IE590 – PREDICTIVE MODELLING – FINAL EXAM

GOPI MANTHENA

1. DATA SOURCE:

2012 CBECS Survey Data. The 2012 CBECS public use microdata file contains untabulated records about individual buildings.

The data used for training our statistical models was obtained from the CBECS Survey Data for 2012. and in specific it deals with Total energy used for space conditioning (i.e., heating, cooling and water heating) in the West South-Central Region as the Response variable.

2. VARIABLE SELECTION AND DATA CLEANING:

The First step was to filter the dataset for the region 'West South-Central' which was given in the objective. This brought the number to rows in the dataset to around 850.

Out of around 1120 variables present the given dataset, 21 variables were chosen based on

- Relevance to the response variable
- Amount of non-missing values in the variable
- Correlations with other variables

Table-1: Predictors Chosen

S.NO	VARIABLE NAME	LABEL	TYPE	UNITS
1	PUBID	Building Identifier	NUMERIC	-
2	CENDIV	Census Division	FACTOR	See Appendix
3	PBA	Principle Building Activity	FACTOR	See Appendix
4	SQFT	Square Footage Area	NUMERIC	Square feet
5	WLCNS	Wall Construction Material	FACTOR	See Appendix
6	RFCNS	Roof Construction Material	FACTOR	See Appendix
7	GLSSPC	Percentage of Exterior Glass	NUMERIC	Percentage
8	NFLOOR	Number of Floors	NUMERIC	-
9	FLCEILHT	Floor to Ceiling Height	NUMERIC	Feet
10	YRCONC	Year of Construction Category	FACTOR	See Appendix
11	GOVOWN	Government Owned	FACTOR	See Appendix
12	MONUSE	Months in Use	NUMERIC	-
13	WKHRS	Hours Open for Week	NUMERIC	Hours
14	PCTERMN	Number of Computers	NUMERIC	-
15	LAPTPN	Number of Laptops	NUMERIC	-
16	HDD65	Heating Degree Days (base 65)	NUMERIC	Days
17	CDD65	Cooling Degree Days (base 65)	NUMERIC	Days
18	PUBCLIM	Building America Climate Region	FACTOR	See Appendix
19	NGUSED	Natural Gas Used	FACTOR	See Appendix
20	PRUSED	Bottled gas/ LPG/ Propane Used	FACTOR	See Appendix
21	OWNTYPE	Building Owner	FACTOR	See Appendix

GOPI MANTHENA

All the Details of the Factor values for the Categorical Data are given in the Appendix at the end of this project.

Number of floors (NFLOOR) has a category of 994 for 15 to 25 floors. This is converted to 20 which is the average number of floors to make the models simpler. Similarly, the category 995 is converted to 28.

Example:

A Lot of variables were redundant. Energy related variables are in KWh and BTU, so only the ones with BTU were selected. Some variables like Year of construction had two variables in the dataset (YRCONC – categorical and YRCON – numeric). All these redundancies were also considered during the selection of variables.

The response variable:

The response variable – Total energy used for space conditioning (i.e., heating, cooling and water heating) was a summation of the following variables in the dataset:

S.NO	VARIABLE NAME	LABEL	UNITS
1	MFHTBTU	Major Fuel Heating Use	thousand BTU
2	MFCLBTU	Major Fuel Cooling Use	thousand BTU
3	MFWTBTU	Major Fuel Water Heating Use	thousand BTU
4	ELHTBTU	Electricity Heating Use	thousand BTU
5	ELCLBTU	Electricity Cooling Use	thousand BTU
6	ELWTBTU	Electricity Water Heating Use	thousand BTU
7	NGHTBTU	Natural Gas Heating Use	thousand BTU
8	NGCLBTU	Natural Gas Cooling Use	thousand BTU
9	NGWTBTU	Natural Gas Water Heating Use	thousand BTU

The variables for fuel oil and district heating were not considered because of a lot missing data in them. Only a few buildings (rows) among these variables (fuel oil and district heating) had considerable values, so in order to not induce error in the modelling, such buildings (rows) were removed before summing the variables mentioned in the above table.

The final response variable is then divided by 1000 to make it 10^6 BTU

The final response variable is:

S.NO	VARIABLE NAME	LABEL	UNITS
1	TOTALEN	Total Energy used for space conditioning	10^6 BTU

3. EXPLORATORY DATA ANALYSIS:

The kernel density distribution of the Response variable i.e. Total Energy for space conditioning is depicted in the Figure-1 and Box-plot of the response variable in Figure-2. It is clearly evident from the distribution and the boxplot that the response variable is not normally distributed. It is skewed towards the side of the lower values.

