Hawaii Airbnb Smart Pricing and Market Segmentation System

Jingyi Chen, Zhanyuan Jiang, Jiaqi Li

December 17, 2025

Abstract

This project analyzes Airbnb pricing in Oahu using a sequence of predictive models and market segmentation techniques. We first establish baseline performance using standard linear and tree-based models, which serve as reference points for later comparisons. Building on these benchmarks, we investigate whether incorporating market segmentation can further improve predictive accuracy. We compare two segmentation-based strategies: cluster-then-predict and cluster-as-features and find that the latter delivers more reliable performance. By incorporating cluster membership directly into a global nonlinear model, we achieve the best balance between predictive accuracy and stability.

1 Background

Airbnb has become one of the most widely used platforms for short-term rentals in the world, allowing hosts to list properties and set nightly prices while guests search for accommodations that match their preferences and budget. In this process, pricing plays a central role in host revenue and guest decision-making.

Airbnb does provide a default Smart Pricing tool to assist hosts in setting competitive prices. The Hawaii Airbnb market, however, has highly differentiated pricing patterns due to geographic location and seasonal tourism. Ocean-front views, access to beaches, amenities such as pools and AC, and proximity to tourist areas significantly influence price dispersion. Unfortunately, these localized factors are not fully captured by Airbnb's generic Smart Pricing algorithm.

Hence, to better serve Hawaii hosts, this research aims to build a localized pricing system that identifies data-driven sub-markets and provides more transparent price recommendations. As part of this research, we envision a dashboard where hosts can see their market segment and obtain a data-driven nightly price estimate tailored to Hawaii's unique market conditions.

2 Data Description

2.1 Data Resource

The dataset used in this project is obtained from Inside Airbnb, a publicly accessible platform that provides detailed, regularly updated snapshots of Airbnb activity across global cities. Inside Airbnb offers quarterly datasets for the past year for each region, allowing time-specific and market-level analyzes.

The Hawaii Islands comprise several distinct regions, but this study focuses exclusively on Oahu, which offers the largest concentration of Airbnb listings and sufficient data for reliable analysis.

2.2 Feature Engieering

The original dataset obtained from Inside Airbnb contains a wide range of listing-level information, including property characteristics (e.g., room type, accommodates, bathrooms, bedrooms), host attributes, review scores, geographic coordinates, and a textual amenities list.

To better capture pricing drivers in the Oahu market, we engineered additional features using the Google Maps Distance Matrix API, computing each listing's driving distance and travel time to Honolulu International Airport and Waikiki Beach. We also parsed the amenities text into structured

indicators (e.g., pool, AC, ocean view) and created an amenities count variable. The target variable was transformed using log(price), categorical attributes were one-hot encoded, and missing values were handled through imputation with corresponding missing-value flags.

3 Methodology

3.1 Baseline Pricing Model

To establish a clear benchmark for evaluating subsequent models, we first fit a set of standard predictive models directly on the full set of cleaned and engineered features. At this stage, the goal is not aggressive optimization, but rather to assess how well common modeling approaches perform when the Airbnb market is treated as a homogeneous population, without explicitly accounting for market segmentation or structural heterogeneity. These baseline models provide reference points for judging whether additional modeling structure is warranted. An overview of the baseline models and their out-of-sample performance is reported in Table 1.

Model	Type	Tuning Strategy	Test R^2	Test $RMSE$
OLS	Linear	No regularization	0.620	0.679
Ridge with CV (α=3.16)	Linear	5-fold CV to select α	0.620	0.679
Random Forest with RandomizedSearchCV	Nonlinear	5-fold CV	0.828	0.457
LightGBM with RandomizedSearchCV	Nonlinear	5-fold CV	0.861	0.410
CatBoost with RandomizedSearchCV	Nonlinear	5-fold CV	0.867	0.402
XGBoost with RandomizedSearchCV	Nonlinear	5-fold CV	0.874	0.392

Table 1: Overview of the Baseline Models

3.1.1 Linear Models: OLS and Ridge Regression

We begin with linear models, including Ordinary Least Squares (OLS) and Ridge regression. Both models are trained on the same feature set, with Ridge introducing ℓ_2 regularization to stabilize coefficient estimates in the presence of multicollinearity. In practice, both OLS and Ridge achieve similar out-of-sample performance, with test R^2 values around 0.62. This indicates that regularization improves coefficient stability but does not meaningfully enhance predictive accuracy. Overall, linear models appear insufficient to capture the complexity of Airbnb pricing dynamics.

