AirBnB Rent prediction

Amir Khuzin Zukhra Khadeeva Andreas Coclet
{a.khuzin, z.khadeeva, a.andreas}@innopolis.university

1 Introduction

Airbnb is a two-sided marketplace that sought to match people that owned real estate properties with people interested in renting short-term lodging. The service provides travelers with unique accommodations and hosts with lucrative income opportunities. But there's one problem and it is the inconsistent pricing. This can lead to accommodations being overpriced or under priced. Addressing this issue is crucial for optimizing revenue and improving occupancy rates.

In our project we aim to identify the key factors influencing Airbnb pricing by analyzing and extensive dataset of global accommodations. By understanding what drives pricing variability and how it impacts revenue, we want to develop a model that assists hosts in pricing their properties competitively and realistically. This will benefit both hosts and guests in this system.

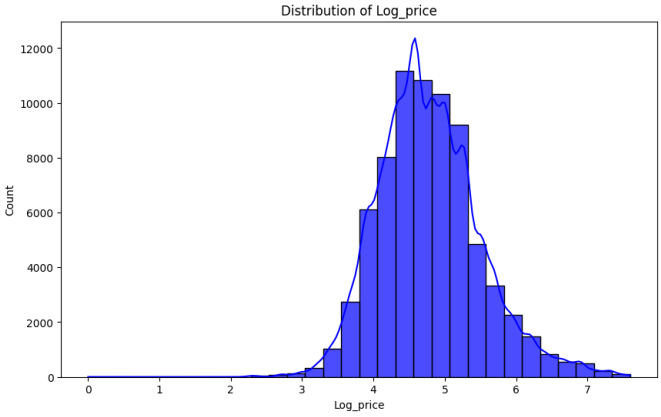

Figure 1. Distribution of Log Price. This histogram shows the distribution of log-transformed prices, indicating a wide range of pricing across different listings.

1.1 Motivation

Inconsistent pricing not only affects individual hosts but also impacts overall perception and reliability of Airbnb as a platform. A data-driven approach to pricing can provide significant benefits, including: increased revenue, improved occupancy rates, better guest experience etc.

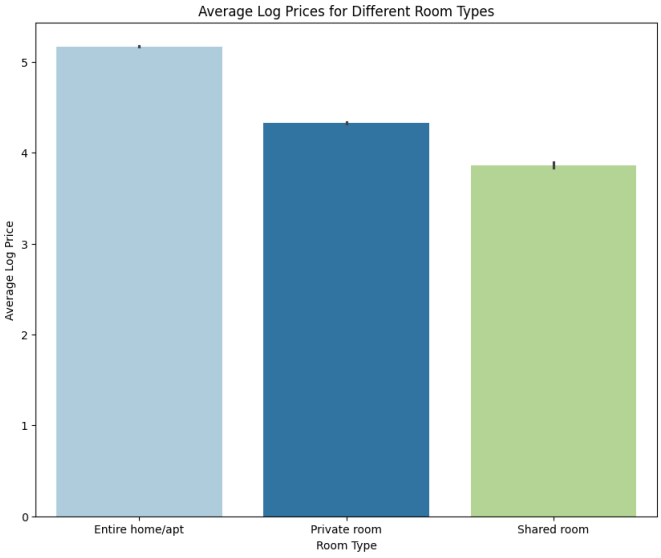

Figure 2. Average Log Prices for Different Property Types. This bar plot highlights the average log prices for various property types, demonstrating significant differences in pricing.

1.2 Data overview

Our dataset comprises 74,111 rows and 29 columns, including 11 numeric features and several categorical features such as property type, room type, and cancellation policy.

The dataset contains missing values across several columns, which were handled through imputation techniques.

The average log_price varies significantly across different property types and room types, indicating the importance of these features in determining the listing price.

For instance, properties with more accommodations, bedrooms, and bathrooms generally have higher prices, with correlations of 0.57, 0.47, and 0.35 respectively with the target variable log_price.

The popularity of property and room types varies significantly by city, affecting pricing strategies. For example, entire homes/apartments are more popular in certain cities, commanding higher prices on average.

