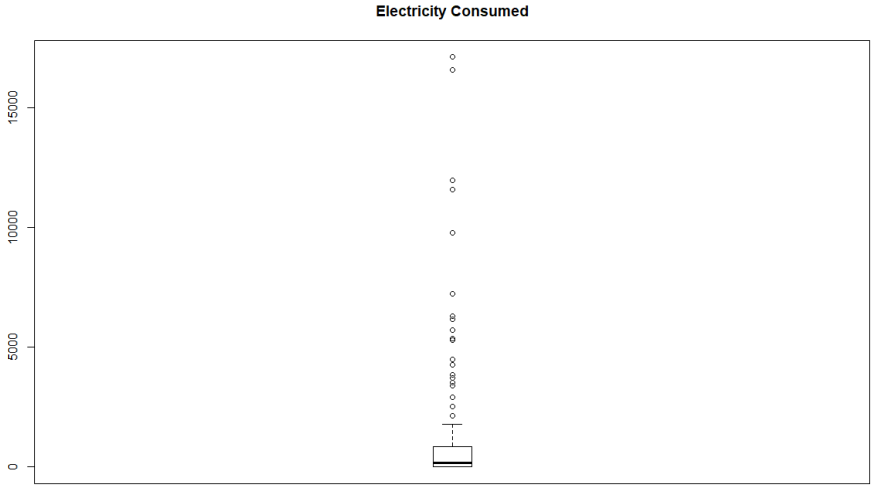
PREDICTIVE MODELING – IE 590
GAGANDEEP SINGH KHANUJA
0029971620

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

Data Pre-processing:

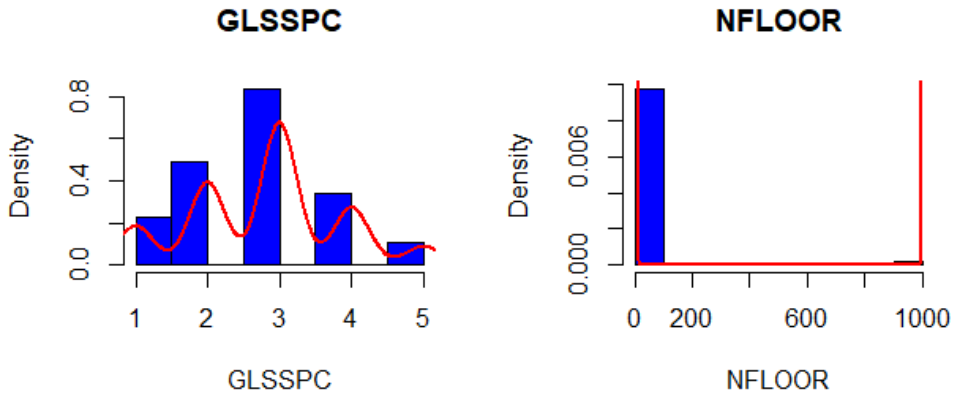
1. The problem statement clearly mentioned me to evaluate and predict the Electricity consumed in the New England region based on the following Space Conditioning Parameters like Heating, Cooling and Water Heating.
2. The dataset had 6720 observations of various regions out of which only 319 observations belonged to New England region.
3. The dataset was well scrutinized, and it was found out that it had a lot of unimportant observations (with respect to the parameters asked) and missing values.
4. On using various techniques, these missing values were dropped out from the dataset with only 148 observations left in the dataset.
5. The system has been finally fitted on 148 observations with 56 predictors.
6. The response variable i.e. Electricity consumed is the **summation** of the three parameters i.e. Electricity Heating use, Electricity Cooling use and Electricity Water Heating use.
7. There were two **outliers** in the dataset that have been removed from the dataset as the Electricity dataset is highly skewed. It was observed that, the skewness had a very big impact on the RMSE values. The skewness of the dataset has been visualized via Boxplots.



