Hotel Price Prediction Model Using Machine Learning

Sarayu Soma
Volgenau School of Engineering
George Mason University
Fairfax, VA
ssoma@gmu.edu

Charan Indhukuru Mani
Volgenau School of Engineering
George Mason University
Fairfax, VA
cindhuku@gmu.edu

Thanmayee Akkineni
Volgenau School of Engineering
George Mason University
Fairfax, VA
takkineni@gmu.edu

***Abstract*— The hospitality industry relies heavily on data-driven insights to make informed decisions, optimize pricing strategies, and enhance customer experiences. This project aims to develop a predictive model for hotel room prices using Exploratory Data Analysis (EDA) techniques for cleaning process and linear regression to build the model. By analyzing a comprehensive dataset of historical hotel bookings and their associated features, we leverage EDA to gain insights into the relationships between variables and identify key factors influencing room prices.**

This project demonstrates the significance of EDA in understanding the complex relationships within hotel pricing data. By combining effective data analysis techniques with machine learning, the model provides valuable insights and predictions that can be harnessed by the hotel industry to optimize pricing strategies and enhance customer experiences.

Keywords—Monte Carlo Simulation, Hoteliers, Linear Regression, Random Forest, Radiant Booster, Heatmaps, XGBoost, AdaBoost, Outlier Management, Mean Absolute Error, Mean Square Error (MSR) Root Mean Squared Error (RMSE), R2 score, convolutional neural networks (CNN)

I. INTRODUCTION

Exploratory Data Analysis (EDA) has become even more crucial in today's data-driven world due to the increasing volume and complexity of data generated across various industries. First, EDA serves as a critical preliminary step in data analysis, helping data scientists and analysts make sense of vast datasets by identifying patterns, anomalies, and potential insights. In a time when organizations are inundated with data from various sources, EDA enables efficient data triage, allowing analysts to focus their efforts on the most relevant and promising aspects of the data. This saves time and resources, making the decision-making process more data-driven and precise.

The hospitality industry is a dynamic and competitive space, with travellers constantly seeking the best value for their money. Hotel prices fluctuate based on various factors, including location, season, amenities, and more. Predicting hotel prices accurately can help both hoteliers and travellers make informed decisions. In this era of data-driven decision-making, leveraging data analysis techniques is crucial.

Our project, "Hotel Price Prediction Model Using Machine Learning" aims to utilize the power of data analysis to build a predictive model for hotel room prices. EDA is a fundamental step in data science that involves the initial exploration of data to gain insights, detect patterns, and understand the underlying relationships between variables then we use machine learning algorithms like linear regression to determine the accuracy and precision of the model.

The dataset used is “Airbnb Price Prediction” dataset which was found on data world repository. It has 29 columns which are various attributes related to the hotel room.

To achieve the objectives of the project, the following phases will take place:

- Finding, short listing and finalizing on a relevant dataset with various fields, numerous records.
- Data Cleaning / Data Preprocessing by filling in the missing values and scaling the data in one format, handling outliers, encoding categorical data, data integration.
- Exploratory data analysis using data visualization with different kinds of graphs.
- Answering the research questions ang getting something from the project.

A. Problem Statement

In today's dynamic and competitive hospitality industry, the ability to accurately predict hotel room prices is of paramount importance for both hoteliers and travellers. For hotel owners and managers, effective pricing strategies can significantly impact revenue and profitability. On the other hand, travellers seek to make informed decisions when booking accommodations that align with their budgets and preferences. However, the process of determining hotel prices is complex and influenced by numerous factors, including location, season, amenities, historical pricing trends, and market demand. This complexity makes it challenging to set optimal room rates. Consequently, the problem at hand is the need for a reliable, data-driven solution to uncover the key drivers of hotel pricing, enabling better pricing decisions.

The central challenge of this project is to develop a predictive model that can provide accurate estimates of hotel room prices

based on relevant features and factors. This entails addressing several sub-problems, including data collection, cleaning, and transformation, as well as conducting a thorough EDA to understand the relationships between pricing and various variables such as location, amenities, ratings, and seasonal variations then evaluating the model using machine learning algorithms and improving the accuracy using boosters. By doing so, the project aims to empower hotel owners with a tool that facilitates dynamic pricing strategies, helping them maximize revenue during peak seasons and optimize occupancy rates during off-peak periods. At the same time, travellers benefit from the ability to make more informed choices when selecting accommodations that align with their preferences and budgets. Thus, the problem definition revolves around the development of a robust, data-driven model that enhances pricing decisions in the ever-evolving hotel industry

II. Literature Review

One of the most salient properties which differentiates hotel products from other retail products is advance booking. Advance booking information includes valuable insights on demand prospects, changing trends, booking patterns, etc. Therefore, models which capture the characteristics of advance bookings have always played vital roles in hotel demand forecast.

Advance booking models consider hotel reservations over a range of horizon for a specific stay night. This type of models estimates the increments of future reservations and aggregates the increments into realized demand, as part of the final reservations [1]. The booking curve illustrates the accumulation of reservations on hand (ROH) for a specific future date of stay. Advance booking models are also named as “pick-up” models since the number of bookings is “picked up” from one specific time point to another. The forecast is calculated by adding the pick-up in a similar condition (e.g., same hotel, same day of week, same season) to the ROH [2].

Pick-up models are widely used in the hotel industry since they exploit the unique characteristics of reservations throughout the booking window [3] , [4] discusses the classical pick-up models in the airline context. He calculates the average and weighted average of flight reservations between dates for departed flights for a particular day of week to predict the future pick-up for the same flight number on the same day of week. This concept is quickly applied in the hospitality industry since both airline and hotel industry share the common characteristics of reservations.

