Summary Report on Price Prediction Models for Apartment Rentals in Sydney, Australia

Student: Thi Mai Pham
GitHub link: https://github.com/thimaipham/2.CEU_DA3_Assignment2

1. Introduction

This report outlines the development and comparison of multiple regression models to predict the rental prices for small to mid-size apartments hosting 2-6 guests. The goal is to assist a company in pricing their new apartments in Sydney not yet on the market.

The raw dataset comprises 25,480 observations and 76 columns, including some columns that are unnecessary or difficult to use during the modeling process. After cleaning part the cleaned data includes: 13773 observations and 38 columns

2. Data Feature Engineering

Certain important columns are encountering the common issues: missing values, mixing data type. I started by refining the dataset:

- Regex to extract useful data
- Change type of data
- Removing unnecessary columns
- Extract time data
- Convert binary categorical variables
- Change some categorical variable by One hot coding

The dataset utilized for model training encompasses a wide range of features, including basic accommodation attributes, host-related information, reviews, availability, and room types. The target variable for our models is the rental price (price). Prior to model training, data was preprocessed to ensure a clean and usable dataset.

3. Model Selection and Training

In my endeavor to construct a robust price prediction model for small to mid-size apartment rentals, I developed four distinct models, each chosen for its unique strengths and suitability to our dataset's characteristics:

3.1. Ordinary Least Squares (OLS) Regression: OLS serves as a fundamental approach for regression analysis, offering a clear view of the linear relationships between the target variable and predictors. We built OLS models with increasing complexity—from basic variables to understand how different feature sets impact model performance:

- Model 1: `basic_vars` focuses on basic features to establish a baseline performance
- Model 2: `basic_vars` + `host_related` + `time_related` introduces host and time-related variables to examine if host characteristics and tenure impact price predictions.
- Model 3: `all_features` leverages the full spectrum of available data to maximize predictive accuracy, testing the hypothesis that a more detailed feature set can provide better predictions.

Key results: The complex model, which includes all specified features, yields the best performance in terms of both lower RMSE and higher R^2 . For all models, the RMSE is higher on the training set compared to the testing set. This is somewhat unusual, as it is typically expected the model to perform better on the data it was trained on.

3.2. CART (Classification and Regression Trees): As a single decision tree model, CART was chosen for its simplicity and interpretability. It allows for an intuitive understanding of how individual features affect the rental price, making it a valuable baseline to compare against more complex ensemble methods.

Results: The CART model displays an RMSE of nearly 2.293 on the training set, indicating perfect predictions. However, the test RMSE is substantially high at 308.459, and the R2 is negative (-1.145), pointing to severe overfitting. This model has learned the training data to an extent where it fails to make reliable predictions on new data, a common pitfall when using a single decision tree.

3.3 Random Forest Regression: The choice to progress from simple linear models to a complex Random Forest was driven by the desire to capture non-linear relationships and interactions between features that linear models may miss. Given its ensemble approach, using multiple decision trees to mitigate overfitting while capturing nonlinear relationships, Random Forest was selected to potentially improve prediction accuracy beyond what linear models can achieve.

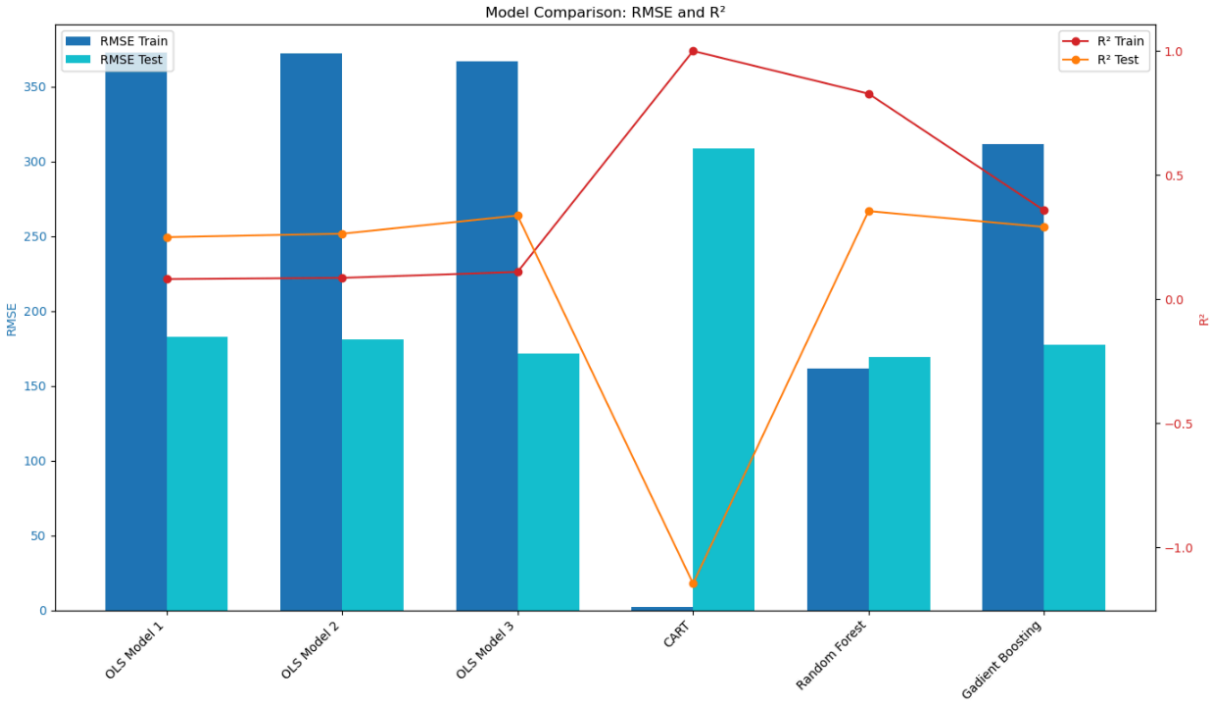
Key results: The Random Forest model markedly outperformed the OLS models, with an RMSE of 169.134 on the test set. This indicates a higher prediction accuracy and a better capacity to capture the variance in rental prices. The model's ensemble approach, leveraging multiple decision trees, helps in addressing the overfitting issue observed in the CART model and captures complex relationships more effectively than linear models. However, the diffience of RMSE and R^2 score on the test set and training set might lead to an light overfiting concern.

