Airbnb Rental Price Prediction – NYC

Data Science Case Study Report

Introduction

Airbnb is a global online marketplace that connects travelers with local hosts offering unique accommodations. Founded in 2008, it has revolutionized the hospitality industry by enabling individuals to monetize their living spaces. In New York City, one of Airbnb's largest markets, hosts face the challenge of setting competitive yet profitable rental prices amidst a dynamic market environment.

This project aims to develop a machine learning model that predicts optimal listing prices based on various features such as property characteristics, location, availability, and review scores.

Abstract

This case study explores the development of a predictive model using machine learning to estimate Airbnb rental prices in New York City. The dataset contains 27,392 listings with 23 features including host details, property attributes, geographic coordinates, and guest reviews.

Exploratory Data Analysis (EDA) was conducted to understand feature distributions, detect outliers, and identify correlations. Data preprocessing included handling missing values, encoding categorical variables, and engineering new features like price per person and distance to city center.

Several models were evaluated:

- Linear Regression
- Lasso & Ridge Regression (with regularization)
- Random Forest Regressor

- XGBoost Regressor

Hyperparameter tuning was performed using RandomizedSearchCV and GridSearchCV. Among all models, **XGBoost achieved the best performance with a test R^2 score of 98%**, indicating strong predictive accuracy.

Tools and Technologies Used

- **Programming Language**: Python
- **Libraries Used**:
  - `pandas`, `numpy` – For data manipulation
  - `matplotlib`, `seaborn` – For visualization
  - `scikit-learn` – For preprocessing, model training, and evaluation
  - `xgboost` – For implementing the XGBoost regressor
- **Evaluation Metrics**:
  - R^2 Score
  - Root Mean Squared Error (RMSE)

Steps Involved in Building the Project

1. Data Loading and Initial Inspection

The dataset was loaded using `pandas`. It contained 27,392 rows and 23 columns. Basic statistical summaries and missing value analysis were performed.

2. Exploratory Data Analysis (EDA)

- Analyzed numerical and categorical feature distributions
- Studied correlations between variables
- Visualized geographic distribution of listings
- Conducted univariate and bivariate analysis to understand feature impact on price

Key findings:

- Most properties are located in Manhattan and Brooklyn.
- Entire home/apt is the most common room type.
- Review scores are generally high, indicating positive guest experiences.

3. Data Preprocessing

- Dropped irrelevant columns (`host_id`, `host_name`, etc.)
- Handled missing values:
  - Removed `square_feet` due to high missing percentage
  - Imputed mean for review scores and median for other numerical features
- Engineered new features:
  - `price_per_person`: Normalizes price by capacity
  - `total_sleeping_capacity`: Based on number of beds
  - `distance_to_center`: Euclidean distance from NYC center
- Encoded categorical variables using Label Encoding

4. Outlier Detection and Treatment

- Detected outliers using the IQR method
- Cap extreme values to upper/lower bounds
- Ensured no outliers remained after treatment

5. Train-Test Split and Feature Scaling

- Scaled features using `StandardScaler`
- Split dataset into 80% training and 20% testing sets

6. Model Training and Evaluation

Linear Regression

- Test R^2 = 86.16%
- RMSE = 32.26

Lasso & Ridge Regression

- Similar performance to Linear Regression
- Slight improvement in regularization

Random Forest Regressor

- Achieved high training R^2 (99.30%) but showed signs of overfitting (test R^2 = 94.90%)
- Identified key features: `price_per_person`, `room_type`, `accommodates`

XGBoost Regressor

- Best performing model with test R^2 = **98.00%**
- Showed balanced performance on both training and test data
- Key features: `room_type`, `accommodates`, `price_per_person`

7. Hyperparameter Tuning

- Used `RandomizedSearchCV` and `GridSearchCV` for tuning
- Final XGBoost model used:
    - `max_depth=5`, `learning_rate=0.01`, `n_estimators=500`

Conclusion

This project successfully developed a robust machine learning model to predict Airbnb rental prices in New York City. Through extensive EDA and preprocessing, the dataset was cleaned and enriched with meaningful features. Multiple regression models were tested, with **XGBoost delivering the highest accuracy** and generalization capability.

Key insights include:

- Features like `room_type`, `accommodates`, and `price_per_person` are the most influential predictors.
- Ensemble methods significantly outperform linear models in capturing complex relationships within the data.

- Accurate pricing predictions can empower hosts to make informed decisions and optimize revenue.

For production deployment, **XGBoost is recommended** due to its superior performance, followed closely by Random Forest for interpretability. Linear models remain suitable for simpler use cases where explainability is critical.

End of Report

Prepared by: Saida Mansoor

Date: June 9, 2025