By leveraging these insights, our project aims to develop a robust pricing model that can accurately predict the optimal price for Airbnb listings, ensuring better revenue management for hosts and improved satisfaction for guests.

2 Business and Data understanding

We address the inconsistency in pricing Airbnb accommodations, which leads to listings that are either overpriced and remain vacant or underpriced, resulting in potential

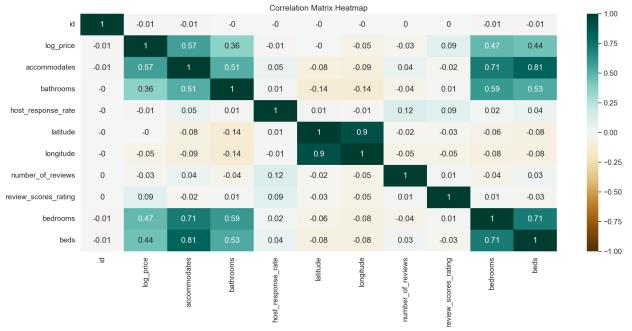

Figure 3. Correlation Matrix Heatmap. The heatmap visualizes the correlations between different numerical features and the target variable, log_price. Higher correlations with accommodates, bedrooms, and beds indicate their strong influence on pricing.

revenue loss for hosts. The current manual and often intuitive approach to pricing does not effectively account for the multitude of factors that influence accommodation prices.

Our objective is to develop a data-driven model that identifies key factors influencing Airbnb pricing, it aims to optimize revenue and improve occupancy rates for hosts. To achieve this, we utilize an extensive dataset of global Airbnb listings, encompassing various features such as property type, location, amenities, and host characteristics.

By analyzing these factors, we seek to uncover patterns and correlations that impact pricing. This report provides a comprehensive overview of our approach, including data preprocessing, feature engineering, and model development, to offer actionable insights for hosts. Neglecting to align the pricing strategy with data-driven insights can result in missed opportunities and financial inefficiencies, making this project crucial for enhancing revenue optimization strategies in the short-term rental market.

Our analysis relies on key statistics and data visualizations to support our findings and recommendations.

2.1 Terminology

2.1.1 Business Terminology.

- **Airbnb**: An online marketplace that connects people who want to rent out their homes with those looking for accommodations.
- **Listing**: An individual rental property available on Airbnb.
- **Occupancy Rate**: The percentage of available rental units that are occupied at a given time.
- **Revenue Optimization**: The process of adjusting prices to maximize income from rental properties.
- **Short-term Rental Market**: A segment of the rental market that offers accommodations for a short duration, typically less than a month.
- **Dynamic Pricing**: A pricing strategy where prices are adjusted based on real-time supply and demand conditions.
- **Booking Window**: The time frame between when a guest books a rental and the start of their stay.
- **Cancellation Policy**: The rules set by the host regarding the conditions under which a guest can cancel a reservation and receive a refund.
- **Review Rating**: A score given by guests based on their stay, reflecting the quality of the property and the host's service.
- **Cleaning Fee**: An additional charge imposed by the host for cleaning the property after a guest's stay.
- **Amenities**: Features provided by the rental property, such as Wi-Fi, parking, or a swimming pool.
- **Property Type**: The classification of rental properties, such as apartment, house, or villa.
- **Location**: The geographical area where the rental property is situated, impacting its attractiveness and price.
- **Seasonality**: Fluctuations in demand and prices due to seasonal factors, such as holidays or weather changes.
- **Competitive Analysis**: The process of evaluating similar listings in the area to set competitive prices.
- **Market Trends**: Changes and patterns in the short-term rental market that influence demand and pricing.
- **Minimum Stay Requirement**: The shortest duration a guest can book a property, as set by the host.
- **Check-in/Check-out Policy**: The rules and timings related to when guests can arrive and depart from the rental property.