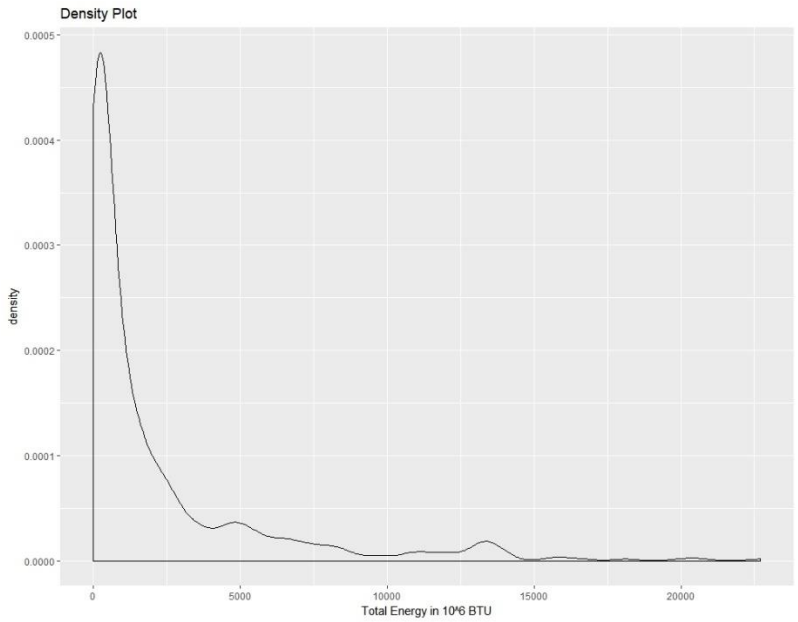


Figure-1: Kernel Density Distribution of Total Energy

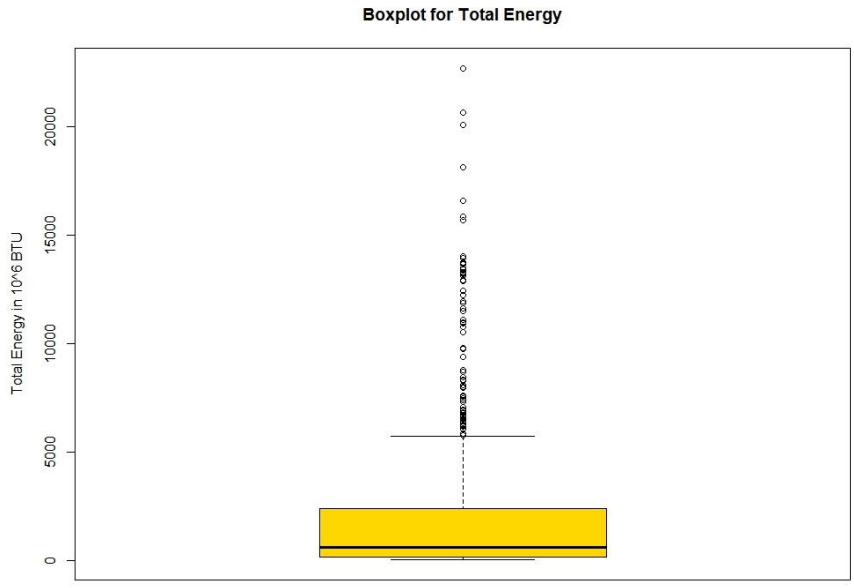


Figure-2: Box-plot to show the variation in Total Energy

The density plot and box-plot are plotted for the response variable with respect to the two different climate regions present in the data (i.e., Mixed-humid region and Hot-Humid region) in figure 2. As you can observe the Total energy is still skewed in both the regions but it is comparatively less for hot-humid regions. This is also logical true because it takes more energy for space heating if the outside temperatures are high.

