Implementing Big Data Analytics for tourist budget planning in New York City by matching tourist preferences through predictive modeling of Airbnb rental prices

*

Sujan Neupane
Herald College Kathmandu
University of Wolverhampton
WLV ID: 2058939
Kathmandu, Nepal
neupanesujan420@gmail.com

Nischay Shakya
Herald College Kathmandu
University of Wolverhampton
WLV ID: 2059157
Kathmandu, Nepal
nischayshakya@gmail.com

***Abstract*—The popularity of Airbnb has increased significantly for tourists looking for affordable and convenient accommodations. The emergence of sharing platforms has revolutionized the hospitality industry as the living costs while traveling lowers significantly. However, the goal of tourists still remains to find the best value for the price they pay. The seemingly endless choices found on Airbnb sites can be overwhelming and a hassle for consumers. This paper proposes an approach for a predictive model to match tourist preferences with rental price ranges, allowing for more effective budget planning. New York City (NYC) Airbnb listings from 2022 have been analyzed and trained under various models to predict rental prices based on multiple features. The research aims to find the factors that influence the prices the most. Various price prediction approaches have been studied to find the best models. Different supervised machine learning models have been implemented and compared using various performance evaluation techniques. The project provides statistics-based suggestions to lower rental prices.**

***Index Terms*—Airbnb Price Prediction, Big Data Analytics, Logistic Regression, Random Forest, XGBoost, LightGBM, Classification, Price Binning and Categorization**

I. Background of the Study

Today's modern and sophisticated technology has diversified the tourist experience vastly. The consumption behavior of users has vastly changed over the past decade due to the sharing economy. The rise of the internet has allowed for better communication and for the rise of online service-providing marketplaces such as Airbnb (Shen, et al., 2020). Airbnb is an online platform that allows hosts to rent their living spaces to travelers who are looking for a place to stay. The living space could range from their homes and apartments to hotels, cottages, and penthouses. With these advanced approaches replacing the traditional ways of traveling and lodging, the options for tourists to find the perfect place to stay has increased. The search for the most affordable and budget-friendly places in cities like New York as per their requirements has been made possible by Airbnb. An attractive rental price that is reasonable for both the host and tenant is the key point of price prediction (Chin, et al., 2021). However, setting such prices for rentals and finding such rentals both are frustrating and time-consuming tasks that can often lead to disappointment.

A. Problem Statement

The tourism industry has been revolutionized by services like Airbnb by providing freedom of choice to tourists. But with vast choices comes the hassle of picking the right one over thousands of rentals. There are over 40,000 Airbnb listings in New York City alone (Inside Airbnb, 2022). The emergence of sharing platforms has allowed consumers with an appetite for sharing to get lodging facilities at much cheaper prices. The price-conscious middle class benefits hugely because of the business model. However, affordability must also come with convenience to create a win-win situation for all the parties involved (Chandra, et al., 2022). There are multiple features that could differentiate the living experience like location, cleanliness, number of bedrooms, amenities, environment, and more. Apart from physical features, factors such as trust, security, host-guest interaction, reviews, and more connected to the sentiment of customers influence the perception of sharing economy. The challenge is to find the appropriate price for the accommodation experience (Bandara, et al., 2022). It is difficult even for the hosts to set the correct prices for their rentals. Similarly, it is even more challenging for tourists to choose the listings that meet their preferences and budget.

B. Aims/Objectives

The primary aim of the project is to use big data analytics to assist tourists wanting to visit New York City with effi-

cient budget planning by predicting prices for Airbnb rentals while also matching their preferences. The objective of this study is to analyze multivariate relationships between features and price and model the data with different classification algorithms. We seek to develop a predictive model that can match tourist preferences with rental prices, allowing for more effective budget planning. This approach can help tourists choose the best rentals based on their preferences and budget, while also helping hosts set appropriate rental prices.

C. Contributions of the work connected with Methodology

The methodology followed the various approaches as researched in the related works. The data preprocessing part is implemented using research on various Python libraries and visualization techniques. The selected works in the related works section contributed to the understanding of feature selection and extraction methods that were used. Based on them, the best machine learning models were identified and used on the dataset. Due to a lack of fundamental research in previous work, as mentioned in the related works, for a classification approach for price prediction, classification models were built and explored. Different evaluation metrics were discovered and used with a supervised learning approach to evaluate the performance of the selected models. Thus, the related works contributed to the methodology section with the ideas of data processing, features and model selection, and model building and evaluation.

D. Organization of the Report

The report discusses related work on price predictive models for Airbnb rentals in NYC and many other areas. Various factors and features affecting the price of rentals are discussed in the section. Next, the methodology section describes the data preprocessing, visualization, feature engineering, and model training and prediction algorithms used in this project. The results and discussion section describes the findings and presents insights from the dataset along with a detailed evaluation of the performance of the predictive models used in accordance with the methodology. Finally, the conclusion section summarizes the contributions of this work and discusses the criteria for tourists to lower Airbnb rental prices.

II. Related Works

The tourism industry has overturned its traditional business models to new ways with the growth of sharing platforms such as Airbnb that provide peer-to-peer (P2P) accommodation. The economic, socio-cultural, and environmental ramifications in the areas of tourism and hospitality have been both positive and negative (Hall, et al., 2022). The increasing popularity of Airbnb also makes it a difficult task for the rental owners to place a reasonable price on their rental properties that is fair but also profitable. Similarly, the customers also need to know if the offered price is reasonable, in their budget, and covers all their requirements. The paper aims to forecast Airbnb prices across New York City using various machine learning approaches and predictive modeling techniques to find the dependent and independent variables along with the correlations between them, despite the insignificant amount of knowledge of the property (Brahmaih, 2020).

