Abstract

The hospitality and short-term rental market has seen exponential growth, with platforms like Airbnb disrupting traditional lodging. However, pricing strategies remain a complex challenge for hosts, while guests often struggle to identify fair market value. Effective decision-making relies on understanding the multifaceted factors that influence listing prices, from location and amenities to host reputation.

This project addresses this critical need by designing, developing, and evaluating a comprehensive, interactive data science dashboard built with **Streamlit**. This report documents the complete lifecycle of the project, from data ingestion to the deployment of a multi-functional web application. We utilized the **Sydney Airbnb dataset**, comprising thousands of listings and user reviews, to perform rigorous Exploratory Data Analysis (EDA) and develop predictive models.

The core of the project involved a three-pronged modeling strategy to analyze the data from every angle:

1. **Regression:** To predict listing prices (price) using **Linear Regression** and **Random Forest Regressors**. The models utilized a diverse feature set including location coordinates, room type, amenity density, and host tenure.
2. **Classification:** To predict "Superhost" status (host_is_superhost) using **Logistic Regression**, analyzing the impact of response rates and aggregated review scores.
3. **Clustering:** To segment listings into distinct market categories using **K-Means Clustering** and **PCA** (Principal Component Analysis) for dimensionality reduction.

Our analysis revealed that while traditional linear models struggled to capture price variance The R-squared value is approximately 0.15., ensemble methods like Random Forest performed better ,The R-squared value is approximately 0.22. by capturing non-linear relationships between location and price. Furthermore, unsupervised clustering successfully identified four distinct market segments. The final application empowers users with advanced tools, including a Price Predictor, Geospatial Mapping with Folium, and Superhost Classifier, bridging the gap between complex data science and actionable real estate insights.

Table of Contents

1. Introduction

1.1. Project Background and Problem Statement

The peer-to-peer accommodation market is highly volatile. Unlike hotels with fixed standard rates, Airbnb prices fluctuate based on dynamic factors such as location desirability, property type, guest capacity, and host reputation.

- For Hosts: Setting a price too high leads to vacancies; setting it too low results in revenue loss. Furthermore, achieving "Superhost" status is a key revenue driver, but the exact criteria—balancing response rates with review scores—can be opaque.
- For Guests: It is difficult to gauge if a property is priced fairly relative to the market average or if a specific location justifies the premium.

Currently, stakeholders often rely on intuition rather than data. The problem this project solves is the lack of accessible, data-driven tools that can process historical listing and review data to provide accurate price estimations and market segmentation for the Sydney region.

1.2. Project Objectives (Aim)

The primary aim of this project is to bridge this gap by designing, building, and deploying a comprehensive, interactive data science dashboard to analyze and predict Airbnb listing dynamics.

The specific objectives are:

1. **To Collect and Prepare** the Sydney Airbnb dataset (Listings.csv and Reviews.csv), handling complex encoding (latin1) and cleaning dirty data formats (currency symbols, percentages, and text blobs).
2. **To Perform** a thorough **Exploratory Data Analysis (EDA)** to visualize price distributions, seasonal trends in reviews using time-series decomposition, and geographical hotspots using interactive maps.
3. To Implement and Evaluate a suite of machine learning models for three distinct tasks:
    - Regression: To predict the exact listing price based on property characteristics.

- Classification: To predict if a host qualifies as a "Superhost".
- Clustering: To group listings into meaningful market segments (e.g., "Budget", "Premium").

4. To Deploy these models into an interactive, user-friendly Streamlit application featuring advanced tools for prediction and geospatial visualization.

1.3. Scope and Limitations

This project covers the complete data science workflow, from data ingestion to a functional web application.

Limitations:

- **Model Performance**: The baseline regression models achieved low R^2 scores (<0.30), indicating that price is heavily influenced by qualitative factors (e.g., interior design quality, ocean views, luxury furniture) that are not captured in the tabular dataset.
- **Class Imbalance**: The classification task for Superhosts faced significant class imbalance (fewer Superhosts than regular hosts), resulting in a model biased toward the majority class (High Accuracy, but 0.00 F1-Score for Superhosts).
- **Geographic Scope**: The analysis is strictly limited to the provided dataset's location (Sydney) and does not generalize to other markets.