2.1.2 ML Terminology.

- **Regression**: A supervised learning technique used to predict continuous values, such as the price of an Airbnb listing based on its features.
- **Classification**: A supervised learning technique used to predict categorical outcomes, such as whether a listing will be booked or not.
- **Feature Engineering**: The process of selecting, modifying, and creating new variables (features) that enhance the performance of ML models.
- **Training Data**: The subset of data used to train ML models, containing input-output pairs.
- **Validation Data**: A subset of data used to tune model parameters and prevent overfitting by evaluating model performance.
- **Test Data**: A subset of data used to assess the final performance of the model after training and validation.
- **Overfitting**: A modeling error that occurs when the ML model captures noise in the training data, performing well on training data but poorly on new, unseen data.

- **Underfitting**: A modeling error that occurs when the ML model is too simple to capture the underlying patterns in the data, leading to poor performance on both training and test data.
- **Cross-Validation**: A technique for assessing how the results of a statistical analysis will generalize to an independent dataset, typically by partitioning the data into subsets and training/testing the model multiple times.
- **Hyperparameters**: Settings that define the model architecture and learning process, such as learning rate or number of trees in a random forest, which need to be specified before training.
- **Model Evaluation Metrics**: Measures used to evaluate the performance of ML models, such as Mean Absolute Error (MAE) for regression or Accuracy for classification.
- **Feature Importance**: A technique to determine the significance of individual features in predicting the target variable.
- **Normalization**: A preprocessing step that scales features to a standard range, often 0 to 1, to ensure equal contribution to the model.
- **ROC Curve**: A graphical plot that illustrates the diagnostic ability of a binary classifier system, plotting the true positive rate against the false positive rate.
- **F1 Score**: A measure of a test's accuracy that considers both precision and recall, providing a single metric that balances both aspects.
- **Data Preprocessing**: The process of cleaning and preparing raw data for ML, involving steps like handling missing values, encoding categorical variables, and normalizing data.
- **Exploratory Data Analysis (EDA)**: An approach to analyzing data sets to summarize their main characteristics, often using visual methods, before applying more formal modeling techniques.

2.2 Scope of the ML Project

2.2.1 Background. Airbnb is a leading online marketplace that connects people looking to rent out their properties with those seeking short-term accommodations. Currently, Airbnb faces challenges in optimizing pricing strategies for different listings, considering varying demand across locations and seasons. The organization has collected substantial data on past bookings, pricing, customer reviews, and property features, which can be leveraged to enhance pricing strategies and improve occupancy rates.

2.2.2 Business Problem. The primary business problem is to optimize the pricing of Airbnb listings to maximize revenue while maintaining high occupancy rates. This involves dynamically adjusting prices based on various factors, such as location, seasonality, property features, and market trends, to attract more bookings and enhance overall profitability.

2.2.3 Business Objectives. The main objective of the project is to increase revenue for Airbnb hosts by developing an intelligent pricing strategy that optimizes rental prices in real-time. Additional related business questions include:

How does location influence booking rates and optimal pricing?

What impact do amenities and property features have on the booking decisions of guests?

How do seasonal trends affect occupancy rates and pricing strategies?

Can dynamic pricing reduce the cancellation rates and increase the length of stays?

2.2.4 ML Objectives. The ML goal for this project is to develop a predictive model that can determine the optimal pricing for Airbnb listings. The model will use historical booking data, property characteristics, location details, and market trends to forecast demand and adjust prices accordingly. Specific ML objectives include:

Predicting the booking likelihood for a given property at different price points.

Identifying the key features that significantly influence booking rates and prices.

Developing a dynamic pricing algorithm that can be integrated into the Airbnb platform to automate price adjustments in real-time.

Evaluating the model's performance using metrics such as Mean Absolute Error (MAE) and comparing it against current pricing strategies to ensure improvements in revenue and occupancy rates.

2.3 Success Criteria

2.3.1 Business Success Criteria. The primary business success criterion for the ML application is to increase the overall revenue for Airbnb hosts through optimized pricing strategies. This success can be measured by a targeted increase in average monthly revenue per listing by at least 15% within six months of implementing the ML-driven pricing model. Additionally, maintaining or improving the occupancy rate to at least 75% across listings is another critical business success metric. These criteria ensure that the pricing strategy not only boosts revenue but also keeps properties consistently booked.