Airbnb has expanded from a small start-up business in the field of short-term renting to a global powerhouse. With such growth, also came challenges and concerns, mostly about the pricing of the rentals. Airbnb allows its hosts to take full responsibility for setting the expected price of their items independently. This task is reasonably difficult when considering many different factors, especially with the pricing models available in the market being either not free or inaccurate. The authors concluded that the hosts can't keep the same price for all dates to ensure appropriate prices that compete with the current market (Mahyoub, et al., 2023). The importance of pre-processing and data visualization along with exploratory data analysis (EDA) is more clearly reflected in data dealing with rental prices with all the features and factors that affect the target variable. The selection of the correct model also depends on those analyses (Dhillon, et al., 2021).

It was clear from past research that quantitative pricing was not enough to train the model properly and make the models robust enough to predict Airbnb prices. To fix the problems, careful consideration should be given to EDA to first make datasets more suitable. The use of cross-validation and random search are suggested in the paper to tune every single parameter in every single model. XGBoost, with a 0.6321 R2 score, gave the best results for Airbnb rental price prediction in Amsterdam (Liu, 2021). Similar research which predicted the prices on two New York City datasets from 2019 and 2022, also concluded XGBoost to be the best model for predicting Airbnb rental prices. They compared machine learning techniques namely Support Vector Machines, K-Nearest Neighbors, Decision Tree, Random Forests, and XGBoost. The study results suggest that machine learning models can be applied to predict Airbnb rental prices accurately (Lektorov, et al., 2023).

Proper pricing is the major factor for successful bookings. The paper focuses on one of the main markets for Airbnb, Beijing, and uses XGBoost and a three layers Neural Network model for price-prediction modeling. The results showed the former performing better than the latter. The author also briefly mentioned the impact that the Beijing Airbnb market faced due to the COVID-19 pandemic. The accuracy of the model is proved to be significantly influenced by outliers as well. Moreover, hosts are provided with recommendations about how to increase prices by exploring the influence on price by adding important amenities like Internet, TVs, Elevators, and services like cancellation policy and instant booking (Yang, 2021).The deficiencies that price predicting models face can mainly be summed up to the feature attributes being poor, the lack of fundamental research on the sector of correlation between the rental price and the point of interest

(POI) around the rentals, and the lack of depth on rental text information research. A multi-source information embedding (MSIE) model has been proposed that combines statistical, text, and spatial features to predict the rental prices of Beijing and Shanghai by forming a multi-source rental representation (Jiang, et al., 2022).

The dynamic nature of the tourism industry demands constant upgrades on even the most powerful algorithms used for rental price prediction such as Regression Tree, Support Vector Regression, and XGBoost. A three-layered deep neural network (DNN) is proposed that results in two outputs i.e., the minimum and maximum prices host can set based on the rental information. The proposed framework was built using the Rio de Janeiro Airbnb dataset and had an accuracy of 74.43% and outperformed the existing models with an R2 value of 0.8104. The paper also identified the top factors that should be included in rentals to optimize profits (Thakur, et al., 2022). It has been shown that multi-modality data performs better in price forecasting in comparison to single-type data. In the research, numeric, text, and map data have been used for price prediction by combining customer reviews, house features, and geographical data as input by incorporating the into natural language processing and machine learning algorithm frameworks. The paper has also concluded an effective approach for Exploratory Data Analysis (EDA) to be Principal Component Analysis (PCA) for the New York City dataset. The paper is in favor of exotic algorithms such as DNN and XGBoost to perform better than models like Support Vector Regression and Linear Regression which are linear in nature (Peng, et al., 2020). Alternatively, machine learning algorithms like KNN and Decision tree have been analyzed to be the best ones to use with the data. The accuracies of those algorithms are 0.91 and 0.94 respectively which are significantly higher than models such as Logistic regression and Naïve Bayes. On the other hand, the best model for predicting price listing is ARIMA model. Data wrangling and hypothesis testing were also used and the model successfully predicted the price of Airbnb rental listings in NYC for the forthcoming years (Garlapati, et al., 2021).

As seen from the majority of the papers, XGBoost is the way to go for the most accurate results in approaches using regression. Our project, however, will use classification models to categorize the prices of Airbnb rentals in NYC into 5 different price ranges from budget to ultra-luxury. The relationship between the features and price will be analyzed with and without the outliers. Different visualization techniques are to be utilized to do so. Furthermore, feature selection techniques will be applied to separate the most important features that manipulate the price. Finally, different machine learning techniques are to be used and different evaluation metrics to analyze the best one for the dataset used.

III. Methodology

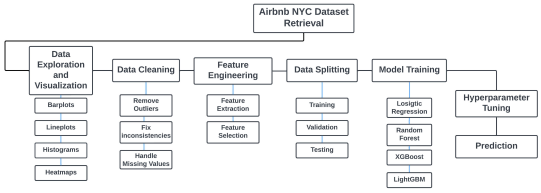

Fig. 1. Project methodology

A. NYC Airbnb Dataset Retrieval

The 2022 Airbnb dataset for New York City was retrieved from the Inside Airbnb website. The dataset consists of multiple files, among which the file listings.csv has been utilized in this study. The dataset contains rental information for five boroughs in NYC where Airbnb rentals are operating actively. The selected file has 75 columns and 41533 rows. The listings.csv file was uploaded to hadoop distributed file systems (HDFS) and retrieved as a PySpark data frame in Jupyter Notebook for further processing.

B. Data Exploration and Visualization

Not all of the 75 columns in the dataset were useful. Each column was explored one by one, and columns like id, listings_url, scrape_id, etc. were dropped as they don't contain any relevant information. Similarly, the columns with a high percentage of null values were also dropped. The dataset had a variety of data types like an object, float, integer, etc. The selected columns underwent further data visualization, exploration, and cleaning. The following bar plot represents the percentage of null values in columns.