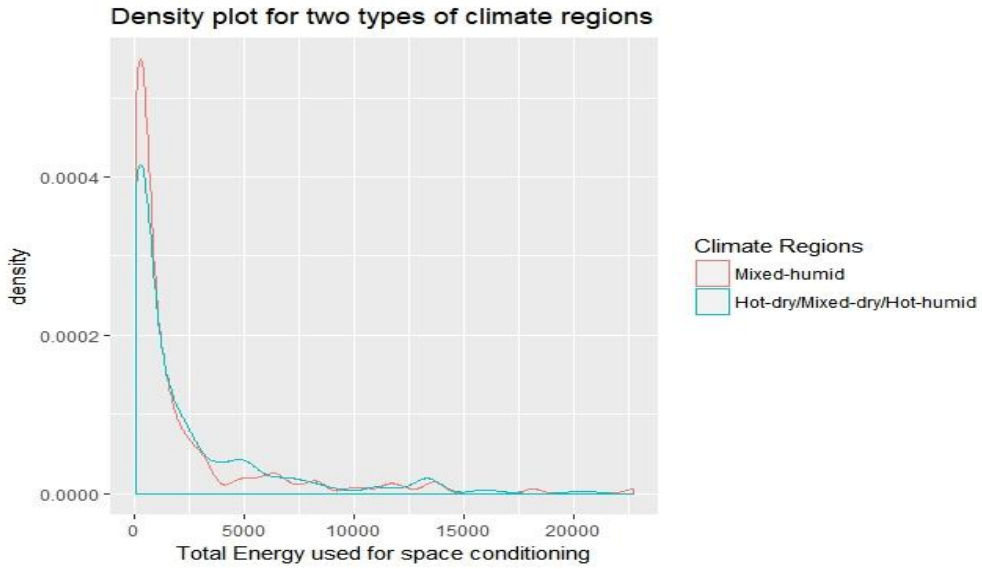


Figure-3: Kernel Density Distribution with respect to the two climate regions

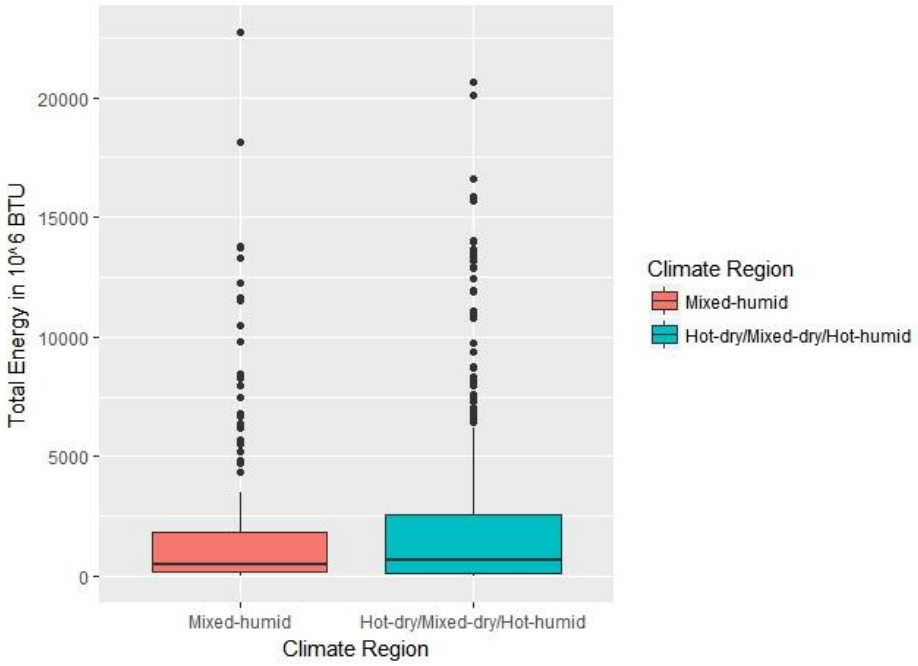


Figure-4: Box-plots of Total Energy with respect to the two climate regions

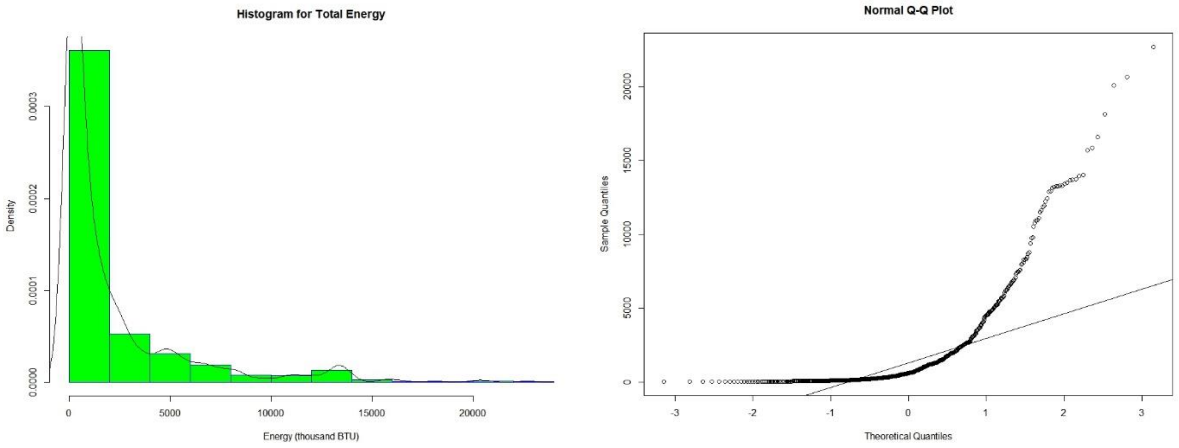


Figure-5 Histogram and Normal-QQ Plot for the response variable

Both these plots in Figure-5 reinstate the fact that the response variable is not normal. The normal qqplot in the above figure shows that the points do not lie along the plotted normal line.

A Violin Plot is used to visualize the distribution of the data and its probability density. This chart is a combination of a Box Plot and a Density Plot that is rotated and placed on each side, to show the distribution shape of the data. Figure-6 below shows the violin plot of Total Energy with respect to the two climate regions.

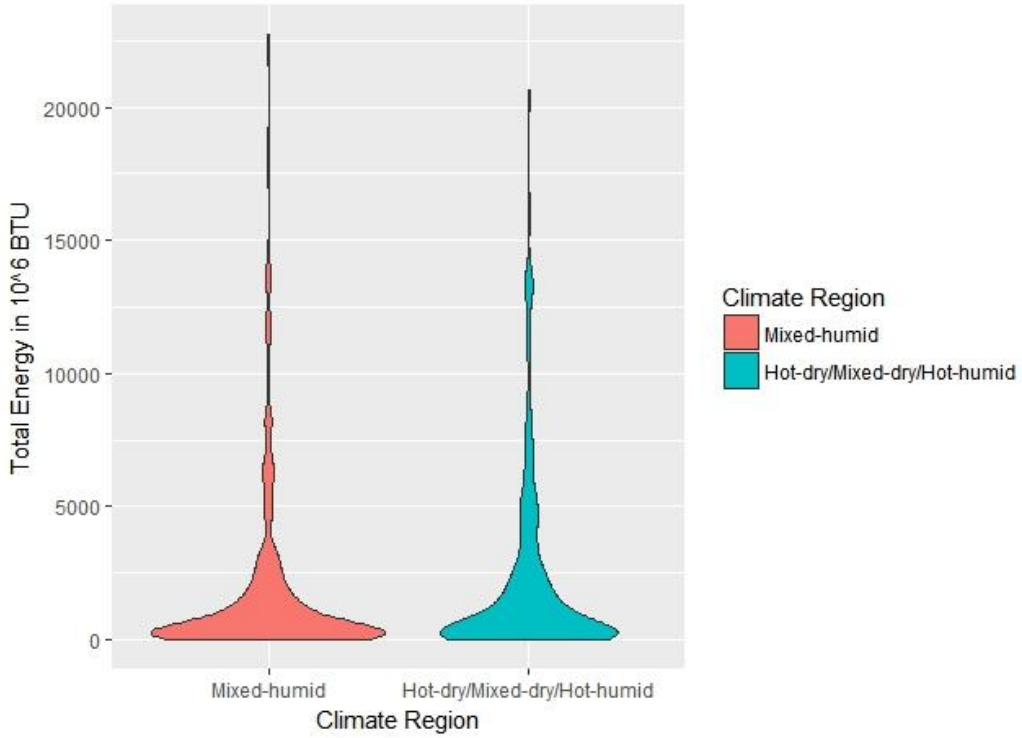


Figure-6: Violin Plot of Total Energy with respect to the two different climate regions