2.3.2 ML Success Criteria. To meet the business objectives, the ML success criteria must include specific performance metrics for the predictive pricing model. The minimum acceptable performance for the model is defined as follows:

Prediction Accuracy: The model should achieve a Mean Absolute Error (MAE) of less than $10 per night in predicted optimal prices compared to actual booking prices. Booking

Likelihood Prediction: The model should accurately predict booking likelihood with an accuracy of at least 85%. Model Robustness: The model should maintain consistent performance across different locations and seasons, as measured by a variance in MAE of less than 5% across different subsets of data. These criteria ensure that the model provides reliable and actionable pricing recommendations that align with business goals.

2.3.3 Economic Success Criteria. The economic success of the ML project will be evaluated through key performance indicators (KPIs) that measure the financial impact of the optimized pricing strategy. The main KPIs include:

Revenue Increase per Listing: A measurable increase in revenue per listing, targeted at a minimum of $500 per month, after the implementation of the ML model. Cost Efficiency: Reduction in the manual effort required for price adjustments, estimated to save at least 10 hours of work per host per month, which translates to cost savings on operational expenses. Return on Investment (ROI): The project should achieve a positive ROI within the first year, with a target ROI of at least 150%. By setting these economic success criteria, the project aligns the ML application with tangible financial benefits, ensuring that the investment in machine learning technology translates into significant economic gains for Airbnb and its hosts.

2.4 Data collection

2.4.1 Data collection report. The dataset used for this analysis was sourced from Kaggle, specifically the Airbnb dataset which includes comprehensive data on Airbnb listings. Below is a detailed description of the data collection aspects: Format: The dataset is provided in CSV format, a common and versatile format for tabular data.

Data Types:

- **Numerical Features:**
  - *id*: Unique identifier for each listing
  - *price*: Continuous numerical values
  - *log_price*: Continuous numerical values
  - *accommodates*: Discrete numerical values
  - *bathrooms*: Discrete numerical values
  - *bedrooms*: Discrete numerical values
  - *beds*: Discrete numerical values
  - *number_of_reviews*: Discrete numerical values
  - *review_scores_rating*: Continuous numerical values between 0 and 100
  - *latitude*: Continuous numerical values
  - *longitude*: Continuous numerical values
- **Categorical Features:**
  - *property_type*: (Apartment, House, Other)
  - *room_type*: (Entire home/apt, Private room, Other)
  - *bed_type*: (Real Bed, Futon, Other)
  - *cancellation_policy*: (strict, flexible, Other)
  - *city*: (NYC, LA, Other)
  - *instant_bookable*: (true, false)
  - *host_has_profile_pic*: (true, false)
  - *host_identity_verified*: (true, false)
  - *cleaning_fee*: (true, false)
- **Text Features:**
  - *zipcode*: ZIP code of the listing
  - *host_response_rate*: Host response rate (as percentages, e.g., '100%')
  - *amenities*: Free-form text describing amenities
  - *description*: Free-form text describing the property
  - *name*: Name of the listing
  - *neighbourhood*: Name of the neighbourhood
  - *thumbnail_url*: URL of the listing's thumbnail image
- **Datetime Features:**
  - *first_review*: Date of the first review
  - *last_review*: Date of the last review
  - *host_since*: Date the host started

Quantity of Data: The dataset comprises 74,111 rows and 29 columns.

2.4.2 Data version control report. Initial raw dataset downloaded from Kaggle and then cleaned and prepossessed and restored in the same filename.

Changes in Data:

- Handled missing values using imputation strategies.
- Encoded categorical features using one-hot encoding and label encoding.
- Standardized numerical features.
- Applied TF-IDF vectorization to text features.
- Extracted and transformed date features.

Backup and Storage:

- Each version of the dataset is backed up and stored in a version-controlled repository.
- Historical data versions are archived with appropriate version tags.

Access and Modification Control:

- Access to the data is restricted to authorized personnel.
- Modification of the data is tracked and logged using version control systems (e.g., Git).
- Any changes to the data are documented in a change log, specifying the nature of the change, the reason for the change, and the individual responsible.