3.4. Gradient Boosting Regression: Gradient Boosting was included for its advanced ensemble technique, which sequentially corrects the mistakes of previous trees, potentially offering superior predictive performance. It was anticipated to provide insights into the nuanced dynamics of the rental market through its iterative error correction process.

Key results: The Gradient Boosting model shows a significant improvement in RMSE on the test set compared to the training set, indicating a better balance between bias and variance than the models which were observed earlier. The Gradient Boosting model showed competitive performance with an RMSE of 177.357 on the test set and an R2 of 0.291. While it managed to mitigate some of the overfitting seen in the CART model through sequential improvement, its performance still lagged behind the Random Forest model.

4. Findings & Discussions:

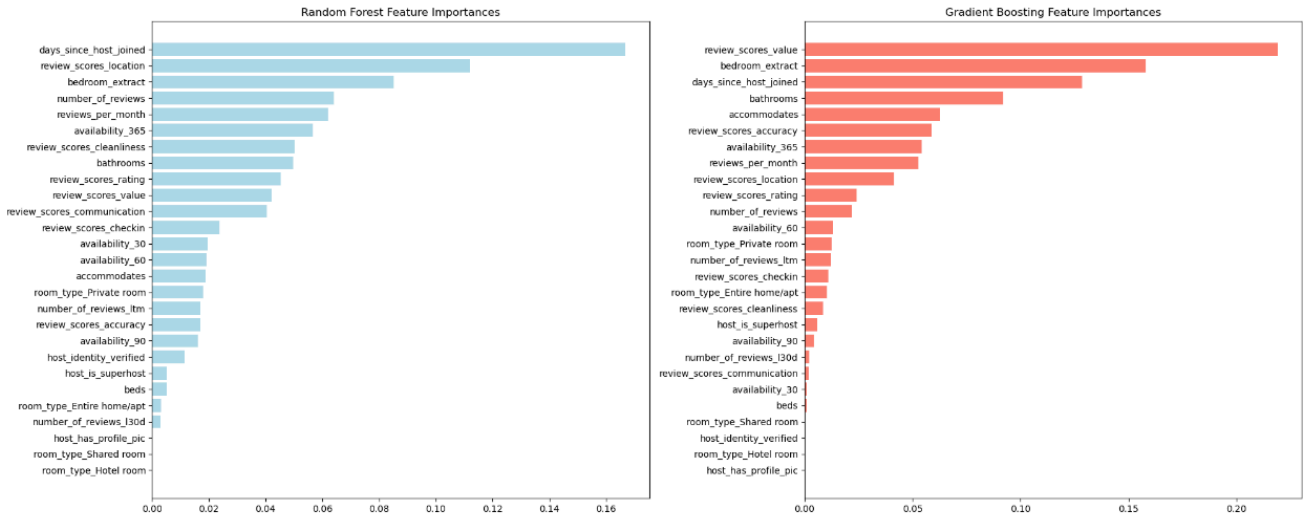
4.1. Model Selection:



- **RMSE (Training vs. Testing):** All models, except CART, show a reasonable balance between training and testing RMSE values. CART, however, shows a significant overfitting with an extremely low training RMSE compared to the testing RMSE.
- **R^2 (Training vs. Testing):** Similar to RMSE, the R^2 values indicate how well the models fit the data. The CART model demonstrates almost perfect fit on the training data but performs poorly on the testing data, further indicating overfitting. The other models show more balanced R^2 values, though still indicate room for improvement in model fit and generalization.
- **General Performance:** The Random Forest and Gradient Boosting models show a better balance between training and testing performance, suggesting they are more robust and generalize better than the OLS and CART models.
- **Model Selection:** Based on these results, Random Forest model appears to be the most promising model, offering a good balance between accuracy and generalization. Gradient Boosting also shows good potential.
- **Overfitting Concerns:** The CART model's performance highlights the importance of controlling for overfitting, especially when using single decision trees.

4.2. Feature Importance:

Due to the selection of Gradient Boosting and Random Forest model, the below plot will show top importance variables which might cause impact on the price.



From the plot, it can be observed that in both the Random Forest and Gradient Boosting models, `day_since_host_joined`, `review_score_value`, `review_score_location`, and `bed_room` contribute the most to the predictions. This implies that the age of the property (assumed based on the day the host joined), the review score, and the number of bedrooms are all significant factors influencing the price of Sydneysmall and mid-size apartments hosting 2-6 guests in both models.

Conclusion and Key Insights

These insights suggest that a property's established reputation, guest satisfaction with value and location, and the number of bedrooms are critical in setting rental prices for Sydney's small to mid-size apartments for 2-6 people. Property managers are advised to focus on these areas to enhance the attractiveness and competitiveness of their listings.

In conclusion, both the Random Forest and Gradient Boosting models prove to be valuable for this prediction task. Nevertheless, discrepancies observed in RMSE and R^2 between training and testing data raise concerns about potential overfitting. Hence, it is imperative to enhance the models using additional techniques such as tree pruning and cross-validation.