Figure-7 shows the variation of Total energy with respect to various type of buildings. From the figure, we can infer that the inpatient health care centers and shopping malls consume a lot of energy for the purpose of heating, cooling and water heating. This is also logically true because of their vastness in size and number of operating hours, they consume more energy.

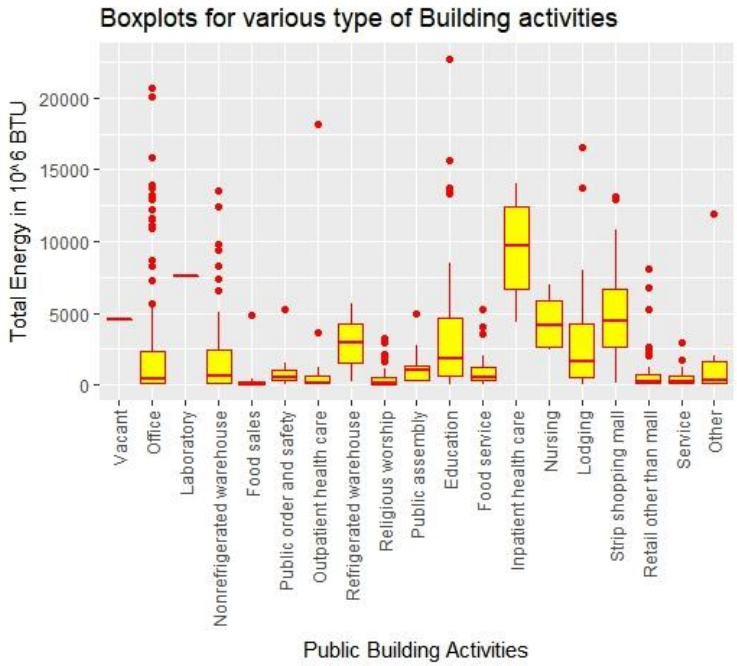


Figure-7: Boxplot to show the Total Energy for various Building types

Figure-8 is again the Total Energy boxplot with respect to wall and roof materials. From the figure, we can infer that 'Plastic, rubber or synthetic sheeting' type of roof material has the highest energy consumption and also highest variation. Similarly, in the wall category, 'windows and vision glass' has the highest consumption.

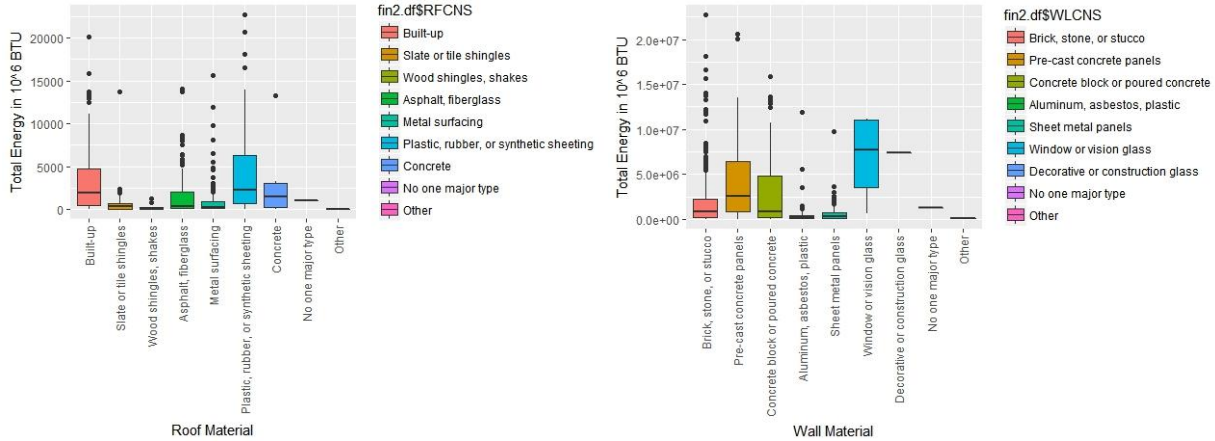


Figure-8: Boxplots of Total Energy with respect to Roof Material and Wall Material

The Total energy is plotted with respect to different Owner types and in both climate regions in Figure-9. The trend in both sets of histograms in the two climatic regions follow a similar trend showing the validity of the data and also that 'Partnership, LLC, LLP' has the highest consumption in both regions.