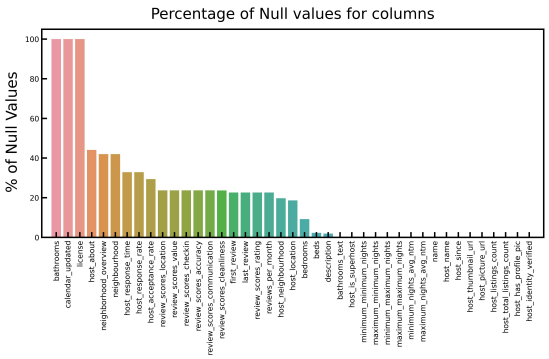

Fig. 2. Percentage of null values

C. Data Cleaning

The target variable, price, was heavily skewed to the right. It made it difficult to directly analyze the relationship between the mean price and other variables. As a result, the median of the original price column and a new price column with outliers were eliminated using the interquartile range were used for analytical purposes.

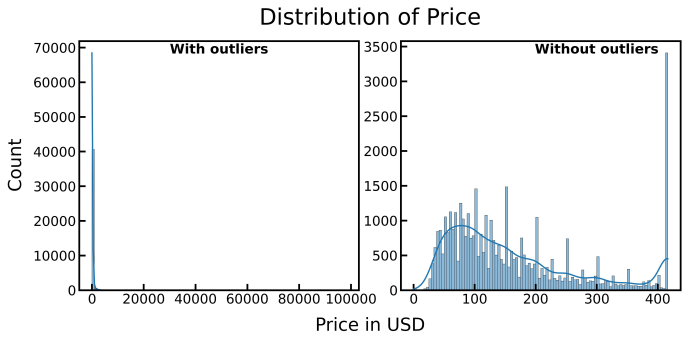

Fig. 3. Skweness in the price

In several columns, some of the rows contained N/A, nan, and other special characters such as $ and %. To ensure uniformity, such special characters were removed, and N/A values were replaced with nan. The null values were then substituted by the mode for categorical columns if the percentage of null values was relatively low. Some columns had f and t values that were replaced with True and False respectively. The relationship between the columns with null values and the price column was investigated first. Then, after filling in the null values, the updated column's relationship with the price column was investigated to ensure that the original relationship was still maintained after imputing the null values. Some of the columns also had too much noise. Such columns were also dropped.

D. Feature Engineering

i. Feature Extraction

Some of the columns had to undergo transformations to extract important relevant information. The property_type column had values like Private room in a condo, Entire guest suite, Private room in a religious building, etc. Such values were transformed into specific categories like Apartment, House, Hotel, Hostel, Bed and Breakfast, and Other. Similarly, the bathroom column had many unique values like 0 baths, 1 shared bath, 4.5 shared baths, etc. Some rentals also offered multiple private and shared bathrooms. Few also offered half-bathrooms.

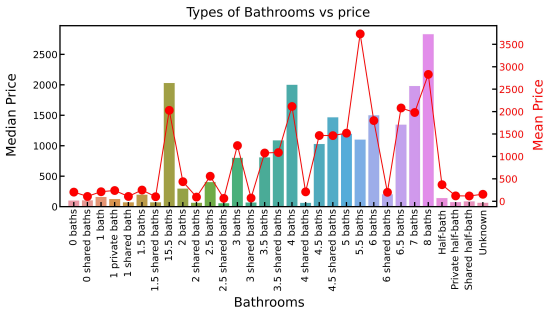

Fig. 4. Variations in bathroom types and median price

Two new columns were extracted using the bathroom column. A new column for the number of available shared baths and another column for the number of private baths was created. Then, the original bathroom column was dropped.

The amenities column had an excessive number of unique values. Some rentals had too many amenities as compared to other rentals. It made modeling the data difficult as each amenity in this column needed to be dummy encoded which would have resulted in too many new columns. Instead, each amenity was classified into one of the following seven categories: safety, kitchen, bathroom, bedroom, entertainment, climate control, and wifi. Finally, these categories were one-hot encoded to create new columns that contained the number of corresponding amenity categories present in the rentals.

Several types of reviews were present in the dataset. Some reviews were for the cleanliness of the rental, communication with the host, location, etc. Each of these review columns had values ranging from 0 to 5. Initially, such continuous values had a random association with the price column. As a result, values in each of the review columns were transformed into separate categories by binning all values from 0 to 5 at intervals of 0.5, yielding a total of 10 categories. The newly engineered features had a strong linear relationship with the price column.

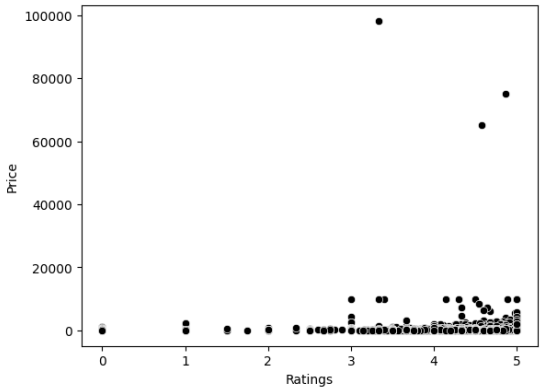

Fig. 5. Scatterplot between ratings and price

The ordinal categorical columns were label encoded. Columns such as review_scores_cleanliness, review_scores_communication, review_scores_checkin, review_scores_location, review_scores_value, Ratings, and instant_bookable were label encoded, whereas neighbourhood_group_cleansed had no order and thus underwent one-hot/dummy encoding.