Boxplot of Electricity Consumed
(Before removing 2 Outliers)

8. Data transformation techniques like **log transformation** have been implemented on the response variable (i.e. Electricity) in the model. But no significant effect can be seen on the RMSE values when it was fitted on the models.

9. **PCA** Analysis was also done on the selected 56 variables, since my objective was to profile the variables. The biplots obtained for my variables were not in a good resolution; R does not process those plots well for many variables. Based on the plots obtained I could conclude that my variables did not form explicitly profiles, since most were Categorical Variables.
10. **Density** plots were used for visualization of the dataset. The plots were used to visualize the concentration of the dependent variables. The histograms helped me in visualizing this data in a better way and show me the frequency distribution of the dataset.



11. **Encoding** was used to create dummy variables as the dataset is categorical and the results were compared. It was found out that there was no change in the rmse values or r square values when the dataset was in the numeric format or in the encoded format. The snippet has been attached in the code for your reference, so that you could implement and check it, if at all you would like to check.

Models Fitted

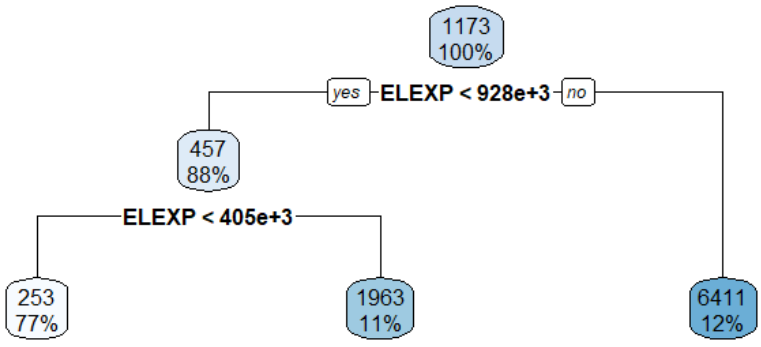
The statistical and Machine learning models that were used in the execution of this dataset are as follows:

1. Linear Regression.
2. Decision Trees.
3. Random Forest(RF)
4. Multi Additive Regression Splines(MARS) (Unpruned)
5. Multi Additive Regression Splines(MARS) (Pruned)
6. Bayesian Additive Regression Trees(BART)
7. Support Machine Vector(SVM)
8. Neural Networks.

Model Building

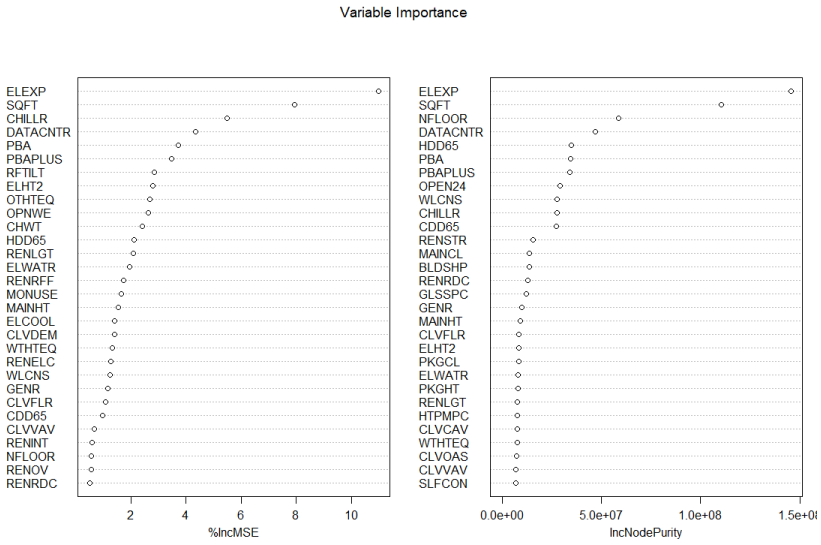
1. **Linear Regression:** - General linear model (GLM) with all the variables left after data cleaning is run on the training set to find the significant parameters. It has been found out that the Linear Models perform the worst of all the models. The RMSE of the LM is 3036.202 (Out Sample). Since the data is categorical it is expected that the Linear Models and the GAM Model will perform worse. The plots for GAM Model was also, plotted and it was found out that step.gam couldn't be used for smoothing as the plots showed linearity. Hence, these models were dropped off from evaluation and consideration.