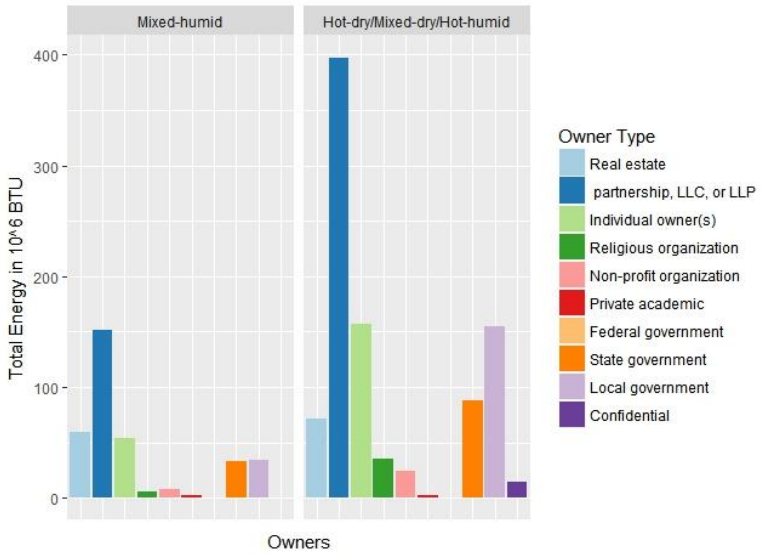


Figure-9: Histogram of Total Energy with various owner types and in both climate regions

Figure-10 also shows a similar histogram with respect to the Principle Building Activity. This graph makes logical sense that Office places and Education buildings consume the most amount of energy in both the climate regions and also, the trend of energy consumption with respect to building types is the same in both climatic conditions.

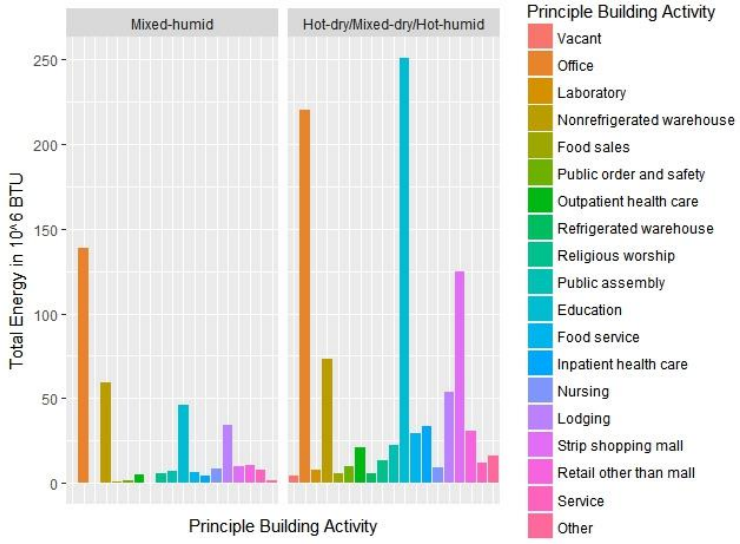


Figure-10: Histogram of Total Energy with various building activities and in both climate regions

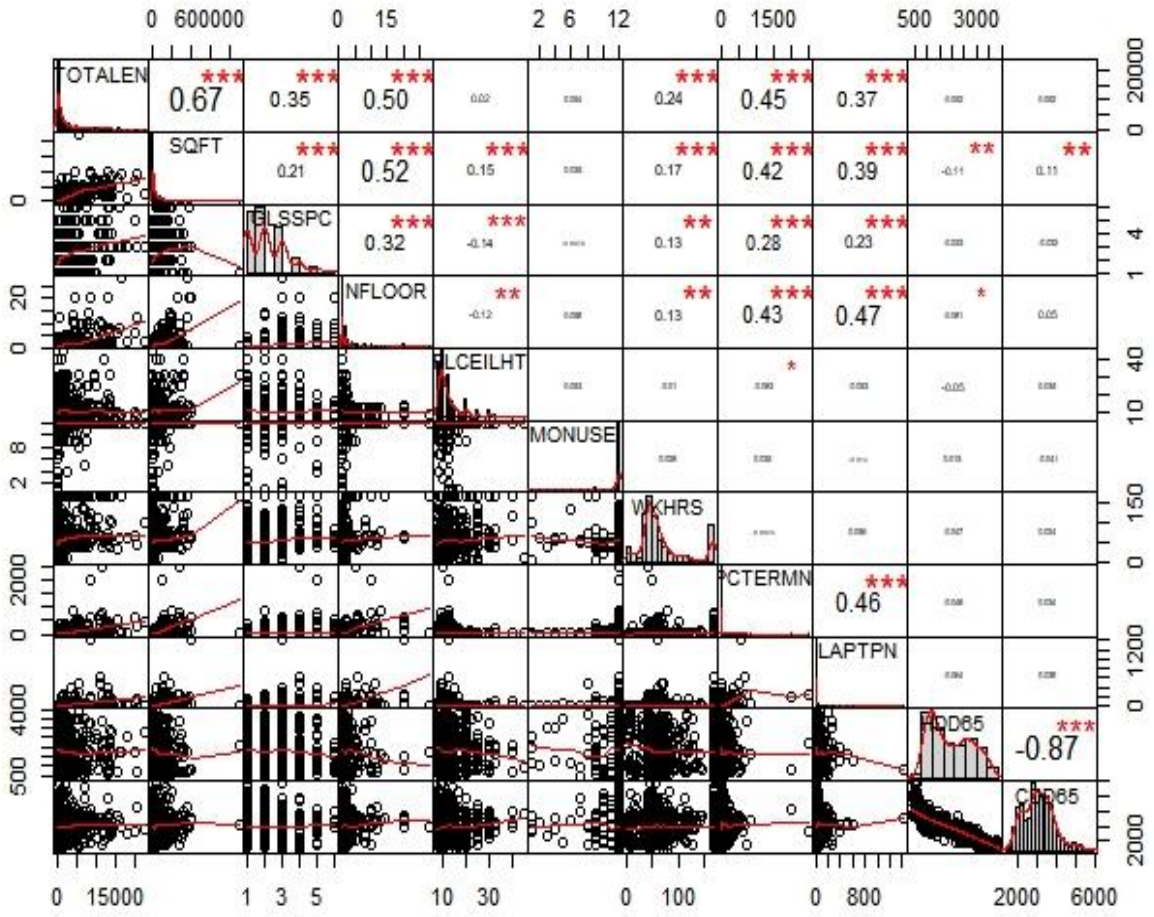


Figure-11: Performance analytics of variables used in modelling Total Energy used for space conditioning