Finally, the price column was engineered such that the continuous values in the price column were converted to ordered categories for performing classification task. Prices for Airbnb rentals ranging from $0 to $50 were assigned the Budget category. The rentals with price between $51 to $200 and $201 to $350 were categorized as Standard and premium

respectively. Likewise, the rentals with price between $301 to $500 were classified as Luxury. Finally, the rentals with a price of more than $501 were assigned to the Ultra-Luxury category. As a result, the dataset ended up with five classes.

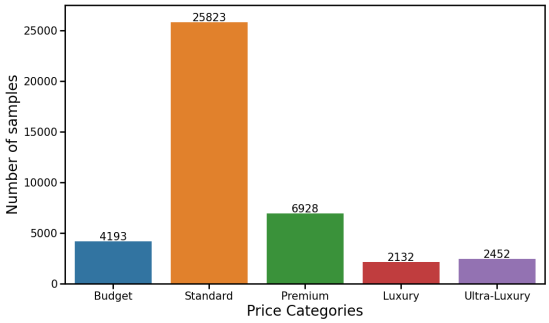

Fig. 6. Number of samples in each class

ii. Feature Selection

75 columns were effectively reduced to 35 columns after the preprocessing and data cleaning. The 35 columns were put through feature selection to pick the most significant columns for modeling purposes and eliminate those that were ineffective. Three feature selection techniques were implemented: Univariate Feature Selection, Recursive Feature Elimination, and Tree-Based Feature Selection. Each technique assigned an importance value to the features. To evaluate the overall importance of each feature, the average importance of all three feature selection techniques was calculated and the top 17 features were selected.

E. Data Splitting

Using sklearn's train_test_split function, the dataset was divided into two parts: the training set and the testing set in an 80:20 ratio. There were 33222 rows and 17 columns in the training set. Similarly, there were 8306 rows in the test set. Each model was trained using cross-validation, which automatically divided the original training set into training and validation sets in an 80:20 ratio. As a result, the original data was divided into three separate parts with a 60:20:20 ratio for training, validation, and testing.

F. Model Training and Hyperparameter Tuning

The following classifiers have been implemented.

- Logistic Regression
- Random Forest Classifier
- LightGBM Classifier
- XGBoost Classifier

Additionally, each classifier was tuned across a wide variety of hyperparameters using RandomizedSearchCV to further improve the model. The following table contains the performance of each model on test dataset measured across multiple metrics.

Model	Precision	Recall	F1 Score	AUC Score	Accuracy
Logistic Regression	0.64	0.67	0.61	0.839	0.67
XGBoost	0.67	0.7	0.66	0.86	0.69
LightGBM	0.68	0.7	0.66	0.859	0.69
Random Forest	**0.69**	**0.71**	**0.67**	**0.868**	**0.71**

TABLE I
MODEL PERFORMANCE METRICS

The Logistic Regression model performed the worst as it is not able to capture nonlinear or monotonic relationships in the data (Hong, et al., 2020). Similarly, two gradient boosting models, XGBoost and LightGBM, performed better than Logistic regression by achieving approximately 0.69 each in test accuracy. The Random Forest model performed best with an approximate 0.71 test accuracy. Each model except Logistic Regression was overfitting with the training data, which is why a wide variety of hyperparameters and regularization techniques were implemented to reduce overfitting and improve generalization on the validation dataset.

G. Prediction

After training and optimizing each model, predictions were made on the test dataset. The confusion Matrix and ROC Curve for each model is discussed in the results section.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The dataset was obtained from Inside Airbnb New York City Listing. The dataset was imported into HDFS and then converted to PySpark dataframe. The dataset was observed and as with any project involving data analytics, we started off by performing exploratory data analysis (EDA). Python libraries such as PySpark was used for data handling and manipulation and matplotlib and seaborn were used for visualization.

- Python version: 3.10.8
- PySpark version: 3.4.0
- Matplotlib version: 3.6.2
- Seaborn version: 0.12.1

The project was implemented on Ubuntu operating system using Jupyter Notebook on a Anaconda virtual environment.

B. Analysis and Discussion of the Findings

After data preprocessing and cleaning, the 35 reduced columns underwent further multivariate analysis and feature selection to filter out unnecessary columns.

First, the median prices of Airbnb rentals were compared amongst the boroughs of New York City. The heatmap below generated using latitude and longitude columns shows that the median price in Manhattan is the highest. Apart from the 5 boroughs, there is also a column dealing with the price in 223 neighborhoods in NYC. The importance of the columns is to be analyzed.

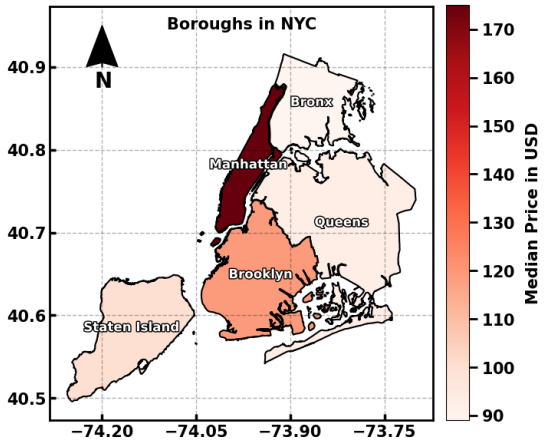

Fig. 7. Heatmap showing median price of airbnb rentals in five New York City boroughs

The observation of the relationship between Response Time for Hosts with Median Price clearly showed that hosts that respond within an hour for rental inquiries have the highest median price. Additionally, the rentals with hosts' response time being unknown had the lowest median price. It indicates that the potential tourists are not likely to choose rentals whose hosts response time is unknown, which ultimately drives the demand and price for such rentals down.

