing (80%) data set. To maintain the same ratio between the target variables, feature engineering and the removal of useless columns are done.

IV. Evaluation Model and Results

A. Price

We used machine learning techniques to forecast the price of Airbnb, Random Forest Regressor, which had the RMSE 52 and R2 score of 0.53.

```
#Fit best params to whole dataset
rdf = RandomForestRegressor(**best_params)
rdf.fit(X_train_scaled, y_train)

y_train_pred = rdf.predict(X_train_scaled)
y_test_pred = rdf.predict(X_test_scaled)

r2_train = r2_score(y_train, y_train_pred)
r2_test = r2_score(y_test, y_test_pred)

RMSE_train = np.sqrt(mean_squared_error(y_train, y_train_pred))
RMSE_test = np.sqrt(mean_squared_error(y_test, y_test_pred))
```

Fig. 7.

```
RMSE for Train:  49.23818671201351
R2 Score Train:  0.5955058044715371
RMSE for Test:  52.769933658890174
R2 Score Test:  0.5361251631503927
```

Fig. 8.

B. Availability

We used machine learning techniques to forecast the availability of Airbnb, which had the Accuracy 18 and ROC AUC 0.53.

V. Conclusion

A. Price

The pricing of the hotels in New York that are connected to Airbnb was finally predicted, and after utilizing a few machine learning models, the regression model was determined to be having an R2 score of 9.30 and an RMSE value of 0.624.

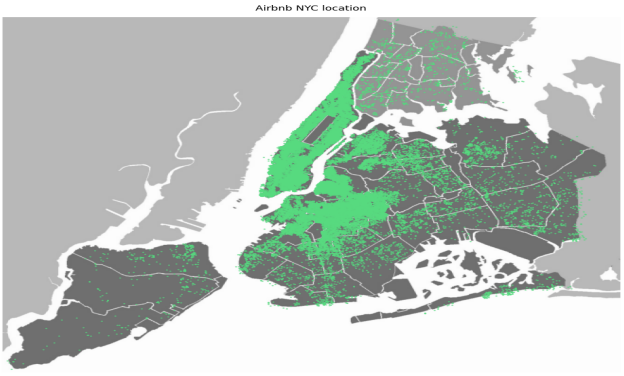

Fig. 9. By using the latitude and longitude we can generate the folium map for Availability

```
    print("Accuracy:", results[1])
    print(" ROC AUC:", results[2])

[184]  ✓ 0.0s

...  Accuracy: 0.18058490753173828
      ROC AUC: 0.5328026413917542
```

Fig. 10.

B. Availability

The Availability of the hotels in New York that are connected to Airbnb was finally predicted, and after utilizing a few machine learning models, the regression model was determined to be having an Accuracy of 0.18 and a ROC AUC of 0.53.

VI. Future Work

Processing data from every month of the year or extending the study to additional locations are two ways that this research may be improved. Since new listings are frequently added, it is possible to track each one using the time series data and determine the monthly progress and statistics. Also frequently removed from Airbnb are listings.Any trends could be discovered by looking at the months prior to each listing that vacates.It is also possible to simulate seasonality effects to see how prices change over the holiday season.

REFERENCES

[1] A. Liaw and M. Wiener, "Classification and regression by randomForest", vol. 2, 3, pp. 18-22, 2002.

[2] Li, Yang, et al. "Price Recommendation on Vacation Rental Websites." Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2017.

[3] A. Zhu, R. Li and Z. Xie, "Machine Learning Technique used to Prediction of New York Airbnb Prices", 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), 2020. Available: 10.1109/ai4i49448.2020.00007

[4] S. Yang, "Learning-based Airbnb Price Prediction Model", 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), 2021. Available: 10.1109/ecit52743.2021.00068

[5]] R. botsman and R. Rogers, "product service system", what's mine in yours .Newyork, pp. 106-108, 2010.

[6] Ang Zhu, Rong li, Zehao Xie ," Machine Learning Prediction of New York Airbnb Prices ", 2017.

[7] https://www.sciencedirect.com/science/article/abs/pii/S1447677019301950.

[8] https://www.altexsoft.com/blog/hotel-price-prediction/

[9] Short-term prediction of parking availability in an open parking lot De Gruyter, DOI:10.1515/jisys-2022-0039

[10] The availability of smoking-permitted accommodations from Airbnb in 12 Canadian cities, DOI:10.1136/tobaccocontrol-2016-053315

[11] Location of Airbnb and hotels: the spatial distribution and relationships, DOI:10.1108/TR-10-2020-0476

[12] Price indicators for Airbnb accommodations Springer November 2022Quality & Quantity DOI:10.1007/s11135-022-01576-6

[13] Machine Learning Based Quantitative Pricing for US Airbnb Renting Program September 2022 DOI:10.1007/978-981-19-5727-7_34

[14] Predicting Airbnb Rental Prices Using Multiple Feature Modalities December 2021

[15] An ensemble machine learning framework for Airbnb rental price modeling without using amenity-driven features March 2023International Journal of Contemporary Hospitality Management DOI:10.1108/IJCHM-05-2022-0562