Even though the Figure-11 is very clumsy and has a lot of details in it, it can be seen through magnification that the relationship between the predictors is non-linear. Therefore, analyzing the response variable using linear regression and generalized linear models like lasso and ridge may not give us any potential statistically significant relationships or inferences. Therefore, no linear models were used in this project.

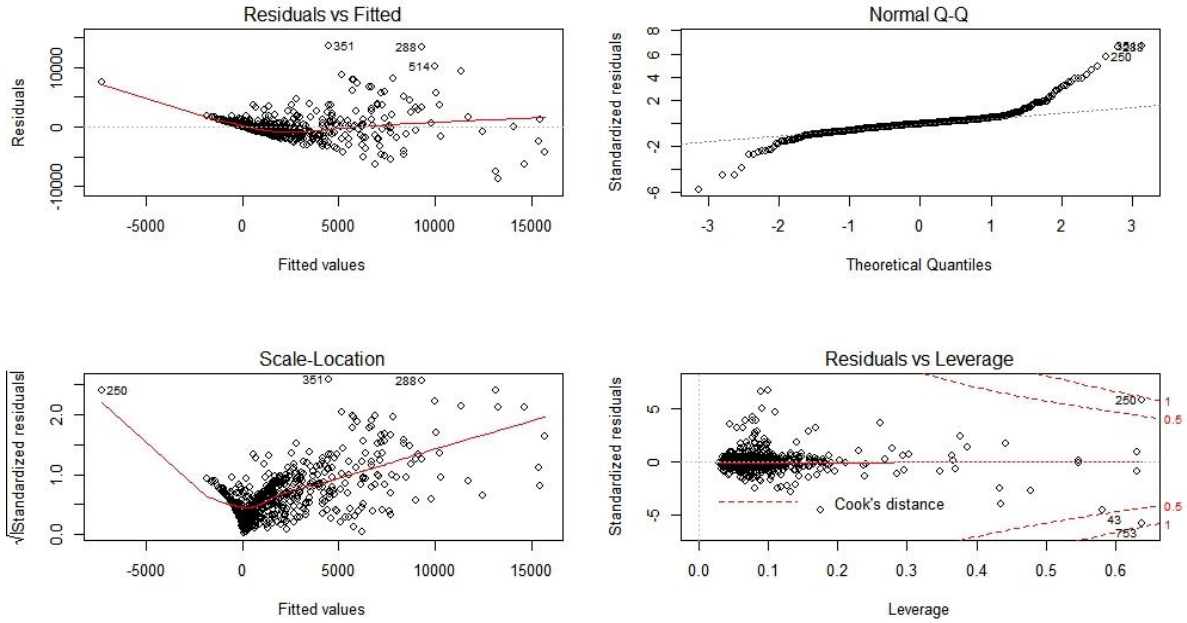


Figure-12: Plots based on a fitted Linear model – to show that the linear assumptions were not satisfied

The residual or error terms are very useful for checking the model assumptions. The assumptions made in a linear model are:

- The errors are independent
- The Y- values can be expressed as a linear function of the Predictors
- The variation of observations around the regression line (the residual SE) is constant (homoscedasticity)
- For a given value of X, the Y values or the error terms are normally distributed.

The plot of a linear model built with all the predictors and the response variable of this dataset will produce the diagnostic plots of the model fit as shown in Figure-12.

The first plot in Figure-12 shows that the variation is not constant or linear. We can also see the red line in the plot is not linear failing to satisfy the assumptions.

In the QQ plot in Figure-12, we can see that the data is not along the linear line.

We can also see the non-linearity clearly both the other diagnostic plots as well.

Therefore, Figure-11 and Figure-12 clearly show us the non-linearity in the dataset and hence Ordinary Linear Regression, Lasso Regression and Ridge Regression were not built on this dataset for model comparisons.