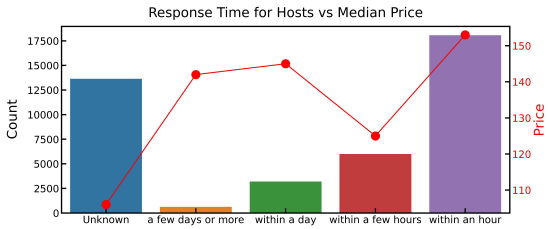

Fig. 8. Host response time vs median price of five boroughs in NYC

It can be seen that there is the greatest number of Airbnb rentals in Manhattan. Since this borough is the major tourist attraction region in New York, it could be why this borough has the highest median price as compared to other boroughs. If potential traveler wants to save money on rentals, they should avoid Manhattan and Brooklyn because their costs are generally higher.

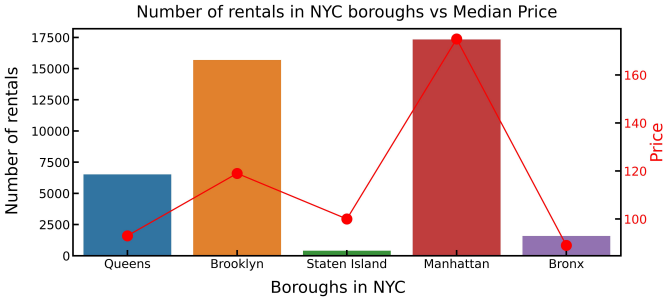

Fig. 9. Number of rentals in NYC boroughs vs median price

The median rental price hosted by super hosts was slightly higher than that hosted by non-super hosts.

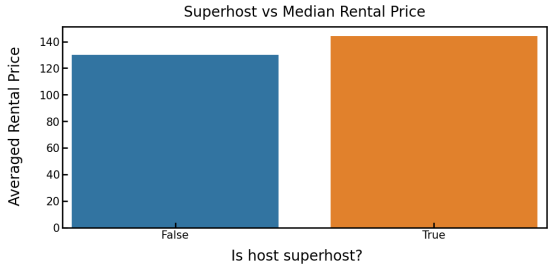

Fig. 10. Median price for superhosts

Similar results are seen for hosts whose identity has been verified. If a host's identity is verified, or if a host is an experienced, it will increase confidence for potential tourists to select such rentals, ultimately driving up the price and demand of such rentals. Therefore, such findings suggest keeping the columns for modeling purposes.

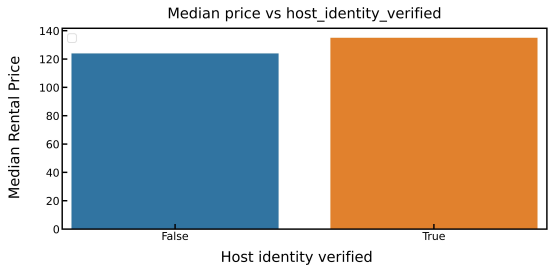

Fig. 11. Host identity verified vs median price

The barplot below shows the median prices of different property types across the 5 boroughs in NYC. The hotel price had the highest median price in Manhattan. The barplot indicates that, on average, apartments and hotels tend to be more expensive as compared to other room types. Furthermore, each rental tends to be cheaper in Bronx and Staten Island whereas the rentals in Brooklyn and Manhattan tend to be much more expensive. It could be due to the fact that these broughs tend to attract more tourists that drives up the demand.

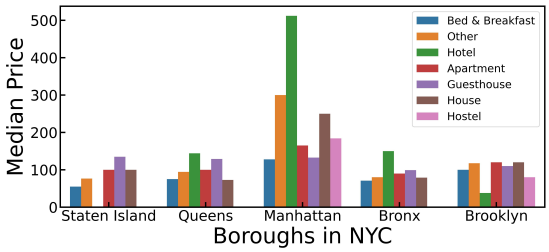

Fig. 12. Property type vs median price

The boxplot below shows the price distribution of different room types across the 5 boroughs of NYC. It can be seen that the price of entire homes and apartments are more expensive in all the boroughs. The boxplot also indicates that the price distribution is very skewed to the right across all room types in all broughs in New York City. Each room type is the most expensive in Manhattan as compared to other boroughs. Bronx and Staten Island have the least expensive room types, which could be due to the fact that these broughs attract the least number of tourists in general.

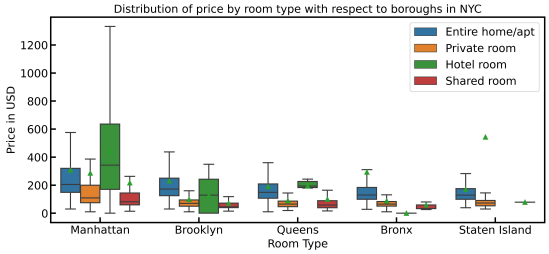

Fig. 13. Room type vs median price

It is seen that the price increases with the number of accommodates allowed in a rental. The price without the outliers is taken as some prices are highly influenced by expensive properties in expensive neighborhoods. The relationship between the number of accommodates and the rental price is more clearly visible with the outliers removed.

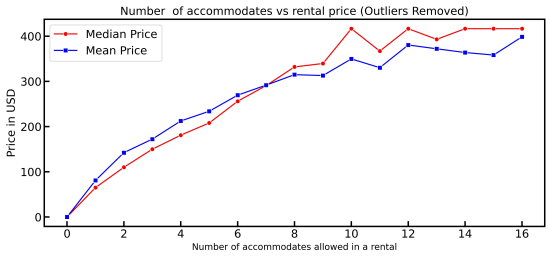

Fig. 14. Number of accomodates vs rental price

It can be clearly seen that the median price of an Airbnb rental increases linearly with the increase in the number of bathrooms. Shared bathrooms tend to be cheaper as compared to private bathrooms. It indicates that tourists are more likely to pay more for privacy and pay much less should they have to share their bathroom with others.

