Nightly Room Price Prediciton for Florence: Comparison of ML Models

Abdullah Duman

Introduction

The home-based tourism industry is developing day by day and it seems that it will continue to develop. Airbnb constitutes one of the most basic building blocks of this industry. This application offers travelers a wide variety of places to stay and homeowners who want to rent their homes for short periods of time to earn extra money. The pricing offered by these homeowners to their customers varies widely depending on the location of the house and the number of taps in the bathroom. In this project, based on Airbnb data, we will try to predict how much a company operating in Florence can rent its houses that can accommodate 2-6 people. We will try to find the answer to this question by creating a model with Machine Learning algorithms with the cross-sectional data obtained from the Airbnb application. In the data set, nightly prices of houses, room types, location, number of bathrooms, etc. There are data such as. The ultimate goal of this project is to try to obtain the best prediction by evaluating the results obtained from Machine Learning models with this data and feature engineering methods.

Data Managing and EDA

At the outset of our analysis, Florence Airbnb data was obtained from http://insideairbnb.com/get-the-data.html and uploaded onto a GitHub repository. This data was subsequently accessed via a Jupyter notebook for further processing. Following this initial exploration, insights into the structure and distributions of the dataset were uncovered through exploratory data analysis (EDA). Feature engineering techniques were then applied to extract meaningful insights from the raw data, including categorization and addressing missing values. A cautious approach was taken when removing rows with NA values to mitigate potential biases in sample selection. Recognizing the potential for missing values to alter underlying patterns and introduce sampling biases, the decision was made to discard all colums with NA values and rows with more than 70% missing values after further investigation. Emphasis was placed on both numerical and categorical variables in our analysis, encompassing factors such as accommodation capacity, room type, property type, and neighborhood. The aim was to incorporate as many pertinent variables as possible, including multiple variables that measured similar metrics (e.g., maximum nights and minimum nights). Additionally, extreme values that could potentially skew predictive models were addressed.

Creating the Models

Machine learning models, such as Random Forest, Ordinary Least Squares Regression (OLS), LASSO Regression, and Gradient Boosting Machines (GBM), were employed to forecast Airbnb prices. Each model underwent training using a blend of numerical and categorical features, targeting the listing price as the outcome variable.

Model performance was assessed using cross-validation techniques, scrutinizing the influence of various factors on price predictions. Additionally, feature importance methods from the scikit-learn library were employed to identify the most significant factors influencing price predictions. While the feature importances provided by the best estimator in scikit-learn are specific to tree-based models and are computed based on the frequency of feature usage in data splitting during training, permutation importance can be applied to any model, evaluating feature significance by assessing the performance change when feature values are randomly shuffled. Moreover, scikit-learn offers a function for generating partial dependence plots for a trained model. This functionality permits users to specify the features of interest and generates plots illustrating the marginal effect of each feature on the predicted outcome, while averaging across the values of other features.

Analyzing the Findings

As a result of the analysis, as can be seen in Figure 1, it was seen that the number of bathrooms was the most important determinant of nightly house prices in Florence. It can easily be said that more bathrooms mean more prices. In addition, it has been observed that the number of reviews and number of accommodations play an important role in pricing. Since OLS is more sensitive to non-linear regressions, unlike Random Forest and GBM, ln(price) was also evaluated as an alternative in OLS and LASSO models. Therefore, RMSE values were calculated for these 4 models using price and ln(price) values. Obtained RMSE results showed that Random Forest and GBM gave better results than OLS and LASSO. However, it cannot be said that the difference is significant. In addition, in the RMSE analysis conducted according to apartment size, it was seen that small apartments had higher RMSE values although they had lower price estimates.

