Data Analysis 3: Assignment 2 - Summary Report

Github Repo: https://github.com/MarcellM01/Data-Analysis-3/tree/main/Assignment_2

Objective

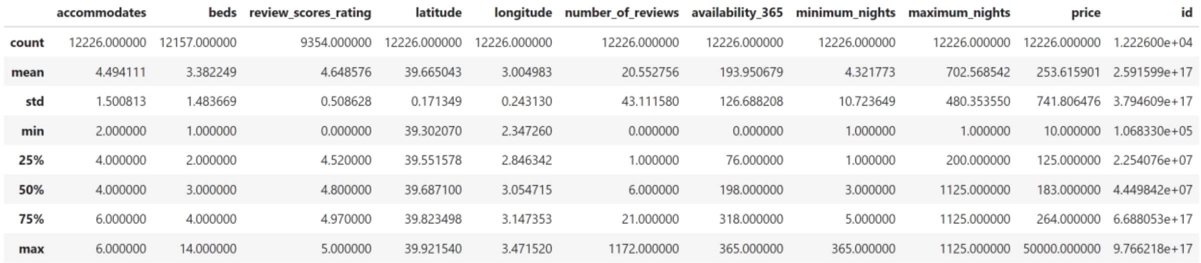
The primary objective of this project was to develop a predictive model for Airbnb rental prices in Mallorca, leveraging machine learning techniques to analyze the dataset containing 18,832 listings. The client requested the focus to be placed on apartments with a possible occupant number of 2-6 as this is the category they wish to target. Our goal was to identify the most effective model based on predictive accuracy and applicability for real-world decision-making, given the clients request. Through discovery it is also the aim to identify the most desirable characteristics of apartments in this region so they can best allocate their funds when buying properties and thus improve their margins.

Data Preparation

The dataset that serves as the basis of this whole operation was sourced from the official Airbnb website, this is relevant as this is the platform that the client has pre-selected as a target to list their properties on. The dataset was refined to focus on listings accommodating 2 to 6 guests, resulting in 12,226 entries for analysis. Key features such as 'accommodates', 'price', and 'amenities' were retained and cleaned for analysis, in total 16 columns where retained. This process also involved converting the 'price' from string (quite a jumble of characters) to a numeric format and filtering out listings with non-viable prices or rare property types.

Exploratory Data Analysis

EDA focused on uncovering the distributions of our numerical variables, as well as the different subdivisions in our categorical variables. This here has revealed a wide range in rental prices, with significant outliers at the high end. The decision to normalize price distribution by removing extreme outliers was critical to improving model performance. The whole entire process was run without this step, and the RMSE was a tenfold of what it is currently. The decision was also made to include hotel rooms as with Airbnb, these are often miscategorized apartments, and where the client to diversify by purchasing a boutique hotel with a small number of rooms, this information may be relevant (can still be excluded at client's request). Lastly, a survey of the available neighbourhoods is also present here.



	accommodates	beds	review_scores_rating	latitude	longitude	number_of_reviews	availability_365	minimum_nights	maximum_nights	price	id
count	12226.000000	12157.000000	9354.000000	12226.000000	12226.000000	12226.000000	12226.000000	12226.000000	12226.000000	12226.000000	1.222600e+04
mean	4.494111	3.382249	4.648576	39.665043	3.004983	20.552756	193.950679	4.321773	702.568542	253.615901	2.591599e+17
std	1.500813	1.483669	0.508628	0.171349	0.243130	43.111580	126.688208	10.723649	480.353550	741.806476	3.794609e+17
min	2.000000	1.000000	0.000000	39.302070	2.347260	0.000000	0.000000	1.000000	1.000000	10.000000	1.068330e+05
25%	4.000000	2.000000	4.520000	39.551578	2.846342	1.000000	76.000000	1.000000	200.000000	125.000000	2.254076e+07
50%	4.000000	3.000000	4.800000	39.687100	3.054715	6.000000	198.000000	3.000000	1125.000000	183.000000	4.449842e+07
75%	6.000000	4.000000	4.970000	39.823498	3.147353	21.000000	318.000000	5.000000	1125.000000	264.000000	6.688053e+17
max	6.000000	14.000000	5.000000	39.921540	3.471520	1172.000000	365.000000	365.000000	1125.000000	50000.000000	9.766218e+17

Feature Engineering

Feature engineering focused on converting amenities into binary variables and categorizing 'property_type' and 'room_type'. This conversion is crucial as it is required for the regression models to make accurate predictions. For the amenities, due to their sheer number, a random assortment of the top amenities was selected and subsequently converted. As for property type and room type, all items where converted. The dataset was also divided into two sub-sections, training and precitions in a 70%-30% split. For the analysis, 3 groups were created, "predictors_small", "predictors_medium" and "predictors_large". The "predictors_large" ended up being used, as this one yielded the lowest RMSE, probably due to the number of predictors assisting in the uncovering of more nuanced associations.

Predictors Used

Basic Variables (Part of: Small, Medium, Large)

This group forms the foundation of our analysis, focusing on essential aspects like the capacity of accommodations, bed count, host quality, and neighborhood specifics. These variables offer a snapshot of the listing's basic attributes, crucial for understanding its appeal and positioning in the market.

House Types (Part of: Small, Medium)

Reflecting the diversity of accommodations, this category encompasses a range of property types from traditional homes and apartments to unique stays like chalets, cottages, and boutique hotels. It provides insights into market trends and preferences, highlighting the variety of options available to travelers.