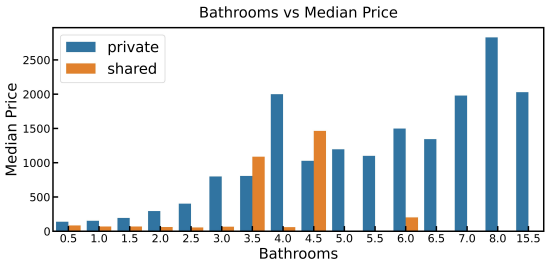

Fig. 15. Number of bathrooms vs rental price

Rentals with more beds tend to cost more than rentals with fewer beds. Potential tourists may be satisfied with rentals that provide more features, such as a larger number of beds, making them more willing to spend more money given their satisfaction. Some rentals had 17, 19, 21, and 42 beds, but there was only one of them, so their median and mean prices were the same. These are the outliers. They were not removed for analytical purposes to better understand the distribution of data and its relationship to price.

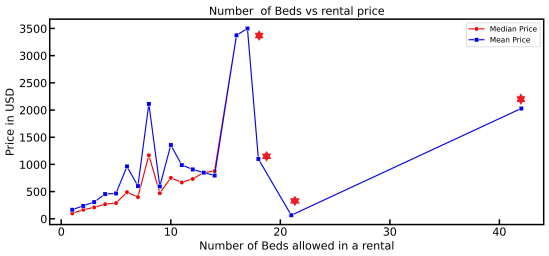

Fig. 16. Number of beds vs rental price

The amenities are first grouped into various categories. The graph above shows that the rental prices increase with respect to the increase in the number of amenities. Additionally, the amenities grouped as such are created into new columns as they will be useful when modeling. It can be clearly seen that, in general, the rise in the number of amenities tends to increase the price of rentals. Potential tourists might be willing to pay more for additional amenities and vice versa.

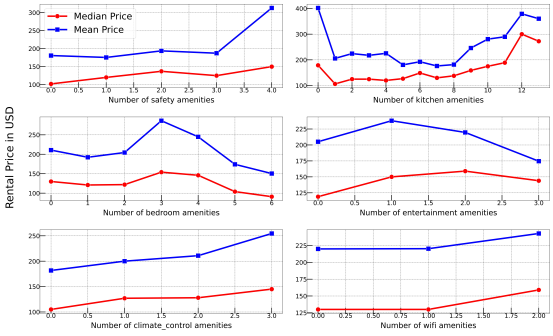

Fig. 17. Number of various ameneties vs rental price

It can be clearly seen that the median price of a rental tends to have a strong positive linear relationship with the number

of bedrooms offered in the rental. Rentals with more than 4 bedrooms tend to cost more than $500 whereas the median price of rentals with 14 bedrooms offered has almost $2000 price. It shows that tourists are willing to pay more if a rental offers multiple bedrooms.

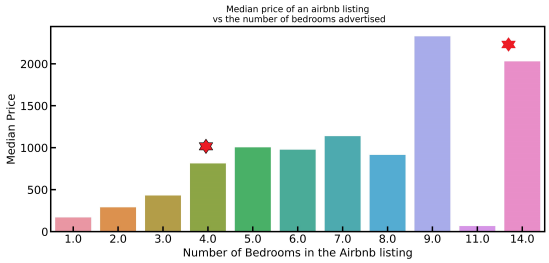

Fig. 18. Number of bedrooms vs rental price

The ratings given by customers were categorized with a difference of 0.5. The lowest rating a rental could receive is 0 and the highest is 5. It is clear that as the ratings of rentals increase, the median price increases as well. Similarly, as all the expensive places might not get a good rating, the graph justifiably has some outliers. Similarly, other review columns also indicated a strong linear relationship with the median rental price.

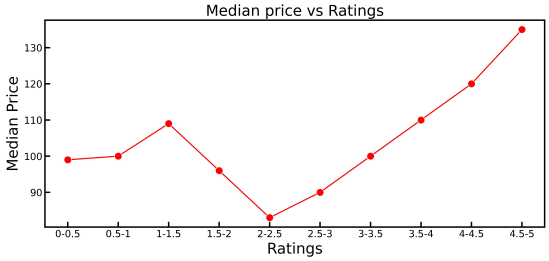

Fig. 19. Ratings vs rental price

If a rental can be instantly booked, it drives up the demand and price. Such property can be analyzed from the barplot above. The rentals that were instantly bookable had a higher median price as compared to the rentals that were not instantly bookable.

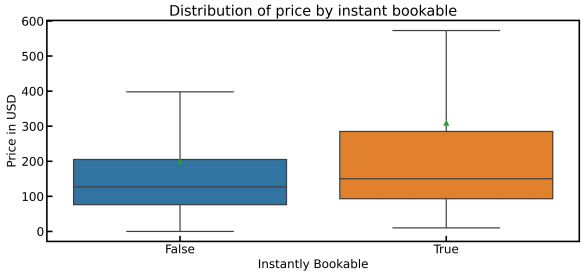

Fig. 20. Instantly bookable vs rental price with respect to room type

The boxplot below explains how the rentals that have an instant booking feature tend to have higher prices than rentals that do not have the feature. Furthermore, Manhattan and Brooklyn had the most instantly bookable rentals as compared to other boroughs. This could have also resulted in Manhattan and Brooklyn having higher rental prices than other broughs.

