Northeastern University

EAI 6000 - FINAL PROJECT

Group 5

Prof. Sergey Shevchenko

May 16th, 2022

Submitted By :-

Akash Raj Member

Abhigna Ramamurthy

Durga Bhanu Nayak

Harshit Gaur

Jeseeka Shah

Introduction

The Project covers analysis of the vital data set "**Airbnb Business Case**". Visitors and hosts have been using Airbnb since 2008 to expand travel possibilities and give a more unique, personalised way of experiencing the world. Use **LIME** and **SHAP** for model interpretability and predictive modelling to develop market-specific forecasts with various variables. The data used in this research offers information on listing activity and indicators in the United States in 2019. It offers over 800,000 listings and 27 key characteristics, enough to learn more about hosts, geographical availability, and the analytics needed to make projections and draw conclusions.

Airbnb has established their market in accommodation business successfully since 2008 and has provided hosts and visitors with attractive options to visit places across the globe. Using the data collected by Airbnb for their listings, we have employed ML models that predict listing price based on user inputs over the UI. In this process we have used regression models like Linear Regression, Decision Tree, Random Forest, and XGBoost algorithms to calculate the price from the selective input variables and have used LIME and SHAP for model interpretability. For this project, we have limited the scope of the training data for United States listings only.

Problem Statement

Airbnb has stood the test of time and has proved to be a go to site for travel planning. With their newly introduced features of the Airbnb App, we can expect a lot of growth from the company. The dataset is taken from their official website *insideairbnb.com.* This provides a great platform to analyze the factors that contribute to the listings' prices. The project objective is to help customers and hosts get a sense of Airbnb listing prices based on the important features that define the listing property. This is aimed at Airbnb too as they can optimize their search models and service better to their customers.

Exploratory Data Analysis

Before advancing to the EDA, we have considered to clean and pre-process the data. The data was taken from the official website *insideairbnb.com*. After careful observation and discussion, we have decided to only take 23 features and 210,617 listing into use. This consist of the data that belongs to United States of America only.

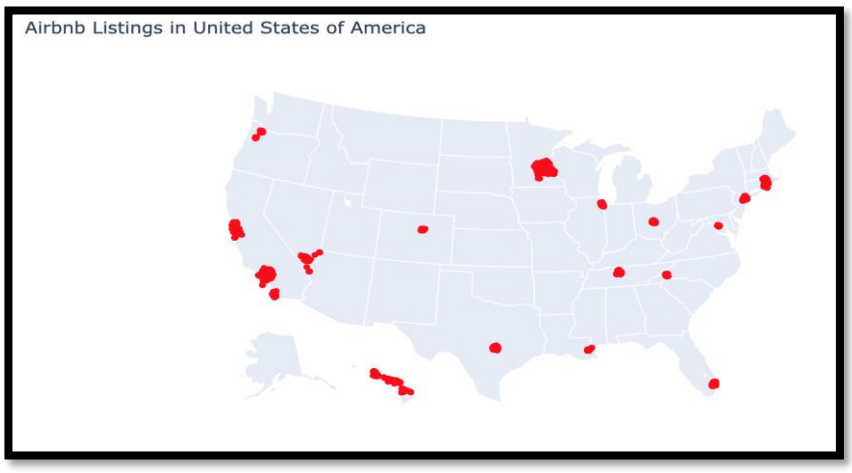

Figure1: shows the major cities that has the highest Airbnb in USA.

The objective here is to predict the listing of prices that are based on the input given by the user on the dashboard that is created and checking the prediction of the values Streamlit UI. To advance toward the main part basic EDA was done as follows.

1. Data Pre-processing: -

Out of the vast 74 attributes list from the extracted dataset, only 23 features are selected to avoid redundancy. Removed the records with missing values as we had a large dataset to work with already. Attributes like amenities, host_response_time categorical variables are converted to integer. True/False categorical variables like host_has_profile_pic, host_identity_verified, instant_bookable are converted to Boolean values. Percentage of host_response_rate converted to floating value and dummy variables for room_type is created for easy analysis. Outliers were removed as they proved to be influential and could increase risk of wrong analysis outcomes.

2. Exploratory data analysis through data visualization: -

As shown the bar graph clearly shows the preferences of the customer at Airbnb. The customers majorly preferred taking single bedrooms with single bed. Indicating that these are people either on business trip or cost-efficient people.

