Session: 2025-2026 Dept: CSE Project No.:_________ Date of Evaluation:___________

SUMMER INTERNSHIP
B. TECH 2nd YEAR PASSING STUDENTS

AIRBNB NYC LISTINGS

Summer Internship Report

Submitted to

Sharda University

In partial fulfillment of the requirements of the award of the

Degree of Bachelor of Technology

in

Computer Science Engineering

by

Khushi Dua (2023339290)

Nishant Tomar (2023216390)

Under the mentorship of

Mr Nishant

Assistant Professor

Department of Computer Science Engineering
School of Engineering & Technology

Sharda University

Greater Noida

DECLARATION OF THE STUDENT

We hereby declare that the project entitled is an outcome of our own efforts under the guidance of Mr. Nishant . The project is submitted to the Sharda University for the partial fulfilment of the Bachelor of Technology Examination 2023-24.

We also declare that this project report has not been previously submitted to any other university.

Khushi Dua [Roll number: 2301010438 , System ID: 2023339290]

Nishant Tomar [Roll number:2301010609 , System ID: 2023216390]

Session: 2025-2026   Dept: CSE   Project No.:_________ Date of Evaluation:___________

CERTIFICATE

This is to inform that Khushi Dua and Nishant Tomar of Sharda University has successfully completed the project work titled AIRBNB NYC LISTINGS in partial fulfilment of the Bachelor of Technology Examination 2023-2024 by Sharda University.

This project report is the record of authentic work carried out by them during the period from 19 May 2027 to 27 June 2027.

Khushi Dua [Roll number: 2301010438 , System ID: 2023339290]

Nishant Tomar [Roll number:2301010609 , System ID: 2023216390]

Mr. Nishant

Assistant Prof.

DR. Sudeep Varshney

HOD,CSE,Sharda University

ABSTRACT

This project explores the application of machine learning techniques to predict Airbnb listing prices using structured and unstructured data. With the growing popularity of short-term rental platforms like Airbnb, accurately pricing a property has become essential for hosts to stay competitive and maximize revenue. The aim of this study is to develop a data-driven model that can estimate listing prices based on various features such as room type, location, number of reviews, availability, and guest sentiment extracted from textual reviews.

Three regression models were implemented and compared: Linear Regression, Random Forest, and XGBoost Regressor. After thorough preprocessing and feature engineering—including log transformations, one-hot encoding, and sentiment analysis using TextBlob—these models were trained and evaluated using standard metrics such as R² score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The XGBoost model outperformed the others, achieving the highest prediction accuracy and demonstrating strong capability in handling complex, non-linear relationships in the data.

Exploratory Data Analysis (EDA) and feature importance techniques revealed that factors like room type, location, and review sentiment have significant influence on pricing. Visualizations and performance metrics validated the model's reliability. However, the project also faced limitations such as reliance on static data, basic sentiment analysis tools, and geographic restrictions. These constraints highlight opportunities for future enhancements, including real-time data integration, web-based deployment, and advanced natural language processing methods.

Overall, this project showcases the effectiveness of combining machine learning with real-world datasets to solve practical business problems. It provides a strong foundation for developing intelligent pricing systems that could be beneficial for Airbnb hosts and property managers aiming to make data-driven decisions.

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. Apart from the efforts of myself, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project. We would like to show my greatest appreciation to **Mr. Nishant**. We can't say thank you enough for her/his tremendous support and help. We feel motivated and encouraged every time we attend her meeting. Without her encouragement and guidance this project would not have materialized. The guidance and support received from all the members who contributed and who are contributing to this project, was vital for the success of the project. We are grateful for their constant support and help. Besides, we would like to thank the authority of Sharda University for providing us with a good environment and facilities to complete this project. Finally, an honourable mention goes to our families and friends for their understandings and supports on us in completing this project. Without helps of the particular that mentioned above, we would face many difficulties while doing this.

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Problem Definition

Airbnb has rapidly transformed the global hospitality industry by offering an online platform that connects property owners with travelers seeking short-term rentals. The platform hosts millions of listings in cities and rural areas across the world. Each listing contains detailed information such as price, availability, location, reviews, amenities, and host details. With such a large and diverse dataset, Airbnb offers a rich opportunity for data-driven exploration and prediction. However, despite the abundance of information, both hosts and guests face several challenges in decision-making, primarily due to the complexity and variability of pricing, availability, and listing quality.

From a host's perspective, determining an optimal pricing strategy is far from straightforward. Prices are affected by several dynamic factors, including location, demand, availability, time of year, type of accommodation, number of reviews, and customer sentiment. Without structured analysis or predictive support, many hosts rely on personal intuition or guesswork, which can lead to inconsistent pricing and reduced booking rates. Similarly, guests often struggle to identify whether a listing offers good value. The lack of transparency and standardization in pricing can result in suboptimal booking decisions, impacting guest satisfaction and platform trust.

This project aims to address these issues by leveraging the power of data science and machine learning. The central problem we tackle is how to predict the nightly price of an Airbnb listing using available features. By analyzing Airbnb's public datasets—specifically listings, calendar, and reviews data—we aim to uncover the most significant factors influencing listing prices and build models that accurately estimate these prices. The project involves extensive exploratory data analysis (EDA) to identify trends and relationships among variables, followed by the application of supervised learning algorithms such as Linear Regression, Random Forest, and XGBoost.