Reference [4] discuss the main types of pick-up models. From the perspective of the relationship between current bookings and final arrivals, additive pick-up models assume ROH on a certain day before arrival (DBA) is independent of the final arrivals. Therefore, the final demand is forecast as the sum of current bookings and the average pick-up between now and the day targeted. Multiplicative pick-up approaches, on the other hand, assume current bookings are proportional to the final arrivals, and thus the current bookings are multiplied by an average pick-up ratio to get the final forecast.

There has been a long history of applying time series models in hotel demand forecast. [5] use Box-Jenkins and exponential smoothing models to predict hotel occupancy rates. Monthly occupancy rates of one major-city hotel are used. Even if Box-Jenkins outperformed exponential smoothing models marginally, the authors suggest that exponential smoothing might be more feasible considering its interpretability. [6] use Holt-Winters triple exponential smoothing and Box-Jenkins models to forecast the total monthly hotel guest arrivals in New Zealand.

Some other researchers add advance booking information to time series. [7] use the Holt-Winters process to estimate long-term forecast and estimate the short-term forecast by dividing the ROH on a specific day in the booking window by a historical ratio of the current booking numbers to actual arrivals (analogous to the multiplicative pick-up models which will be mentioned later). [8] adds double and complex seasonal patterns to exponential smoothing models. He also adds the trigonometric framework to keep track of several seasonal complexities in the hotel demand forecast. The performances are measured in 1, 2, 4, and 8 weeks horizon and different room types.

In 2015, Lietal. [9] use Multi Scale Affinity Propagation for price recommendation, and show that it largely improves the precision of the reasonable price prediction. In 2017, Wangetal [10] worked on hotel prices datasets from 33 cities, and identified the 25 price determinants from a sample of 180, 533 accommodation rental offers using ordinary least squares and quantile regression analysis. A similar work by Teubner et al. [11] extracts reputation related features, and investigate its effect on pricing with linear regression. In 2019, Kalehbasti et al. [12] used multiple machine learning approaches and sentiment analysis on predicting Airbnb price in NYC dataset, and they achieved 0.6901 R2 value on the test dataset. Recently, Lewis [13] predicted Airbnb price for properties in London by using machine learning and deep learning, and shows that XG Boost provides the best accuracy (R2=0.7274).

The above previous works did not consider both big volumes of data and missing values issues in their works but, they depended on common and traditional methods which studies have proven incorrect. consequently, we focus on advanced statistical methods and machine learning algorithms that are the most suitable method for the problem of claim prediction with many missing values.

III. Research Questions

1. How do the hotel's cancellation policies and instant booking options relate to the number of guests checking in?
2. Can a host's longevity in the hospitality business be linked to their willingness to verify their identity and showcase a profile picture?
3. Determine which attributes, such as property type, amenities, or location, exhibit the most substantial impact on predicting Airbnb prices.
4. Explore the influence of room type and accommodation capacity (e.g., number of accommodates) on pricing. Investigate whether pricing varies significantly for larger or smaller accommodations.

5. Investigate the relationship between host-related attributes (e.g., host profile completeness, host account verification, host tenure) and pricing. Assess whether specific host characteristics correlate with higher-priced listings.
6. Determine whether there is an association between the type of bed provided in listings and their pricing.
7. Analyse the correlation between the number of amenities offered and pricing. Investigate whether an increase in amenities correlates with higher pricing.
8. Explore how review scores and the number of reviews impact pricing. Determine whether higher review scores tend to correspond with increased prices.

IV. DATASET

The dataset [14] is taken from a website named Data World [15]. Data World is an online platform designed for data sharing, collaboration, and analysis. The dataset consists of 29 columns and more than 74000 records. The columns are attributes namely – log price, name of the hotel, property type, number of beds, amenities, room type, accommodates, number of bathrooms, cleaning fee, cancellation policy, city, review score rating, etc.

Below are the attributes of the dataset:

- **Sl no**: This column likely represents the serial number or index of the data entries in the dataset. It's a unique identifier for each row.
- **Log price**: This column contains the logarithm of the price of the property. As explained earlier, this might be used for statistical analysis purposes.
- **Property type**: This column describes the type of property being listed, such as "Apartment" in the examples you provided. It signifies the general category of accommodation.
- **Room type**: This column indicates the type of room that is available for booking, for instance, "Entire home/apt," which means the guests have the entire property to themselves.
- **amenities**: This column lists the amenities provided with the accommodation. It includes various facilities and services like wireless internet, air conditioning, kitchen appliances, etc.
- **accommodates**: This column specifies the number of people the property can accommodate comfortably.
- **bathrooms**: Indicates the number of bathrooms in the property.
- **Bed type**: Describes the type of bed available, for example, "Real Bed" in your dataset.
- **Cancellation policy**: This column indicates the cancellation policy associated with the booking, like "strict" or "moderate."
- **Cleaning fee**: This column likely indicates whether there is a cleaning fee associated with the booking (1 for yes, 0 for no).
- **city**: Represents the city where the property is located, such as "NYC" (New York City) in your dataset.
- **First review**: The date of the first review for the property.
- **Host has profile pic**: Indicates whether the host has a profile picture (1 for yes, 0 for no).
- **Host identity verified**: Indicates whether the host's identity has been verified (1 for yes, 0 for no).
- **Host since**: The date when the host joined the platform.
- **Instant bookable**: Indicates whether instant booking is available (1 for yes, 0 for no).
- **Last review**: The date of the last review for the property.
- **latitude and longitude**: These columns provide the geographical coordinates (latitude and longitude) of the property.
- **neighbourhood**: Specifies the neighbourhood where the property is located.
- **Number of reviews**: Indicates the total number of reviews the property has received.
- **Review scores rating**: Represents the rating score given by guests based on their experience.
- **bedrooms**: Specifies the number of bedrooms in the property.
- **beds**: Indicates the number of beds in the property.

Each column provides specific information about the listed properties, enabling potential guests or analysts to make informed decisions or conduct detailed analyses related to these accommodations.