2. System Design and Methodology

2.1. Technology Stack

The project was developed using the **Python** ecosystem, leveraging a modern stack for data science and web deployment.

- **Web Framework:** `Streamlit` was chosen for its ability to rapidly build and deploy interactive data applications without extensive frontend code.
- **Data Manipulation:** `pandas` and `numpy` were used for all data loading, cleaning, transformation, and feature engineering.
- **Machine Learning:** `scikit-learn` was the primary library for building preprocessing pipelines (`StandardScaler`, `LabelEncoder`, `PCA`) and implementing models (`LinearRegression`, `RandomForestRegressor`, `LogisticRegression`, `KMeans`).
- **Data Visualization:** `seaborn` and `matplotlib` were used for static statistical plots (boxplots, heatmaps), while `folium` (via `streamlit_folium`) was used for interactive geospatial mapping.
- **Time Series:** `statsmodels` was used for seasonal decomposition of review data.

2.2. System Architecture & Data Flow

The application follows a modular pipeline approach:

1. **Data Loading:** The app loads Listings.csv and Reviews.csv using a cached function load_and_preprocess_data() to optimize performance and prevent reloading large files on every interaction.
2. **Preprocessing:** Data is cleaned by converting price strings to floats, imputing missing values, and merging review data.
3. **User Interface:** The sidebar allows users to navigate between "EDA", "Map", "Prediction", and "Clustering" pages, triggering different analysis modules.
4. **Model Inference:** On-the-fly training and prediction occur based on user inputs, providing real-time feedback.

2.3. Data Ingestion and Preprocessing

Data quality was a significant challenge. Key preprocessing steps included:

1. **Encoding Handling:** The raw CSV files required encoding='latin1' to be read correctly, bypassing standard UTF-8 decoding errors common in older datasets.
2. **Currency & Format Conversion:**
   - The price column contained $ and , symbols (e.g., "$1,200.00"). A Regex replacement pattern was applied to strip these characters and convert the feature into a numerical float.
   - host_response_rate and host_acceptance_rate contained % symbols, which were removed and converted to floats (0.0 to 100.0).
3. **Date Processing:** The host_since column was converted to datetime objects to calculate host_tenure. The reviews dataset's date column was processed to perform time-series decomposition.
4. **Missing Value Imputation:**
   - Numerical columns (e.g., bedrooms, bathrooms, accommodates) were filled with the **median** to be robust against outliers.
   - Categorical columns (e.g., host_response_time, neighbourhood) were filled with the **mode** (most frequent value).

2.4. Feature Engineering

New features were created to enhance model performance:

- **host_tenure**: Calculated as the number of days between the current date (datetime.now()) and the host_since date. This serves as a proxy for host experience and platform trust.
- **amenities_count**: The raw text of amenities (e.g., "{TV, Wifi, Pool, Kitchen}") was parsed to count the total number of items, creating a numerical proxy for "luxury" or "value add".
- **Label Encoding**: Categorical variables like room_type, neighbourhood, city, and district were encoded into integers for machine learning compatibility.
- **Standardization**: Numerical features were scaled using StandardScaler to ensure distance-based algorithms (like K-Means and Logistic Regression) performed correctly and weren't biased by large magnitudes (e.g., price vs bedrooms).

3. Exploratory Data Analysis (EDA)

3.1. Key Findings from Visual Analysis

The "Exploratory Data Analysis" page of the dashboard presents a detailed visual analysis of the dataset using Seaborn and Matplotlib.

- **Price Distribution:** The target variable price is highly right-skewed. The histogram shows that the majority of listings are budget-friendly ($50 - $200), while a small, long tail of luxury properties ($1000+) significantly inflates the mean. This suggests that median price is a better indicator of the "typical" Airbnb than the mean.
- **Geographical Hotspots:** Using Folium heatmaps, we observed that listings in coastal areas (e.g., Bondi, Manly) and the CBD command significantly higher prices than those in the Western suburbs. This confirms the hypothesis that "location" is a primary price driver.
- **Room Type Analysis:** Boxplots revealed that "Entire home/apt" listings have the highest median price and significant variance, whereas "Shared rooms" are the cheapest and most consistent in pricing.
- **Seasonality:** Time-series decomposition of the reviews data revealed clear seasonal peaks, likely correlating with summer holidays and tourism influxes, alongside a general upward trend of platform growth over the years.
- **Correlation Analysis:** The correlation matrix highlighted that accommodates (capacity) and bedrooms have the strongest positive correlation with price. Surprisingly, number_of_reviews showed a weak or negative correlation with price, suggesting cheaper listings get booked (and reviewed) more frequently than expensive ones.