In addition to numerical features, we explore the role of guest reviews in influencing listing success. Using Natural Language Processing (NLP), we extract sentiment from textual feedback and investigate whether positive or negative sentiments correlate with higher prices or better occupancy rates. The review dataset provides qualitative insights that complement the quantitative data, offering a more comprehensive understanding of guest satisfaction and its impact on pricing.

Furthermore, calendar data is analyzed to detect seasonal patterns in booking behavior and price fluctuations. This allows for temporal modeling, enabling hosts to adjust pricing strategies dynamically based on demand cycles, holidays, or local events. Insights gained from such analysis can lead to more competitive pricing and improved host earnings.

In summary, the core problem this project addresses is the lack of systematic, data-driven tools for price prediction and performance analysis in the Airbnb ecosystem. Through a combination of data cleaning, visualization, sentiment analysis, and predictive modeling, we aim to deliver actionable insights and build a solution that benefits both hosts and guests. The project not only enhances our technical understanding of machine learning applications but also contributes practical value to one of the world's largest peer-to-peer marketplaces.

1.2. Hardware Specifications

To efficiently process large datasets and perform machine learning operations, the following hardware specifications were required:

- **Processor (CPU):** Intel Core i5 (8th Gen or above) / AMD Ryzen 5 or equivalent — for fast data processing and smooth execution of ML models.
- **RAM:** Minimum 8 GB (16 GB recommended) — essential for handling large DataFrames and parallel computations.
- **Storage:** 256 GB Solid-State Drive (SSD) — ensures fast read/write speeds, especially when loading large datasets or saving models.
- **Graphics Card (GPU):** Optional. An NVIDIA GPU (GTX 1050 or better) is helpful but not mandatory, as the project does not involve deep learning.
- **Operating System:** Windows 10 / Ubuntu 20.04 LTS — both environments support required libraries and tools without compatibility issues.

This configuration was sufficient to support data exploration, feature engineering, model training, and visualization tasks throughout the project.

1.3. Software Specifications

The software stack was designed using widely adopted, open-source tools in Python's data science ecosystem. The following tools and libraries were essential:

- Programming Language: Python 3.10+ — selected for its simplicity, versatility, and strong ecosystem for data science and machine learning.
- Development Environment:
  - Jupyter Notebook — for interactive coding, documentation, and visualization.
  - Google Colab — used when additional memory or GPU acceleration was required.
- Libraries Used:
  - Data Handling: pandas, numpy
  - Visualization: matplotlib, seaborn, plotly
  - Machine Learning: scikit-learn, xgboost
  - Text Analysis (optional): nltk, TextBlob
  - Model Evaluation: r2_score, mean_squared_error, mean_absolute_error
  - Preprocessing & Utilities: sklearn.preprocessing, datetime, re

These tools enabled seamless data cleaning, modeling, evaluation, and presentation of results.

1.4. Motivation

In recent years, data has become one of the most valuable resources in the world. The ability to extract meaningful insights from data and turn them into actionable outcomes has become essential across industries. The hospitality industry, particularly platforms like Airbnb, generates an enormous volume of data related to property listings, customer behavior, pricing patterns, and reviews. This wealth of information holds significant potential for analytical exploration and predictive modeling.

The core motivation behind this project lies in the intersection of real-world problem-solving and academic learning. As students pursuing a degree in Computer Science with a specialization in Artificial Intelligence and Machine Learning, we were eager to apply theoretical concepts to a real-world dataset. Airbnb's publicly available data presented an ideal opportunity to analyze diverse features such as property type, location, price, availability, customer sentiment, and more. Through this project, we aimed to reinforce our understanding of data preprocessing, visualization, feature engineering, and machine learning models like Linear Regression, Random Forest, and XGBoost.

From a practical standpoint, this project addresses common challenges faced by both Airbnb hosts and guests. Many hosts struggle to determine an optimal pricing strategy for their listings, often relying on guesswork or comparing with similar properties. Inaccurate pricing may lead to reduced bookings or missed revenue opportunities. On the other hand, guests frequently face difficulties in evaluating whether a listing offers good value for money, especially when there are hundreds of listings in the same location. By building a price prediction model and identifying key pricing factors, we aim to provide data-driven support for better decision-making on both sides.

Another significant motivation was to work with a dataset that contains both structured data (like prices, dates, ratings) and unstructured data (like text reviews). This enabled us to explore natural language processing (NLP) techniques such as sentiment analysis, which added depth to the analysis and introduced us to more advanced AI applications.

We were also inspired by our curiosity. As users of Airbnb ourselves, we often wondered why similar-looking properties had drastically different prices or why some listings were booked months in advance while others remained available. This project gave us a chance to investigate these questions systematically and uncover patterns that are not immediately visible.

Finally, this project prepares us for industry-level roles in data science and machine learning. Employers value hands-on experience with real datasets, and working on a project like this helped us build a portfolio-worthy piece that demonstrates our technical skills, critical thinking, and problem-solving abilities.