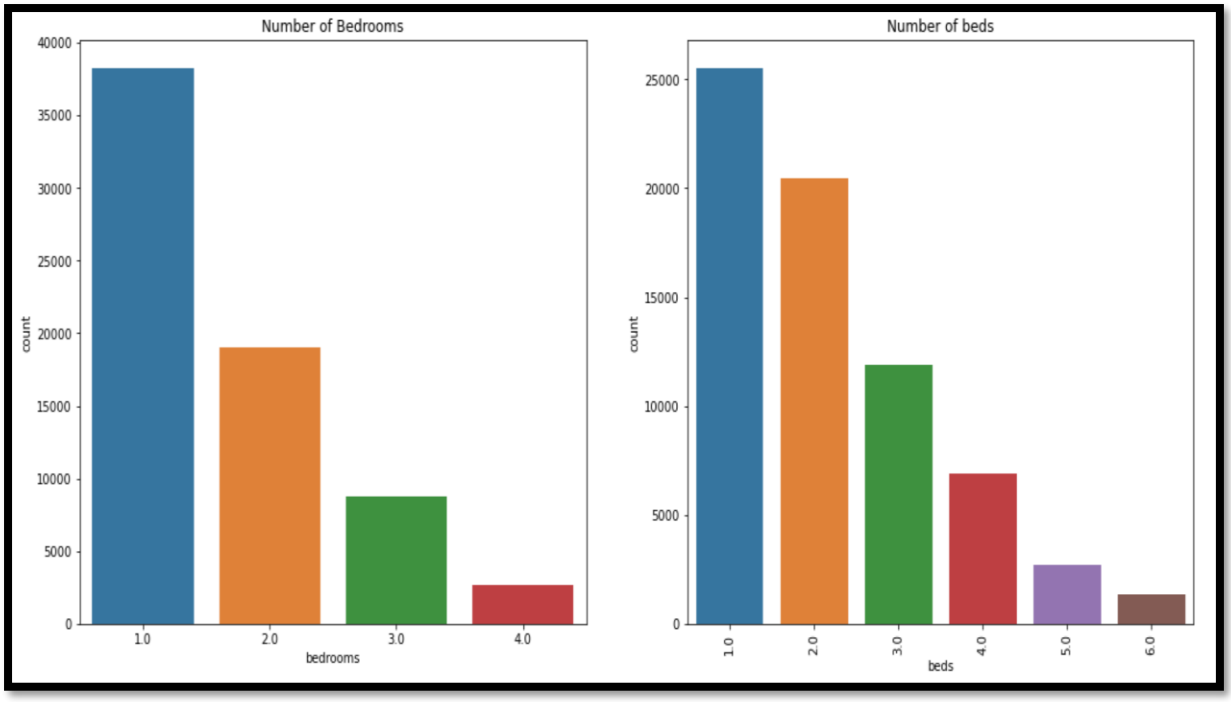

Figure 2: showing most of the customers preference on booking

Secondly, after selection of the layout of room and number of beds. The focus is shifted to price that are related to these preferences. The price vs. density histogram below gives us a range of prices that most customers has chosen. The highest being between the range of $50 to $200 per night. And the on average maximum people considered taking up to $100.

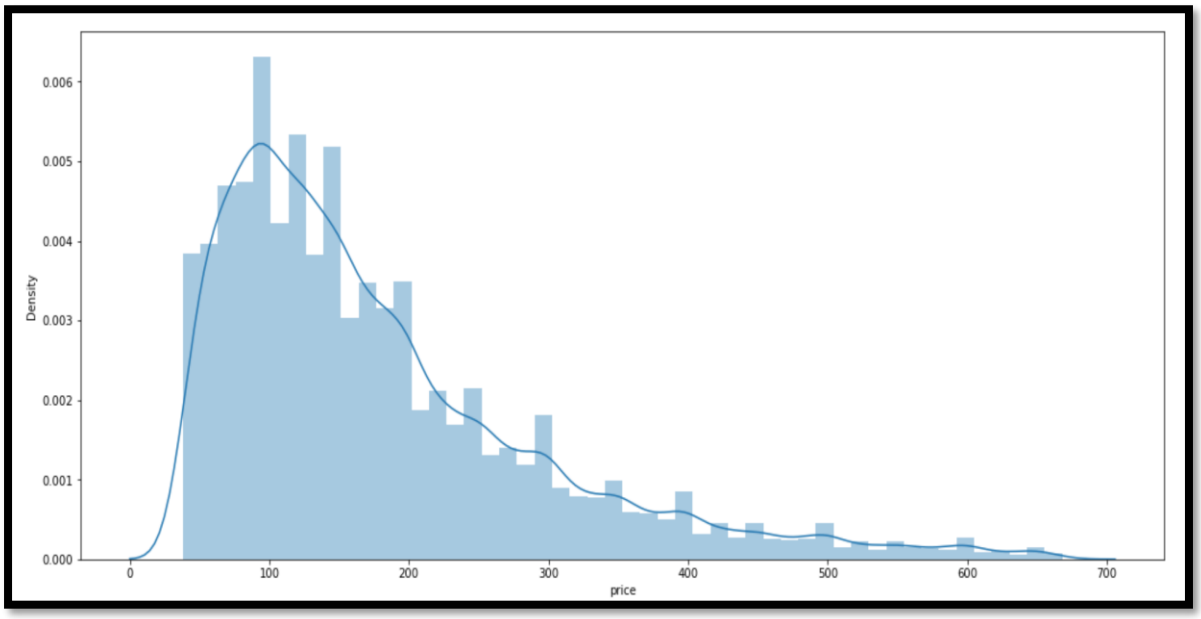

Figure 3: showing the prices that maximum people paid for single night at Airbnb

Moving forward, we wanted to know regarding the average price did they pay per property. The highest in the listing was the private room in the boat listing to $1400. And lowest being the shared room in a dom.