3.1.2 Tree-Based Models

To capture nonlinear relationships and feature interactions, we next evaluate several tree-based models, including Random Forest, LightGBM, CatBoost, and XGBoost[1]. All tree-based approaches substantially outperform linear models, confirming the presence of strong nonlinear effects in the data. Among them, XGBoost delivers the best performance, achieving a test R^2 of approximately 0.87, representing a large improvement over the linear baseline. The remaining tree-based models serve as useful robustness checks and achieve competitive performance, but do not surpass XGBoost.
Overall, these baseline results suggest that while increased model flexibility yields large gains in predictive accuracy, further improvements are unlikely to come from algorithmic complexity alone. This motivates the exploration of models that explicitly incorporate market structure and segmentation in later sections.

3.2 Market Segmentation

3.2.1 Clustering in Original Feature Space

Although several baseline models (e.g., XGBoost) achieve strong predictive performance, they rely on a single global pricing relationship across all listings. In practice, Airbnb pricing on Oahu is highly location-dependent: beachfront properties near Waikiki, airport-adjacent listings, and inland residential units often follow distinct pricing mechanisms. To capture such heterogeneity, we apply unsupervised clustering to partition the market into data-driven sub-markets.

Clustering is performed on a subset of interpretable features capturing geographic accessibility[2] and listing characteristics, including latitude, longitude, driving distances to Honolulu International Airport and Waikiki Beach, review-based quality measures, and accommodation-related attributes (capacity, bedrooms, bathrooms). All features are standardized prior to clustering. Using Agglomerative Clustering and silhouette scores to select the number of clusters, we find that a coarse segmentation with $k = 3$ provides a good balance between within-cluster cohesion and between-cluster separation (silhouette score = 0.524). A robustness check using K-means yields broadly consistent partitions, suggesting that the identified sub-markets are not sensitive to the specific clustering algorithm. Figure 1 visualizes the resulting clusters on the latitude–longitude map, confirming clear geographic differentiation.

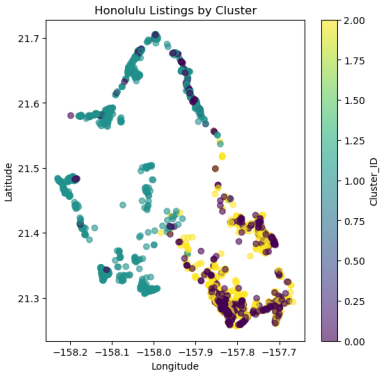

Figure 1: Geographic Clusters

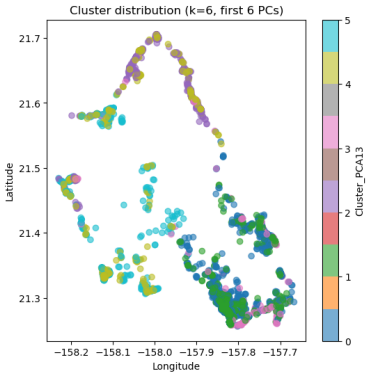

Figure 2: PCA-Based Clusters

3.2.2 PCA-Based Analysis and Supplementary Clustering

To explore whether additional latent structure exists beyond these coarse groupings, we further apply Principal Component Analysis (PCA) to the same standardized feature set. The first four principal components explain approximately 90 % of the total variance, indicating that much of the information in the original features can be summarized in a low-dimensional, decorrelated representation.
Clustering in the PCA-transformed space yields a different silhouette[3] profile, with the optimal value attained at $k = 6$ (silhouette score = 0.61). As illustrated in Figure 2, these clusters exhibit finer spatial separation compared to clustering in the original geographic space. This does not imply that PCA-based clustering is superior; rather, PCA changes the notion of similarity by emphasizing directions of maximal variance and removing feature correlations. Taken together, the results suggest that the Oahu Airbnb market exhibits multi-scale structure: a small number of intuitive, geographically interpretable segments, alongside more nuanced subdivisions revealed in a variance-driven representation.
Motivated by these findings, we subsequently consider incorporating segmentation information into pricing models, either by adding cluster membership as features or by fitting separate models within each identified sub-market.

4 Results

4.1 Baseline and Segmentation-Aware Models

To evaluate whether explicitly modeling market heterogeneity improves pricing performance, we compare two segmentation-aware strategies: *cluster-then-predict* and *cluster-as-features*. Both approaches build upon the baseline tree-based models introduced earlier, which already outperform linear models.

4.2 Cluster-Then-Predict

In the cluster-then-predict approach, we first partition listings into sub-markets using agglomerative clustering based on geographic and listing characteristics. Separate predictive models are then trained within each cluster. Due to computational constraints, we focus on two representative tree-based methods: Random Forest (RF) and LightGBM (LGBM).

Table 2 reports the aggregated out-of-sample performance across clusters. While cluster-specific models achieve strong in-sample fit, overall test performance is comparable to the best global baseline models, with LightGBM achieving a test R^2 of 0.840.

