IE 590 – Predictive Modelling Final Exam

**NAME : Mihir Bhatia**

**PUID: 0029923247**

1. **Describe how you fit the model. If you used data transformations, you need to clearly discuss it.**

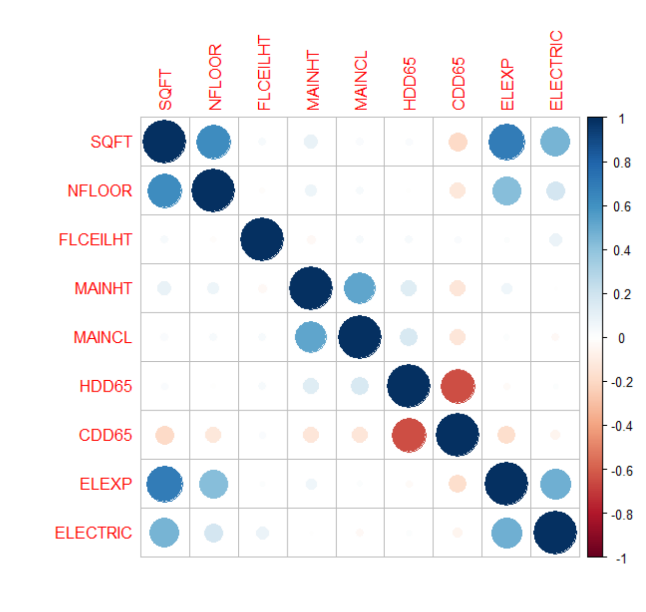
The following steps were taken to get to the final model:

## Data Preprocessing

Firstly, data pertaining to the pacific region was filtered out. Secondly, the dataset was explored and variables that were repetitive and obviously did not contribute to predicting the electricity usage from space conditioning were not considered. Lastly, variables with where the NA values were too high were removed. Moreover, the *complete.cases()* function was used to filter out rows without N/A values.

## Exploratory Analysis

Since the dataset consisted mostly of categorical data and a few variables that were numeric, it didn’t make sense to use Principal Component Analysis for variable selection because PCA would have an issue of representing the distances between variable categories in factorial space. Also, since the data was categorical, correlation plots for the categorical variables wouldn’t make much sense.But the corrplot() for numeric variables was plotted:

There was a high negative correlation between CDD65 and HDD65 (which makes sense because – higher the number of cooling degree days – lower the heating degree days tend to be).

From the above corrplot we also observe a high correlation between electricity expenditure, square feet and number of floors which is obvious taking building size into consideration.

## Data Transformation

The column $ELECTRIC was obtained by adding the values of electricity usage for heating, cooling and ventilation. A log transform was applied on the target variable $ELECTRIC. This was done because. Logarithmic Transformation was applied because of the following:

1. Certain parametric models (GAM, Linear Model) dealt poorly with this skewed data and gave low predictive performance in comparison to other models – although this wasn’t the
2. The range of the ELECTRIC data went from 0 to approximately 127 million with a median of 48,341.
3. The number of outliers of considerably reduced by using the log transformation.

## Models Applied

1. **Linear Model**

The linear model was run to check the important variables that may linearly correlated. But the RMSE value obtained from this model was very high.

1. **CART & Random Forests**

Since we were dealing with categorical data, the tree based models performed relatively well with Random Forests being the best

1. **GAM**

The GAM function is a combination of multiple smooth functions. The GAM didn’t perform very well.

1. **MARS**

Since MARS is combination of several splines that stitched together, MARS performed well for this mix of numeric and categorical data (although this was unexpected) and was able to give the second lowest RMSE value (after random forests)

