Model Comparison & Results Interpretation

The aim of this project is to propose a data-driven solution, by using machine learning to predict rental price Airbnb Amsterdam listings. I have followed the phases of IBM Methodology for Data Science how to make a machine learning project from scratch. I started with cleaning the data and getting insights from it. I followed by doing feature engineering and feature selection. This is probably the most important step in a machine learning project as it helps in increasing accuracy, avoiding over-fitting, and reduce the training time of the model. I plotted the feature importance to find out the most important features in predicting the house prices by correlation heatmap and the SelectKBest, f_regression. I used Random Forest Regressor, Linear Regression, Decision Tree Regressor, Lasso, and Support Vector Regressor(SVR) as the model for regression with R^2 and (MAE) median absolute errors as the evaluation metric.

For price prediction, I started with linear regression and Lasso models with the features are selected by the SelectKBest. Then I reduced the features which I got from feature engineering to my base model and I was not able to improve the overall accuracy for both regression models. Then I also evaluated the performance of tree-based regressors as well as SVM regressors on this data. I saw that the tree-based regressors were performing better while the SVM-based regressors took a lot of time to train and their performance was not as good as the random forest regressor.

Conclusion

In this study, I used different machine learning algorithms to predict Airbnb's Amsterdam listing price and rating. Two different datasets were used to train the models. The transformed dataset contains 25 columns, whereas the corr dataset contains 5 columns by using a correlation heatmap.

- After rigorously testing all of the models defined above, the model that consistently performed the best was the Random Forest Regressor with the transformed dataset. Out of the five, random forest typically reported r-square (R2) score in the (0.34 - 0.44) range with 25 features and the (35-37) range with 5 features, around €24 median absolute errors and in the (39 - 43) range RMSE score,
- Despite nothing has changed on tuned Linear Regressor reported r-square (R2) score as 0.43 with 25 features(transformed dataset) and 0.35 with reduced correlated features(corr dataset) and €24.5 median absolute errors with almost 40 RMSE score.
- Decision tree reported the lowest r-square (R2) score in the (-0.1 ~ 0.40) range by the transformed dataset with €24.5 median absolute errors and in the (41- 55) RMSE score,
- Lasso reported the same r-square (R2) score with Linear Regressor as 0.43 with by tuning transformed dataset whereas the corr dataset reported r-square (R2) score around 0.35, and €24.36 median absolute errors and 40 RMSE scores.
- The last applied model, the support vector regressor(SVR) reported an r-square (R2) score in the (38-42) range by transformed dataset whereas the corr dataset reported an r-square (R2) score in the (0.34 - 0.36) range with €24 median absolute errors and reported an RMSE score as 41 like previous 2 models.

Comparing the spread of each model's errors against the test data confirms the Random Forest Regressor on the transformed dataset as the most accurate of the five.