4. METHODOLOGY AND MODEL FITTING:

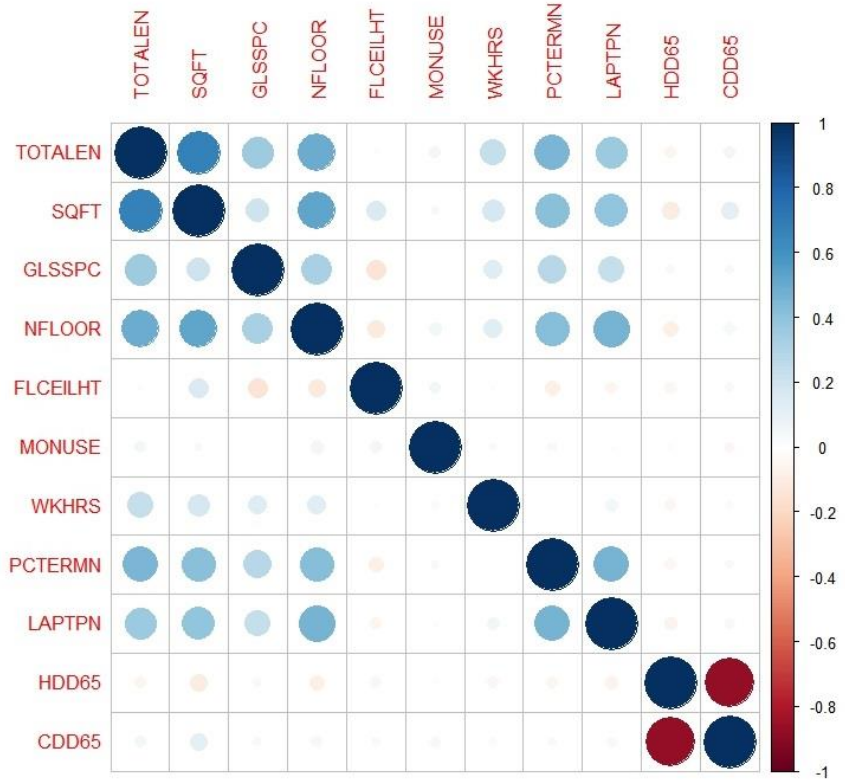


Figure-13: Correlation Plot for the selected variables

Based on the correlation plot the variable 'CDD65' is found be highly correlated with 'HDD65', Hence 'CDD65' which is the Number of cooling degree days (base 65) is removed from the models.

Trained the data with the following supervised learning models to investigate the sensitivity of Total Energy used for space conditioning.

- Classification and Regression Trees (CART)
- Random Forest (RF)
- Multi-variate Regression Splines (MARS)
- Bayesian Additive Regression Trees (BART)
- Generalized Additive Model (GAM)
- Support Vector Machines (SVM)
- Neural Net

Since Random Forest was the best model obtained in this project, a brief overview about Random Forest model is given below:

Random Forest is an ensemble, data-miner which uses 'deep' (unpruned) decision trees as base learners (Breiman, 2001). It is a modification of applying bagging (bootstrap aggregating) to multiple classification and regression trees (CART), and averaging the predictions of the approximately uncorrelated trees to yield the final estimate. Random Forest model was unable to show any clear patterns in the data through variable importance plots and did not show any significant improvement in performance in comparison to generalized linear models.

5. MODEL SELECTION:

A 10-fold cross validation using train and test data is performed on the models given below and their average In-sample RMSE, average Out-of-sample RMSE and average R^2 values are tabulated below:

Table-2: Performance Metrics of Various Models

S.NO	MODEL	In-Sample RMSE	Out-of-Sample RMSE	R^2 (Fit)
1	MARS (Unpruned)	2476.96157	1861.24126	0.716567385
2	MARS (pruned)	2516.36016	1894.32717	0.70642004
3	Random Forest	2065.50108	981.592335	0.921196094
4	CART	2377.09244	1966.25146	0.683262604
5	BART	2261.95932	726.700728	0.956205045
6	Random Forest (important variables)	2087.36135	1027.21692	0.913687028
7	SVM	2478.09599	1803.34514	0.733968559
8	Neural Nets	3391.62782	3496.18733	-1.2408E-05
9	GAM	2992.69577	1969.11229	0.682809957

We can clearly see from Table-2 that Random Forest has the least average Out-of-sample RMSE among all the models built and also has a decent In-sample RMSE and R^2.

Another Random forest model was built just by considering the important variables in the model which is also mentioned in Table-2 above. This did not perform better than the regular Random Forest model.

Figures showing the variation in the above tabulated values in each cross-validation is shown below in Figures-