4. Modeling Methodology and Evaluation

A three-pronged modeling approach was adopted to gain a holistic understanding of the market.

4.1. Task 1: Regression (Cost Prediction)

- **Objective:** To predict the nightly listing price.
- **Input Features:** host_response_rate, accommodates, bedrooms, amenities_count, minimum_nights, latitude, longitude, room_type, neighbourhood, city, district, host_tenure.
- **Models:** Multiple Linear Regression (Baseline) and Random Forest Regressor.
- **Results:**
    - **Linear Regression:** $R^2 = 0.15$. The model struggled to fit the data, indicating the relationship between features and price is highly non-linear.
    - **Random Forest Regressor:** $R^2 = 0.22$. This model outperformed the baseline by capturing non-linear interactions (e.g., the interaction between location coordinates and room type), though significant variance remains unexplained.
- **Key Drivers:** Feature importance analysis identified longitude, latitude, and room_type as the most critical predictors of price.

4.2. Task 2: Classification (Superhost Prediction)

- **Objective:** To predict if a host is a "Superhost" (host_is_superhost).
- **Input Features:** review_scores_rating, review_scores_accuracy, review_scores_cleanliness, host_response_rate, host_acceptance_rate, host_tenure.
- **Model:** Logistic Regression.
- **Results:** The model achieved a seemingly high accuracy but a **F1-Score of 0.00** for the positive class.
- **Analysis:** This highlights a severe "Class Imbalance" issue. The model biased itself toward the majority class (non-Superhosts) because the dataset contains far more regular hosts. As a result, it predicted "No" for almost every instance. This finding suggests that future iterations must use resampling techniques like SMOTE to balance the training data.

4.3. Task 3: Clustering (Market Segmentation)

- **Objective:** To discover hidden "listing archetypes" using unsupervised learning.
- **Input Features:** price, accommodates, review_scores_rating, availability_365.
- **Model:** K-Means Clustering with PCA (Principal Component Analysis) for 2D visualization.
- **Results:** The "Elbow Method" suggested $K=4$ as the optimal number of clusters. The algorithm successfully segmented the market into:
    1. **Cluster 0:** Standard/Budget Rentals (Moderate price, moderate rating).
    2. **Cluster 1:** High-Volume/Premium (Higher price points, professional hosts).
    3. **Cluster 2 & 3:** Variations based on capacity and specific review profiles (e.g., highly rated small units vs. lower rated large units).

6. Conclusion

This project successfully achieved its primary objectives. We processed a complex, real-world dataset and built a functional multi-page dashboard.

The **Exploratory Data Analysis** provided critical insights into the geographical price disparity in Sydney. The **Modeling phase** revealed that while price prediction in the peer-to-peer market is difficult (low R^2), ensemble methods like Random Forest significantly outperform linear baselines. The **Clustering analysis** successfully identified four distinct market archetypes, providing a new layer of business intelligence.

Finally, the **Streamlit Application** democratizes this data, allowing non-technical users to leverage machine learning for pricing decisions and market analysis.

7. Future Work

While this project provides a robust foundation, future iterations could expand its capabilities:

1. **Advanced Modeling:** Implement **XGBoost** or **Gradient Boosting** to improve the regression R^2 score, as suggested by the preliminary results indicating Random Forest's superiority over Linear Regression.
2. **Handling Imbalance:** Apply **SMOTE (Synthetic Minority Over-sampling Technique)** to the Superhost classification task to fix the low F1-score and improve recall for identifying actual Superhosts.

3. **NLP on Reviews:** Use Natural Language Processing (Sentiment Analysis) on the reviews text data to extract features like "Guest Sentiment," which likely correlates strongly with price and booking frequency.
4. **External Data:** Integrate real-time data such as distance to landmarks (Opera House, Bondi Beach) or public transport hubs via APIs to improve prediction accuracy.
5. **Cloud Deployment:** Deploy the app to Streamlit Community Cloud for public access.