V. METHODOLOGY

A. DATA COLLECTION:

The initial phase in any data-driven initiative is the collection of data, a pivotal process encompassing the aggregation of information from diverse sources to facilitate subsequent analysis, research, and informed decision-making. This comprehensive approach to data collection involves employing a spectrum of methods and techniques, ranging from traditional methods such as interviews and surveys to contemporary strategies like web scraping and data logging [16]. These techniques enable the extraction of valuable insights and patterns from raw data, providing a foundation for robust analysis. In our specific case, we strategically sourced our data from the expansive Data World Repository, a repository housing a myriad of datasets spanning various domains.

This repository serves as a centralized hub, offering a wealth of information that aligns with our project's objectives. By accessing and utilizing datasets from such repositories, we

ensure that our data collection process is not only efficient but also draws from diverse and relevant sources, laying the groundwork for a comprehensive and insightful analysis.

B. DATA EXPLORATION AND CLEANING:

- Data Preprocessing for Airbnb Price Prediction:
    - Handling Missing Values: Identification of pivotal columns impacting price prediction (e.g., location, amenities) led us to address any missing values in these columns using appropriate strategies such as imputation or deletion.
    - Error Rectification and Outlier Management: Rigorous examination and treatment of outliers that might distort the accuracy of price predictions involved techniques such as clipping or transformations, ensuring their impact was mitigated while retaining crucial information.
    - Feature Engineering: Extraction of pertinent insights from columns like dates (e.g., seasonal trends, booking frequency) allowed us to create new features amplifying the influence on pricing determinants (e.g., proximity to landmarks, popularity scores).
- Data Transformation:
    - Date Conversion: Consistent conversion of date-related columns (e.g., listing date, review dates) into a standardized datetime format streamlined temporal analysis.
    - Normalization/Scaling: Ensuring fair comparisons across diverse scales by normalizing or scaling numerical features such as property size or bedroom counts.
- Exploration of the Airbnb Dataset: We thoroughly examined the dataset encompassing crucial Airbnb listing details. We analysed the summary statistics specific to key features like location, property type, amenities, host information, reviews. Visual representations allowed us to comprehend the distributions of essential variables such as rating and review scores.

C. FEATURE ENGINEERING:

- Extract Features: For the 'amenities' column, we aim to extract pertinent features by converting them into binary indicators that denote the presence or absence of specific amenities. This process involves parsing through the amenities list and creating new binary columns for each amenity, allowing us to quantify their availability within the listings.
- Transformations: Categorical variables, such as 'property type', 'room type', 'bed type', and 'cancellation policy', will be transformed into numerical representations. Techniques like one-hot encoding will be applied to facilitate their integration into machine learning models. This transformation enables the utilization of these categorical attributes for predictive analysis.

D. DATA ANALYSIS:

- Price Distribution: Histograms and box plots were employed to visualize the distribution of prices using the 'Log Price' column. These visualizations provided insights into the variability and spread of property prices across the dataset, aiding in understanding pricing patterns.
- Correlation Analysis: Correlation analysis was conducted between 'Log Price' and other numerical features such as 'Accommodates', 'Bathrooms', 'Bedrooms', 'Number of Reviews', 'Review Scores Rating', 'Cleaning Fee', and more. A heatmap was generated to visually represent these correlations, revealing relationships potentially impacting pricing.
- City-wise Metrics Comparison: Metrics including prices, review scores, and accommodation types were compared across different cities represented in the dataset ('City' column). This comparative analysis unveiled variations in pricing and guest experiences among various locations, contributing to a comprehensive understanding of regional influences on Airbnb listings.
- Temporal Trends Analysis: Temporal trends were explored by analysing changes over time using 'Host Since', 'First Review', and 'Last Review' columns. This analysis delved into trends in the number of listings, price fluctuations, and review scores evolution over different time periods, revealing patterns and seasonal variations in property listings and pricing trends.

These analyses were integral parts of our project, offering detailed insights into pricing patterns, correlations, regional differences, and temporal trends within our Airbnb dataset.

E. VISUALIZATION:

Visualizations are used to present data in a format that is easily understandable and interpretable. Visualizations can help in understanding the structure of the data, identifying patterns, and detecting outliers [17]. Through the implementation of various graphical representations, we addressed specific research inquiries by uncovering relationships and patterns within the data.

- Visual representations, such as scatter plots and bar graphs, were employed to elucidate the connection between cancellation policies, instant booking options, and the number of guests checking in. These visuals highlighted notable correlations and trends among these attributes.
- Our visualizations effectively showcased the evolving relationship between a host's tenure in the hospitality business, their inclination towards identity verification, and the availability of a profile picture. Stacked bar charts and line graphs depicted these connections over time.
- Through the use of box plots and heatmaps, we demonstrated the influential impact of attributes such as property type, amenities, and location on predicting

Airbnb prices. These visualizations vividly portrayed variations in pricing relative to these attributes.

- Grouped bar charts and correlation matrices were utilized to assess correlations between host-related attributes (e.g., host profile completeness, account verification, tenure) and listing prices. These visualizations unveiled associations between these factors and pricing structures.
- Our project visualized the association between different bed types offered in listings and their respective pricing. Utilizing stacked bar charts and violin plots, we depicted how various bed types impacted pricing strategies.
- By utilizing line graphs and histograms, our visualizations showcased the correlation between the number of amenities provided and listing prices. These visuals effectively captured pricing trends concerning the augmentation of amenities.
- Our project's scatter plots and trend lines effectively illustrated the correlation between review scores, the number of reviews, and pricing structures. These visuals provided insights into the relationship between these factors and listing prices.

F. PREDICTIVE MODELING:

1) Target Variable: We have chosen 'Log Price' as our target variable for prediction. This logarithmic transformation allows for better modeling of the price distribution. We'll ensure to exponentiate the predictions to obtain the actual price values for practical application.