In conclusion, this project was driven by a combination of academic curiosity, practical value, personal interest, and future career development. It allowed us to explore one of the largest and most dynamic datasets in the hospitality industry and apply modern data science tools to extract insights that matter.

1.5. Objectives

The primary objective of this project is to apply data analysis and machine learning techniques to understand and predict Airbnb listing prices based on various features. The project focuses on turning raw data into meaningful insights and building predictive models that can assist stakeholders in decision-making.

Key Objectives:

- **To perform exploratory data analysis (EDA):** Understand trends and patterns in the Airbnb dataset, including pricing, availability, and host-related attributes.
- **To identify key factors affecting listing prices:** Determine which features (e.g., location, room type, number of reviews, amenities) have the most significant impact on pricing.
- **To develop predictive models:** Build and evaluate machine learning models such as Linear Regression, Random Forest, and XGBoost to estimate the price of a listing.
- **To evaluate model performance:** Use performance metrics like R², Mean Squared Error (MSE), and Mean Absolute Error (MAE) to assess prediction accuracy.
- **To analyze review sentiment (optional):** Apply Natural Language Processing (NLP) to extract sentiment from user reviews and examine its influence on pricing and ratings.
- **To explore seasonal and temporal effects:** Analyze the calendar data to identify how pricing and availability change over time (e.g., peak seasons, weekends, holidays).

  These objectives were designed to align with academic goals, practical applications, and skill development in machine learning and data analytics.

1.6. Contributions

The project was a joint effort, with tasks divided based on interest and skillsets. Below are the key individual contributions:

- **Khushi Dua:**
  - Led data cleaning and preprocessing of the Airbnb datasets.
  - Conducted in-depth exploratory data analysis (EDA) and feature engineering.
  - Developed and evaluated machine learning models for price prediction.
  - Performed sentiment analysis on guest reviews using NLP techniques.
  - Created key visualizations and drafted the main sections of the final report.
- **Nishant Tomar:**
  - Assisted in data merging and supported EDA tasks.
  - Contributed to model development, testing, and parameter tuning.
  - Helped interpret model results and improve visual outputs.
  - Co-edited the report and contributed to preparing the presentation materials.
- **Joint Contributions:**
  - Collaborated on defining project goals and workflow.
  - Reviewed analysis results and ensured accuracy and completeness.

- Worked together on final formatting, documentation, and submission.

Summary

This project explored the application of data analysis and machine learning to a real-world dataset obtained from Airbnb, a widely used online rental platform. The goal was to analyze listings in a specific region and predict the nightly price of a property based on a variety of features such as location, amenities, host attributes, and review scores. By doing so, we aimed to uncover patterns in pricing and availability that could benefit both hosts and guests.

The project began with extensive data cleaning and preprocessing. The dataset included missing values, inconsistent formats, and categorical features that required transformation. Once cleaned, exploratory data analysis (EDA) was conducted to gain initial insights into the distribution of listings, pricing trends, correlations among variables, and regional variations. Visualizations such as heatmaps, bar plots, scatter plots, and availability calendars were used to illustrate trends and support deeper analysis.

Following EDA, several machine learning models were implemented to predict listing prices. These included Linear Regression, Random Forest, and XGBoost. Various performance metrics, including R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), were used to evaluate model accuracy. Among the models tested, ensemble-based algorithms like XGBoost showed the best results, providing more robust predictions by capturing non-linear relationships and interactions between features.

In addition to numerical features, the project also included basic Natural Language Processing (NLP) to analyze user reviews. Sentiment scores were extracted using text analysis libraries to understand the tone of customer feedback. These sentiment scores were then correlated with overall listing performance and price, offering another perspective on what influences customer behavior.

The calendar dataset allowed for temporal analysis, revealing seasonal pricing trends, booking availability, and fluctuations around weekends and holidays. This time-based insight can be valuable for hosts who wish to dynamically adjust pricing or manage availability for optimal performance.

The project also emphasized the importance of feature selection and data transformation. Certain features, such as room type, number of reviews, and availability over the year, were found to be strong predictors of price. Log transformations were applied to skewed data (like price and number of reviews) to improve model performance.

Overall, this project provided hands-on experience in working with real-world datasets and solving practical problems using machine learning. It combined technical skills such as data wrangling, visualization, and modeling with strategic thinking around pricing and user experience in the sharing economy. The insights generated from this analysis could help Airbnb hosts price their listings more effectively and assist guests in making better-informed booking decisions.

This end-to-end project highlighted the power of data in driving intelligent decisions and served as an excellent capstone to apply the AI and ML skills gained during our academic journey

2. LITERATURE SURVEY

2.1 Related Work Summary

Airbnb’s rise as a dominant platform in the short-term rental market has led to a growing interest in analyzing its data using machine learning and data science techniques. Several studies have been conducted in recent years to understand the factors that influence listing prices and to develop models that can accurately predict them.

Researchers have widely used datasets provided by **Inside Airbnb**, which offer detailed information about listings, calendars, hosts, and reviews. These datasets have served as the foundation for many machine learning projects focusing on price prediction, occupancy rate analysis, and review sentiment classification.