2. **Decision Trees**: - RPART regression trees builds classification or regression models of a very general structure using a two-stage procedure; the resulting models can be represented as binary trees. Typically, you will want to select a tree size that minimizes the cross-validated error. The RMSE (In Sample) for rpart comes out to be 1767.566 and RMSE (Out Sample) is 1891.546 which is very high compared to others. This is because the decision tree changes when the dataset is perturbed a bit. This reduces the robustness of the classification algorithm to noise and isn't able to generalize well to future observed data. This can undercut confidence in the tree and hurt the ability to learn from it.



Decision Trees Plot

3. **Random Forest** is used as it is one of the most accurate learning algorithms available. It produces a highly accurate classifier and runs efficiently on large databases. It reduces overfitting by averaging several trees and reduces variance by using multiple trees which avoids the inclusion of a classifier that doesn't perform well between training and testing data. The significant variables given by Random Forest are: ELEXP, SQFT, DATACNTR, NFLOOR, WLCNS, CHILLR, HDD65, CDD65, PBAPLUS, PBA, OPEN24. The RMSE value for random forest (Out Sample) is 1247.883 which is the lowest and hence this shows that it learns from the training data and helps predict on the test data very well.



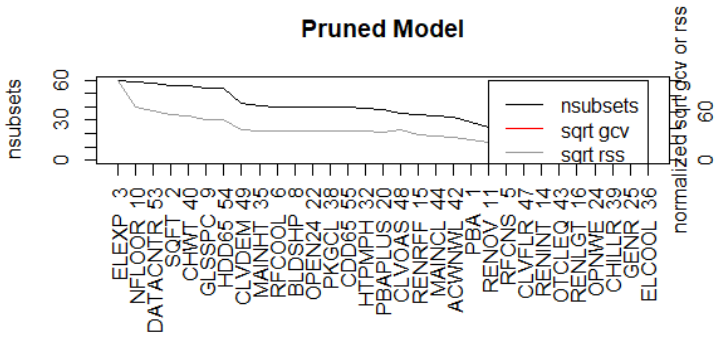
Variable Importance Plot: Random Forest

4. **Multi Additive Regression Splines**(MARS): MARS is non-parametric regression technique which builds linear models which allows for `piecewise continuous linear' models to represent linearities in the data.

   There are two types of MARS models:

4A. **Pruned MARS**

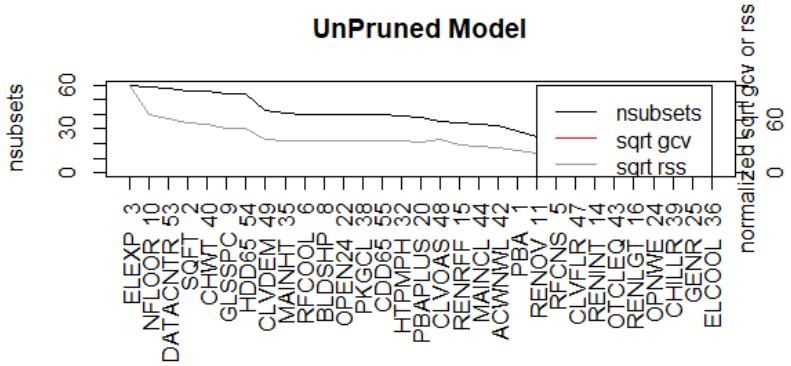
   Variable Selection in MARS (Unpruned): - Based on the Variable Importance plot obtained in the un-pruned MARS, following variables have been selected ELEXP, DATACNTR, HDD65, NFLOOR, CHWT, BLDSHP, SQFT, GLSSPC, MAINHT, RFCOOL.