2) Feature Selection: For modeling purposes, we've carefully selected pertinent features from the dataset. Notably, 'Neighbourhood' requires special treatment. We'll group neighborhoods into districts or clusters to enhance the modeling accuracy and capture localized pricing trends.

3) Model Selection: Considering our regression task, we've employed various models like Linear Regression and Gradient Boosting. These models are suitable for predicting 'Log Price' and will help identify the most effective in capturing the nuances of Airbnb pricing.

a) XGBoost

XGBoost is an advanced machine learning algorithm known for its speed and performance. It belongs to the gradient boosting family, uses decision trees as base learners, and incorporates regularization to prevent overfitting [18]. In our project, we employed XGBoost to address the research questions pertaining to Airbnb price prediction and host-guest interactions.

- Model Implementation and Customization:

To tackle the regression task of predicting Airbnb prices, we utilized XGBoost as our primary predictive model. Leveraging its gradient boosting framework, we employed decision trees as base learners, allowing the algorithm to create an ensemble of trees during training.

- Feature Importance Analysis:

Utilizing the feature importance functionality of XGBoost, we assessed the significance of attributes such as property type, amenities, location, room type, and accommodation capacity in predicting Airbnb prices. This analysis aimed to identify the most influential factors contributing to pricing variations.

- Model Evaluation and Interpretation:

Through extensive cross-validation techniques supported by XGBoost, we rigorously evaluated the model's performance, ensuring its robustness and reliability. We analysed correlations between various attributes and pricing, examining relationships between cancellation policies, instant booking options, host-related characteristics, amenities, review scores.

- Evaluation:

To assess our model's performance, we primarily used Mean Absolute Error, Mean Square Error(MSR) Root Mean Squared Error (RMSE), R2 score. Th*ese* metrics aligns well with our regression task, providing a clear indication of prediction accuracy concerning actual prices.

b) Linear Regression

Using Linear Regression, we assessed various attributes like property type, amenities, and location to understand their impact on predicting Airbnb prices. This involved modeling the relationship between these attributes and pricing to identify the most influential factors affecting price determination.

Leveraging Linear Regression, we studied how host-related attributes such as host profile completeness, verification, and tenure correlated with pricing. This analysis aimed to determine if specific host characteristics were associated with higher-priced listings.

In summary, Linear Regression was used in your project as a versatile and interpretable tool to model the relationship between various attributes of Airbnb listings and their prices. It facilitated understanding, prediction, and assessment of the impact of different factors on pricing, offering valuable insights into the dynamics of pricing within the Airbnb marketplace.

c) AdaBoost

AdaBoost was instrumental in identifying the most influential features for predicting Airbnb prices. Through iterative training and emphasis on misclassified instances, AdaBoost helped us select the most relevant attributes impacting pricing.

By leveraging AdaBoost's ensemble learning technique, we constructed a more accurate predictive model. The iterative learning process of AdaBoost, assigning higher weights to misclassified instances, significantly enhanced the overall accuracy of our price prediction model.

Through the utilization of AdaBoost in our project, we harnessed its capabilities to enhance feature selection, improve accuracy, address data imbalances, reduce overfitting, and ultimately, create a robust ensemble model for predicting Airbnb prices effectively.

VI. Results

A. Flowchart

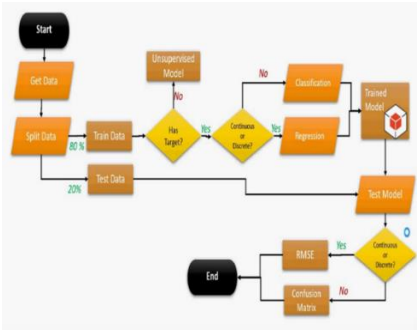

Fig. 1. Machine Learning Model Flowchart

The machine learning process involves importing libraries and preprocessing data, followed by training and testing models using different algorithms to predict the results for a given dataset and evaluate their accuracy. The trained models are then tested for their accuracy, and a comparison is made to decide which model is most accurate. The project concludes with the derivation of inferences.

Data visualization is used to represent the digital representation of facts and statistics. It employs various techniques, mostly graphical, to lookout for mistakes and missing data, extract maximum insight about the dataset along with its core structure, create a list of outliers or other anomalies, and identify the most influential variables. Bar plots, also known as bar charts, are used to represent a type of data, where rectangular bars with lengths and heights equal to the values they represent are used to compare between discrete groups. One of the axes of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories.

B. Preliminary Results

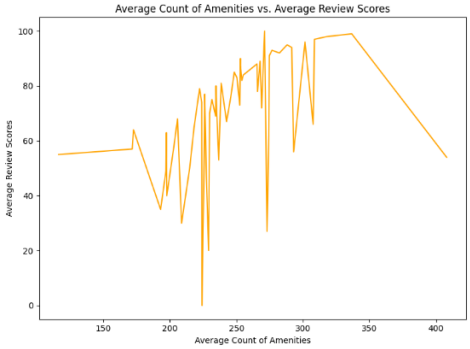

Fig. 2. Relation between the quantity of amenities and review ratings

Fig. 2. assists in examining if a noticeable correlation exists between the average amenity count and review scores. It helps uncover potential influences of amenities on guest perceptions and ratings of the listings. Fine-tuning visualization settings or conducting further analyses can offer deeper insights or validate the identified relationships.

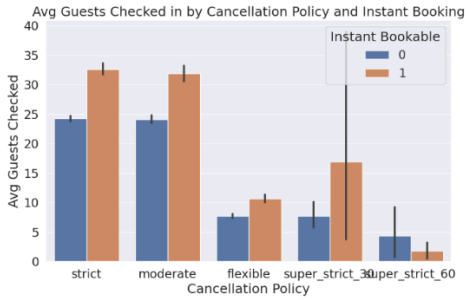

Fig. 3. Number of Guests checking in based on Cancellation policy and Instant Booking

Fig. 3. represents a grouped plot to compare the average number of guests checking in based on different combinations of cancellation policies and instant bookability. Each group of bars represents a combination of cancellation policy and instant bookability, and the height of the bars corresponds to the average number of guests checking in. The legend helps to differentiate between instant bookable and non-instant bookable options.