Model	Test R^2	Test RMSE
RF (cluster-then-predict)	0.815	0.474
LGBM (cluster-then-predict)	0.840	0.441

Table 2: Overall test performance for cluster-then-predict

4.3 Cluster-As-Features

Alternatively, we incorporate cluster assignments directly into the feature space as categorical indicators, allowing a single global model to condition predictions on sub-market membership. This approach avoids training separate models per cluster while still leveraging segmentation information.
Table 3 shows that adding cluster features consistently improves performance across tree-based models. XGBoost with cluster features achieves the best overall performance, with a test R^2 of 0.873 and RMSE of 0.393.

Model	Test R^2	Test RMSE
RF + Cluster Features	0.830	0.454
LightGBM + Cluster Features	0.868	0.401
CatBoost + Cluster Features	0.861	0.410
XGBoost + Cluster Features	**0.873**	**0.393**

Table 3: Performance of cluster-as-features models

Overall, the cluster-as-features strategy provides a better balance between predictive accuracy and computational efficiency, outperforming both global baselines and cluster-then-predict models.

4.4 Interpretation of Price Drivers

To interpret pricing drivers, we analyze both linear and nonlinear models. Ridge regression highlights capacity, host portfolio size, review quality, and location proxies as key linear predictors. To capture nonlinear effects, we further examine SHAP values from tree-based models with cluster features.
SHAP-based feature importance analyses for the tree-based models are reported in the Appendix. Across models, listing capacity (e.g., accommodates and bedrooms), host activity, availability, and review scores consistently dominate model predictions. Cluster indicators contribute to pricing predictions but play a secondary role relative to core listing characteristics.

5 Discussion and Future Work

5.1 Linear versus Tree-Based Models

Results from the baseline analysis show that linear models (OLS and Ridge) achieve similar out-of-sample performance, with test R^2 values around 0.62. Introducing ℓ_2 regularization stabilizes coefficient estimates but does not meaningfully improve predictive accuracy, indicating that multicollinearity is not the primary limitation of the linear specification. Instead, the main constraint is the limited expressive capacity of global linear models when applied to Airbnb pricing.
In contrast, tree-based models substantially outperform linear baselines. Random Forest, LightGBM, CatBoost, and XGBoost all achieve much higher test R^2 values, confirming the presence of strong nonlinear effects and feature interactions in pricing behavior. These gains suggest that modeling flexibility, rather than regularization strength, is the key driver of improved predictive performance.

5.2 Comparing Cluster-Then-Predict and Cluster-as-Features

Motivated by the observed market heterogeneity, we explored two strategies for incorporating clustering information into the pricing models: cluster-then-predict and cluster-as-features.

The cluster-then-predict approach fits separate models within each cluster, aiming to capture segment-specific pricing relationships. However, this strategy fragments the data, reducing effective sample size within each cluster and increasing estimation variance. Although cluster-level models often achieve strong in-sample fit, their out-of-sample performance is unstable, particularly for smaller or more heterogeneous clusters.

In contrast, the cluster-as-features approach augments a global model with cluster membership indicators while preserving the full training sample. This method consistently outperforms cluster-then-predict and yields more stable improvements across all tree-based models. By conditioning predictions on sub-market membership without sacrificing data efficiency, cluster-as-features achieves a more favorable bias–variance tradeoff.

5.3 Why Segmentation Yields Modest but Consistent Gains

Although incorporating cluster information improves predictive performance, the magnitude of the gain is modest. This suggests that much of the segmentation-related information is already implicitly captured by strong tree-based models through geographic variables, capacity measures, and availability constraints.

Rather than introducing fundamentally new information, cluster indicators act as coarse summaries of patterns that tree models partially learn on their own. As a result, clustering refines predictions rather than transforming the pricing mechanism, leading to incremental but consistent performance improvements.

5.4 Interpretability and Feature Effects

Both Ridge coefficients and SHAP analyses highlight a consistent set of price drivers across models. Listing capacity (e.g., accommodates, bedrooms), host portfolio size, availability constraints, and review-related measures emerge as the most influential features.

SHAP results further reveal nonlinear and asymmetric effects that are not captured by linear models. Cluster indicators appear in SHAP rankings but contribute less than core listing and host characteristics, reinforcing the interpretation that segmentation information complements rather than dominates the predictive structure.

5.5 Future Work

Several directions remain for future research. First, more flexible spatial methods—such as spatially varying coefficient models or graph-based clustering—may capture finer-grained geographic effects. Second, incorporating temporal dynamics and seasonality could improve pricing accuracy in highly seasonal markets like Hawaii. Finally, integrating demand-side signals, such as booking probability or occupancy rates, may allow pricing models to move beyond static prediction toward revenue optimization.

References

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[2] Yanfang Liu, Xiang Kong, Xiaoping Liu, and Shixuan Wang. Analyzing spatial variance of airbnb pricing determinants using multiscale geographically weighted regression (mgwr) approach. *Sustainability*, 12(11):4710, 2020.