A number of studies have explored **regression techniques** for price prediction. Basic models like **Linear Regression** and **Ridge Regression** have been used due to their simplicity and interpretability. However, their performance often suffers due to the non-linear and complex nature of real-world data. To address this, more advanced models such as **Random Forest**, **Gradient Boosting Machines (GBM)**, and **XGBoost** have been widely adopted. For example, in a study by Gunter and Önder (2019), XGBoost was found to outperform other models in terms of accuracy when predicting listing prices in European cities.

Other researchers have emphasized the importance of **feature selection and engineering**. Features such as room type, number of guests, location, number of reviews, average review rating, and availability have been identified as significant predictors of price. Additionally, **location-based clustering** using latitude and longitude data has been applied to improve geographic specificity in pricing models.

An emerging area of interest is the use of **Natural Language Processing (NLP)** to extract insights from guest reviews. Studies have shown that **sentiment analysis** can be a valuable feature for price prediction, as listings with positive reviews tend to be priced higher. Most researchers have used tools like **TextBlob** or **VADER** for sentiment scoring, though some recent efforts have started to explore transformer-based models like **BERT** for better context understanding.

Several academic and industry-level projects have also proposed **dynamic pricing systems**, which adjust prices based on factors like booking time, seasonality, and special events. These systems often integrate time-series forecasting with machine learning to adapt prices dynamically rather than relying on static models.

Despite these advancements, there are still gaps in the literature. Many models lack real-time data integration, generalizability across cities, and user-centric deployment features. Furthermore, the explainability of complex models remains a challenge, which can impact trust and usability.

In summary, the related work highlights significant progress in Airbnb price prediction using machine learning, but also points to several opportunities for innovation—particularly in terms of deployment, interpretability, and user interaction.

Summary

The literature on Airbnb data analytics reveals a clear evolution from simple statistical models to sophisticated machine learning and natural language processing (NLP) approaches, all aimed at understanding and predicting listing performance. Early work predominantly applied **hedonic pricing models** and basic **linear regressions** to quantify the impact of property attributes—such as room type, number of bedrooms, and location—on nightly rates. These studies established foundational insights, confirming that factors like centrality of location, host status, and number of reviews significantly influence price.

As datasets grew in size and complexity, researchers turned to **ensemble methods**—notably Random Forests and Gradient Boosting Machines (GBMs). Empirical comparisons consistently demonstrate that these models outperform linear approaches by capturing non-linear interactions and higher-order effects. In particular, **XGBoost** has emerged as a preferred algorithm across multiple studies for its blend of performance and flexibility; it routinely achieves R^2 scores well above 0.70 on hold-out test sets.

Concurrently, the rise of publicly available review text inspired the integration of **sentiment analysis** into pricing models. By converting guest feedback into quantitative sentiment scores (using tools like TextBlob or VADER), researchers have shown that positive review sentiment correlates with higher prices and occupancy rates. More recent efforts employ transformer-based NLP models (e.g., BERT) to capture nuanced contextual sentiments, though these approaches remain less common due to higher computational cost.

Another key theme is **temporal dynamics**: calendar data have been exploited to reveal seasonality, weekend effects, and event-driven price surges. Time-series methods, sometimes combined with regression, enable dynamic pricing frameworks that adjust estimates based on date-specific demand patterns. However, many academic studies still rely on static snapshots rather than continuously updated data streams.

Despite these advances, notable gaps remain. Few models are truly **generalizable** across multiple cities or regions without retraining on local data. **Explainability** of complex models is also underexplored—while feature-importance scores offer some insight, most studies do not leverage formal interpretability tools (e.g., SHAP, LIME). Finally, **deployment and usability** considerations—such as web-based interfaces or API integration for real-time predictions—have received scant attention in academic work.

In conclusion, the literature affirms that combining structured listing features with unstructured review sentiment and temporal availability data leads to the most accurate price-prediction models. Yet there is ample opportunity to enhance generalizability, interpretability, and real-time application, which this project seeks to address by building a modular, extensible framework for Airbnb price prediction and analysis.

3.DESIGN AND IMPLEMENTATION

3.1. Methodology

The methodology adopted for this project follows a structured and practical approach to analyze Airbnb data and predict listing prices. It includes six major phases: data collection, preprocessing, exploratory data analysis (EDA), feature engineering, model development, and evaluation.

1. Data Collection

Data was sourced from the Inside Airbnb open dataset for a specific city. Three main CSV files were used: listings.csv, containing property details; calendar.csv, showing availability and pricing over time; and reviews.csv, which included guest feedback.

2. Data Preprocessing

The raw data required cleaning before analysis. Steps included:

- Removing irrelevant or duplicate columns.
- Handling missing values using imputation or deletion.
- Converting categorical variables into numerical form using encoding techniques.
- Cleaning and formatting price columns, dates, and textual fields.

3. Exploratory Data Analysis (EDA)

EDA helped uncover relationships between features and the target variable (price). Visual tools like histograms, heatmaps, and box plots were used to analyze distributions, outliers, and correlations. Insights gained here guided model selection and feature importance decisions.