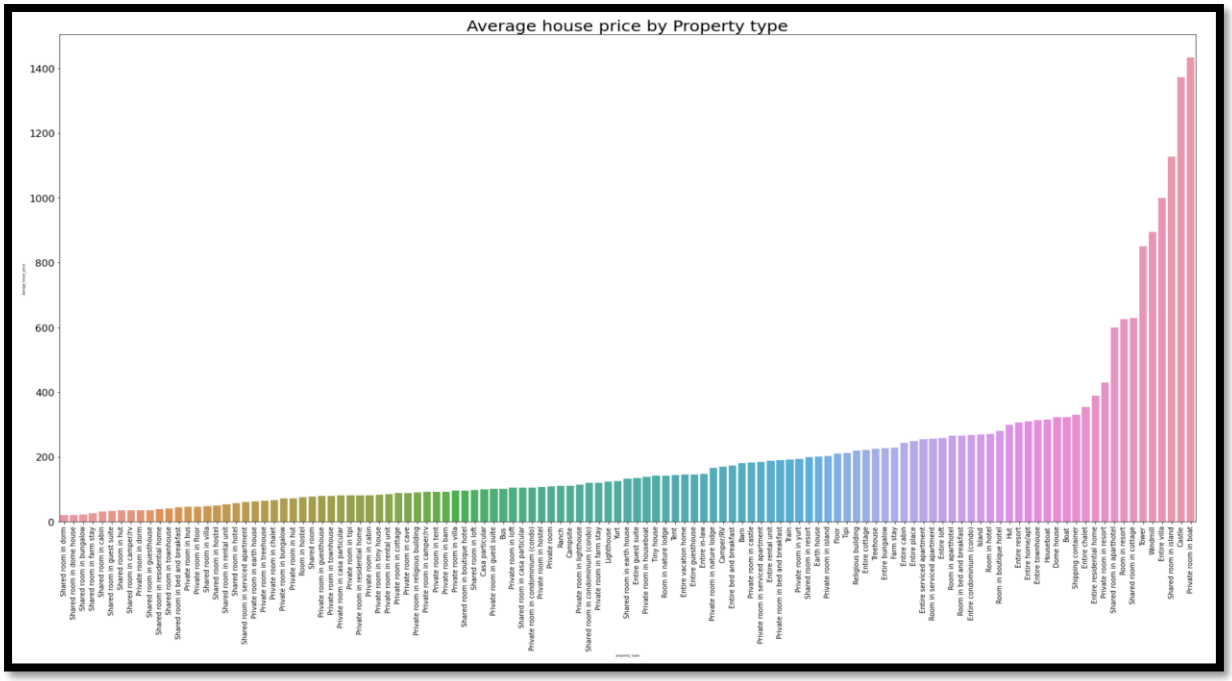

Figure 4: showing the price listing per property

After understanding price listing, we wanted to narrow down our search to analyse the demand in the market. For this we have shown below, boxplot for various room types that are preferred by the customers. Looking at the median and omitting the outlier we have seen that major booking are made in hotel rooms followed by entire home or apartment and least as shared rooms. The customer are willing to pay more for the Hotel rooms with average price as $200 and above compared to the rest.

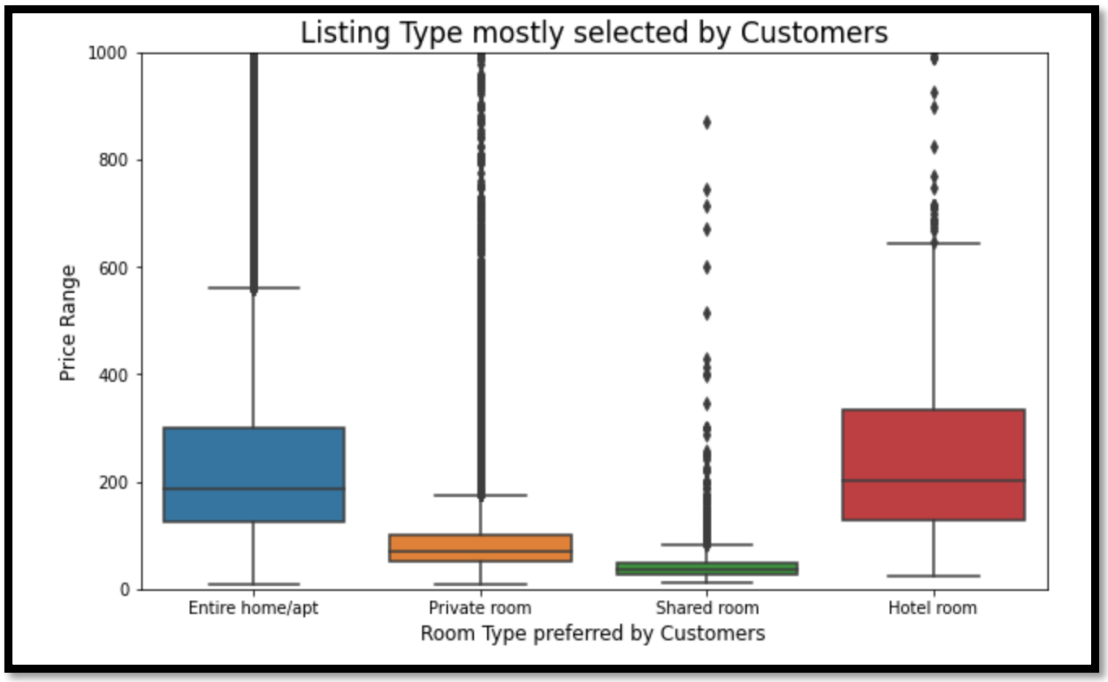

Figure 5: Boxplot indicating the customer choices based on the cost

Furthermore, Correlation matric is used here to determine which price has the highest positive correlation with accommodates, bedrooms, and beds in the respective order. The price of the property listing rises as the above-mentioned elements rise. The scale on the right-hand side shows the range of correlation between the features where 1 being the highest and -1 the lowest.