This structured approach ensures that the dataset remains consistent and any modifications are transparent and reproducible.

2.5 Data Quality Verification

In this section, we assess the quality of the acquired Airbnb dataset. We define expectations for the data and validate its integrity and consistency through exploratory analysis and validation checks.

2.5.1 Data Description. Here's an overview of the dataset:

- **Format**: CSV
- **Quantity**:
  - 74,111 records
  - 29 columns
- **Key Features**: Includes numerical, categorical, text, and datetime data types

2.5.2 Data Exploration. Upon initial exploration, we identified several key points:

- **Numerical Features**:
  - Includes 11 numeric columns
  - Significant presence of missing values in some columns
- **Categorical Features**:
  - Important features include property_type, room_type, bed_type, cancellation_policy, city, among others
  - Presence of categorical values needing encoding for analysis
- **Text Features**:
  - Includes free-form text such as description and amenities
  - Requires preprocessing for meaningful analysis
- **Datetime Features**:
  - Includes dates such as first_review, last_review, and host_since
  - Requires extraction and transformation for temporal analysis

2.5.3 Data Quality Assessment.

Missing Values Analysis. The dataset exhibits missing values across various columns:

- **Total Missing Values**: Calculated to be 3,524 missing values in total
- **Percentage of Missing Values**: Average percentage of missing values across columns is 6.23%

Unique Values Analysis. Each feature's uniqueness provides insight into dataset diversity:

- **Number of Unique Values per Feature**: Varied number of unique values across different features
- **Percentage of Unique Values per Feature**: Reflects diversity and distribution within each feature category

Distribution Analysis of Numerical Features. Exploring distributions helps understand the data's spread:

- **Histograms**: Visualized distributions for log_price, accommodates, bathrooms, bedrooms, and other relevant features
- Highlighted central tendencies and potential outliers

Distribution Analysis of Categorical Features. Understanding categorical distributions aids in segmentation:

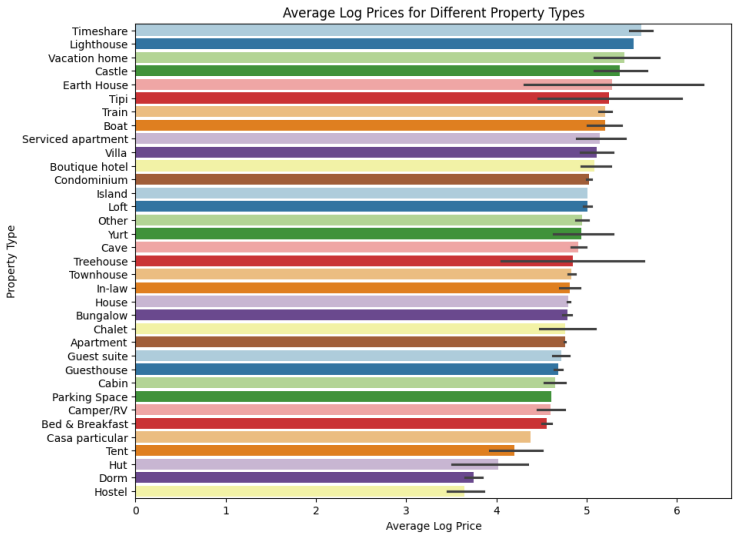

Figure 4. Plot for Average Log Prices for Different Property Types

- **Count Plots**: Plotted distributions for room_type, bed_type, cancellation_policy, city, and others
- Identified popular categories and their prevalence across the dataset

Correlation Analysis. Examining correlations between features and log_price:

- **Heatmap Visualization**: Illustrated correlations using a heatmap to highlight relationships
- Identified significant correlations between features like accommodates, bedrooms, and bathrooms with log_price