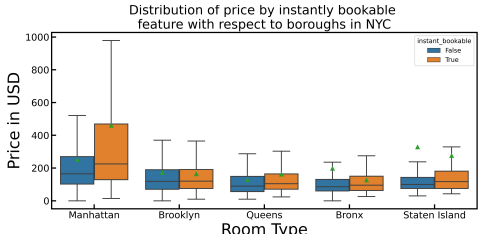

Fig. 21. Instantly bookable vs rental price with respect to boroughs in NYC

C. Analysis of the Findings

After the initial overview of the relationships between various features and rental prices, many columns that proved irrelevant were dropped, and a few additional columns that seemed necessary were added. There were 35 columns left, which underwent three feature selection techniques: Tree-Based Feature Selection, Recursive Feature Selection, and Univariate Feature Selection. Each method assigned an importance value to each feature that was converted to the range of 0 to 1 for averaging across all three methods. Finally, the average importance for each feature was calculated that summed to 1. The barplot below represents the average importance of each feature.

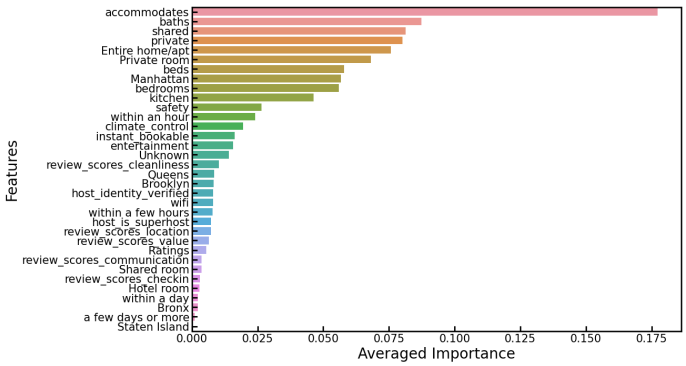

Fig. 22. Averaged importance for features calcualted using three feature selection technques

The features with average importance greater than 0.01 were selected, which resulted in a total of 17 features, excluding price, being selected with overall retained importance equaling 91.33%. The following heatmap showcases the Pearson correlation coefficient between selected features and the price column.

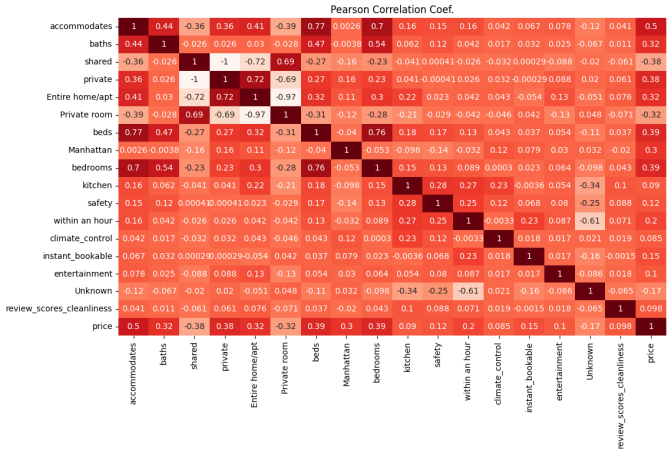

Fig. 23. Pearson correlation coefficient between variables

The Pearsons correlation coefficient measures the strength of linear relationships between variables. The heatmap showcases that some variables are linearly related but most aren't. This further solidifies the methodology implemented in this project to make use of non-linear models like XGBoost and Random Forest, as they are more able to capture non-linear relationships in data than a model like Logistic Regression.

After the target column was categorized using 5 price ranges, the values were label encoded. Four classification models, Random Forest, Logistic Regression, XGBoost, and LightGBM, were trained with the training along with RandomizedSearchCV for optimizing hyperparameters. Due to Logistic Regression's inability to capture non-linear relationships in the data, it performed poorly as compared to the other models. Each model's performance was evaluated on the test dataset across a wide range of classification metrics like accuracy, precision, recall, f1-score, and auc score. Furthermore, a confusion matrix and roc curve were also developed for each model to understand the model's ability to distinguish between different classes.

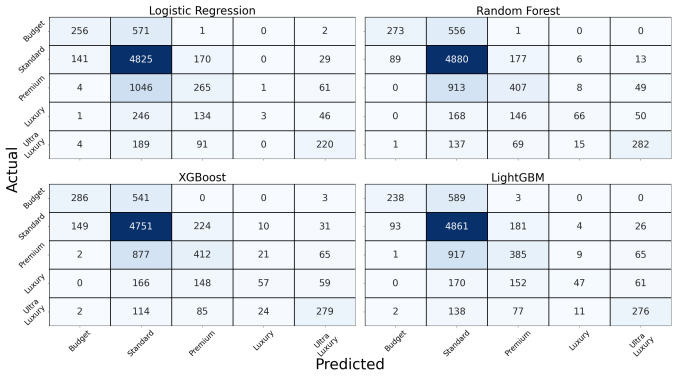

Fig. 24. Confusion matrix

It can be seen that the majority of samples that were correctly classified were from the Standard category. It could have been because the majority of the data belonged to this category. Additionally, across each class, the false positives and negatives are highest for Logistic Regression Model. It could be due to the inability of this model to capture non-linear relationships in the data. Additionally, Random Forest has the highest true positives for all classes among all models. Similarly, it also has very low false positives and negatives as compared to other models. The majority of budget category samples have been misclassified by all models. It could be due to this category having fewer data as compared to other classes. Overall, the Random Forest model has the most true predictions across all classes as compared to other models. Furthermore, the ROC curve for each model has also been developed. The true positive rate is plotted on the y-axis and the false positive rate is plotted on the x-axis for various classification thresholds. The following plot showcases the ROC curve for each model.