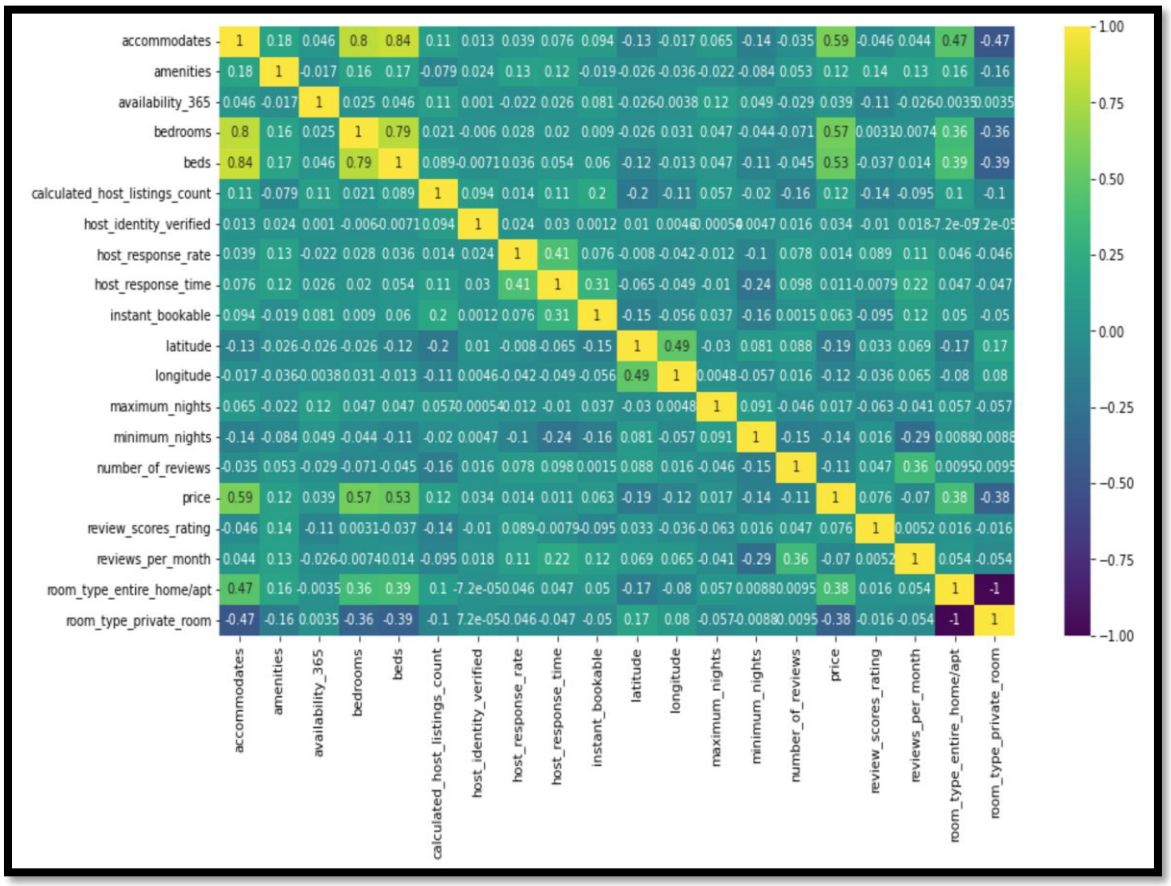

Figure 6: Correlation Matrix between various features

Model Building

We examined the accuracy of four machine learning prediction models in this project, determining their practicality and scope for the project's unique use case. Here, we have taken train and test of the data as 7:3 or can say 70 -30. Looking at the features that are required for modelling and checking if its numerical or categorical in nature. One hot encoding was performed on categorical values and scaled the numerical values using the robust scaler to level up the variations. Following Ml models are used to perform prediction of the prices:-

1. Linear Regression
2. Decision Tree
3. Random Forest
4. XGBoost (Extreme Gradient Boosting)

Here, we have used **Lime and Shap** python libraries for model explainability. SHAP (Shapley Additive explanation) is a model feature influence scoring technique that uses Shapley values. The "average marginal contribution of a feature value over all potential coalitions" is the technical definition of a Shapley value. To put it another way, Shapley values take into account all potential predictions for a given instance based on all conceivable input combinations. SHAP can ensure qualities such as consistency and local accuracy thanks to this thorough methodology. LIME (Local Interpretable Model-agnostic Explanations) constructs sparse linear models around each prediction to describe how the black box model operates in that specific area.
Hyperparameters are the parameters that govern the model architecture, and hyperparameter tuning is the process of determining the ideal model architecture. They are usually addressed before the training method begins. These values represent important features of the model, such as its complexity and learning rate. Models might have several hyperparameters, and the goal is to identify the best configuration for them.

1. Linear Regression :-

A approach for modelling the relationship between one or more independent variables and a scalar dependent variable is linear regression. Simple linear regression is used when there is only one independent variable, whereas multiple linear regression is used when there are numerous input variables. For the linear regression model, I used the default hyperparameter. R2 is 0.42, which is a low number. With an accuracy of 57.48 percent, it can demonstrate that the model is not overfitted. Linear regression shows that bedrooms and accommodations are the most important features.

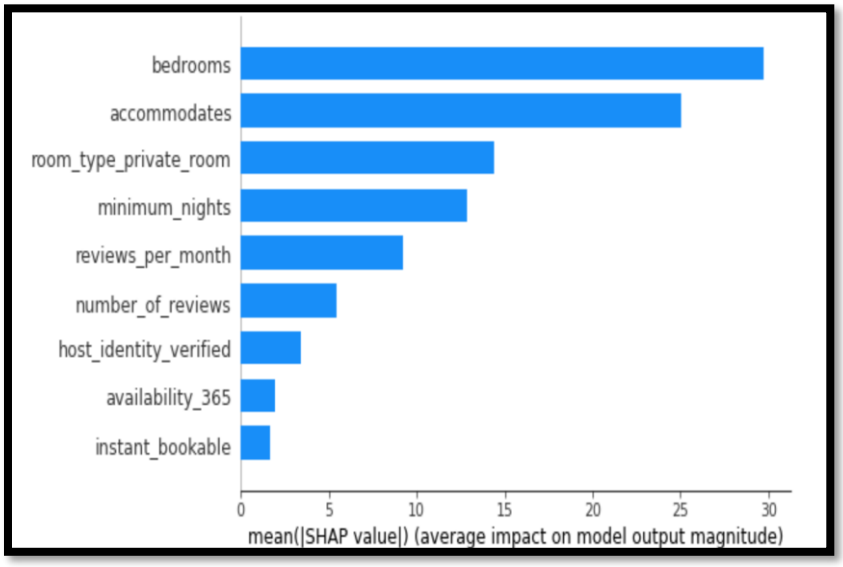

Figure 7: Tread seen for preferences

Below you can see the trend in the original price and predicted price of the data based on the sample that was considered. The predicted has 57.48 percent accuracy as compared to the actual values.