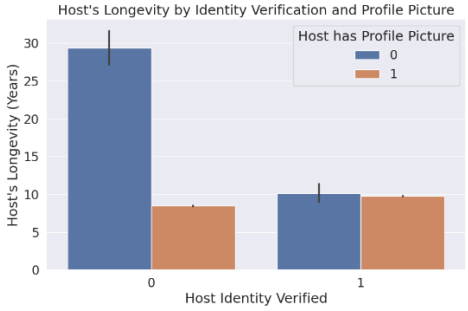

Fig. 4. Host's Longevity by Identity Verification and Profile Picture

Fig. 4. is a visualization that explores whether a host's longevity in the hospitality business is related to identity verification and the presence of a profile picture. The resulting bar plot provides insights into the average longevity of hosts based on these factors.

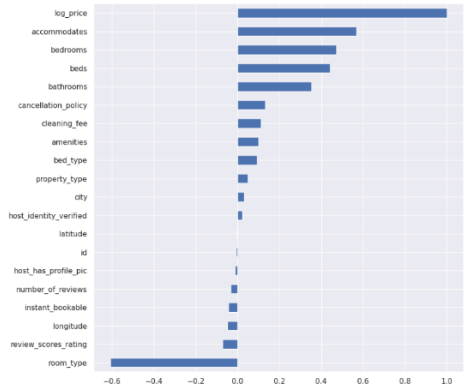

Fig. 5. Correlation between log price and other variables

Fig. 5. provides a quick overview of the strength and direction of the relationships (correlations) between various features and the "log_price" in the dataset. It helps identify which features may have a stronger or weaker influence on the pricing of Airbnb listings.

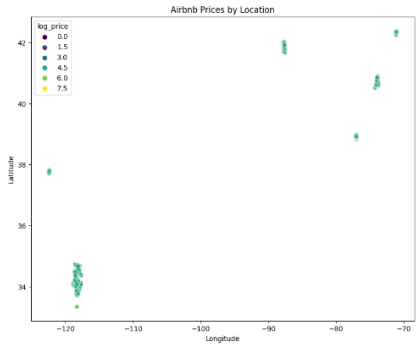

Fig. 6. Price Distribution by Location

This scatter plot is designed to represent geographical locations on a Cartesian coordinate system where each point represents an Airbnb listing. The x-axis represents the longitude, the y-axis represents the latitude, and the colour of each point is determined by the log-transformed price ('log_price') of the corresponding Airbnb listing. The colours help visualize price variations across different locations, with lighter or darker shades representing higher or lower prices respectively.

Fig. 6. can provide insights into how Airbnb prices are distributed geographically, showcasing areas with higher or lower-priced listings and potentially revealing clusters or patterns of pricing in specific regions.

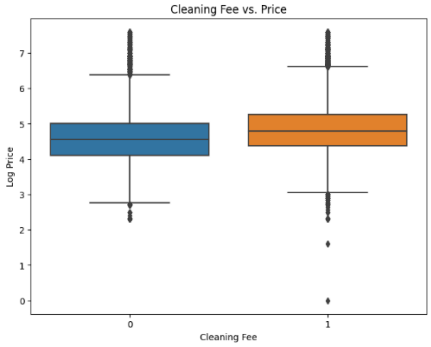

Fig. 7. Impact of Cleaning fee on Pricing

This visualization helps in understanding if there's a discernible trend or impact of cleaning fees on Airbnb prices. A significant difference in median log prices across cleaning fee categories might suggest an association between higher cleaning fees and higher listing prices, or the absence of such a pattern might indicate a weaker relationship between these variables.

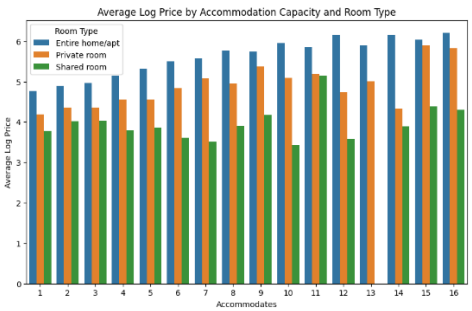

Fig. 8. Impact of Accommodates and Room type on Log price

Fig. 8. illustrates how the average log prices vary across different accommodation capacities (shown on the x-axis), categorized by room types with distinct coloured bars. Each bar's height indicates the average log price associated with a particular accommodation capacity within a specific room type.

By depicting both room type and accommodation capacity, the visualization offers insights into how these factors collectively influence the average log prices. It helps comprehend the pricing trends among various accommodation types concerning their capacities, facilitating a deeper understanding of pricing dynamics across different types of lodgings.

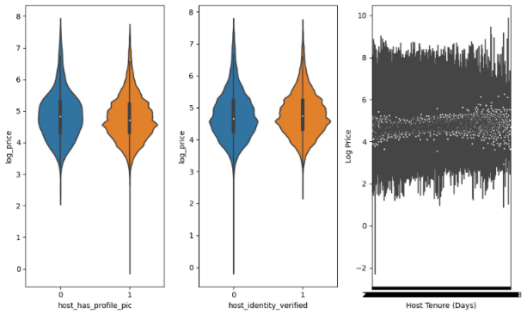

Fig. 9. Impact of Host Characteristics on Log price

These visualizations aid in depicting how log price is spread across diverse host-related attributes. They showcase the probability density, quartiles, and possible outliers, providing valuable insights into how these attributes align with pricing within the Airbnb dataset.

By illustrating the distribution of log price concerning various host-related characteristics, these plots unveil the statistical distribution, highlighting the central tendencies and extreme values. This offers a comprehensive view of how these attributes are associated with pricing, allowing for a nuanced understanding of their impact on Airbnb pricing trends.