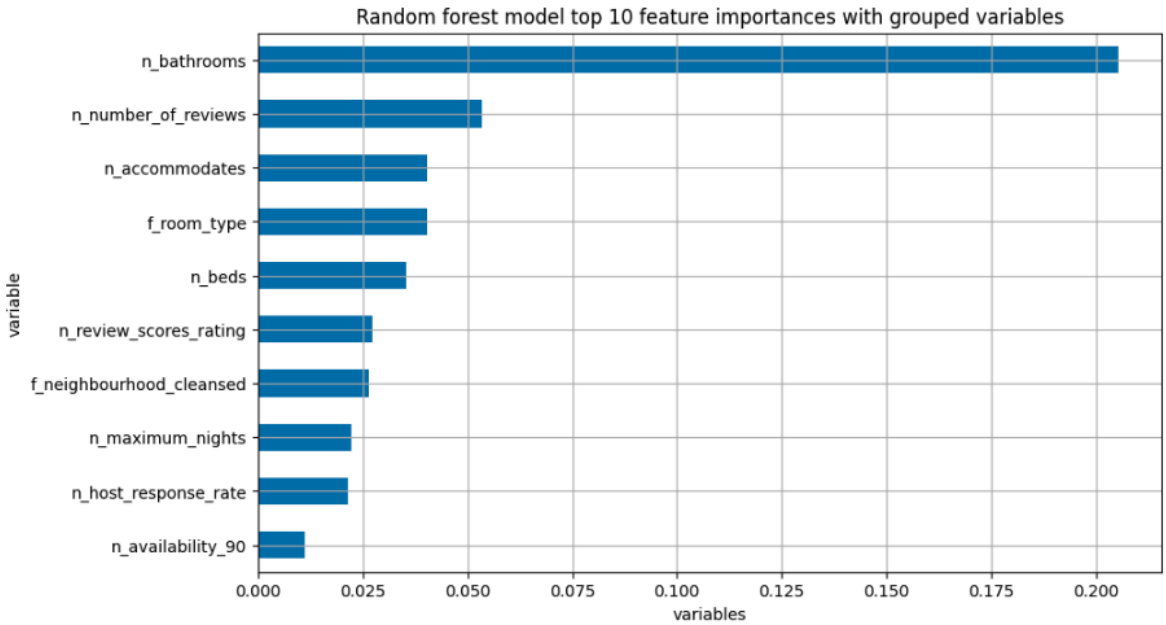

Figure 1(Feature of Importance Plot)

Then, we obtained the partial dependency graph of n_bathrooms, which has the highest feature of importance, as in figure 2, and the graph of another feature, room_type, as in figure 3, and the effect of the number of bathrooms and room type on the price was observed.

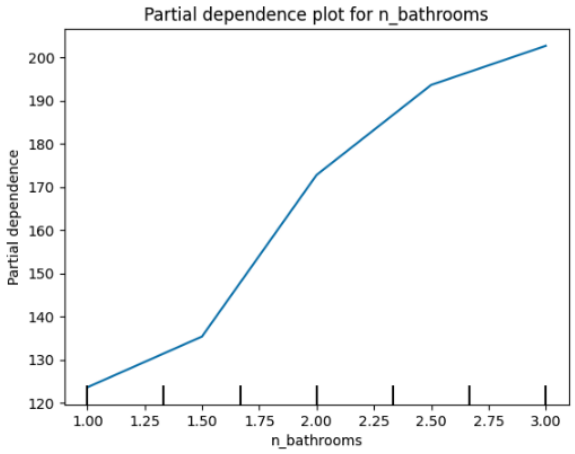

Figure 2 (Partial Dependence Plot for n_bathrooms)

