Predicting Airbnb Success: A Data-Driven Investment Framework for the Los Angeles Market

Ananya Praveen Shetty, Apoorva Shastry

Abstract—The short-term rental market, dominated by platforms like Airbnb, presents a high-risk, high-reward opportunity for real estate investors. Making an uninformed investment can lead to significant financial loss. This project develops a comprehensive machine learning framework to guide investor decisions in the Los Angeles market. We integrate three datasets—Airbnb listings, U.S. Census income, and EPA Walkability scores—enhanced through latent feature engineering for host quality, popularity, and space efficiency. XGBoost regression on the Latent feature set ($R^2 \approx 0.68$) outperformed Base ($R^2 \approx 0.35$) and Enriched ($R^2 \approx 0.52$). A Random Forest classifier achieved a weighted F1-score of 0.65. K-Means and Fractal Clustering identified market segments, and SHAP explained key drivers. This production-ready model predicts nightly prices within ≈$95 RMSE, empowering investors with data-driven insights.

Index Terms—Machine Learning, Regression, Classification, Clustering, Feature Engineering, Data Amalgamation, SHAP, Airbnb, Real Estate Investment

I. INTRODUCTION

The sharing economy has revolutionized real estate, with Airbnb enabling homeowners to monetize properties. However, investment in such volatile markets demands data-driven guidance. This study introduces a framework that predicts property prices, classifies desirability, and identifies success factors. We hypothesize that combining property-level, socioeconomic, and environmental data—augmented with engineered latent features—yields superior predictive accuracy.

II. METHODOLOGY

We followed a data science lifecycle involving data sourcing, preprocessing, feature engineering, modeling, and analysis.

A. Data Sourcing and Amalgamation

Three datasets were integrated: (1) Inside Airbnb listings for Los Angeles, (2) U.S. Census ACS median_income data merged by ZIP code, and (3) EPA Walkability Index joined via census tracts. A fourth scraped dataset (crime data) demonstrated extensibility.

B. Latent Feature Engineering

Three engineered variables captured hidden relationships: host_quality_score (from host_is_superhost and response_rate), popularity_score (from number_of_reviews and review_scores_rating), and space_efficiency_score (ratio of accommodates to bedrooms).

C. Preprocessing Pipeline

A ColumnTransformer handled median imputation, scaling, and categorical encoding. A custom parser standardized mixed-format bathroom data.

III. EXPERIMENTS AND RESULTS

Three experiments—clustering, classification, and regression—validated the hypothesis.

A. Clustering (Answering Q3)

K-Means (k=4) produced interpretable clusters: (0) Budget/Economy, (1) Premium/Luxury, (2) High Value/Undervalued, and (3) Potential Underperformers. Fractal Clustering on the Golden Cluster (price>$200, rating>4.8) revealed Ultra-Luxury and Prime Location sub-clusters.

B. Classification (Answering Q2)

Listings were grouped by desirability_score (price × rating) into three classes. Ensemble models performed best.

Table I: Classification Müller Loop Results

Algorithm	F1-Weighted	Precision	Recall
XGBoost	0.65	0.65	0.65
Random Forest	0.64	0.64	0.64
Logistic Regression	0.63	0.63	0.63
MLP	0.62	0.62	0.62
SVM	0.61	0.61	0.61

C. Regression (Answering Q1)

Regression tests across Base, Enriched, and Latent features confirmed that feature augmentation improves performance. XGBoost achieved the highest R^2 and lowest RMSE.

Table II: Regression Performance (XGBoost)

Algorithm	F1-Weighted	Precision	Recall
XGBoost	0.65	0.65	0.65
Random Forest	0.64	0.64	0.64
Logistic Regression	0.63	0.63	0.63
MLP	0.62	0.62	0.62
SVM	0.61	0.61	0.61

IV. ANALYSIS AND DISCUSSION

A. Feature Importance (Gini)

Top predictive features included popularity_score, median_income, space_efficiency_score, room_type, and accommodates.

B. Model Explainability (SHAP)

SHAP analysis showed popularity positively influences price, while shared rooms depress it.

C. Data Distribution

SMOTE improved SVM slightly, but ensemble models remained robust without resampling.

V. CONCLUSION

This framework improves Airbnb investment prediction accuracy through data amalgamation and latent feature engineering. The final model achieves ≈$95 RMSE, identifies desirability factors, and segments markets. Future extensions include seasonal forecasting and finer-grained amenity data scraping.

REFERENCES

[1] Inside Airbnb. http://insideairbnb.com/. Accessed: Oct. 2025.

[2] U.S. Census Bureau. American Community Survey (ACS) 5-Year Data. https://data.census.gov/. Accessed: Oct. 2025.

[3] U.S. Environmental Protection Agency (EPA). Smart Location Database & Walkability Index. https://www.epa.gov/smartgrowth/. Accessed: Oct. 2025.

[4] S. Lundberg and S. Lee. “A Unified Approach to Interpreting Model Predictions.” NIPS, 2017.

[5] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python.” JMLR, 12, pp. 2825–2830, 2011.

[6] T. Chen and C. Guestrin. “XGBoost: A Scalable Tree Boosting System.” Proc. KDD, 2016.