Predicting Airbnb Prices: Understanding What Drives Short-Term Rental Costs

Abstract

This notebook explores the factors influencing **Airbnb listing prices** using publicly available data from *Inside Airbnb*. The objective is to understand how property characteristics, location, and customer engagement metrics affect nightly rental prices, and to build a predictive model capable of estimating these prices.

Purpose

The project aims to move from raw, unstructured data to a well-defined, interpretable machine learning pipeline that predicts listing prices accurately. By combining **data cleaning**, **exploratory analysis**, and **regression modeling**, the notebook demonstrates how data-driven insights can be used to understand pricing behavior in online marketplaces.

⚙ Methodology

1. **Data Preprocessing:** Clean and transform raw data to handle missing values, inconsistent formatting, and categorical variables.
2. **Exploratory Data Analysis (EDA):** Visualize distributions, correlations, and key variables influencing price.
3. **Model Training:** Implement two regression models –
    - **Linear Regression:** for interpretability and baseline performance.
    - **Random Forest Regressor:** for improved accuracy and non-linear pattern detection.
4. **Evaluation:** Assess models using **Root Mean Squared Error (RMSE)** and **R^2 (Coefficient of Determination)**.

Key Outcomes

- Property **capacity**, **room type**, and **neighborhood** emerged as the strongest determinants of nightly price.
- The **Random Forest Regressor** achieved higher accuracy, demonstrating its ability to capture complex, non-linear relationships.
- The notebook illustrates the full **data understanding lifecycle** – from cleaning and visualization to predictive modeling and interpretation.

Significance

This work highlights the importance of **data understanding** in producing reliable, explainable results. It bridges the gap between statistical modeling and real-world business context — showing how transparent, well-documented data workflows can drive meaningful insights in the sharing economy.

📘 Theory & Background

📊 Understanding Regression Analysis

Regression analysis is a cornerstone of **predictive modeling in data science**, used to estimate a continuous outcome variable (like price) from a set of input features (predictors). In the context of **Airbnb**, regression helps quantify how listing characteristics — such as **room type**, **number of reviews**, or **location** — influence the nightly price of a property.

⚙ Key Concepts

1. Linear Regression

- Assumes a straight-line relationship between predictors and response.
- It's **interpretable and efficient**, allowing us to understand how much each feature contributes to price changes.
- However, it can **struggle with outliers** and non-linear interactions (e.g., price increases not being proportional to the number of guests).

2. Random Forest Regressor

- An **ensemble learning** method that aggregates predictions from multiple decision trees.
- Each tree learns different feature relationships; the combined output reduces variance and improves generalization.
- Excels at modeling **non-linear patterns**, **feature interactions**, and **heterogeneous data** — common in real-world datasets like Airbnb listings.

Theoretical Foundation

Both approaches operate under the umbrella of **supervised learning**, where:

- The model is trained on labeled data (listings with known prices).
- It learns to map input features (e.g., number of bedrooms, location, reviews) to output prices.
- The trained model is then evaluated on **unseen listings** to test how well it generalizes.

This framework mirrors real-world pricing systems — learning from past listings to estimate new ones.

Why This Matters

Understanding these algorithms provides two key advantages:

- **Interpretability:** Linear Regression offers transparency – we can explain *why* a prediction was made.
- **Performance:** Random Forest adds robustness and flexibility, capturing complex relationships that linear models miss.

Together, they showcase the trade-off between **simplicity and accuracy**, a recurring theme in modern data science practice.

Problem Statement

Goal

To **predict the nightly price** of an Airbnb listing based on its observable attributes. The objective is to determine **which features most strongly influence price** and to evaluate **how accurately regression models** can predict these prices.

Context

Pricing on Airbnb is influenced by multiple factors — from tangible property characteristics to qualitative signals such as reviews and location. By understanding these relationships, hosts can make data-driven pricing decisions and travelers can identify fair-value listings.

