Summary Report:Predicting Airbnb Prices in Sydney

Question and the data

The primary objective is to develop a predictive model for estimating the prices of Airbnb listings in Sydney. This model will serve the purpose of aiding a company specializing in small and mid-size apartments for 2-6 guests to establish competitive pricing for their new apartments. The dataset, comprising details on approximately 24,000 Airbnb listings in Sydney, is sourced from Inside Airbnb and will be the foundation for our predictive modeling efforts. Please refer to the provided Technical Report for an in-depth exploration and cleaning of the dataset.

Regression models

I have defined eight regression models to predict the price of Airbnb listings in Sydney. The complexity of the models increases from the first to the last model and measured by the number of predictors used in the model. The models include such key predictors as the number of bedrooms, bathrooms, accommodates, and location. The Table below shows the performance of the models on the cross-validated training and test set.

Table 1: Comparing model fit measures

Model	N predictors	R-squared	BIC	Training RMSE	Test RMSE
(1)	1	0.28	181728	104.65	104.65
(2)	6	0.37	179812	98.00	98.02
(3)	50	0.45	178186	91.49	91.79
(4)	53	0.45	178010	90.86	91.20
(5)	232	0.49	178806	88.00	89.50
(6)	241	0.49	178814	87.77	89.34
(7)	255	0.50	178649	86.89	88.57
(8)	336	0.51	179256	86.32	88.80

All performance measures exhibit the expected behavior. As the number of predictors increases in the model, R-squared improves, and more complex models yield higher Bayesian Information Criterion (BIC). The training Root Mean Squared Error (RMSE) consistently decreases, and the test RMSE initially decreases, eventually demonstrating a gradual increase. The choosen model is the one with the lowest BIC because it is the most parsimonious model that explains the data well.

Random Forests Models and Gradient Boosting Machines

Initially, I conducted an auto-tuning process for the random forest model to identify optimal hyperparameters. The best-performing model, determined through this process, achieved a Root Mean Squared Error (RMSE) of 87. Subsequently, I proceeded to a systematic exploration, varying the number of variables in the model within the range of 8 to 12 and adjusting the minimum node size across values from 1 to 5. The refined model resulting from this experimentation yielded an improved RMSE of 86. The performance metrics of this finalized models are detailed in the table below.

Table 2: Random forest RMSE by tuning parameters

	vars		
nodes	8	10	12
5	86.1	86.1	86.0
10	86.2	86.2	86.1
15	86.3	86.2	86.2

The benchmark random forest model selected by algorithm within the defined hyperparameters is 500 bootstrap samples, 12 variables and minimum node size of 5. For the gradient boosting machine please refer to the Technical Report

Diagnostics on holdout set

To pick up the best model we will perform a test on holdout set of June 2023 prices. The performance of the models on the holdout set is shown in the table below.

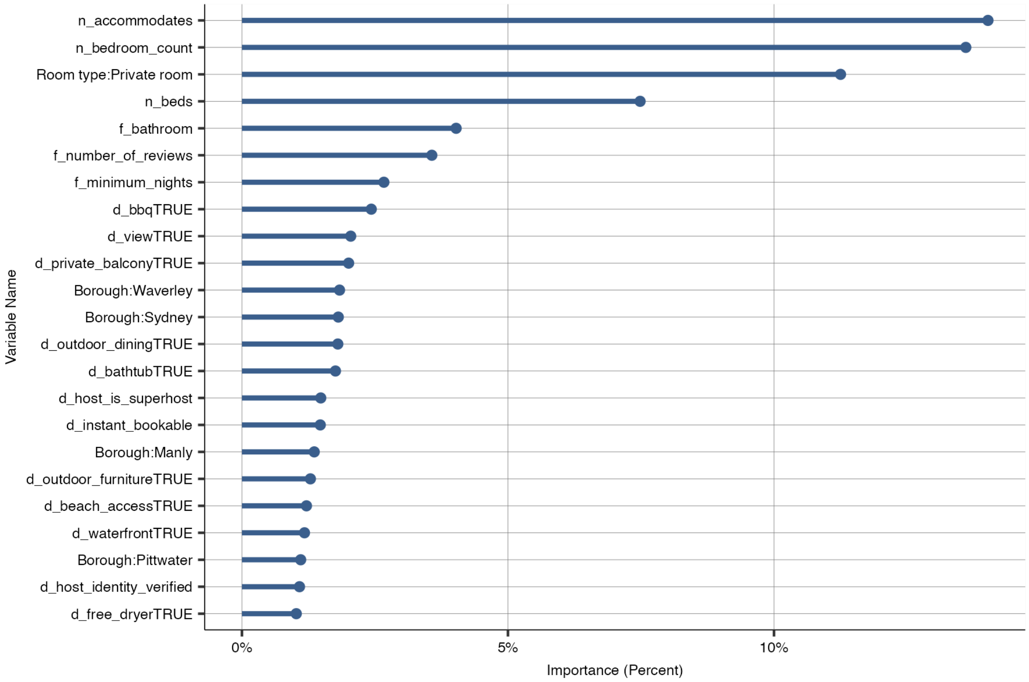
Table 3: Predictive Performance of the models

	Holdout RMSE
OLS	90.3
Random forest	73.8
GBM (basic tuning)	82.3

The random forest model exhibits the lowest Root Mean Squared Error (RMSE) on the holdout set, establishing its superiority as the optimal model for predicting Airbnb prices in Sydney. The marginally lower score compared to the test RMSE could be attributed to the similarity between the holdout set and the training set. The holdout set, representing June 2023 prices, is only marginally different from the training set, which comprises a sample of September 2023 prices, with a mere two-month gap between the two samples.

Feature importance

The next step is to look at the feature importance of the our random forest model. The plot below shows the top most important features.



The most important features are the number of bedrooms, bathrooms, accommodates, and location which is consistent with the reality. Larger accommodations and prime locations consistently emerge as significant contributors, reflecting the intuitive impact these attributes have on price.

Concluding Remarks

Typically, Gradient Boosting Machine (GBM) models outshine Random Forests in predictive tasks; however, in this specific caase, the Random Forest model takes the lead. One plausible explanation for this, is the potential lack of proper tuning in the GBM model. GBM models are known to be sensitive to tuning parameters, and in this case, basic tuning was employed for the speed. Additionally, the properties of the dataset could also be at play. Random Forest models, are known for their robustness, often demonstrate superior performance in scenarios involving small datasets and a limited number of features. The simplicity and stability afforded by Random Forests in such conditions could be a contributing factor to their outperformance over the GBM model in this specific context. While GBM models are powerful, their nuanced tuning requirements and potential sensitivity make them subject to performance variations, particularly when compared to the stability and adaptability of Random Forest models on smaller datasets.