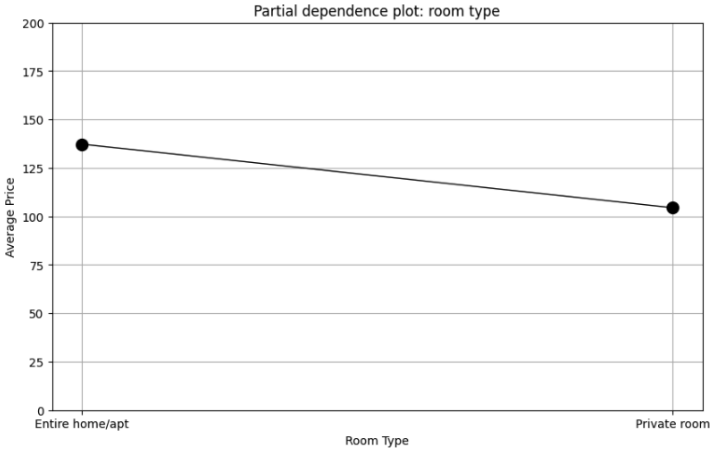

Figure 3 (Partial Dependence Plot for Room Type)

Figure 4, obtained using the SHAP (SHapley Additive exPlanations) algorithm, gave us values close to the feature importance table. However, while the first 2 values are n_bathrooms and n_number_of_reviews, the 3rd value appears differently as n_availability_90. The biggest reason for this is thought to be that while feature importing and often involves evaluating the direct impact of each feature on the model's

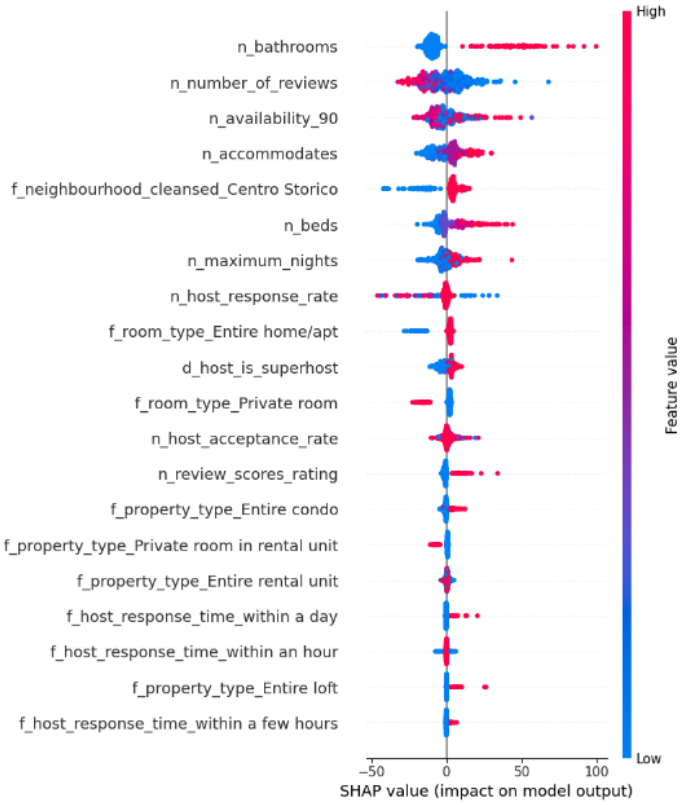

Figure 2 (SHAP Value Plot)

predictions, Shapley values measure the contribution of each feature in a prediction by considering the interactions with other features.

The graph in Figure 5 drawn for OLS, another model to be used in comparisons, shows the residual and predicted value values. It can easily be said that the distribution is dispersed, not accumulated, at point 0. This means that the OLS prediction model also gives more inaccurate results. The value in Figure 6 obtained using the ln(price) value shows that the points accumulate more at point 0, which means that the value obtained using ln(price) can give a more accurate prediction.