Room Types (Part of: Small, Medium)

Differentiating listings by the kind of space offered, such as entire homes, private rooms, or hotel rooms, this classification helps in assessing the impact of room type on pricing and guest choice, catering to different privacy and space requirements.

Reviews (Part of: Small, Medium)

Incorporating feedback metrics, this category captures the number of reviews and overall guest satisfaction scores. These variables are indicative of a listing's reputation and popularity, serving as key determinants of its competitive edge in the marketplace.

Amenities (Part of: Small, Medium, Large)

Covering a broad spectrum of features that enhance the guest experience, from basic comforts like heating and free parking to luxury offerings like pools and garden views. This extensive list of amenities allows for a detailed analysis of how specific conveniences and luxuries influence rental prices and desirability among guests.

Model Development and Selection

OLS (Ordinary Least Squares)

The OLS model, used as the initial benchmark in the analysis, provided a straightforward linear approach to modelling. With an RMSE of 70.26, it offered a baseline level of prediction accuracy without the complexity of more advanced algorithms. OLS models excel in their simplicity and interpretability, their limitation however lies in assuming a linear relationship between the variables and the outcome, which may not always be the case in real-world data.

CART (Classification and Regression Trees)

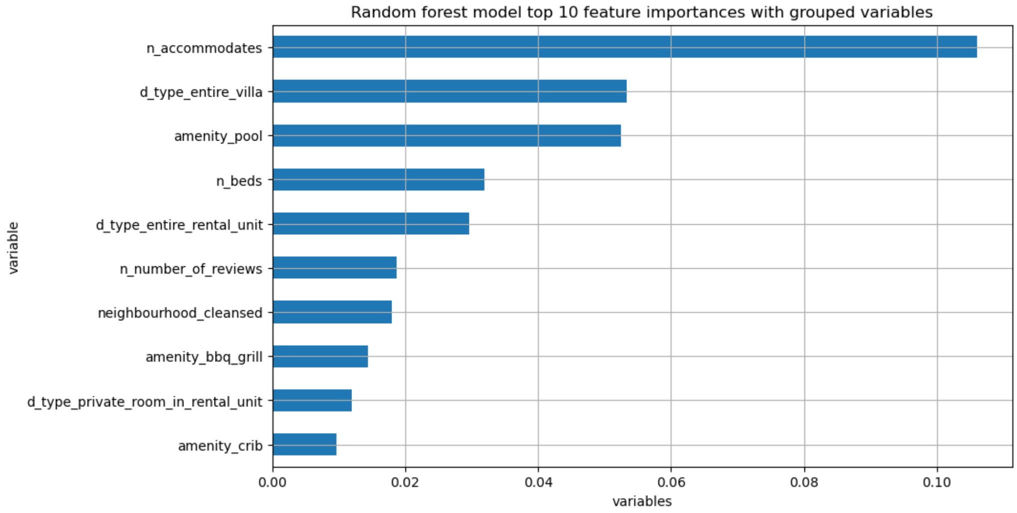
The CART model, while a step up from OLS in terms of handling non-linearity, showed limitations in this context, with an RMSE of 79.47. CART models are more flexible than OLS as they do not assume a linear relationship between features and the target variable. However, they can be prone to overfitting, especially when dealing with complex data structures or when there is a lot of noise in the dataset.

GBM

The GBM model emerged with a very low RMSE of 65.01, signalling its advanced capability to handle the dataset's complexities. GBM models build trees sequentially, with each tree attempting to correct the errors of the previous one, which can lead to a highly accurate model. Their performance is often superior in complex datasets as they combine the predictive power of numerous decision trees through boosting, which can capture complex patterns and interactions between variables.

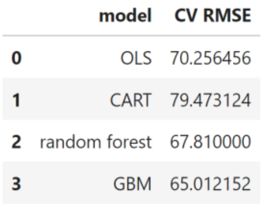
Random Forest

The Random Forest model, an ensemble method that builds upon the concept of CART by creating a multitude of decision trees and aggregating their predictions, showed a marked improvement in prediction accuracy with an RMSE of 67.81. This model's strength lies in its ability to reduce overfitting through its ensemble approach, as it considers the results of many trees to come up with a more stable and robust prediction. The graph clearly shows that the number of guests a property can accommodate is the top predictor of rental price, along with the property being an entire villa and the inclusion of a pool. Other significant factors include the type of rental unit and amenities like a barbecue grill, underscoring preferences for self-contained spaces and features that improve the stay. The importance of the neighborhood and the volume of reviews highlight the impact of location and guest feedback on rental pricing.



Results and Interpretation

The GBM model's superior performance (RMSE of 65.01) underscores the effectiveness of ensemble methods in predictive modelling for complex datasets like Airbnb listings. This suggests that GBM's ability to learn from previous errors and its adaptability through hyperparameter tuning are crucial for capturing the nuanced relationships within the dataset.



	model	CV RMSE
0	OLS	70.256456
1	CART	79.473124
2	random forest	67.810000
3	GBM	65.012152

Recommendation and Conclusion

Given its lowest RMSE and adaptability, the GBM model is recommended for predicting Airbnb rental prices. Its performance indicates a robust model capable of handling the dataset's complexities and variability, making it suitable for deployment in real-world applications.

This project demonstrates the value of a structured approach to predictive modeling, from data cleaning and feature engineering to model selection and evaluation. The success of the GBM model highlights the potential of machine learning techniques in real estate pricing and the importance of methodical model development and evaluation processes.