4. Feature Engineering

Relevant features were selected and new ones were created, such as:

- Log transformation of skewed numerical features like price and reviews.
- Sentiment scores from review texts using basic Natural Language Processing (NLP).
- Calendar-based features like availability trends and seasonal patterns.

5. Model Development

Three models were used for price prediction:

- **Linear Regression** – as a baseline model.
- **Random Forest** – for handling non-linear patterns.
- **XGBoost** – for higher accuracy and improved performance.

Models were trained and validated using a split dataset (train-test split), and hyperparameters were tuned where applicable.

6. Model Evaluation

Model accuracy was measured using:

- **R^2 Score** – to assess model fit.
- **MAE** and **RMSE** – to evaluate prediction errors.

XGBoost performed the best, balancing prediction accuracy and generalization.

This end-to-end methodology ensured clean, structured data handling and effective application of machine learning models to solve the real-world problem of Airbnb price prediction.

3.2. Design

The design of this project is centered around a clear and modular pipeline to predict Airbnb listing prices and analyze related trends using structured and unstructured data. The entire system is divided into several functional components, each responsible for a specific phase of the workflow. This modular design improves flexibility, maintainability, and ease of testing.

System Components

- **Data Collection & Input:** The system begins by importing three main CSV files: listings, calendar, and reviews. These files are read using Python's pandas library and stored in DataFrames for processing.
- **Data Preprocessing:** Raw data often contains missing values, inconsistent formats, and non-numeric features. The preprocessing step involves cleaning the data, handling null values, converting prices and dates into usable formats, and encoding categorical variables (e.g., room type, neighborhood).
- **Feature Engineering:** Important transformations are applied to create or modify features. For example, log transformation is used on skewed fields like price and reviews. Sentiment scores are also derived from guest review texts using TextBlob.
- **Exploratory Data Analysis (EDA):** Using seaborn and matplotlib, visualizations are generated to understand data distribution and relationships among variables. Correlation matrices, scatter plots, and boxplots help identify patterns that influence pricing.
- **Model Development:** Machine learning models such as Linear Regression, Random Forest, and XGBoost are developed using scikit-learn. The data is split into training and test sets, and models are trained to predict listing prices based on selected features.
- **Evaluation & Output:** Model performance is evaluated using metrics like R^2 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Predicted vs. actual prices are visualized for clarity.

Tools and Technologies Used

- **Programming Language:** Python
- **Libraries:** pandas, numpy, scikit-learn, matplotlib, seaborn, xgboost, TextBlob
- **Platform:** Jupyter Notebook / Google Colab

This streamlined design ensures each component—from data cleaning to prediction—is logically organized and easy to adapt or extend in future work.

3.3. Implementation

The implementation phase focuses on executing the planned steps to analyze Airbnb data and build a working price prediction model. Python was used for all coding, with Jupyter Notebook and Google Colab serving as the development environments.

1. Data Import and Cleaning

The project used three key datasets: listings.csv, calendar.csv, and reviews.csv. These were loaded using the pandas library. Initial cleaning involved:

- Removing irrelevant columns and rows with excessive missing values.
- Converting columns like price to numerical format.
- Encoding categorical features such as room_type using one-hot encoding.

2. Exploratory Data Analysis (EDA)

With the cleaned dataset, various visualizations were created using matplotlib and seaborn. These helped us understand pricing trends, popular property types, and relationships between features. Insights from EDA were used to guide feature selection.

3. Feature Engineering

To enhance model accuracy, the following steps were applied:

- Log transformation of skewed features like price and number_of_reviews.
- Sentiment analysis of guest reviews using TextBlob to generate a sentiment score.
- Inclusion of calendar-based features (like availability across dates).

All final features were combined into a single dataset for modeling.

4. Model Training

Three machine learning models were implemented using scikit-learn:

- **Linear Regression** – as a baseline.
- **Random Forest Regressor** – to capture non-linear patterns.
- **XGBoost Regressor** – for best performance and accuracy.

The dataset was split into training and test sets (80/20). Models were trained, tested, and compared using the same evaluation metrics.

5. Model Evaluation

Each model was evaluated based on:

- **R² Score** – indicating model fit.
- **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** – showing prediction accuracy.

XGBoost performed the best, offering accurate and reliable price predictions.

This implementation successfully applied machine learning techniques to real-world data, resulting in a functional system that can predict Airbnb prices and support further data-driven insights.

Summary

The design and implementation phase of this project served as the backbone for transforming raw Airbnb data into a well-structured, interpretable, and predictive machine learning system. Each stage—from initial data exploration to final model evaluation—was carefully structured to ensure accuracy, efficiency, and practical insight.

The project began with importing and understanding three main datasets: listings.csv, calendar.csv, and reviews.csv. These datasets contained a mix of numerical, categorical, temporal, and textual data, offering a comprehensive view of Airbnb listings, their availability over time, and user experiences through reviews.

A modular approach was adopted to handle the complexity of the data. The system design divided the workflow into logical components: data preprocessing, exploratory data analysis (EDA), feature engineering, modeling, and evaluation. Each component was developed using Python, leveraging popular libraries such as pandas, seaborn, scikit-learn, and xgboost. Google Colab and Jupyter Notebook were used as development environments to enable efficient code execution, sharing, and visualization.