Variable Importance Plot: Pruned MARS

4B. **Un-Pruned MARS**

Variable Selection in MARS (Unpruned): - Based on the Variable Importance plot obtained in the un-pruned MARS, following variables have been selected ELEXP, DATACNTR, HDD65, NFLOOR, CHWT, BLDSHP, SQFT, GLSSPC, MAINHT, RFCOOL.



Variable Importance Plot: UnPruned MARS

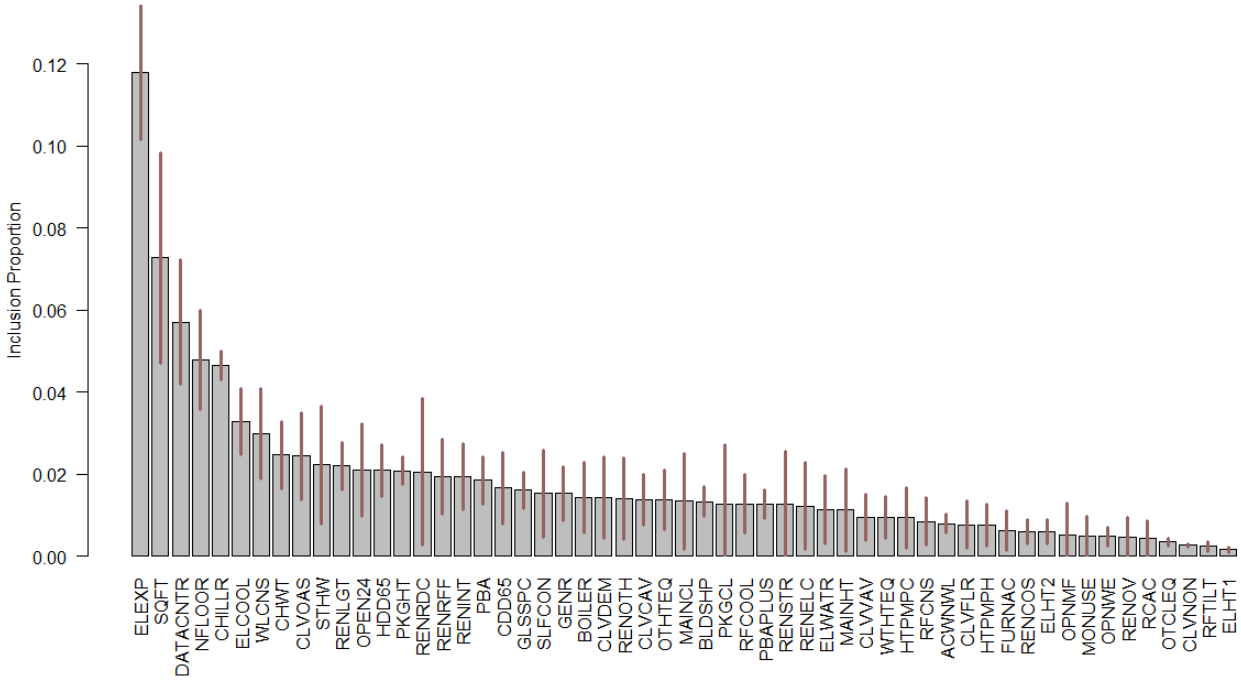
5. Support Vector Machine

Parameters: kernel, cost and gamma. For the kernel, a cross validation was run for linear, radial and polynomial types, gamma and cost constant. Values are mean of the 10 cross validation values.