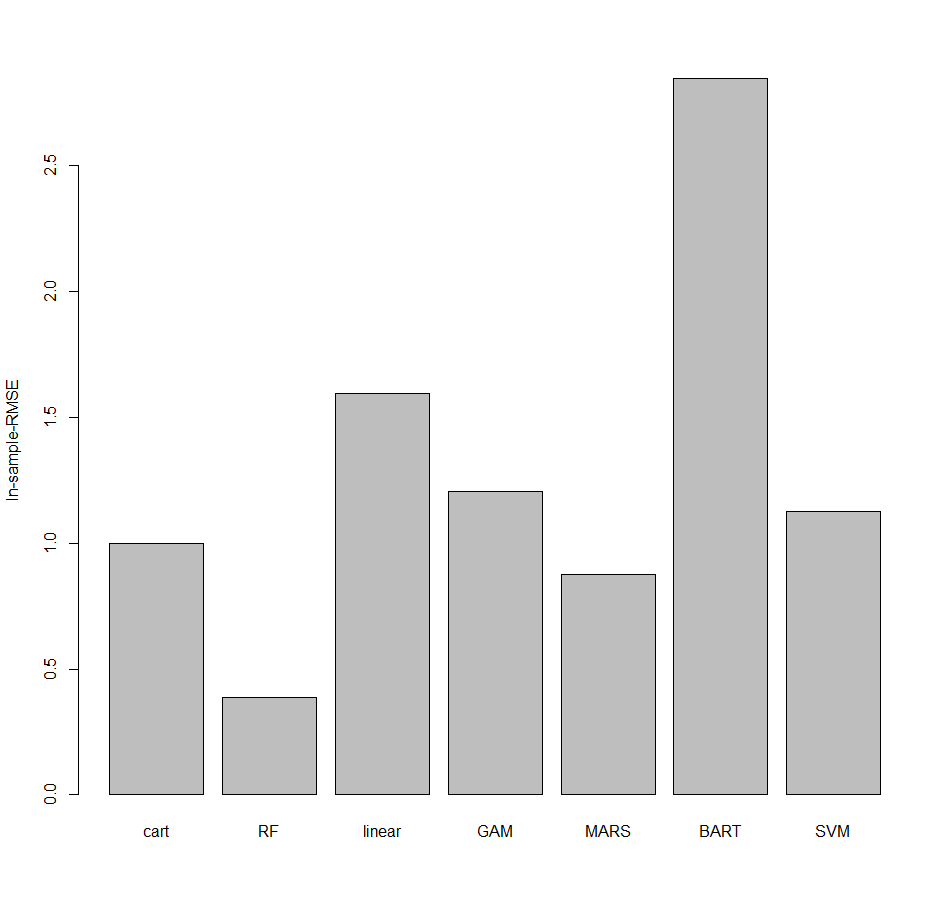
1. **BART**

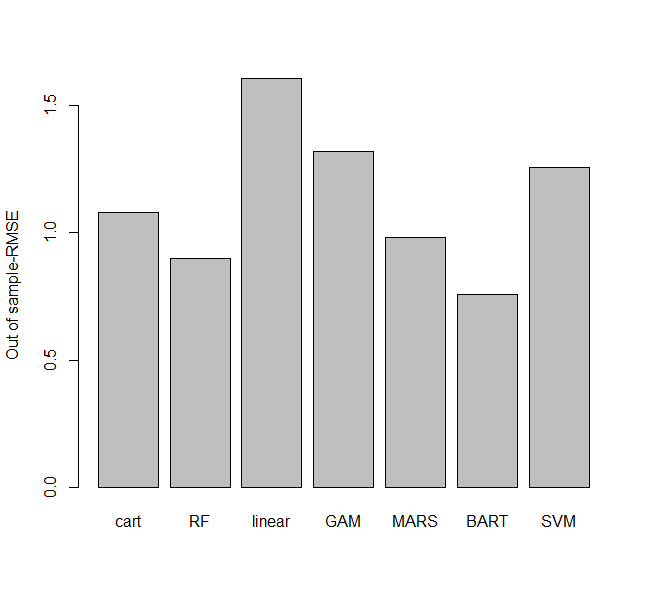
BART model was trained and tested. This model usually performs the best in cases where the dataset is large. In my case, since my dataset has only 751 data points, I think BART behaves in a weird, i.e. the in sample RMSE is much higher than out of sample RMSE (this happened for all folds of the k fold cross validation that was applied)