Preprocessing played a vital role in cleaning the data. Irrelevant features were dropped, missing values were handled, and formats were standardized. Categorical variables were encoded, and skewed numerical features like price and reviews were transformed using log scaling to stabilize variance.

EDA revealed important trends, such as the influence of room type, number of reviews, and location on pricing. These insights guided feature selection and model development. Feature engineering further improved model performance by adding new variables such as sentiment scores (extracted from guest reviews using TextBlob) and availability patterns from the calendar data.

For model training, three regression algorithms were used: Linear Regression, Random Forest, and XGBoost. These models were evaluated using standard metrics like R^2, MAE, and RMSE. XGBoost achieved the best results, capturing complex feature interactions and producing accurate predictions.

The implementation successfully combined structured data analysis and machine learning to solve a real-world problem—predicting Airbnb listing prices. In doing so, it not only demonstrated technical competence in data science but also offered valuable insights that could benefit both hosts and users on platforms like Airbnb.

Overall, the design and implementation stages ensured that the system was well-structured, scalable, and adaptable for future enhancements, such as dynamic pricing or real-time sentiment integration.

4. RESULT AND DISCUSSION

4.1 Result

The project implemented three machine learning models—**Linear Regression**, **Random Forest Regressor**, and **XGBoost Regressor**—to predict Airbnb listing prices. These models were evaluated using key performance metrics: **R^2 Score**, **Mean Absolute Error (MAE)**, and **Root Mean Squared Error (RMSE)**.

Model	**R^2 Score**	**MAE**	**RMSE**
Linear Regression	0.51	41.23	58.90
Random Forest	0.71	29.50	42.75
XGBoost	0.78	25.13	38.22

The **XGBoost Regressor** performed best, achieving the highest R^2 score and the lowest MAE and RMSE. This indicates strong predictive power and minimal error in estimating listing prices compared to the other models. Random Forest also gave fairly good results, outperforming the baseline Linear Regression model.

Key Influencing Features

Using feature importance scores from the tree-based models, the most significant predictors of price were:

- **Room type**
- **Location**
- **Number of reviews**
- **Review ratings**
- **Availability (calendar data)**
- **Sentiment score from guest reviews**

These features had a direct impact on the predicted price, with positive guest feedback and full-time availability often linked to higher pricing.

Visualization & Output

Plots comparing actual vs. predicted prices confirmed that XGBoost predictions were closely aligned with real values, showing a tight distribution around the ideal diagonal line. The error residuals were smaller and more consistent for XGBoost, further supporting its reliability.

In summary, the results confirmed that machine learning can effectively predict Airbnb prices. Models like XGBoost, which combine multiple weak learners and handle feature interactions well, are especially suited for this task. Including both structured data and text-based sentiment scores added depth to the model and improved accuracy.

4.2 Discussion

The results obtained from the implementation of multiple machine learning models provide valuable insights into the pricing dynamics of Airbnb listings. The project successfully demonstrated that data-driven techniques, when applied to real-world datasets, can be used effectively to predict the price of listings with a good degree of accuracy.

Among the models used, **XGBoost Regressor** provided the best performance, achieving an R^2 score of 0.78, which indicates that approximately 78% of the variance in the listing prices could be explained by the model. Additionally, it had the lowest error values in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), making it more reliable for accurate price prediction. This highlights the strength of ensemble-based models that leverage boosting techniques and feature optimization.

One of the key observations during this project was the influence of certain features in predicting price. **Room type**, **location**, **host response rate**, **number of reviews**, and **review ratings** consistently emerged as highly influential variables. Listings offering an entire home or apartment were priced significantly higher than shared or private rooms. Likewise, listings located in central or popular areas had a noticeable pricing advantage compared to suburban or less visited neighborhoods.

The inclusion of **review sentiment analysis** added a unique layer of value. By analyzing the text reviews left by previous guests using tools like TextBlob, the project was able to quantify the emotional tone of feedback and incorporate it as a predictive feature. Listings with highly positive sentiments tended to not only receive better ratings but also command higher prices. This emphasizes the importance of customer satisfaction and perception in influencing both popularity and pricing.

The **calendar dataset** enabled the team to uncover seasonal trends and variations in listing availability. For example, increased availability and pricing around holidays, weekends, or peak tourist seasons were clearly visible. This opens up potential applications for dynamic pricing algorithms that could help hosts adjust prices based on time and demand.

Comparing models also brought out the **limitations of simpler models**. While Linear Regression provided a baseline, its performance was significantly lower due to its inability to model complex, non-linear relationships. Random Forest performed better but was still outpaced by XGBoost in both prediction accuracy and consistency.

Furthermore, the project reinforced the importance of proper **data preprocessing and feature engineering**. Applying log transformations, handling missing values, encoding categories, and reducing data skewness significantly improved model performance. Each step of the pipeline played a crucial role in producing high-quality predictions.

In conclusion, this discussion illustrates that machine learning, supported by thoughtful data analysis and sentiment extraction, can effectively solve real-world problems like Airbnb price prediction. The outcomes not only validate the approach but also open doors to future enhancements, including real-time pricing tools and recommendation systems for both hosts and users.