Inputs (Predictor Variables)

1. **neighbourhood_cleansed** – Location of the listing within the city.
2. **room_type** – Type of property (Entire home, Private room, Shared room, etc.).
3. **accommodates** – Number of guests the listing can host.
4. **bedrooms, bathrooms_text** – Indicators of property size and amenities.
5. **number_of_reviews, review_scores_rating** – Measures of popularity and perceived quality.
6. **availability_365** – Number of days the property is available for booking per year.

Output (Target Variable)

- **price** – A continuous numeric value representing the listing's **nightly price (USD)**.

Research Question

> ***Which listing features most strongly influence Airbnb prices, and how accurately can we predict these prices using regression models?***

This question drives the analytical focus of the notebook — combining **data exploration**, **feature engineering**, and **predictive modeling** to uncover meaningful patterns in real-world Airbnb data.

🧹 Data Preprocessing

🎯 Objective

To **transform raw Airbnb listing data** into a clean, structured, and machine-readable format suitable for analysis and modeling. This step ensures that the dataset is consistent, accurate, and free from missing or misleading values before applying regression algorithms.

⚙️ Steps and Rationale

1. 📂 **Data Loading**
   - Load the dataset (listings.csv or listings.csv.gz) using **pandas**.
   - Validate successful import and inspect dataset dimensions and column names.
2. 💲 **Target Cleaning (price)**
   - Remove currency symbols ($, ,) and convert values to **float** type.
   - Drop rows with invalid or missing price entries.
   - Ensures the target variable is numeric and ready for regression analysis.
3. 🧩 **Feature Selection**
   - Retain only relevant columns such as room_type, accommodates, bedrooms, bathrooms_text, review_scores_rating, and availability_365.
   - Reduces noise and focuses analysis on features that logically affect price.
4. 🛁 **Parsing Text Columns (bathrooms_text)**
   - Extract numeric values (e.g., "1.5 baths" → 1.5).
   - Standardize bathroom counts for consistent numerical interpretation.
5. 🧮 **Handling Missing Values**
   - Use **imputation** (median for numeric features, most frequent for categorical ones) to fill missing values.
   - Prevents data loss from unnecessary row deletions while maintaining statistical balance.

6. **Outlier Treatment**
    - Clip extreme prices beyond the 5th-95th percentile range.
    - Reduces skew and prevents high-priced luxury listings from biasing the model.
7. **Encoding Categorical Variables**
    - Convert non-numeric features (room_type, neighbourhood_cleansed) into numeric format using **One-Hot Encoding**.
    - Enables regression models to interpret categorical attributes mathematically.
8. **Data Splitting**
    - Split the cleaned dataset into **Training (80%)** and **Testing (20%)** subsets.
    - The training set is used for model learning, while the test set evaluates generalization performance.

✅ Outcome

After preprocessing:

- The dataset is fully numeric and free from major inconsistencies.
- Outliers are mitigated, missing values are handled, and categorical features are encoded.
- The data is now ready for **exploratory analysis** and **regression modeling** in the next section.

Data Analysis

This section explores how Airbnb listing prices vary across different features and helps us understand key data patterns before modeling. The goal is to connect *raw data understanding* with *predictive insights* later in the notebook.

What We Examine

- **Distribution of Prices:** To check for skewness, outliers, and typical value ranges.
- **Capacity Features:** How the number of guests (accommodates) or bedrooms affects price.
- **Categorical Factors:** Whether room_type or neighbourhood_cleansed significantly influence price.
- **Correlation Between Variables:** To see which features are related and whether multicollinearity might exist.
- **Feature Importance (Post-Model):** To connect EDA findings with machine learning outputs.