2.5.4 Data Features Transformation. Preparing data for further analysis involved:

- **Handling Missing Values**:
  - Imputed missing values using appropriate strategies for numerical, categorical, and text features
- **Feature Encoding**:
  - Utilized techniques like one-hot encoding, label encoding, and frequency encoding for categorical features
- **Standardization**:
  - Applied standard scaling to numerical features for normalization
- **Text Processing**:
  - Employed TF-IDF vectorization for textual data like description and amenities
- **Date Transformation**:
  - Extracted temporal components (year, month, day) from date features like first_review, last_review, and host_since

2.5.5 Data Quality Validation Using Great Expectations. To validate the dataset's quality and consistency, we utilized Great Expectations:

- **Expectation Suites**:
  - Defined expectations such as completeness, uniqueness, and validity for critical features like id, log_price, accommodates, bathrooms, bedrooms, and others
- **Validation Results**:
  - Confirmed expectations using Great Expectations to ensure data integrity and reliability for downstream analysis

This section comprehensively evaluates the Airbnb dataset, ensuring it meets quality standards for reliable analysis and modeling.

2.6 Data Quality Verification

Data Requirements
Key considerations include:

- **Expected Feature Values:** Define ranges for continuous features (e.g., price, number of reviews) or lists for discrete features (e.g., property types, room types).
- **Data Format:** Ensure data formats adhere to specified standards, such as date formats or textual content.
- **Maximum Number of Missing Values:** Define thresholds for acceptable missing data across features, guiding data cleaning and imputation processes.

2.6.1 Data Quality Verification Report. The data quality verification process evaluates the integrity and reliability of the dataset through various analyses:

- **Completeness:** Missing data statistics per column were examined, revealing areas requiring imputation or further investigation.
- **Accuracy:** Quality checks revealed discrepancies that impact data reliability, particularly in numerical and categorical features.
- **Missing Values:** Strategies like median imputation and default values were used to address missing data systematically.

The dataset, sourced from Airbnb listings, consists of 74,111 rows and 29 columns, predominantly numeric, with categorical, text, and date features also present. Initial exploration revealed significant missing data across several columns, necessitating robust handling strategies.

2.6.2 Statistical Summary and Data Exploration. Basic statistics and exploratory data analysis (EDA) provided insights into the dataset's composition:

- **Numerical Features:** Eleven numeric features were identified, including log_price, accommodates, bathrooms, bedrooms, beds, number_of_reviews, and review_scores_rating.
- **Categorical Features:** Identified categorical features included property_type, room_type, bed_type, cancellation_policy, city, instant_bookable, host_has_profile_pic, host_identity_verified, and cleaning_fee.
- **Text Features:** Included amenities, description, name, neighbourhood, and thumbnail_url, requiring specialized preprocessing such as tokenization and TF-IDF vectorization.
- **Datetime Features:** Utilized first_review, last_review, and host_since, requiring conversion and imputation strategies.

2.6.3 Data Transformation and Preprocessing. Data preprocessing aimed to enhance data quality and prepare features for modeling:

- **Handling Missing Values:** Implemented imputation strategies (e.g., median for numerical features, default values for categorical and text features) to address missing data.
- **Feature Engineering:** Applied transformations like one-hot encoding (for categorical features) and TF-IDF vectorization (for textual features) to enhance model performance.
- **Normalization and Standardization:** Ensured consistency and comparability of features using techniques such asstandard scaling for numeric features.

2.7 Project Feasibility

Here's a summary of our findings and considerations:

2.7.1 Data Overview. Dataset Size: The dataset consists of 74,111 rows and 29 columns, with 11 columns identified as numeric.

Missing Values: Several columns contain missing values, necessitating imputation strategies for data completeness.

2.7.2 Data Exploration. Statistical Summary: Descriptive statistics provided insights into the distribution of numeric and categorical features.

Unique Values: Each feature's uniqueness was analyzed to understand data diversity and potential categorical transformations.

2.7.3 Data Distribution Analysis. Numerical Features: Histograms and box plots were used to visualize the distribution and identify outliers in features such as log_price, accommodates, and bathrooms.

Categorical Features: Count plots illustrated the distribution of categorical variables like room_type and cancellation_policy across different cities.

2.7.4 Correlation Analysis. Correlation Matrix: Pearson correlation coefficients were computed to reveal relationships between numeric features and the target variable (log_price).

