Summary Report: Airbnb Price Prediction Model

Tokyo, Japan

Introduction

The goal centers on developing a machine learning model to price small to mid-size apartments in Tokyo, Japan. Tokyo's distinctive market necessitates adjustments in data acquisition from Inside Airbnb, feature engineering, and sampling strategies. The approach includes employing **three models:** foundational linear regression via **Random Forest, OLS, and LASSO**. Each model will undergo evaluation for its effectiveness in predicting optimal pricing, with a focus on performance comparisons and justification for model selection. This analysis seeks to establish pricing strategies for a company specializing in apartment rentals, demonstrating the adaptability and precision of machine learning techniques within a unique urban context.

Data Preparation

The dataset for the analysis was procured from a substantial Airbnb listing database, focusing on a city with over 10,000 entries. Missing prices were omitted, and the 'price' column was normalized by removing currency symbols and converting to numerical format. Listings were filtered to spotlight properties accommodating 2-6 guests, aligning with the company's service range and excluding 'Hotel' and 'Shared' room types.

Feature Engineering

The selected features included basic variables like 'accommodates', 'beds', 'property_type', 'room_type', 'bathrooms_text', and 'neighbourhood_cleansed'. Reviews and amenities were also factored into the model to enrich the dataset. Interaction terms between 'accommodates', 'property_type', and 'neighbourhood_cleansed' were introduced to capture complex dynamics in the data.

Model Development

Random Forest Regressor

The Random Forest model was refined through grid search, focusing on 'max_features' and 'min_samples_leaf' parameters. The optimal model demonstrated commendable predictive accuracy with a minimized RMSE via cross-validation. Analysis of feature importance highlighted 'accommodates', 'property_type', and 'neighbourhood_cleansed' as key predictive variables.

OLS Linear Regression

An Ordinary Least Squares regression was implemented as a linear benchmark model. This model offered insights into the direct effects of the features on pricing, serving as a comparative standard for other models.

LASSO Regression

The LASSO regression, applied with a cross-validated approach, emphasized its strength in feature selection by imposing penalties on less significant variables, streamlining the predictive model.

Model Evaluation

	model	CV RMSE
2	random forest	17169.860000
0	OLS	29572.754369
1	LASSO	37169.260289

Model performance was primarily assessed by cross-validated RMSE scores. The Random Forest model surpassed the OLS and LASSO models in predictive performance, suggesting a better fit for capturing the dataset's nonlinear patterns and interactions.

Interpretation and Insights

The investigation of the 'accommodates' feature through partial dependence plots unveiled an anticipated positive correlation with listing prices. A detailed analysis of model performance across various subsets, such as apartment sizes, property types, and neighborhoods, provided nuanced insights into price determinants.

Conclusion

The Random Forest model, with its thorough analysis of feature importance and partial dependence evaluations, stands out as the most effective predictive tool for establishing Airbnb pricing strategies. The insights obtained are therefore crucial for making informed pricing decisions for the company's new apartment offerings in Tokyo, Japan. Considering the model's predictions and the absence of historical pricing data for the newly introduced apartments, it is recommended that they be listed at an initial price point of JP¥ 20,000. This recommendation is contingent upon the assumption that this price reflects the model's valuation of similar properties in the dataset.