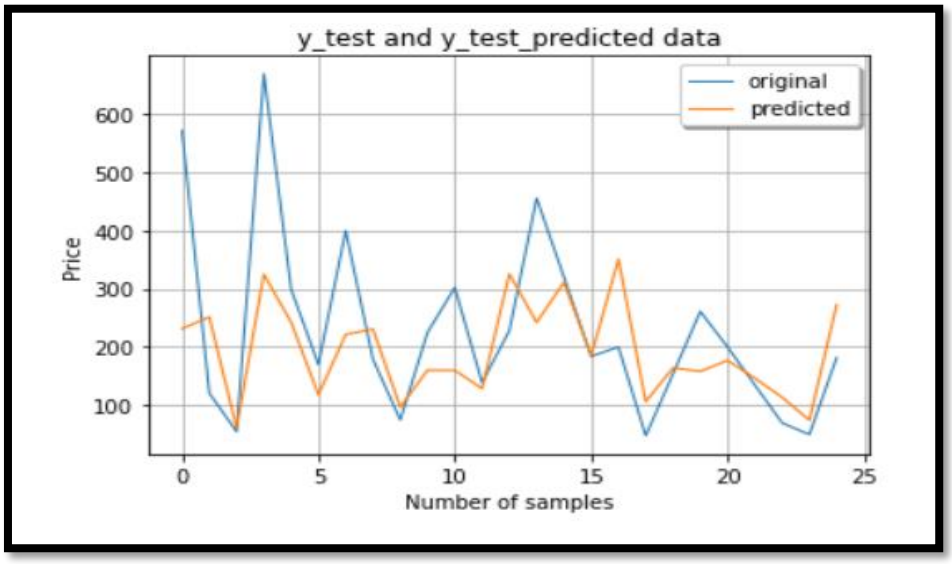

Figure 8: Prediction trend Vs. Original trend

2. Decision Tree :-

A non-parametric supervised learning tool for regression is the decision tree. It's used to build a model that learns basic decision rules from data attributes and forecasts the value of a target variable. The choosing criteria get more advanced as the tree grows deeper, and the model becomes more accurate. Experimented with numerous hyperparameters and found the following combination to be the most effective with random state as 42 and maximum depth as 10. With an accuracy of 58.75 percent, the R2 value on training data is 0.5 and 0.4 on test data, indicating that overfitting is avoided.
Similar to linear regression, bedrooms and accommodations are assigned the greatest feature value.

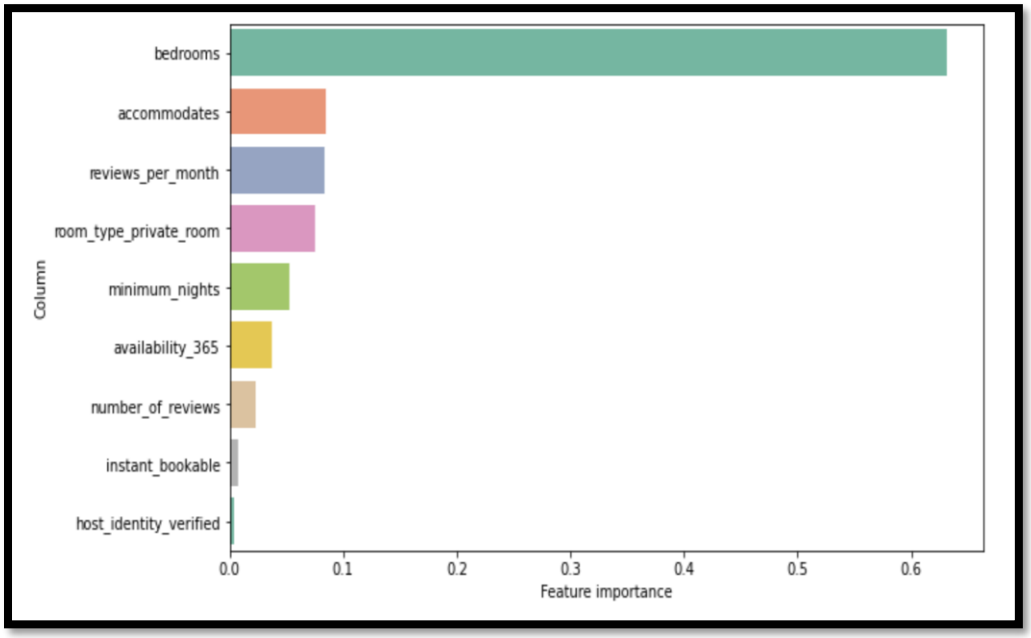

Figure 9: Shows feature importance

Looking at the plot below it clearly indicates that the original and predict values have certain similarity and is better than that of the linear regression model. The likes coincide or has similarity that is around 58.74 percent accurate.