RMSE.train.linear	RMSE.train.radial	RMSE.train.poly
96299.968	2008.185	73936814.737
RMSE.test.linear	RMSE.test.radial	RMSE.test.poly
75303.610	2090.833	12459060.091

With those results, we chose radial kernel. A second parameters tunning was performed to chooses cost and gamma in a model fitter with the radial kernel. The function svm.tune was used with possible cost and gamma:
Cost: 10
Gamma : 0.5

6. **Bart Machine**
   Using the bartMachineCV function, I had the winning model as bartMachine CV win: k: 2 nu, q: 3, 0.99 m: 200, which means, 200 as the number of trees, and hyper parameters alpha= 0:99 and beta = 2, k = 2 and v = 3. The function investigate var importance was used to check the variable importance of the bart models.



Variable Importance Plot: BART Machine

The BART Model on Variable Selection ran only for 2 iterations in the 10 fold cross validation and gave me a heap space error. The results would have been biased hence, even though it can be seen the RMSE values of BART is the least but still I haven't considered it as the Best Model.

```
Error in .jcall("RJavaClassLoader", "[B", "toByte", .jcast(o, "java.lang.Object")) :
  java.lang.OutOfMemoryError: Java heap space
```

7. **Neural Nets**

A 10 fold Cross validation function was fit in the model first to find the best model out of the 4 model, based on the values obtained from this model, tuning parameters were selected.

rmse.train.1	rmse.train.2	rmse.train.3	rmse.train.4
2539.588	2539.588	2026.906	2307.152
rmse.test.1	rmse.test.2	rmse.test.3	rmse.test.4
3120.483	3120.483	2664.28	3146.737

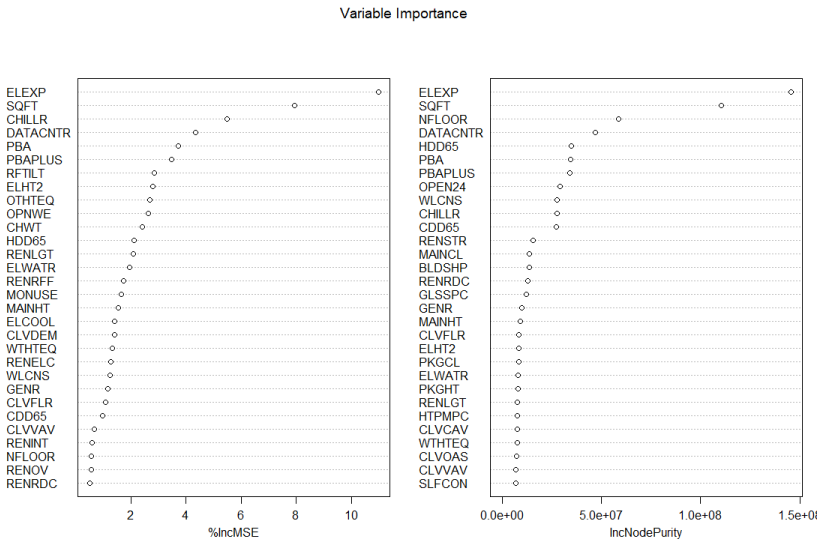
Model 3 was selected as it had the least RMSE for In Sample and Out Sample.

Question 2. Describe the final model, including equations or a figure encapsulating the final model or giving the final model with all parameters specified numerically.

The main criteria of comparison between different models was the RMSE values. The RMSE value shows how a model performs on the test data set after learning from the training dataset. The lower the RMSE the higher is the predictive power of the model.