2.7.5 Data Quality Assessment. Completeness and Validity Checks:

Great Expectations library was employed to validate data completeness (e.g., id not null) and attribute validity (e.g., log_price within expected range).

2.8 Project Plan

Based on the insights gathered from the feasibility analysis, we propose the following project plan:

2.8.1 Data Cleaning and Preprocessing. Handling Missing Values:

Impute missing values using appropriate methods for different data types (e.g., median imputation for numerical, mode for categorical).

Feature Engineering: Transform features using techniques such as one-hot encoding for categorical variables and standardization for numerical ones.

2.8.2 Exploratory Data Analysis (EDA). Data Visualization:

Further explore data distribution and relationships through advanced visualization techniques (e.g., scatter plots, box plots) to refine understanding and identify patterns.

2.8.3 Model Development. Feature Selection:

Use correlation analysis and domain knowledge to select relevant features that influence `log_price`.

Model Selection: Implement machine learning models (e.g., MLPRegressor) to predict `log_price`, leveraging cleaned and transformed data.

2.9 Model Evaluation and Optimization

Performance Metrics: Evaluate model performance using metrics like Mean Squared Error (MSE) on both training and testing datasets.

Hyperparameter Tuning: Optimize model parameters to improve predictive accuracy and generalizability.

2.9.1 Project Validation. Cross-Validation:

Validate model robustness using techniques like k-fold cross-validation to ensure reliability of results.

Result Interpretation: Interpret model outputs and insights gained to derive actionable recommendations for stakeholders.

2.9.2 Documentation and Reporting. Documentation:

Maintain clear and concise documentation throughout the project lifecycle, including code comments and model documentation.

Reporting: Prepare a final project report summarizing methodology, findings, and recommendations for stakeholders and project sponsors.

By following this structured project plan, we aim to deliver a robust and reliable predictive model for Airbnb listing prices, enabling stakeholders to make informed decisions based on data-driven insights.

2.10 Conclusion of Phase 1

Phase 1 has laid a solid foundation by identifying key challenges and insights, setting the stage for subsequent phases to refine and implement our predictive modeling approach effectively

Data Preparation

3 Select Data

3.0.1 . sectionData Sources

The data for this project comes from the following sources:

1. **Airbnb Listings Dataset**: Contains information about Airbnb listings, including features like location, price, number of rooms, availability, and host details.
2. **Supplementary Datasets**: Additional datasets such as neighborhood crime rates, local amenities, and weather conditions may also be considered to enrich the prediction model.

3.0.2 Data Inclusion Criteria.

The primary criteria for including data are relevance to the prediction goals, data quality, and technical feasibility. Specifically, we focus on the following:

3.0.3 Relevance to Prediction Goals.

- Attributes that are directly related to rental prices and can serve as predictors.
- Features that provide context about the listing, such as location, property type, and amenities.

3.0.4 Data Quality.

- Completeness: Attributes with minimal missing values.
- Consistency: Attributes that maintain uniform formats and units.
- Accuracy: Attributes that accurately reflect the reality of the listings.

3.0.5 Technical Constraints.

- Data volume: Managing the size of the dataset to fit into memory and processing capabilities.
- Data types: Ensuring that data types are compatible with the machine learning models used.

3.0.6 Included Data.

Based on the criteria above, the following attributes are selected for inclusion:

3.0.7 Listing Information.

- `listing_id`: Unique identifier for each listing.
- `name`: Name of the listing.
- `host_id`: Unique identifier for the host.
- `host_name`: Name of the host.
- `neighbourhood`: Neighborhood where the listing is located.
- `latitude`: Latitude coordinate of the listing.
- `longitude`: Longitude coordinate of the listing.

- `room_type`: Type of room (e.g., entire home/apt, private room, shared room).
- `price`: Price per night.
- `minimum_nights`: Minimum number of nights required for a booking.
- `number_of_reviews`: Total number of reviews for the listing.
- `last_review`: Date of the most recent review.
- `reviews_per_month`: Average number of reviews per month.
- `availability_365`: Number of days the listing is available in a year.