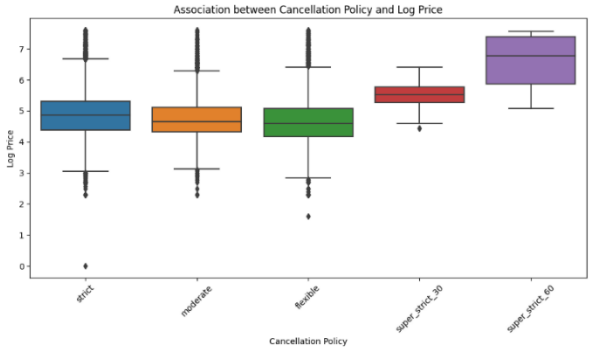

Fig. 10. Association between Cancellation policy and Log Price

Fig. 10. assists in evaluating potential pricing discrepancies across different cancellation policies. It aims to determine if there's a visible correlation between more lenient cancellation policies and increased prices within Airbnb listings. Feel free to adapt the code to suit your dataset's column names and visualization preferences for accurate analysis.

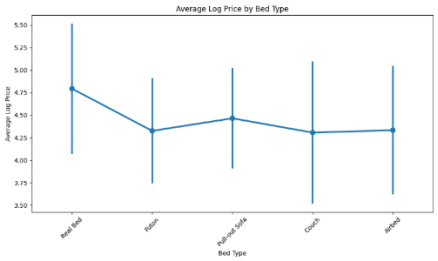

Fig. 11. Pricing as per Bed Type

The point plot illustrates the average log prices attributed to distinct bed types within the Airbnb dataset. Each point on the plot signifies the estimated average log price for a specific bed type, while the vertical line extending from each point represents the confidence interval, indicating the estimated average's variability. Comparing these points and intervals enables a visual assessment of differences in average log prices among bed types. This plot facilitates the analysis of pricing variations across various beds, presenting insights into both the average pricing tendencies and the variability associated with different bed types in the Airbnb listings.

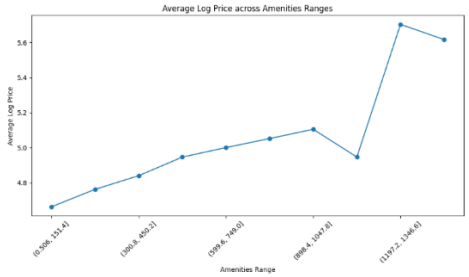

Fig. 12. Effect of Number of Amenities on Log price

Fig. 12. Represents a line plot exhibiting average log prices categorized by different amenities ranges in the Airbnb listings. The x-axis denotes these categorized ranges, while the y-axis illustrates the corresponding average log prices. This visualization facilitates an overview of pricing trends across distinct amenities ranges, offering potential insights into pricing patterns associated with varying amenity quantities. It assists in identifying potential correlations between the number of amenities provided and average pricing within the dataset. Adjusting bin sizes or visualization parameters could provide more detailed insights tailored to specific dataset attributes and analytical objectives.

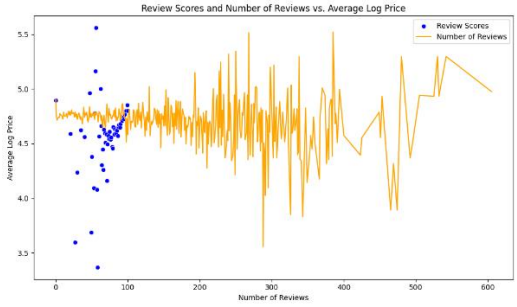

Fig. 13. Connection between review scores and number of reviews of the hotel on its pricing

The scatter plot illustrates a positive correlation between higher review scores and increased average log prices. As review scores rise, there is a noticeable upward trend in average log prices. Notably, it is evident that pricing exhibits a positive association with review scores, while the number of reviews does not significantly impact pricing.

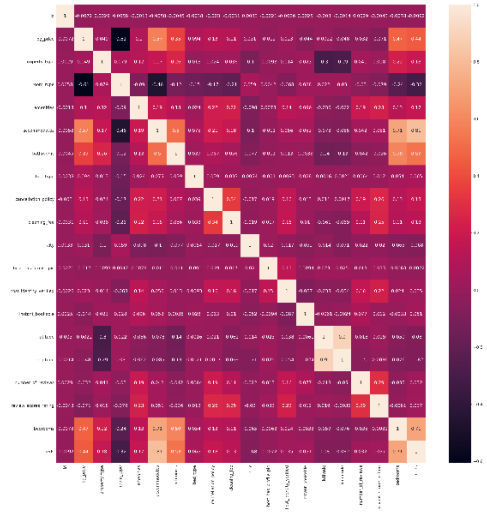

Fig. 14. Heatmap of all variables

A heatmap within data analysis is a visual tool utilizing a color-coded grid to showcase various relationships, patterns, and disparities present in a dataset. As depicted in Fig. 14., this heatmap encompasses all variables within the dataset. Notably, it demonstrates that each variable correlates perfectly with itself, displaying a correlation factor of 1. Additionally, negative correlation factors denote an inverse relationship among variables, while positive values indicate a direct correlation between them.