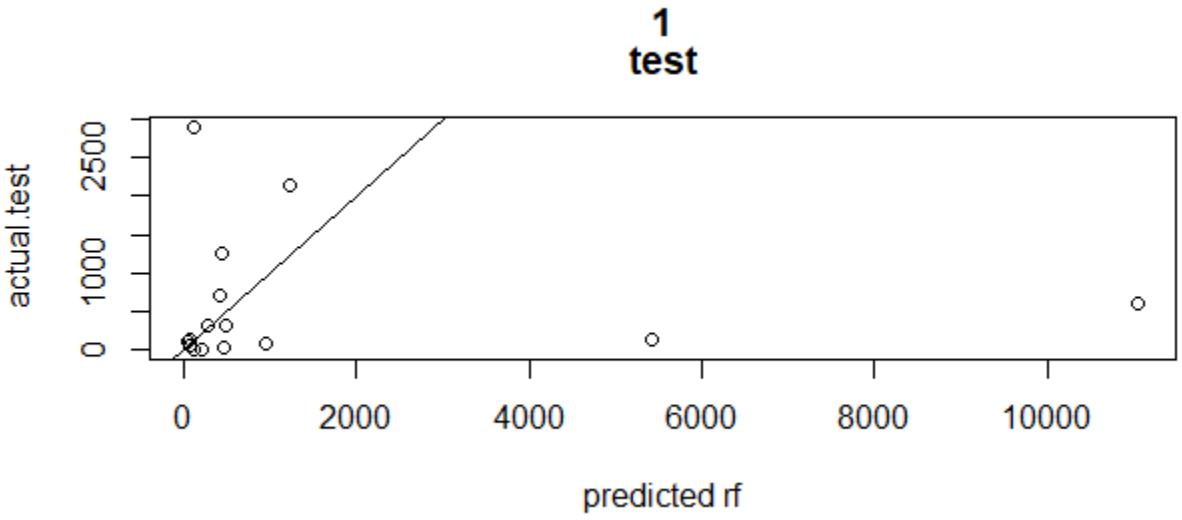
The following were the justification points for model selection:

1. It gave the lowest RMSE = 1247.883
2. It performs better than all the other models because it reduces overfitting by averaging several trees. Also, it reduces variance by using multiple trees which avoids the inclusion of a classifier that doesn't perform well between training and testing data.
3. Independent training of each base classifier on a training set sampled with replacement from the original training set. As the number of trees increase the error decreases. This technique is known as bagging, or bootstrap aggregation. In Random Forest, further randomness is introduced by identifying the best split feature from a random subset of available features.
4. Overall it has good accuracy, stability and interpretability.



Variable Importance Plot: Random Forest

It was found out that the significant variables given by Random Forest are: ELEXP, SQFT, DATACNTR, NFLOOR, WLCNS, CHILLR, HDD65, CDD65, PBAPLUS, PBA, OPEN24.



Actual Vs Predicted Plot: Random Forest

The residual plot shows the goodness of fit of the model as the rmse for the given model is the least. A few outliers are shown in the plot, but it performs better than the rest of the models.

Question 3. Justify why you selected your final model (include model diagnostics, out-of-sample accuracy results etc.).

RESULTS

Using 10 foldk	LM	DT	RF	SVM	NN	MARS(UP)	MARS(P)
RMSE (In Sample)	1418.452	1767.566	700.2868	1042.439	2120.16	1345.495	1431.407
RMSE (Out Sample)	3036.202	1891.546	1247.883	2483.564	2199.28	1746.54	1656.381

In Sample (k fold)