Insights from Visual Analysis

1. **Price Distribution:**
    - Prices are right-skewed; a few luxury listings raise the average.
    - Log-transformation produces a more balanced spread.
2. **Price vs. Capacity:**
    - Price increases with more guests and bedrooms but not linearly.
    - Doubling capacity doesn't double price – diminishing returns are observed.
3. **Room Type Differences:**
    - Entire homes/apartments show the highest median price.
    - Private and shared rooms form lower-cost clusters.
4. **Neighborhood Effects:**

- Central or tourist-heavy neighborhoods show noticeably higher average prices.

5. **Correlation Analysis:**
   - Positive correlation between accommodates, bedrooms, and price.
   - review_scores_rating shows weak direct correlation but helps when combined with other features.
6. **Feature Importance:**
   - Top predictors: location, room type, and capacity features.
   - These align closely with expectations from EDA.

Takeaways

- **Location**, **room type**, and **capacity** are the most powerful predictors of Airbnb price.
- Review quality and availability play secondary roles.
- Patterns observed here support and validate the later regression results.

⚙ Code Implementation

This section implements the predictive modeling workflow to estimate **Airbnb listing prices** using two regression algorithms — **Linear Regression** (interpretable baseline) and **Random Forest Regressor** (non-linear ensemble model).

Modeling Pipeline Overview

1. **Preprocessing:**
   - One-hot encoding for categorical variables (neighbourhood_cleansed, room_type).
   - Median imputation for numeric variables.
   - Unified through a **ColumnTransformer** for consistency.
2. **Models:**
   - **Linear Regression:** establishes a transparent baseline.
   - **Random Forest:** captures complex, non-linear interactions.
3. **Evaluation:**
   - Metrics: **RMSE** (error in dollars) and **R²** (variance explained).
   - Comparison highlights the performance trade-off between interpretability and accuracy.

Results Summary

Model	RMSE ↓	R^2 ↑	Notes
Linear Regression	98.04	0.600	Baseline model – explains ~60% of price variance.
Random Forest Regressor	**87.08**	**0.685**	Captures non-linear relationships – better accuracy.

Interpretation

- The Random Forest Regressor outperforms Linear Regression by reducing error and improving predictive power.
- This indicates **Airbnb pricing patterns are non-linear**, shaped by complex feature interactions.
- Pipelines and transformers make the process **modular, reproducible, and free of data leakage**.

Conclusion

This project explored **Airbnb price prediction** using publicly available Inside Airbnb data to understand what drives short-term rental pricing. Through a structured data science workflow — including preprocessing, exploratory analysis, and model evaluation — we derived both quantitative insights and practical learnings.

Key Findings

- **Primary Drivers:** Neighborhood, room type, and property capacity.
- **Model Results:** Random Forest performed better ($R^2 \approx 0.69$) than Linear Regression ($R^2 \approx 0.60$).
- **Interpretation:** Tree-based models better capture non-linear dynamics in pricing behavior.

Limitations

- Lack of time-based or seasonal features.
- Amenities and textual descriptions not included.
- Spatial proximity (distance to city center, landmarks) not modeled.

Future Work

1. Add **temporal data** for dynamic pricing.
2. Incorporate **text features** (listing descriptions, reviews) using NLP.
3. Use **geospatial analysis** to include distance-based attributes.
4. Test advanced models like **XGBoost** or **Gradient Boosting**.

Final Thoughts

The notebook demonstrates the full **Understanding Data** workflow — from messy raw data to interpretable prediction. It highlights how effective data cleaning, feature engineering, and thoughtful modeling yield **insightful, trustworthy results**.

📚 References & License

📖 References

1. **Inside Airbnb.** (2025). *Get the Data.* Retrieved from https://insideairbnb.com/get-the-data.html
2. **Pedregosa, F., et al.** (2011). *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, 12, 2825-2830.
3. **Hunter, J. D.** (2007). *Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering*, 9(3), 90-95.
4. **Pandas Development Team.** (2024). *pandas: Powerful Python Data Analysis Toolkit.* https://pandas.pydata.org
5. **Breiman, L.** (2001). *Random Forests. Machine Learning*, 45(1), 5-32.

⚖️ License