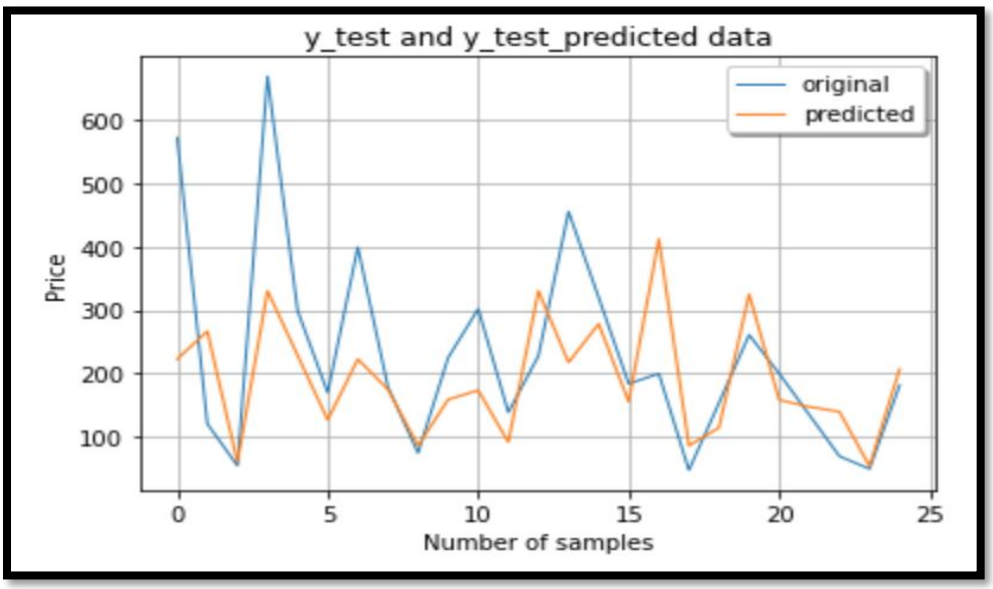

Figure 10: shows the price prediction using decision tree

3. Random Forest:-

Random forest is a decision tree-based modelling and behaviour analysis approach that comprises of numerous decision trees, each reflecting a distinct categorization of data entered the random forest. It's based on the idea that many considerably uncorrelated models (trees) working together as a group will outperform any of the individual models. Experimented with numerous hyperparameters and found the following combination to be the most effective which gave n_estimators as 100 and random state values as 42. With a model accuracy of 58 percent, we achieved R2 values of 0.92 for the training set and 0.42 for the testing set. Again, Bedrooms, monthly reviews, and accommodations are the top three most crucial qualities.

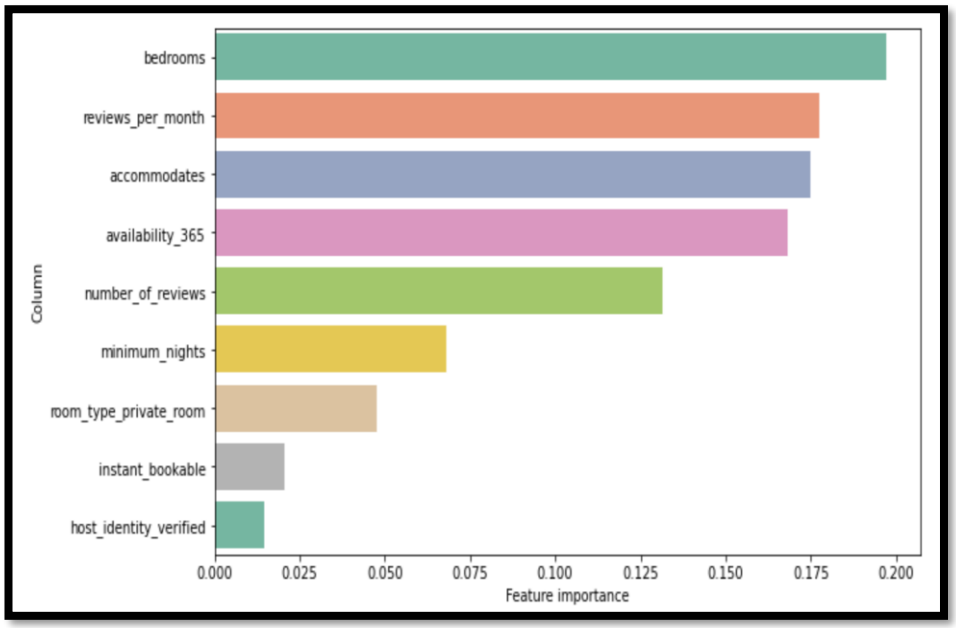

Figure 11: shows feature importance using Random Forest

The below graph shows that random forest gives accuracy to our feed data to around 58 percent which is very close to the accuracy predicted by the decision tree previously.