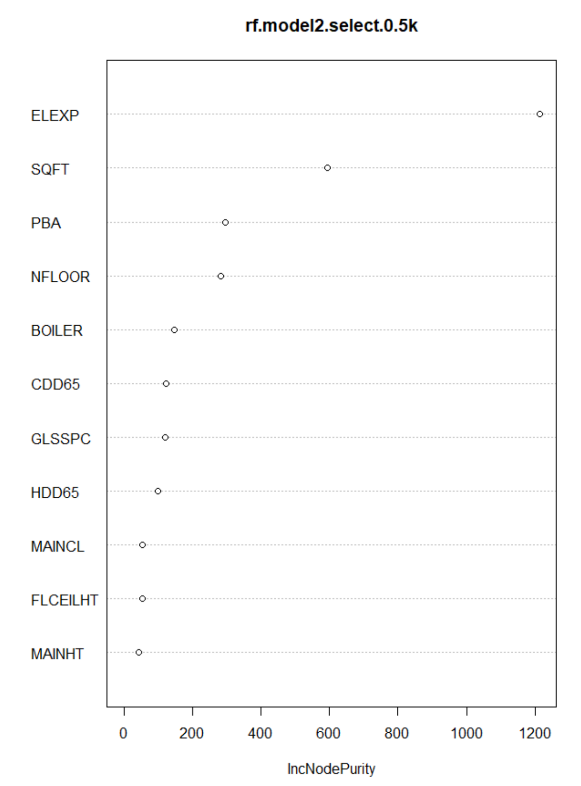
1. **SVM**

SVM was also performed for with all 30 selected variables. But it did not perform well. Thus this model was not tuned further.

Summary of model In sample and Out of sample performance





MARS pruned and unpruned models were run on with all 30 variables in final dataframe, and both didn’t have any significant difference in performance (from paired t-test). Moreover, the RMSE value was higher than that of random forests and hence this model was not tuned further

The best model was found to be random forests from these models. But these models were tested out on either all or a larger no. of important variables. Since random forests was the best performer and the data is mostly categorical, I decided to further go on to tune the random forests model.

After running 10- fold cross validation on all the above models, based on RMSE values of these models. **Random Forests model was selected.**

Using the variable importance plot of random forests, I started removing variables from the 500 tree model and test random forest model for various variable combinations.

I was able to bring it down to 10 variables without significant depreciation in model performance.

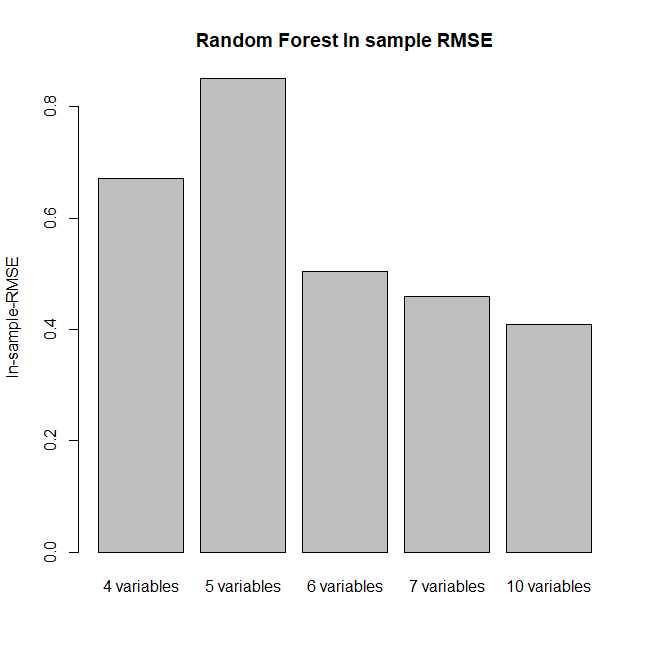
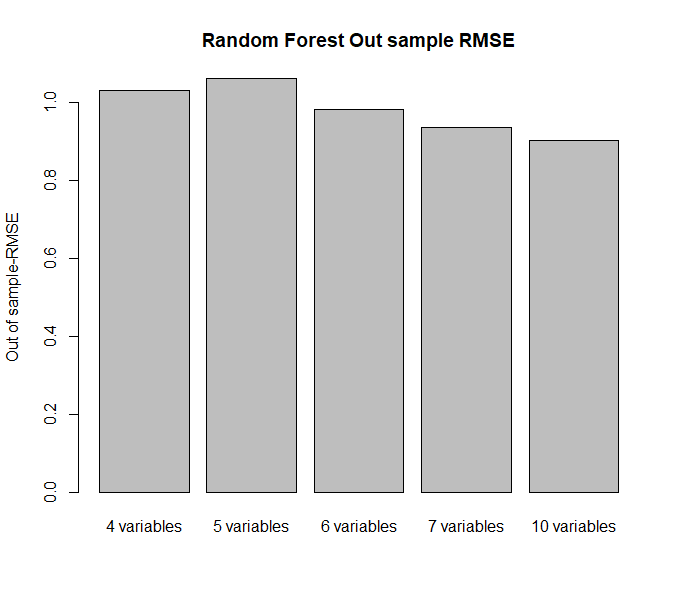
1. **Describe the final model, including equations or a figure encapsulating the final model or giving the final model with all parameters specified numerically.**