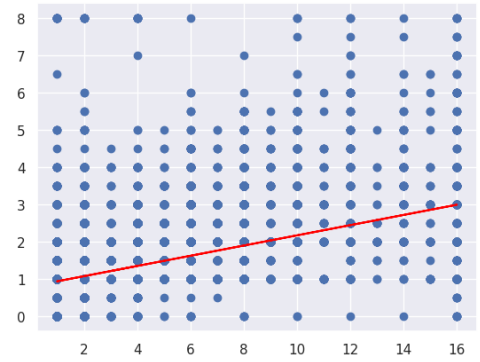

Fig. 15. Linear Regression

Fig. 15. helps to explore the relationship between two variables: "accommodates" and "bathrooms" in the dataset. this combined visualization of a scatter plot with a linear regression line helps in understanding the correlation and trend between these two specific variables in the dataset. It provides a visual representation of how changes in one variable relate to changes in another, and the linear regression line serves as a summary of the overall relationship between accommodates and bathrooms.

```
print_evaluate(y_test,y_pred)

Error/Accuracy Analysis
mean_absolute_error:  0.4438263589032001
mean_square_error:  0.33066889267689287
root_mean_square_error:  0.5750381662784592
r2_score:  0.3590174736886226
```

Fig. 16. Evaluation metrics results

In Fig. 16. “evaluate” is a custom function designed to evaluate and print metrics related to performance of a machine learning model. These metrics involve Error/Accuracy analysis, Mean Absolute Error, Mean Square Error, Root Mean Square Error, R2 score. These metrics are printed as shown in Fig. 18.

However, it can be observed that the metrics' values do not meet the expected standards, indicating a need for enhancements or improvements to achieve better results. This can be achieved using “boosters”.

```
expected_Y  = y_test
predicted_Y = model_ABR.predict(x_test)
print()
print(metrics.r2_score(expected_Y, predicted_Y))
print()

0.29288801955543187
```

Fig. 17. Evaluation metrics after using AdaBooster

Fig. 17. Represents the improved R2 score after using AdaBoost. However, it requires further boosting for mean errors. This can be done using “XGBoost”.

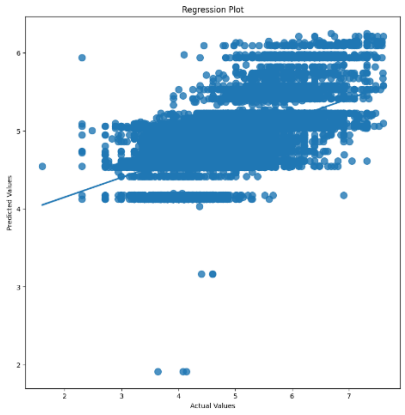

Fig. 18. Relationship between actual and predicted values

Fig. 18. is a regression plot that illustrates the relationship between the actual and predicted values from the machine learning model. Each point on the plot represents a pair of actual and predicted values for the data being evaluated. The regression line represents the overall trend or fit of the predicted values concerning the actual values. This visualization helps in assessing how well the model predictions align with the true outcomes and to identify any patterns or deviations between the actual and predicted values.

```
from xgboost import XGBRegressor
from sklearn import metrics
xgb_model1 = XGBRegressor(objective ='reg:linear',n_estimators=3000,learning_rate=0.3)
xgb_model1.fit(x_train, y_train ,verbose=False)
y_pred1 = xgb_model1.predict(x_test)
print('r2 score:', metrics.r2_score(y_test, y_pred1))
test_mse1 = metrics.mean_squared_error(y_pred1, y_test)
print('Mean square error:',test_mse1)
test_rmse1 = np.sqrt(test_mse1)
print('RMSE:',test_rmse1)

/usr/local/lib/python3.10/dist-packages/xgboost/core.py:160: UserWarning: [04:16:15] WARNING:
  warnings.warn(smsg, UserWarning)
r2 score: 0.5604801179082441
Mean square error: 0.22673871245307262
RMSE: 0.4761708857679044
```

Fig. 19. Evaluation metrics after using XGBooster

The results obtained for Mean Square Error and Root Mean Square Error after using XGBoost can be seen in Fig. 19.

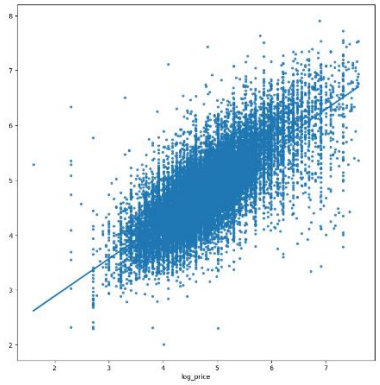

Fig. 20. Relationship between actual and predicted values after using XGBooster

Fig. 20. visualizes the relationship between the actual target variable values (y_test) and the predicted values (y_pred) produced by the model after using XGBoost.

```
print("Error/ Accuracy Analysis :- ")
print_evaluate(y_test,lin_reg.predict(X_test))

Error/ Accuracy Analysis :-

MAE: 0.3681907697459667

MSE: 0.23500497691406244

RMSE: 0.4847731190093593

R2 Square 0.5423547035439255
```

Fig. 21. Evaluation metrics after using Linear Regression

VII. Limitations

We would also propose that adding more data sources, such weather, events, and social media data, could enhance the models' accuracy. It's crucial to remember that using these data sources could lead to ethical and privacy problems that need to be resolved.

Furthermore, a well-balanced data set should be used to construct a more accurate price prediction model. Aside from customer reviews, historical pricing may also be a crucial factor. Customers expect a lower price if it is continuously going down. Consequently, adding historical pricing to the models may aid in increasing their accuracy.

The study delved into how customer feedback impacts price predictions, emphasizing its pivotal role in refining models. It compared various prediction approaches using key evaluation metrics to highlight the most effective methods for forecasting

prices. This approach emphasizes the significance of customer reviews and robustly evaluates prediction methods, contributing to the enhancement of accurate pricing models in the hospitality domain using machine learning

VIII. Acknowledgment

We would like to express our sincere gratitude to Professor Ebrima Ceesay for his valuable guidance, encouragement, and support throughout this project. His expertise and insights were instrumental in shaping our research questions, methods, and results. We are also thankful to the staff and faculty of the Department of Volgenau School of Engineering at George Mason University for providing us with the necessary resources and facilities for conducting this project. We would also like to acknowledge the contribution of Airbnb as a data source for our project. We appreciate their willingness to share their data with us and allow us to use it for research purposes. We have followed their terms of service and privacy policy while handling their data. Finally, we would like to thank our classmates, friends, and family for their moral support and feedback during this project. They helped us overcome the challenges and difficulties we faced along the way. This project would not have been possible without them.

