

Airbnb Success Predictor: An Investor's Guide to the LA Market

Project Description

The project builds a predictive analytics framework to assist real-estate investors in Los Angeles in evaluating Airbnb properties. It estimates nightly prices, predicts listing desirability, and uncovers natural market clusters.

Questions

1. **Regression (Q1):** What is the fair market price for an Airbnb listing given its features and neighborhood context?
 2. **Classification (Q2):** Can we classify a listing as Least / More / Most Desirable for investment potential?
 3. **Clustering (Q3):** Are there natural market segments (K-Means) and how do they compare with our business-defined Golden Cluster (Fractal Clustering)?
-

Experiments

- **Data Amalgamation:** Compared Base → Enriched → Latent feature sets to prove the value of external and engineered features.
- **Algorithm Comparison (Müller Loops):** 7 regression and 5 classification models tested.
- **Clustering Analysis:** K-Means for unsupervised segments and Fractal Clustering for Golden Cluster refinement.
- **Explainability:** Gini importance and SHAP explain model drivers.
- **Productionization:** Best regression model pickled for future use.

Team Name and Members

Team: Airbnb LA Investor Insight

Members: Ananya Praveen Shetty & Apoorva Shastry

Roles

- **Ananya:** Data acquisition & amalgamation, EDA, clustering, dashboarding.
 - **Apoorva:** Model development, feature engineering, explainability, deployment.
-

Data

Sources

- **Data Set 1:** *listings.csv* – Inside Airbnb (listing details).
- **Data Set 2:** *ACSDT5Y2023.B19013-Data.csv* – U.S. Census median household income by ZIP.
- **Data Set 3:** *Walkability_Index.csv* – U.S. EPA walkability index.
- **Scraped (Extra):** AreaVibes crime scores for L.A. ZIP codes.

Discussion on Amalgamation

Feature Set	Best Model	R ²	RMSE
Base	XGBoost	0.5261	\$114.56
Enriched	XGBoost	0.5470	\$112.02
Latent	XGBoost	0.5623	\$110.11

→ Adding income and walkability (+ Enriched) then latent variables (+ Latent) raised accuracy and lowered error.

Exploratory Data Analysis & Visualization

- Histogram of prices (\$10–\$1000), median \approx \$150.
- Room type distribution dominated by “Entire home/apt.”
- Scatter: Price vs Walkability revealed positive trend for premium areas.

Clustering

Techniques

- **K-Means (Euclidean)**: Elbow method $\rightarrow k = 4$.
- **Clusters**: Budget / Premium / High-Value / Underperformer.

Fractal Clustering

Defined business rule cluster first (price $>$ \$200 & rating $>$ 4.8 & Entire home/apt), then K-Means ($k = 3$) within that Golden segment \rightarrow **Ultra-Luxury**, **Golden Value**, **Prime Location**.

Golden Cluster

High-performing listings meeting the above criteria; used to zoom into top-tier market segments.

Answer to Q3: Natural segments exist and mirror business intuition; fractal analysis adds micro-segmentation.

Objective Function: Minimize within-cluster sum of squared errors (SSE).

Latent Variables & Manifolds

Added latent features:

- `host_quality_score = 0.5(superhost)+0.5(response_rate)`
- `popularity_score = log1p(reviews × rating)`
- `space_efficiency_score = accommodates / (bedrooms + 1)`

Improvement: Latent feature set achieved highest R² (0.5623) and lowest RMSE (\$110.11).

Metrics: Regression – R², RMSE; Classification – weighted F1, precision, recall.

Classification

Answer to Q2: Best model = **XGBoost**, weighted F1 = **0.6921**.

- Precision/recall ≈ 0.70 overall; strong for Least and Most Desirable classes.

Müller Loop

Compared to Logistic Regression, Random Forest, XGBoost, MLP, SVM — XGBoost led in F1.

Selected Features

Top drivers from Gini importance (XGBoost): `popularity_score`, `median_income`, `Walkability`, `room_type_Entire home/apt`, `accommodates`.

Regression

Answer to Q1: Best regressor = **XGBoost** (Latent set). R² = 0.5623, RMSE = \$110.11. Model saved as `best_airbnb_price_model.pkl`.

Müller Loop

7 algorithms tested (Linear, RF, XGB, KNN, SVR, MLP, Tree) across 3 feature sets; R² rose Base → Latent.

Distributions of Your Data (EDAV)

- Price distribution right-skewed but filtered (\$10–\$1000).
- Room-type distribution and scatter plots visualize market composition.

Selected Features

Base, Enriched, Latent sets used for comparative testing.

Algorithms for Feature Importance

- **Random Forest/XGBoost (Gini)** – feature selection.
- **SHAP values** – direction of impact (+ or -).

Why: They quantify feature influence on predictions.

Changing Data Distributions

Tree-based models (XGBoost, RF) performed best under original data balance; oversampling benefited minor class algorithms (SVM/MLP).

Best Distribution: Original (50 % slider).

Worst: Aggressive undersampling (loss of information).

Data Narrative & Conclusions

- **Hypothesis Proved:** $R^2 \uparrow$ and $RMSE \downarrow$ with data enrichment and latent features.
- **Q1:** Fair price predictor (XGBoost $R^2 = 0.5623$).
- **Q2:** 3-class XGBoost classifier ($F1 = 0.6921$).
- **Q3:** Clear segments found ($k = 4$) and refined (Golden Cluster → 3 subclusters).

- **Explainability:** Gini and SHAP validated key drivers.
- **Production:** Pickled regressor for investor tooling.

Yes — all three research questions were answered, and the hypothesis was validated.

Ananya and Apoorva's Contribution:

Assignment	Core Task	Ananya Praveen Shetty's Contributions	Apoorva Shastry's Contributions
Assignment 1	Project Plan	Co-authored plan; defined Goals & KPIs ; researched "Guest Experience."	Co-authored plan; defined Business Use Case ; led data acquisition strategy.
Assignment 2	Clustering	Developed Profitability Score ; ran Iteration 1 K-Means; selected initial "Golden Cluster."	Implemented Fractal Clustering (Iteration 2); engineered the final <code>is_golden_cluster</code> target variable.
Assignment 3	Classification	Led Data Amalgamation (Spatial Joins); analyzed Ensemble Models (XGBoost, RF) .	Analyzed Instance/Boundary Models (KNN, SVM) ; established performance baselines.

Assignment 4	Dashboard	Developed Resampling Pipelines (SMOTE); analyzed MLP & SVM model sensitivity.	Built the Interactive Dashboard UI ; analyzed XGBoost & RF model robustness.
Assignment 5	Optimization	Calculated Feature Importance ; re-trained and validated MLP & SVM models on the optimized dataset.	Formed optimization hypothesis; re-trained and validated XGBoost & RF models; analyzed business impact.
Assignment 6	Regression	Engineered ' Host Experience ' & ' Popularity ' latent features; ran linear & Keras regression models.	Engineered ' Space Efficiency ' latent feature; identified the top-performing Random Forest regression model.