Random Forests is an extremely powerful algorithm that can learn data quickly. Moreover, there is also no fear that overfitting will occur, because it takes the average of a large number of trees and uses BAGGING for variable selection for each of the trees. This turned out to be very effective and yielded the lowest RMSE value amongst all the models.

NOTE: The “finalmodel” is a Random Forest model that has been trained on the entire data (i.e. not on training or testing set). This RF model was trained on 10 important trained variables as this yielded the lowest RMSE value.

1. **Justify why you selected your final model (include model diagnostics, out-of-sample accuracy results etc.).**

Summary of random forests models for with 500 trees

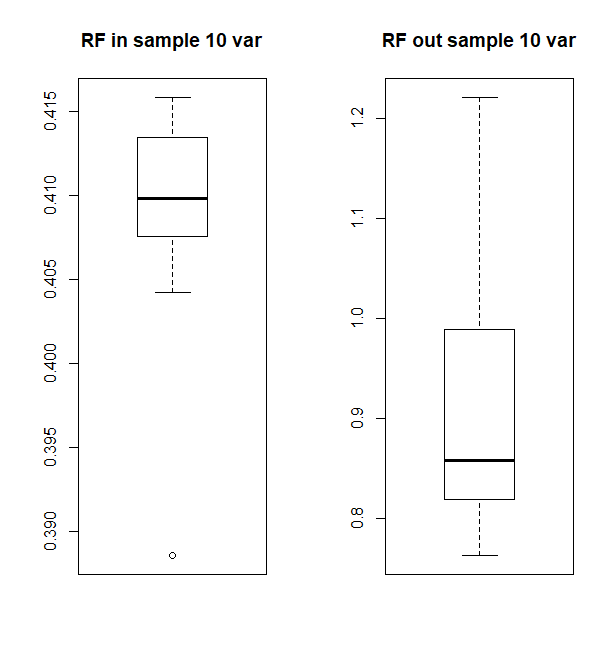


From 10 fold cross validation the final model was selected based on RMSE values. The random forest models have 10 variables

mean(0.4055617

mean(0.91391

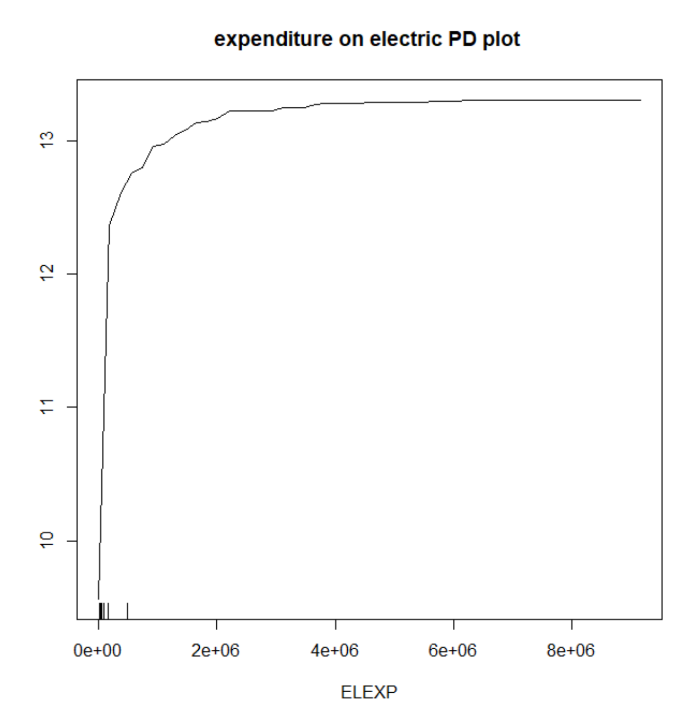
**Boxplots of in-sample and out-of-sample RMSE values for the best model**

