ECE4710J Project 1 Part 2

Yiwen Yang 519370910053

In this part, different regression models were tested on the given data. Finally, random forest regression was chosen to be the best model that I can use to minimize RMSE. In this model, bootstrap without out-of-bag scores (since oob data do influence both train and test RMSE by a lot) is used. The maximum sample size is varied to see whether it will impact on the result; Figure 1 shows that selecting more samples for each tree performs better, so 100% of the test split data are used.

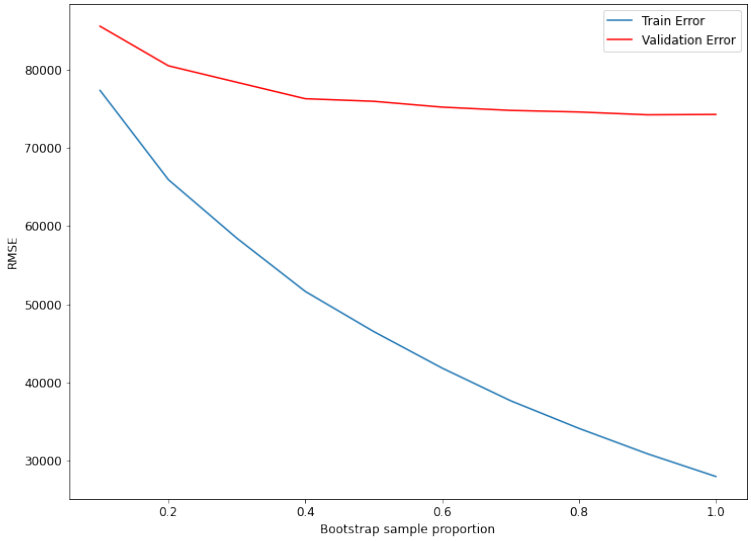

Figure 1: RMSE against bootstrap sample size.

Original data contains too much features among which many are useless, so dimension reduction was performed to choose those with larger influence calculated by entropy change. Then around 30 potentially significant features are selected to fit the model. Since there are categorical features such as "Roof Material", One-hot encoding was performed to make the regression more accurate.