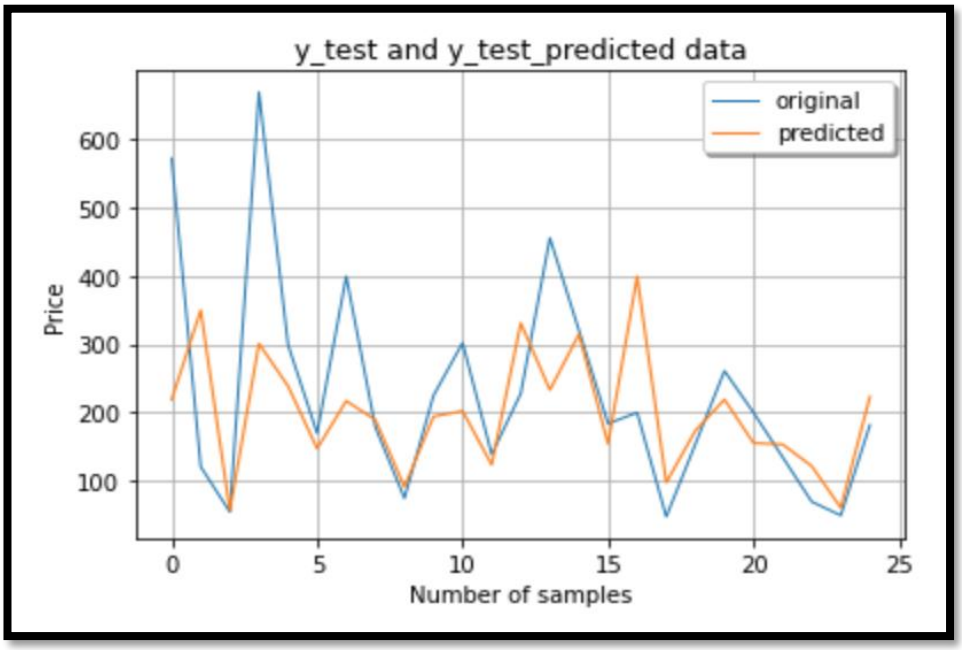

Figure 12: Shoes accuracy of the predicted model and original data values

5. XGBoost (Extreme Gradient Boosting)

XGBoost is an efficient, flexible, and portable distributed gradient boosting toolbox. It creates machine learning algorithms using the Gradient Boosting framework. XGBoost employs parallel tree boosting to solve a wide range of data science problems rapidly and reliably. The same algorithm can handle billions of examples and can solve problems in huge, distributed settings. Experimented with numerous hyperparameters and found the following combination to be the most effective with n_estimators as 100 and random state as 42. R^2 of 0.7 on the training set and 0.44 on the testing set, with an accuracy of 59.97%, the highest of the four models. Bedrooms, room type private room, and accommodations are the most important features.