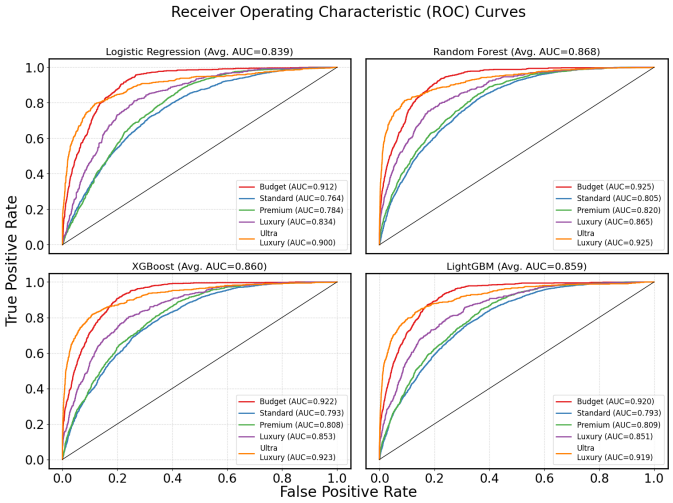

Fig. 25. ROC curve

V. Conclusion

The results showed that a tree-based model, such as Random Forest, can capture non-linear correlations in Airbnb data with the highest degree of accuracy. The findings revealed that the cost of an Airbnb rental in New York City is influenced by a wide range of factors. From how the host behaves with the customers to the reviews and ratings that the rentals get, price is heavily influenced. As one would expect, the study concluded that if the host is more communicative and experienced, the rental is clean, and it has positive reviews and higher ratings, the rentals tend to be more expensive. Similarly, the location of the rentals plays an important role in determining the price, with Manhattan rentals commanding much higher rent prices than those of other boroughs in NYC. Rentals that allow more number of accommodates tend to be much more expensive than the rentals that allow a lower number of accommodates. On the other hand, the amenities category columns did not show a clear linear association with the price, which was surprising because guests are anticipated to pay more for

rentals with more amenities. Even though multiple columns were expected to have a strong linear relationship with the price, it was not the case as shown by the Pearson correlation coefficient. A suggestion for tourists wanting to book cheaper Airbnb rentals in NYC would be to look for rentals that allow fewer accommodates, avoid expensive boroughs like Manhattan, and choose rentals with fewer bathrooms, beds, and average reviews.

ACKNOWLEDGMENT

We would like to thank Mr. Basudeo Shrestha for guiding us while completing this report. We would also like to thank Herald College Kathmandu and its staff for being very supportive towards students. This work coudn't have been completed without the support of our beloved tutor.

REFERENCES

[1] Bandara, H., Charles, J. & Lekamge, L. S., 2022. Using Sentiment Analysis to Explore the Accommodation Experience in the Sharing Economy through Topic Modeling. Colombo, IEEE.
[2] Brahmaih, K., 2020. Predicting Airbnb Listing Price Across NewYork, s.l.: Letterkenny Institute of Technology.
[3] Chandra, Y. U., Gunawan, C., Tandrian, F. B. & Seanbert, S., 2022. Performance Factors Analysis of the Digital Sharing Economy. Semarang, IEEE.
[4] Chin, T. A., Kiat, T. M., Pheng, H. S.& Lai, L. Y., 2021. Enhancing Consumer Repurchase Intention towards Airbnb. Dubrovnik, IEEE.
[5] Dhillon, J. et al., 2021. Analysis of Airbnb Prices using Machine Learning Techniques. NV, IEEE.
[6] Garlapati, A. et al., 2021. Price Listing Predictions and Forthcoming Analysis of Airbnb. Kharagpur, IEEE.
[7] Hall, C. M. et al., 2022. Airbnb and the sharing economy. Current Issues In Tourism, pp. 1-11.
[8] Hong, X., Wei, H. & Gao, J., 2020. Nonlinear Logistic Regression Model Based On Simplex Basis Function. Glasgow, IEEE.
[9] Inside Airbnb, 2022. New York City. [Online] Available at: http://insideairbnb.com/new-york-city/
[10] Jiang, L. et al., 2022. A Multi-Source Information Learning Framework for Airbnb Price Prediction. Orlando, IEEE.
[11] Lektorov, A., Abdelfattah, E. & Joshi, S., 2023. Airbnb Rental Price Prediction Using Machine Learning Models. Las Vegas, IEEE.
[12] Liu, Y., 2021. Airbnb Pricing Based on Statistical Machine Learning Models. Stanford, IEEE.
[13] Mahyoub, M., Ataby, A. A., Upadhyay, Y. & Mustafina, J., 2023. AIRBNB Price Prediction Using Machine Learning. Baghdad & Anbar, IEEE.
[14] Peng, N., Li, K. & Qin, Y., 2020. Leveraging Multi-Modality Data to Airbnb Price Prediction. Chongqing, IEEE.
[15] Shen, L., Liu, Q., Chen, G. & Ji, S., 2020. Text-based price recommendation system for online rental houses. Big Data Mining and Analytics, 3(2), pp. 143-152.
[16] Thakur, N., Jain, R., Mahajan, A. & Islam, S. M. N., 2022. Deep Neural Network based Data Analysis and Price Prediction framework for Rio de Janeiro Airbnb. Mumbai, IEEE.
[17] Yang, S., 2021. Learning-based Airbnb Price Prediction Model. Hangzhou, IEEE.

VI. APPENDIX

The GitHub repository hosting the Jupyter notebook containing the HDFS data retrieval and PySpark code can be found at the following **link**.