Airbnb Listing Price Prediction Analysis Report

Overview

This report analyzes the Airbnb dataset to predict listing prices using ANN machine learning techniques. The dataset contains 12,805 listings with 23 features, including property details, amenities, and location information. The goal is to understand the factors influencing Airbnb listing prices and build predictive models.

Dataset Exploration

Key Features

- **Size**: 12,805 listings
- **Features**: 23 columns including:
  - `id`, `name`, `rating`, `reviews`, `host_name`, `address`, `amenities`, `price`
  - Property features: `bathrooms`, `beds`, `guests`, `bedrooms`, etc.
- **Target Variable**: `price` (price per night in local currency)

Missing Values

- `host_name`: 8 null values
- `checkin`: 800 null values
- `checkout`: 2,450 null values

Data Cleaning

- Converted numerical columns (`rating`, `reviews`, `price`, etc.) to numeric types, dropping rows with invalid entries.
- Final cleaned dataset: 8,566 rows.

Exploratory Data Analysis (EDA)

Price Distribution

- The price distribution is right-skewed, with most listings priced below 10,000 units.
- Outliers exist, with some listings priced as high as 190,796 units.

Correlation Analysis

- **Positive Correlations**:
    - `price` correlates with `bathrooms` (0.37), `beds` (0.28), `guests` (0.35), and `bedrooms` (0.39).
- **Negative Correlations**:
    - `reviews` shows weak negative correlations with most features.

Visualizations

- A histogram of `price` (limited to 0–100,000 for clarity) shows the majority of listings are concentrated in the lower price range.

Key Insights

1. **Price Drivers**:
    a. Larger properties (more bedrooms, bathrooms, and beds) tend to have higher prices.
    b. The number of guests a property accommodates also positively influences price.
2. **Rating and Reviews**:
    a. `rating` shows a weak positive correlation with `price`, suggesting higher-rated listings may command slightly higher prices.
    b. `reviews` has little impact on price, indicating popularity (measured by reviews) does not strongly affect pricing.
3. **Missing Data**:
    a. The `checkout` feature has significant missing values (2,450), which may require imputation or exclusion in modeling.

Modeling Approach

Preprocessing

- **Handling Missing Values**: Impute or drop rows with missing `checkin` and `checkout` data.
- **Feature Engineering**: Extract useful information from categorical features like `amenities` and `address`.
- **Scaling**: Normalize numerical features for models sensitive to scale (e.g., neural networks).

Models to Consider

1. **Neural Network (Sequential Model)**:
    a. Suitable for capturing complex patterns in the data.
    b. Requires careful tuning of layers and hyperparameters.

Evaluation Metrics

- **Mean Squared Error (MSE)**: Measures average squared difference between predicted and actual prices.
- **R^2 Score**: Indicates the proportion of variance in prices explained by the model.

Recommendations

1. **Feature Engineering**:
    a. Extract and encode amenities (e.g., count of amenities or binary flags for specific amenities).
    b. Parse `address` to derive location-based features (e.g., city, neighborhood).
2. **Outlier Handling**:
    a. Investigate and potentially cap extreme price values to improve model robustness.
3. **Model Interpretability**:
    a. Use Random Forest to identify key price drivers (e.g., property size, location).
4. **Advanced Techniques**:
    a. Experiment with gradient boosting models (e.g., XGBoost) for potentially better performance.

b. Consider clustering listings to identify price segments.

Conclusion

The dataset provides a solid foundation for predicting Airbnb listing prices. Initial analysis highlights property size and capacity as key price influencers. Further preprocessing and model tuning are needed to improve predictive accuracy and uncover deeper insights. The next steps include feature engineering, model training, and validation to build a robust price prediction system.