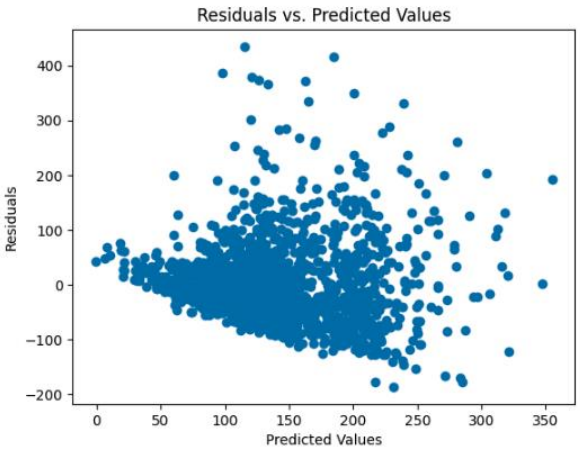

Figure 4 (level price Residuals and Predicted values)

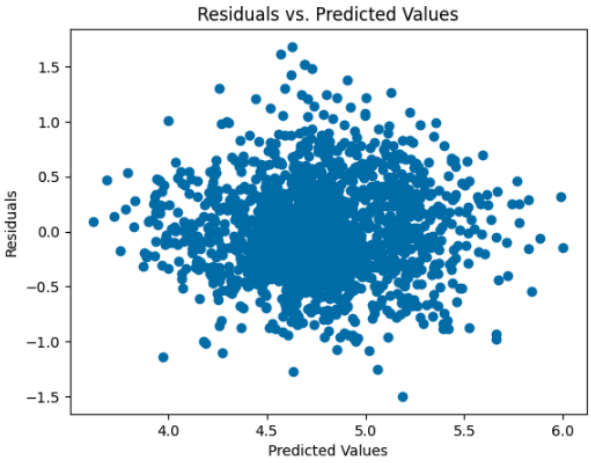

Figure 3 (ln(price) Residuals and Predicted Values)

Comparison of the Models

The results obtained for the models, both cross-validated RMSE (Root Mean Squared Error) and RMSE on holdout/test data, can be seen in Figure 7. While Random Forest and GBM give better results than OLS and LASSO, Random Forest gives better results for level price values and GBM gives better results for ln (price) values.

	Model	CV RMSE (Level Prices)	Test RMSE (Level Prices)	CV RMSE (Log Prices)	Test RMSE (Log Prices)
0	OLS	69.408654	67.642552	0.425960	0.406626
1	LASSO	69.245943	68.302988	0.459388	0.452637
2	Random Forest	66.816975	65.406494	0.400065	0.383598
3	GBM	66.096982	66.301386	0.395622	0.378350

Figure 5 (Model Comparison Table)

Additionally, it can be easily seen that the test RMSE values are lower than the train RMSE values. However, this situation should not be considered normal. Because test RMSE values are generally expected to be higher. The main reason for this is thought to be overfitting of the model to the train set.

Conclusion

Finally, the project can be summarized as like that we firstly commenced our analysis using Airbnb data for Florence, which encompasses numerous qualitative variables along with a handful of quantitative ones. Employing exploratory data analysis (EDA) and feature engineering techniques, we leveraged our domain expertise to craft estimation models. Initially, we opted for a Random Forest model and meticulously scrutinized to identify the key factors influencing prices. Subsequently, we gauged the overall performance of the base Random Forest model against three alternative models: OLS, LASSO, and GBM.

In terms of overall performance, GBM emerged as the frontrunner for lowest RMSE across cross-validated samples and Random Forest gave the best result for the test dataset. Interestingly, while room type proved to be the most influential factor in London's data, the number of bathrooms took precedence in Florence. Similarly, the second crucial factor in London was the number of bathrooms, whereas in Florence, it was the number of reviews. Moreover, the number of accommodate emerged as the 3rd most important variable for both London and Florence. Notably, certain neighborhoods in both London and Florence exhibited tendencies to either elevate or diminish accommodation prices.

For this project, it was preferred to drop missing values. However, other methods can also be used. Especially having a pattern of missing values may cause us to find incorrect predictions. Additionally, since there were many variables in the data set, the analysis was made by selecting among them. You can proceed deeper by selecting different variables.

Thus, we have completed an informative and predictive project for hosts who want to do Airbnb business in Florence and for guests who want to visit Florence. I think we have shown that data analysis with high predictions can be made by employing data-driven methods and combining Machine Learning algorithms.