Summary

The result and discussion phase of the project provided valuable insights into both the effectiveness of machine learning models for Airbnb price prediction and the importance of key listing features. By implementing multiple regression algorithms—namely Linear Regression, Random Forest Regressor, and XGBoost Regressor—the project aimed to identify the most accurate and interpretable method for predicting listing prices based on a wide range of features extracted from the Airbnb dataset.

From the results, it became clear that **XGBoost Regressor** outperformed the other models in terms of accuracy and reliability. It delivered the highest R² score (0.78) and the lowest MAE and RMSE values, indicating that it was better at capturing the underlying patterns in the data and minimizing prediction errors. The Random Forest model followed closely behind, while Linear Regression served as a helpful baseline but was less effective due to its limited ability to handle non-linear relationships.

The **importance of different listing features** was also highlighted through model interpretation tools. Factors such as room type, location, number of reviews, review ratings, and availability had a strong influence on the price. Listings with entire homes, prime locations, positive reviews, and high availability were consistently priced higher than others. These findings aligned with real-world expectations and confirmed the logical relevance of the data.

In addition, the project successfully integrated **sentiment analysis** using guest reviews. Text data was processed using TextBlob to extract polarity scores, which were then used as additional input features in the models. The presence of positive sentiment correlated with higher ratings and prices, reinforcing the importance of customer satisfaction and online reputation in the short-term rental market.

Visualizations created during the exploratory and evaluation phases provided a clearer understanding of feature distributions, correlations, and the alignment of predicted vs actual prices. These visuals helped validate the model's performance and demonstrated the impact of feature transformations and selection.

Finally, the discussion reinforced the importance of a well-structured data science workflow. Each phase—data cleaning, feature engineering, model selection, and evaluation—contributed to achieving strong and meaningful results. The practical insights drawn from the model outcomes have applications not just for academic purposes, but also for real-world usage, where Airbnb hosts could leverage such tools to set competitive prices and improve listing performance.

This summary concludes the result and discussion section, encapsulating the success of the modeling approach and setting the stage for final conclusions and future enhancements.

5.CONCLUSION

5.1. Conclusion

This project successfully demonstrated how machine learning and data analysis techniques can be applied to a real-world dataset like Airbnb to predict listing prices and uncover meaningful insights. By leveraging data from multiple sources—listings, calendars, and reviews—the project created a comprehensive model that not only predicts prices with accuracy but also explains the factors that influence them.

The core objective was to develop a predictive system that could help understand what makes certain listings more expensive than others. This goal was achieved through a structured approach involving data cleaning, preprocessing, exploratory data analysis, feature engineering, and model implementation. Among the models used, **XGBoost Regressor** emerged as the most accurate, outperforming both Linear Regression and Random Forest. Its ability to handle non-linear relationships, outliers, and complex feature interactions made it the most suitable model for this use case.

One of the key takeaways from the project is the **importance of feature selection**. Attributes such as room type, location, review scores, number of reviews, and listing availability were shown to have the greatest influence on pricing. This knowledge is beneficial not only for model accuracy but also for Airbnb hosts who may want to optimize their listings based on these factors.

Additionally, incorporating **textual data analysis** through sentiment scoring of reviews added a new dimension to the predictive capability. Listings with positive review sentiment generally had higher ratings and were able to command better prices. This suggests that hosts should focus not only on physical attributes of their property but also on guest experience and satisfaction.

Visualizations created during the exploratory and evaluation phases supported the findings and made it easier to interpret model performance and data trends. These visuals also reinforced the real-world applicability of the solution, making the system easy to understand even for non-technical stakeholders.

Overall, this project showcased how a combination of structured and unstructured data, when analyzed through modern machine learning techniques, can result in actionable insights. It also demonstrated the effectiveness of tools like Python, pandas, scikit-learn, XGBoost, and TextBlob in building robust predictive systems.

In conclusion, the project met its objectives effectively, offering a reliable, interpretable, and scalable solution for Airbnb price prediction. The system has potential for future expansion, including features like real-time prediction, automated pricing recommendations, or integration with external data like tourism trends and public events.

5.2. Limitations

While the project achieved its goal of predicting Airbnb listing prices using machine learning models, there are several limitations that should be acknowledged. These limitations help identify areas where the system could be improved in future iterations.

1. Geographic and Dataset Limitations

The data used was specific to a single city, meaning the model is limited in scope. Factors influencing Airbnb pricing can vary greatly across different regions, so applying this model elsewhere without retraining could reduce its accuracy.

2. Static Data Usage

The project used a one-time snapshot of Airbnb data. As a result, it does not account for ongoing market changes, seasonality, or economic factors. Pricing trends during holidays, events, or travel booms were not dynamically captured.

3. Incomplete or Noisy Data

Several fields in the dataset contained missing or inconsistent values, especially user-generated content like reviews and descriptions. Though preprocessing helped clean the data, residual noise may have affected model performance.

4. Basic Sentiment Analysis

Review sentiment was calculated using **TextBlob**, which provides only basic analysis. It doesn't account for nuances, sarcasm, or multilingual reviews. More advanced NLP tools like transformer-based models could provide deeper insights but were not included due to time and resource constraints.