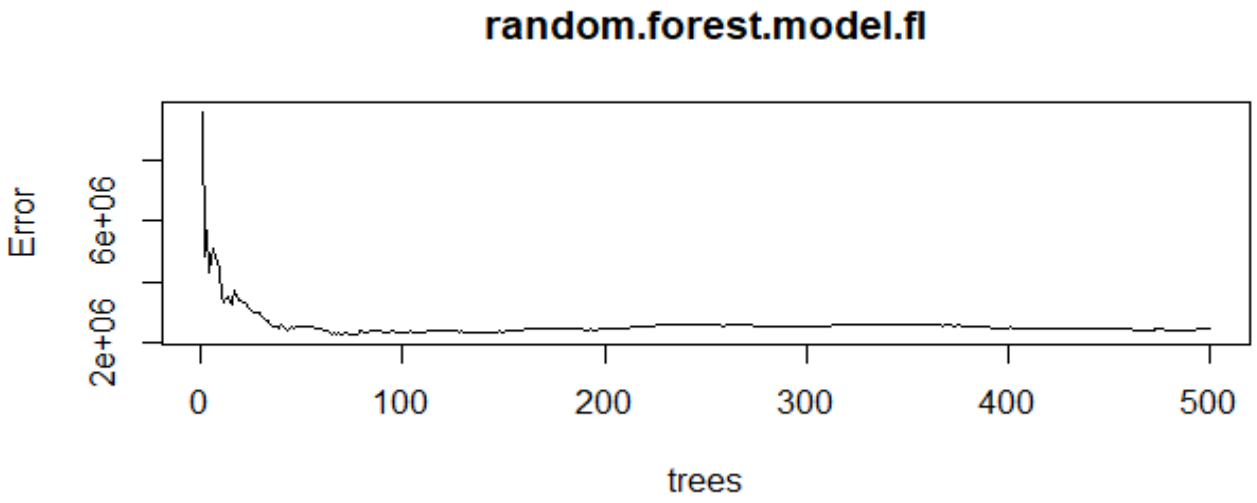
LM(IS)	DT(IS)	RF(IS)	SVM(IS)	NN(IS)	MARS(Unpruned)	MARS(Pruned)
1499.86	1943.68	706.17	281.57	2536.06	1384.13	1471.47
1452.43	1362.64	735.91	2903.56	2729.64	1395.50	1507.45
1248.59	1632.29	633.86	230.65	1562.04	1243.95	1342.90
1408.92	1855.73	704.34	282.16	2341.41	1407.57	1529.41
1521.82	1742.60	729.59	275.22	2600.24	1378.99	1453.24
1500.28	1908.01	720.95	277.03	2089.57	1402.80	1516.63
1538.94	1910.98	685.11	3007.16	2319.28	1432.40	1519.19
1425.94	1869.05	719.10	281.56	2314.25	1353.21	1424.99
1160.32	1654.62	691.34	2620.73	1559.77	1105.39	1168.56
1427.43	1796.04	749.06	264.75	1813.93	1351.02	1380.23

Out Sample (k fold)

LM(OS)	DT(OS)	RF(OS)	SVM(OS)	NN(OS)	MARS(Unpruned)	MARS(Pruned)
2064.41	1055.11	1011.26	1199.17	788.60	1379.92	1191.65
2370.34	1222.01	631.64	2953.48	2687.51	1211.01	1087.90
3961.11	3519.50	2536.49	4925.70	3570.79	2624.99	2742.32
4291.19	1874.03	900.87	1196.26	1647.95	1413.56	1348.50
1553.52	1152.04	783.94	1975.59	1830.07	1589.79	1497.92
4493.35	1310.19	643.98	1759.02	1391.73	1294.56	959.01
1280.89	1234.53	1129.69	1715.57	1744.61	651.21	874.91
2898.81	1780.42	765.02	1343.86	1370.55	1848.17	1712.50
4326.21	3366.91	2652.67	4767.56	3453.72	3404.02	3255.78
3122.18	2400.71	1577.92	2999.43	2410.95	2048.20	1893.33