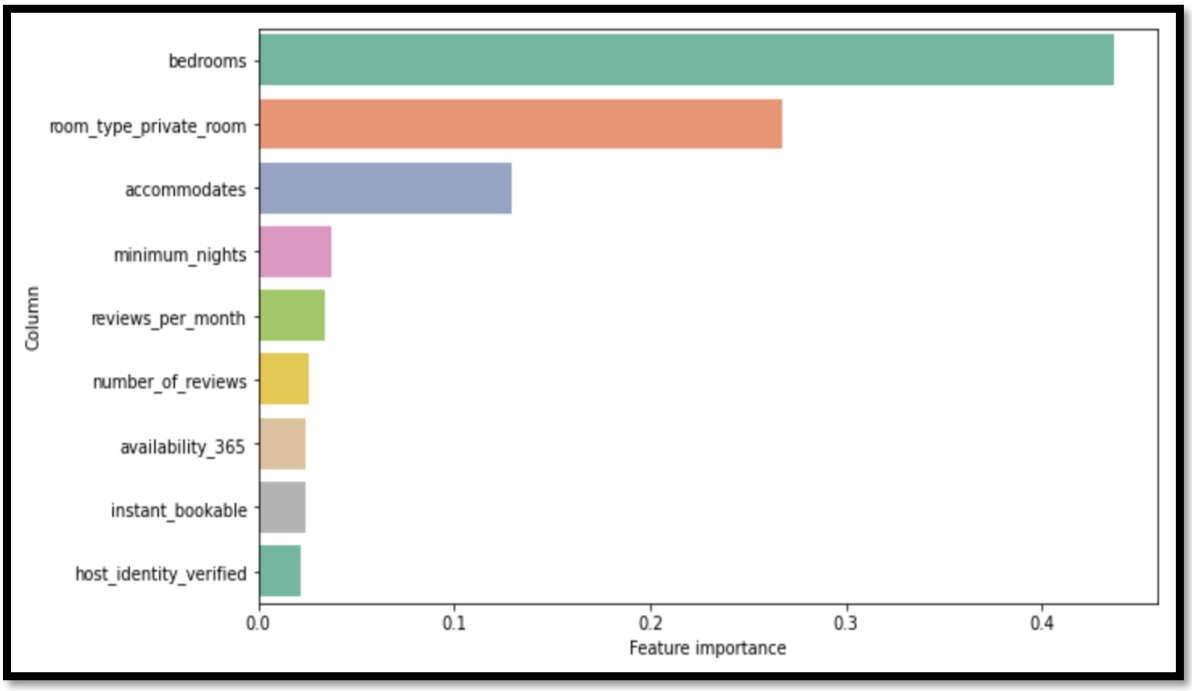

Figure 13: Showing the preference based on XGBoost model

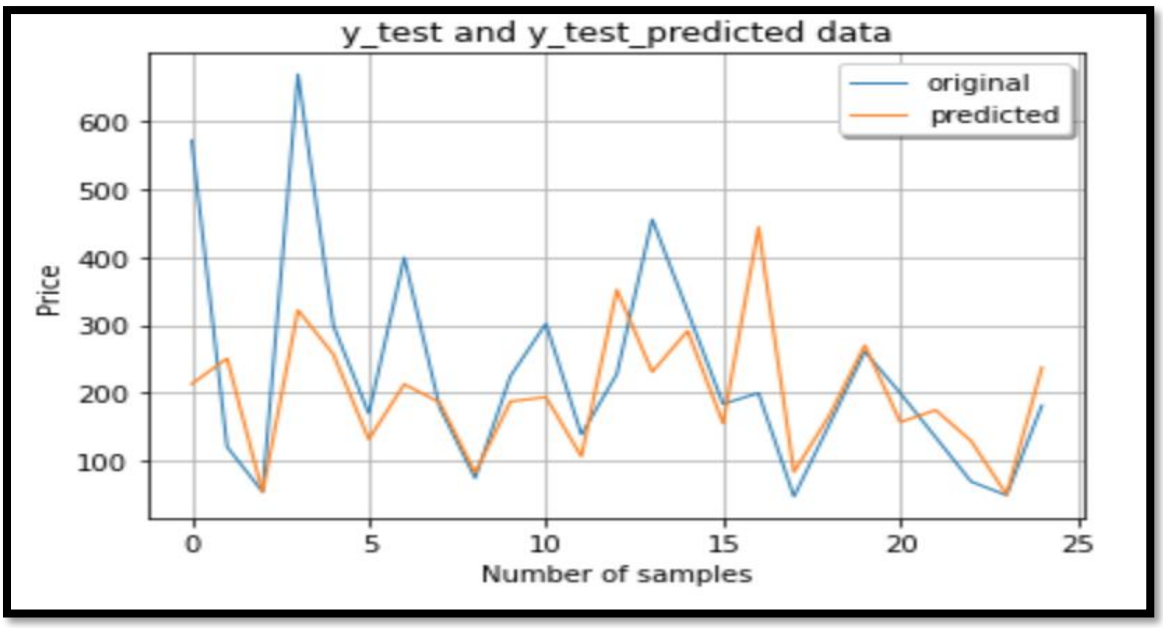

Figure 14: Shows predicted value vs. original value

Model Comparison: -

To compare the performance of the models, the Root Mean Squared Error (RMSE) was chosen. Given the least RMSE value, XGBoost is shown to be the best model to perform among all the models. Decision Tree performs almost identically to XGBoost.

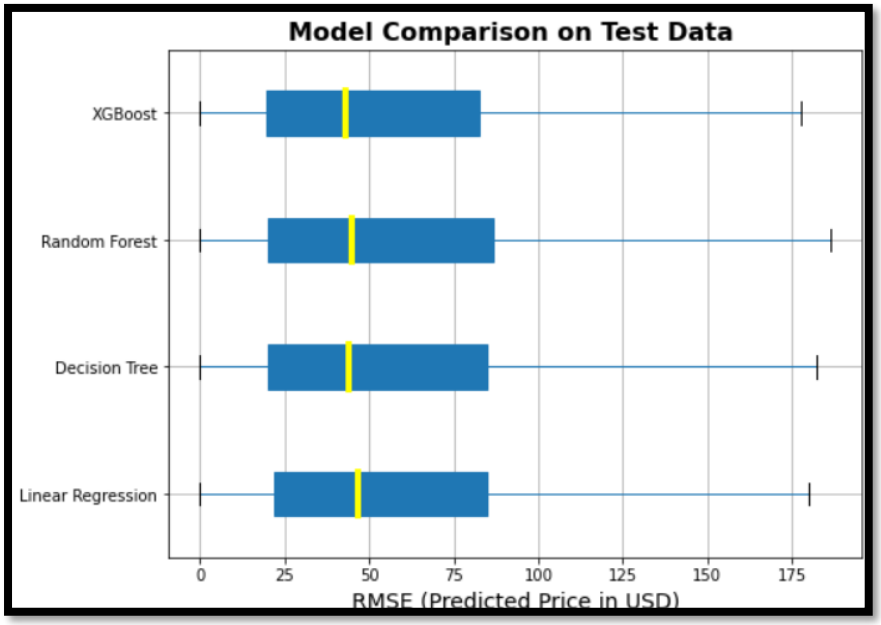

Figure 15: Comparison between various models

Streamlit UI Application

The final model was deployed using a Streamlit UI, which is a Python framework for developing Machine Learning and Data Science web apps. Using Streamlit, we can quickly create web apps and deploy them. Streamlit enables you to create apps in the same way as you would write Python code. Working on the interactive cycle of coding and watching outcomes in the web app is made simple using Streamlit (Streamlit • The fastest way to build and share data apps, 2022).

The user can obtain the predicted price from the model using the UI. The sidebar panel has user inputs for feeding in the requirements, and the final price along with other visualizations can be obtained in the main panel.

Link to the app - https://eaiproject.herokuapp.com/

A **YouTube link** to the presentation of the project can also be found here: https://www.youtube.com/watch?v=GQsjwYs4GX4

Figure 16: Streamlit UI application screenshot.

Conclusion: -

We analyze the Airbnb dataset with 210,617 listings and 23 attributes in this research. To preprocess the data, we go through many data science procedures such as Data Collection, Data Cleaning, Data Manipulation, Data Visualization, and Data Reduction. We created four models to forecast listing prices. Linear Regression, Decision Tree, Random Forest, and XGBoost are the four methods. On the test set, the XGBoost and Random Forest models had the highest R2 scores (0.44 vs. 0.42). To determine how characteristics impact pricing at the local and global levels, we employ LIME, SHAP, and Feature Importance. We've concluded that the price is largely affected by three features: bedrooms, accommodations, and room type.

References: -

Radečić, D. (2022, March 21). *LIME vs. SHAP: Which is Better for Explaining Machine Learning Models?* Medium. https://towardsdatascience.com/lime-vs-shap-which-is-better-for-explaining-machine-learning-models-d68d8290bb16

SHAP and LIME: Great ML Explainers with Pros and Cons to Both. (n.d.). Shap & Lime. https://blog.dominodatalab.com/shap-lime-python-libraries-part-1-great-explainers-pros-cons

Streamlit • The fastest way to build and share data apps. (2022). Streamlit • The fastest way to build and share data apps. https://streamlit.io/