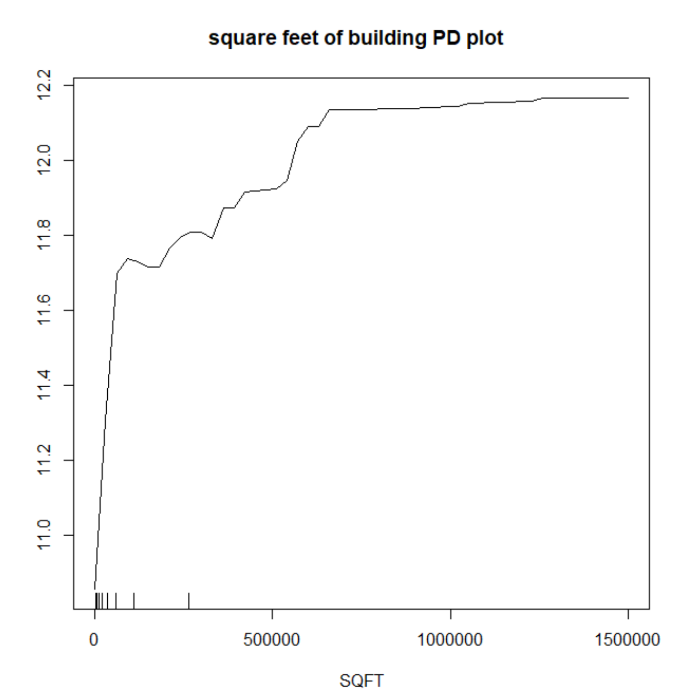


There are no outliers in the RMSE values that are distorting the RMSE mean values that final model selection has been based on.

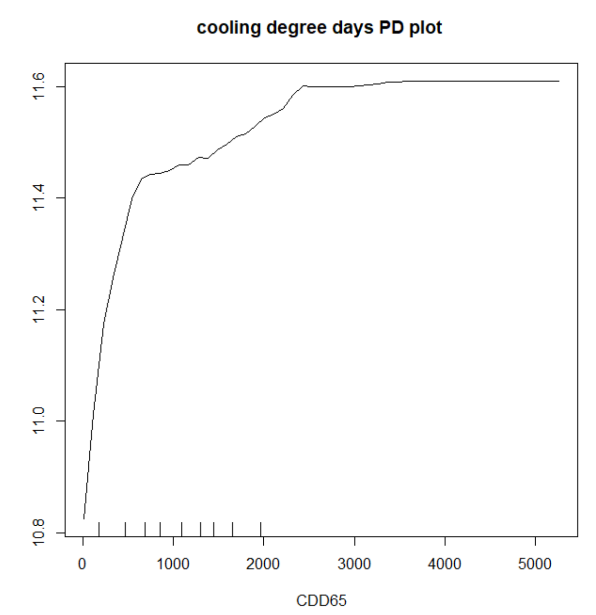
1. **Include a section on model inference**

Model inference can be done using partial dependence plot generated from Random Forests





From the electricity expenditure PD plot, we find that higher electricity is directly proportional to electricity usage for space conditioning. Also, with increase in square footage of the building there is an increasing trend of electricity usage.



Increase in cooling degree days and no. of floors also increase electricity usage, but in the case of cooling degree degree days, not much change seems to occur on the higher side.

NOTE: There were other partial dependence plots, but interesting trends are not visible in them.