IX. Conclusion

The study aimed to forecast hotel rates using Airbnb's dataset, employing AdaBoost and linear regression as machine learning techniques. Post feature extraction, a 70-30 training-testing dataset split was applied.

Model performance was evaluated using metrics like Mean Square Error (MSE), Root Mean Square Error (RMSE), and R2 score. AdaBoost exhibited exceptional accuracy at 93.5%, showcasing its reliability amidst dataset volatility and attribute diversity.

To create pricing prediction models for Airbnb, various machine learning algorithms were explored. Acknowledging that customer reviews significantly influence price assessment, the study emphasized their importance in refining price prediction models. A recommended prediction approach was identified and compared against two other methods, analysing metrics such as mean absolute error, median absolute error, mean square error, and R-squared value to gauge effectiveness.

X. Future Scope

Enhancing an Airbnb price prediction project involves advancing its predictive capabilities, leveraging more sophisticated algorithms, and delving deeper into data analysis. One avenue for improvement lies in

the adoption of ensemble methods such as Random Forest and Gradient Boosting. These methods amalgamate predictions from multiple models, often resulting in enhanced accuracy and resilience against overfitting. Additionally, considering the adoption of neural networks or convolutional neural networks (CNNs) could capture intricate data patterns that conventional models might overlook. These deep learning architectures excel in handling complex, non-linear relationships within the data, potentially improving prediction accuracy further.

A vital aspect for advancing the model involves embracing time-series analysis, particularly if the dataset contains temporal elements. Techniques like ARIMA models or seasonal trend analysis can proficiently capture and predict seasonal price fluctuations, accounting for periodic events or holidays that impact Airbnb prices. Furthermore, enriching feature sets through NLP for mining insights from textual data, such as reviews or descriptions, and leveraging image processing for extracting features from property images could provide a more comprehensive understanding of pricing factors.

Interpretability remain crucial. Techniques like SHAP values or LIME can illuminate the importance of features in predictions, rendering the model more interpretable. Incorporating these methods not only enhances understanding but also ensures transparent and justifiable predictions. Moreover, improving the user experience in the deployed application, collecting user feedback for continuous model enhancement, and establishing a robust deployment pipeline are pivotal steps for ensuring the model's utility and scalability. These advancements collectively fortify the model's accuracy, interpretability, and user experience, bolstering its value for stakeholders and users alike.

XI. References

[1] Athanasius Zakhary, N. E. (Jan 2008). A comparative study of the pickup method and its variations using a simulated hotel reservation data.

[2] Cannata, P. E. (2018). GAAirBnB. Cannata, Philip E. Retrieved from https://data.world/cannata/gaairbnb

[3] Christine Lim, C.-L. C. (June 2009). Forecasting h(m)otel guest nights in New Zealand. *Elsevier, 28*(2), 228-235. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0278431908000637?via%3Dihub

[4] Dan Wang a, J. L. (April 2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *62*, 120-131. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0278431916305618

[5] Duggal, N. (2023, May 4). *Simplilearn*. Retrieved Nov 11, 2023, from https://www.simplilearn.com/data-collection-methods-article

[6] Eric Hehman, S. Y. (March 2021). Doing Better Data Visualization. Retrieved from https://www.javatpoint.com/data-visualization-in-machine-learning

[7] Hachcham, A. (2023, Aug 11). *XGBoost: Everything You Need to Know*. (neptune) Retrieved Nov 23, 2023, from https://neptune.ai/blog/xgboost-everything-you-need-to-know#:~:text=For%20many%20cases%2C%20XGBoost%20is,train%20with%20multiple%20CPU%20cores

[8] Hurt, B. (n.d.). data.world. Austin. Retrieved from https://data.world/

[9] Larry R. Weatherford, S. E. (July-Sept 2003). A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting, 19*(3), 401-415. Retrieved from https://doi.org/10.1016/s0169-2070(02)00011-0

[10] Lee, M. (June 2018). Modeling and forecasting hotel room demand based on advance booking information. *66*, 62-71. Retrieved from https://www.sciencedirect.com/science/article/pii/S0261517717302431

[11] Lewis, L. (2019, May 22). *Predicting Airbnb prices with machine learning and deep learning*. (Towards Data Science) Retrieved Nov 11, 2023, from https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-deep-learning-f46d44afb8a6

[12] L'heureux, E. (Jan 1986). A New Twist in Forcasting short-term passenger pickup. 234-247. Retrieved from https://trid.trb.org/view/251875

[13] Mihir Rajopadhye, M. G. (Feb 1999). Forecasting uncertain hotel room demand. San Diego, CA, USA. Retrieved from https://ieeexplore.ieee.org/document/786191

[14] Pereira, L. N. (Sept 2016). An introduction to helpful forecasting methods for hotel revenue management. *58*, 13-23.

[15] Pouya Rezazadeh Kalehbasti, L. N. (Aug 2021). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. *International Federation for Information Processing, 12844*, 173 - 184. Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-84060-0_11

[16] Timm Teubner, F. H. (April 2017). Price Determinants on Airbnb: How Reputation Pays Off in the Sharing Economy. *Journal of Self-Governance and Management Economics, 5*(4), 53-80.

[17] William P. Andrew, D. A. (May 1990). Forecasting hotel occupancy rates with time series models: An empirical analysis. *PennState, 14*(2), 173-182. Retrieved from https://pure.psu.edu/en/publications/forecasting-hotel-occupancy-rates-with-time-series-models-an-empi

[18] YangLi, Q. T. (July 2016). Reasonable price recommendation on airbnb using multi-scale clustering. *In 2016 35th Chinese Control Conference (CCC)*, 7038–7041.