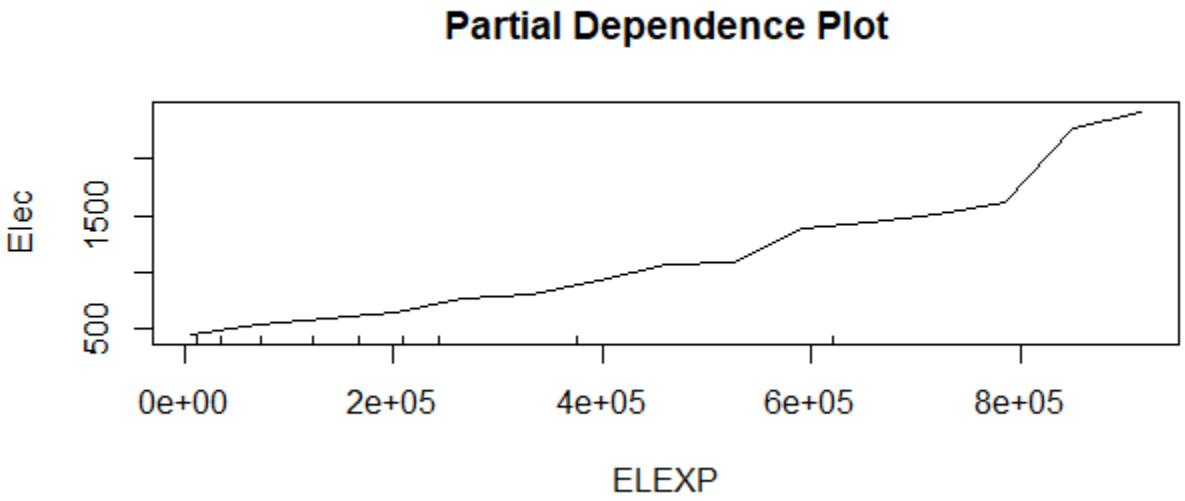
Question 4

Model Inference



From the above graph of Rmse vs Number of trees we can observe that the model performance increases as the number of trees increase.

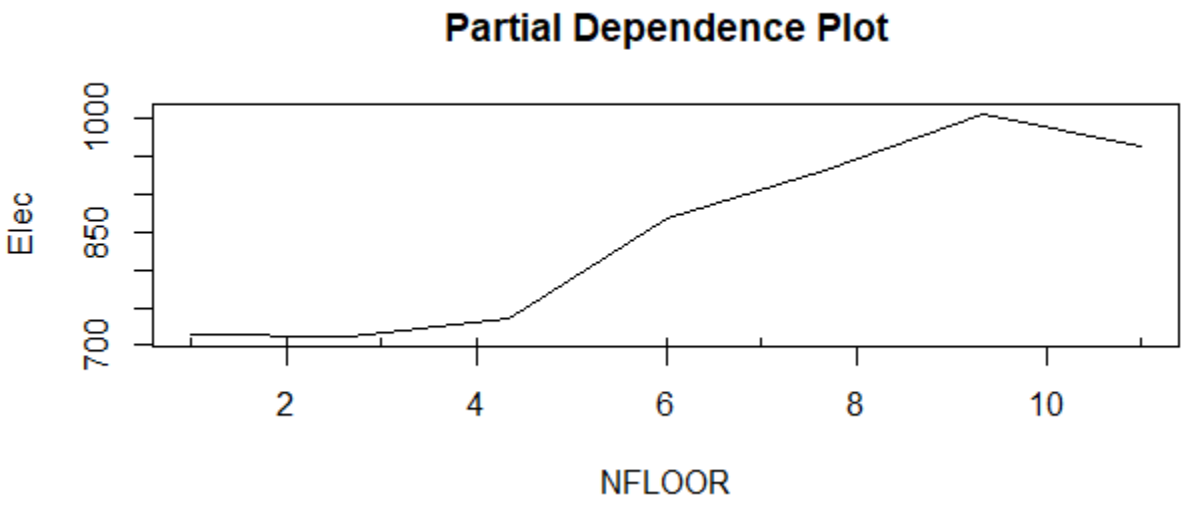
Partial Dependence Plot : Electricity Expenditure



As it can be seen there is a direct relation of Electricity Consumption with the amount of Electricity Expended. Partial Dependency plot shows us that there is gradual increase in the expenditure per 1000

btu units of Electricity consumed. It is one of the most significant parameters that has been used to predict the consumption of the Electricity.

Partial Dependence Plot: No of Floors



As it can be seen, as the number of floor increases the Electricity consumption also increases, hence there is direct correlation between the two parameters. Similarly, this process was carried out for all the parameters and it could be made out that **Random Forest** was the best model out of the all.