3.0.8 Host Information.

- `host_since`: Date when the host started listing on Airbnb.
- `host_response_time`: Average response time of the host.
- `host_response_rate`: Response rate of the host.
- `host_is_superhost`: Whether the host is a superhost.

3.0.9 Booking and Review Information.

- `review_scores_rating`: Overall rating score from reviews.
- `review_scores_cleanliness`: Cleanliness rating score.
- `review_scores_communication`: Communication rating score.
- `review_scores_location`: Location rating score.
- `review_scores_value`: Value rating score.

3.0.10 Excluded Data. The following attributes are excluded due to irrelevance, poor data quality, or technical constraints:

3.0.11 Irrelevant Attributes.

- `listing_url`: URL of the listing, irrelevant for prediction.
- `scrape_id`: Identifier for the data scraping session, not useful for prediction.
- `last_scraped`: Date of the last data scrape, not relevant for analysis.
- `description`: Text description of the listing, difficult to standardize and use effectively in modeling without NLP techniques.

3.0.12 Poor Data Quality.

- `neighbourhood_group`: Largely missing or inconsistent values.
- `license`: Many listings do not provide this information.

3.0.13 Justification for Data Selection.

- **Listing and Host Information**: These attributes are crucial as they provide fundamental details about the listing and host, which are essential for predicting rental prices.
- **Booking and Review Information**: These attributes reflect customer feedback and demand, which are important indicators of listing performance and price justification.

3.1 Clean Data

To ensure the data quality meets the requirements for machine learning, several cleaning steps were performed:

3.1.1 Handling Missing Values.

- **Drop Columns with Excessive Missing Values**: Columns with more than 50% missing values were dropped.
- **Impute Missing Values**: For numerical columns with missing values, we used the median or mean to fill in the gaps. For categorical columns, the mode was used.

3.1.2 Removing Duplicates.

- Duplicate rows were identified and removed to ensure that each listing is unique.

3.1.3 Outlier Detection and Treatment.

- Outliers in numerical columns, such as price, were identified using the IQR method and capped at the 1st and 99th percentiles to reduce their impact.

3.1.4 Standardizing Formats.

- Dates were converted to a standard format (YYYY-MM-DD).
- Text data was cleaned to remove special characters and whitespace.

3.1.5 Impact on Analysis Results.

- **Improved Model Performance**: Cleaned data reduces noise and helps machine learning algorithms learn better, leading to improved prediction accuracy.
- **Reduced Bias**: Handling missing values and outliers reduces the risk of bias in the model.

3.2 Construct Data

Constructive data preparation involved creating new attributes and records to enhance the dataset:

3.2.1 Derived Attributes.

- **Price per Room**: Created a new attribute by dividing the total price by the number of rooms.
- **Host Experience**: Calculated the number of years a host has been active by subtracting the host's start date from the current date.

3.2.2 New Records.

- **Zero Purchase Records**: Created records for listings with zero bookings in the past year to account for listings that may be new or have low visibility.

3.2.3 Transforming Values.

- **Categorical to Numerical**: Converted categorical variables like `room_type` into numerical codes using one-hot encoding.

3.3 Standardize Data

To ensure that the data fed into the modeling stage is consistent, the following standardization methods were used:

3.3.1 Normalization.

- **Min-Max Scaling**: Applied Min-Max scaling to numerical attributes to bring all values into the range [0, 1].
- **Standard Scaling**: Used StandardScaler to normalize attributes with a standard distribution.

3.3.2 Data Merging.

- **Merge Data Containers**: Combined data from multiple sources into a single dataframe, ensuring all relevant attributes are included.

3.3.3 Uniform Schema.

- **Schema Development**: Ensured all datasets follow a uniform schema, with consistent attribute names and data types.

3.4 Conclusion of Phase 2

The data preparation process involved selecting relevant attributes, cleaning data to ensure quality, constructing new features, and standardizing the data. These steps were critical to ensure that the dataset is ready for the machine learning modeling phase, ultimately leading to more accurate and reliable predictions for Airbnb rental prices.

References