CAPSTONE PROJECT: SUGGESTING RENTAL PRICES FOR AIRBNB PROPERTIES

Prepared and Excuted by - Group (1)

MEET THE TEAM

Mayur Patel

Siddhartha Majumder

Shubham Gupta

TABLE OF CONTENTS

PROJECT OVERVIEW

To suggest appropriate listing prices to property owners for rental properties in Antwerp, Belgium.

VISION

Ensure rental prices are competitive yet reflective of the value offered, balancing affordability with quality.

MISSION

Reflect the costs associated with type of property the lister is providing and services in the rental price.

VALUES

Price rentals transparently, with clear breakdowns of costs and fees, and consider community impact when setting prices.

DATA UNDERSTANDING

01. Teamwork

Collaborate with stakeholders to gather and understand data requirements.
Coordinate with data sources to access and integrate data from different tables.

02. Timeline

Establish timelines for data gathering, cleaning, transformation, and modeling phases.
Align with project milestones and deliverables.

03. Goals

Define clear goals such as predicting listing prices accurately, optimizing for specific metrics, or understanding market trends.

04. Results

Evaluate model performance using appropriate metrics
Present insights and findings to stakeholders, highlighting predictors with significant impact on listing prices.

DATA QUALITY AND EXPLANATION

01. Variable Analysis

To check number of unique values.
To calculate the percentage of missing values.
Identifying smallest and largest values in each continuous variable to spot outliers or data entry errors.

02. Categorial Variable Analysis

To count the unique categories or levels within each categorical variable.
By calculating the percentage of missing values for each categorical variable.

03. Identifying and Fixing Data Anomalies

By using box plots, histograms, or statistical methods to detect outliers.
Checking missing values in both continuous and categorical variables.
Verifying that each variable is stored in correct data type.

04. Quality Explanation

Maintaining accuracy, we can ensure data values are correct and liable.
Completeness minimizes missing data to avoid biasness.
Consistency ensures uniformity in data.
Relevance focuses on analysis and meaningful data points.

VARIABLE PROFILING AND CHECKING RELATIONSHIPS

01. Price vs. Bedrooms

There is a positive relationship, indicating that listings with more bedrooms tend to have higher prices.

02. Price vs. Accomodates

There is a positive relationship, showing that listings that accommodate more people tend to have higher prices.

03. Price vs. Beds

There is a positive relationship, suggesting that listings with more beds tend to have higher prices.

Summary of Findings

Highly Correlated Predictors - adjusted_price is almost perfectly correlated with price, so it can be removed to avoid multicollinearity.

Moderately Correlated Predictors: bedrooms, accomodates, and beds show a moderate positive correlation with price, indicating that they are good predictors.

Weakly Correlated Predictors: latitude, longitude, and host_id have weak correlations with price.

MODELLING AND INSIGHTS

Comparison of Regression Models

MODEL	MSE	R-SQUARED	MAE	RMSE
Linear Regression	7.43E+24	-2.60E+25	2.66E+12	2.73E+12
Regression Tree	3.07E-05	0.9999	8.73E-05	0.0055
Random Forest Regressor	3.29E-05	0.9999	1.11E-04	0.0057
Gradient Boosting	0.0045	0.9844	0.0434	0.0668
XGBoost Regressor	9.71E-05	0.9997	0.0044	0.0099
LightGBM Regressor	0.0003	0.999	0.0104	0.0166

MODELLING AND INSIGHTS

Linear Regression

The model exhibits poor performance with high MSE, negative R-squared, MAE, and RMSE. Potential causes include overfitting, underfitting, incorrect data assumptions, and outlier impact.

Regression Tree

The regression tree model shows high accuracy with low MSE, MAE, and RMSE, indicating precise predictions. An R-squared value near 1 (99.99%) reflects an excellent fit to the data. Overall, the model is reliable for predictions based on the dataset.

Random Forest Regressor

The Random Forest Regressor shows high accuracy with low MSE, MAE, and RMSE values, indicating precise predictions. It exhibits an excellent fit with an R-squared value near 1, capturing almost all variability. Overall, the model is reliable and robust for predictions based on the dataset.

Gradient Boosting Regressor

The Gradient Boosting Regressor shows good performance with low MSE, MAE, and RMSE values. It has a high R-squared value of 98.43%, indicating a strong fit to the data and effective prediction of the target variable.

XGBoost Regressor

The XGBoost Regressor shows:
High Accuracy with low MSE, MAE, and RMSE.
Excellent Fit with a 99.97% R-squared value.
Reliability for precise predictions based on the dataset.

LightGBM Regressor

The LightGBM Regressor shows strong performance with low MSE, MAE, and RMSE values, a high R-squared value (99.89%), and efficiency in handling large datasets and complex models.

MODEL COMPARISON WITH PYCARET

1. Setup and Model Selection

Leveraged PyCaret for streamlined model evaluation and comparison.
Selected diverse regression models including Decision Tree, Gradient Boosting, Random Forest, Extra Trees, XGBoost, and LightGBM.

2. Model Evaluation with Cross-Validation

Conducted 5-fold cross-validation to assess models based on R-squared scores.

Results:
Decision Tree: R2 Score = 0.9327
Gradient Boosting: R2 Score = 0.9642
Random Forest: R2 Score = 0.9478
Extra Trees: R2 Score = 0.9290
XGBoost: R2 Score = 0.9715
LightGBM: R2 Score = 0.9730

3. Insights

Top Performers: LightGBM and XGBoost demonstrate the highest R-squared scores, indicating robust predictive capabilities.
Gradient Boosting and Random Forest: Also exhibit strong performance, with scores close to the top models.
Decision Tree and Extra Trees: While slightly lower, still provide respectable predictive accuracy.

4. Conclusion

Recommendation: Proceed with LightGBM or XGBoost for their superior performance metrics.
Next Steps: Fine-tune the selected models and explore ensemble methods to potentially boost predictive power further.

CONCLUSIONS

1. Consistency in Key Predictors: Both models concur that price_per_night and minimum_nights are the most crucial factors influencing listing prices, emphasizing their significance in setting rental rates.

2. Geographic Factors: Although longitude is a common feature in both models, LightGBM emphasizes that latitude is a key predictor. This indicates that geographic location, including both latitude and longitude, significantly influences pricing.

3. The “room_type_Entire home/apt“ variable in the Random Forest model suggests that providing an entire home or apartment affects pricing, but to a lesser extent than price_per_night and minimum_nights.

4. The model includes one-hot encoded categorical variables for diverse property types and room configurations, considers host experience, guest interactions, and satisfaction metrics, utilizes geographical clustering for regional pricing insights, enriching the dataset for accurate listing price predictions.

5. To refine models and enhance accuracy for real-world deployment, utilize feature engineering, model refinement, ensemble strategies, data quality preprocessing, and interpretability.

THANK YOU

Open the floor for questions and discussions