[3] Peter J Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

A Additional Figures

Figure A.1–A.4 present SHAP summaries for tree-based models with cluster-as-features. These figures complement the discussion in Section 4.

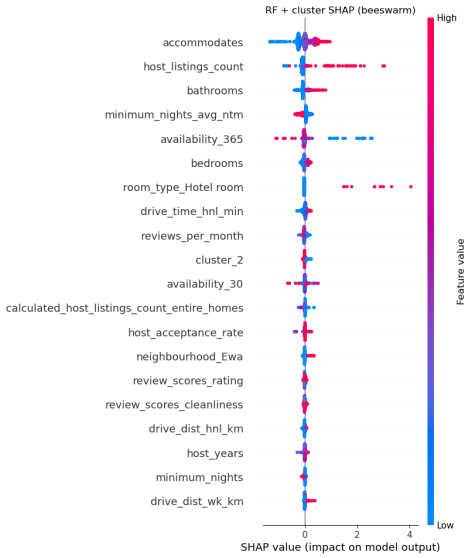

Figure A.1: SHAP summary for Random Forest with cluster-as-features.

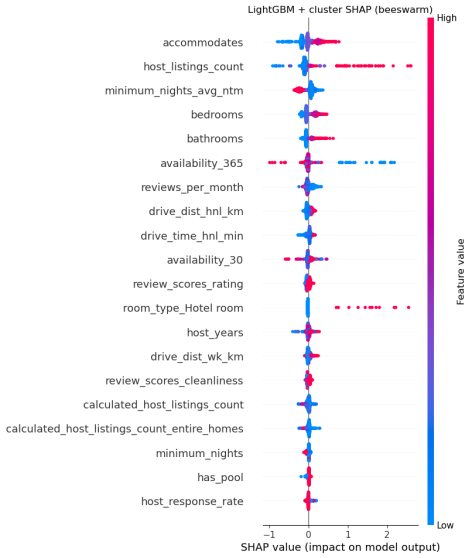

Figure A.2: SHAP summary for LightGBM with cluster-as-features.

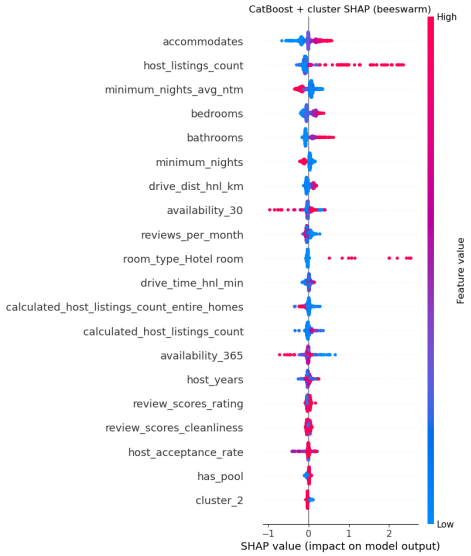

Figure A.3: SHAP summary for CatBoost with cluster-as-features.

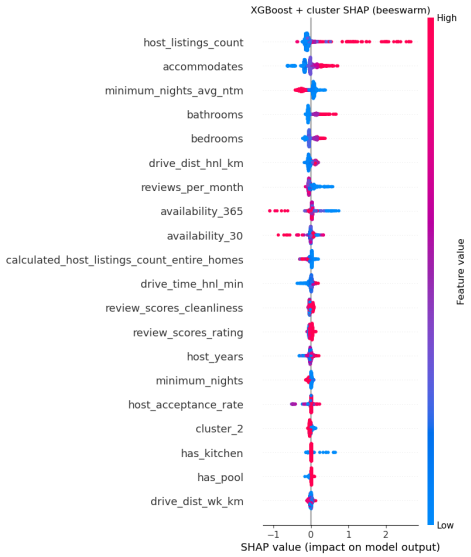

Figure A.4: SHAP summary for XGBoost with cluster-as-features.

B Additional Tables

Model	Cluster	Test R^2	Test RMSE	Train R^2	Train RMSE
RF	0	0.681	0.349	0.921	0.204
LGBM	0	0.778	0.291	0.983	0.094
RF	1	0.513	1.083	0.979	0.265
LGBM	1	0.545	1.047	1.000	0.025
RF	2	0.771	0.428	0.944	0.205
LGBM	2	0.799	0.401	1.000	0.010
RF	3	0.885	0.621	0.964	0.356
LGBM	3	0.879	0.638	0.992	0.170
RF	4	0.588	0.418	0.927	0.185
LGBM	4	0.602	0.411	0.975	0.108
RF	5	0.677	0.519	0.919	0.261
LGBM	5	0.776	0.432	0.998	0.039

Table B.1: Cluster-level performance for the cluster-then-predict strategy