5. Model Complexity

While **XGBoost** gave the best performance, it is not easily interpretable. Its internal decision-making is complex, which may make it harder for non-technical users to understand how the predictions are generated.

6. No Live Integration

The current system runs in an offline environment (Jupyter Notebook/Colab) and does not support real-time prediction or web-based deployment. Integration with live Airbnb platforms or automation pipelines would require additional development.

In conclusion, while the project effectively demonstrates a machine learning approach to Airbnb price prediction, addressing these limitations could improve its robustness, flexibility, and real-world applicability in future versions.

5.3. Future Scope

While this project successfully demonstrated the use of machine learning to predict Airbnb listing prices, there are several potential directions for future development that could enhance its accuracy, flexibility, and real-world applicability.

1. Real-Time Data Integration

Currently, the model is based on static historical data. In the future, integrating real-time or regularly updated data through Airbnb APIs or web scraping could allow dynamic predictions that reflect market fluctuations, seasonal trends, or local events.

2. Web-Based Interface

The current implementation is notebook-based, making it less accessible to general users. A future enhancement could involve developing a web application using frameworks like **Streamlit**, **Flask**, or **Django**, where users can enter property details and receive instant price estimates.

3. Enhanced Text Analysis

Guest reviews were analyzed using basic sentiment tools like TextBlob. Future versions can include advanced NLP models like **BERT** or **spaCy**, which would offer deeper insights into review sentiment and guest satisfaction, contributing to more accurate predictions.

4. Multicity or Global Expansion

At present, the dataset is limited to a single city. Expanding the model to handle multiple cities or countries would improve its versatility. It could either be a generalized global model or a set of city-specific models depending on available data and regional factors.

5. Pricing Recommendations

Beyond prediction, the model could evolve into a **dynamic pricing assistant**, suggesting optimal listing prices based on factors like demand, availability, and guest preferences. Incorporating variables such as holidays, weekends, or event dates could enhance this capability.

6. Better Explainability

Future improvements could include the use of explainable AI techniques such as **SHAP** or **LIME** to help users understand how specific features influence pricing. This would make the model more transparent and trustworthy for hosts and users alike.

In conclusion, the current system provides a strong foundation for Airbnb price prediction, and with further enhancements, it has the potential to become a robust, real-time, and user-friendly tool for data-driven decision-making in the short-term rental market.

SUMMARY

This project aimed to predict Airbnb listing prices using machine learning models by analyzing real-world data from various sources including listings, calendars, and guest reviews. Through systematic steps involving data preprocessing, feature engineering, exploratory data analysis, and model implementation, a reliable predictive system was developed. The project demonstrated that machine learning, especially advanced models like **XGBoost**, can effectively capture patterns in the data and provide accurate price predictions.

The initial phase focused on understanding the data, handling missing values, and transforming categorical and numerical variables into a form suitable for modeling. Sentiment analysis was also applied to guest reviews to include a qualitative measure of customer experience. These features were used to train and test three machine learning models: **Linear Regression**, **Random Forest**, and **XGBoost Regressor**. Among them, XGBoost showed the best results in terms of R^2 score and prediction error, indicating it was the most effective model for this use case.

The analysis revealed that several factors—such as **room type**, **location**, **number of reviews**, **review ratings**, and **availability**—play a significant role in determining listing prices. Listings with better guest feedback and high availability were more likely to be priced higher, reflecting real market behavior. The use of sentiment scores from reviews also added depth to the model, highlighting the impact of customer satisfaction on pricing.

Despite its success, the project faced certain limitations, including reliance on static data, geographic restriction to one city, and the use of basic text analysis tools. These constraints provide opportunities for future enhancement, such as integrating real-time data, deploying the system as a web application, and using more advanced natural language processing techniques.

The future scope also includes expanding the model's applicability to multiple cities or countries, adding dynamic pricing recommendations, and improving model transparency using explainable AI techniques.

In summary, this project established a solid foundation for intelligent price prediction in the Airbnb ecosystem. It combined data science techniques with real-world datasets to develop a system that is both functional and insightful. With further development, this model can evolve into a practical tool for Airbnb hosts and property managers seeking data-driven pricing strategies.

6. REFERENCES

i. Airbnb. (n.d.). *Inside Airbnb: Adding Data to the Debate*. Retrieved from http://insideairbnb.com

ii. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

iii. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

iv. McKinney, W. (2010). *Data structures for statistical computing in Python*. Proceedings of the 9th Python in Science Conference, 51–56.

v. Loper, E., & Bird, S. (2002). *NLTK: The Natural Language Toolkit*. Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics.

vi. TextBlob. (n.d.). *Simplified Text Processing*. Retrieved from https://textblob.readthedocs.io

vii. Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90–95.

viii. Waskom, M., et al. (2021). *Seaborn: Statistical Data Visualization*. Journal of Open Source Software, 6(60), 3021.

ix. Jupyter Project. (n.d.). *Jupyter Notebook*. Retrieved from https://jupyter.org

x. Google Colab. (n.d.). *Collaboratory by Google*. Retrieved from https://colab.research.google.com