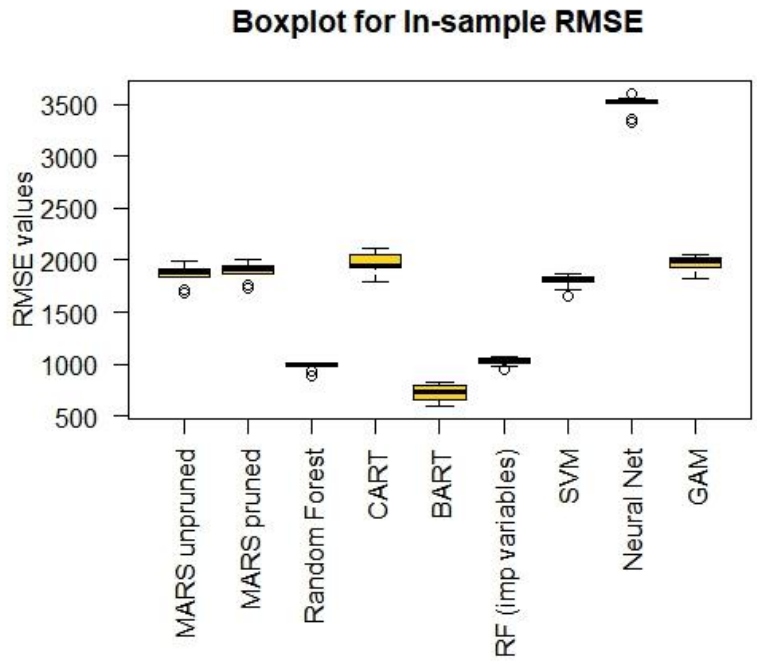


Figure-21: Boxplot for In-sample RMSE values in each cross-validation loop

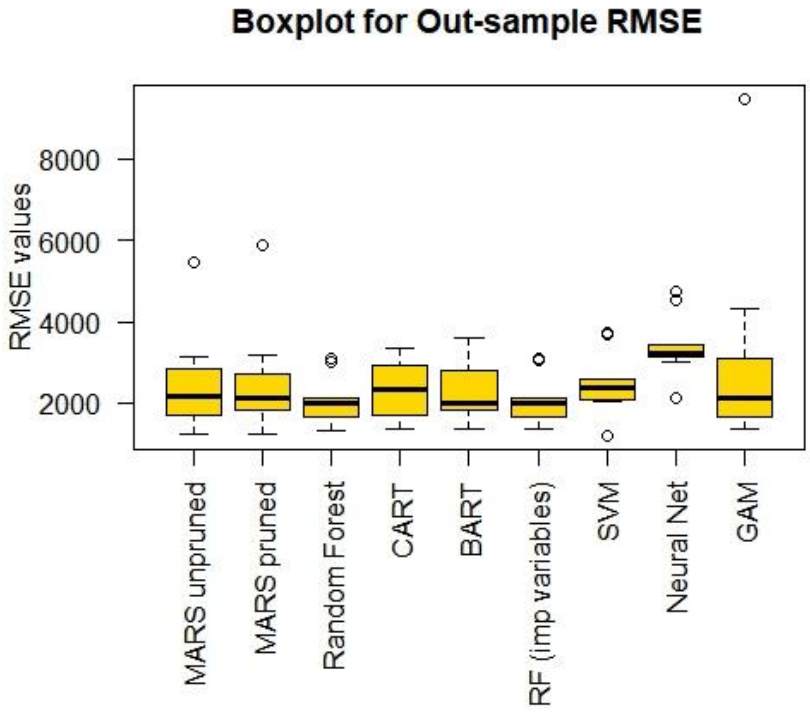


Figure-21: Boxplot for Out-sample RMSE values in each cross-validation loop

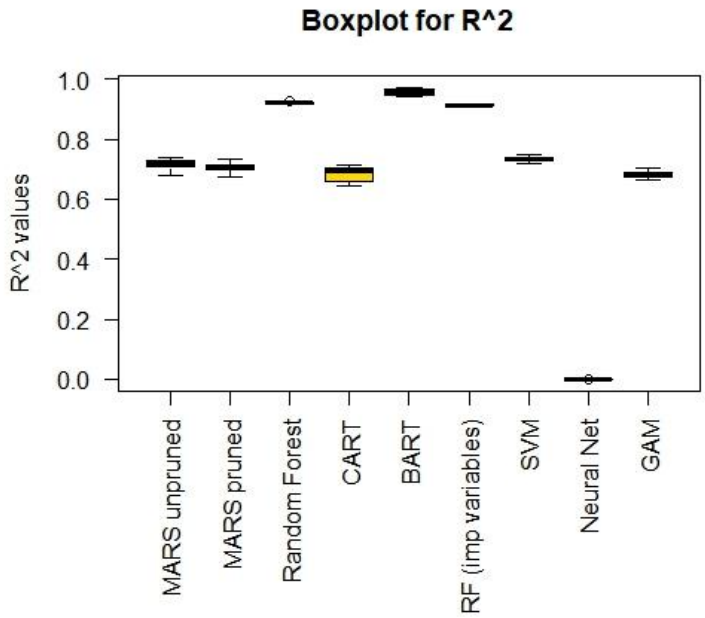


Figure-21: Boxplot for R^2 values in each cross-validation loop

6. MODEL INFERENCES AND PARTIAL PLOTS

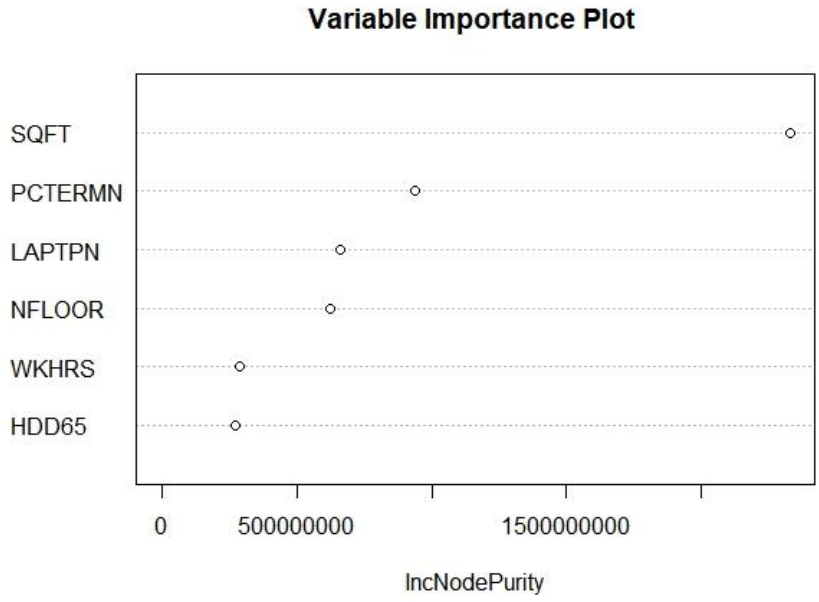


Figure-14: Variable Importance Plot for the Random Forest Model

From Figure-14, we can see that the Total energy used for space heating has 'Square foot area' as the most important predictor. This also makes logical sense that a building with larger area will consume more energy.

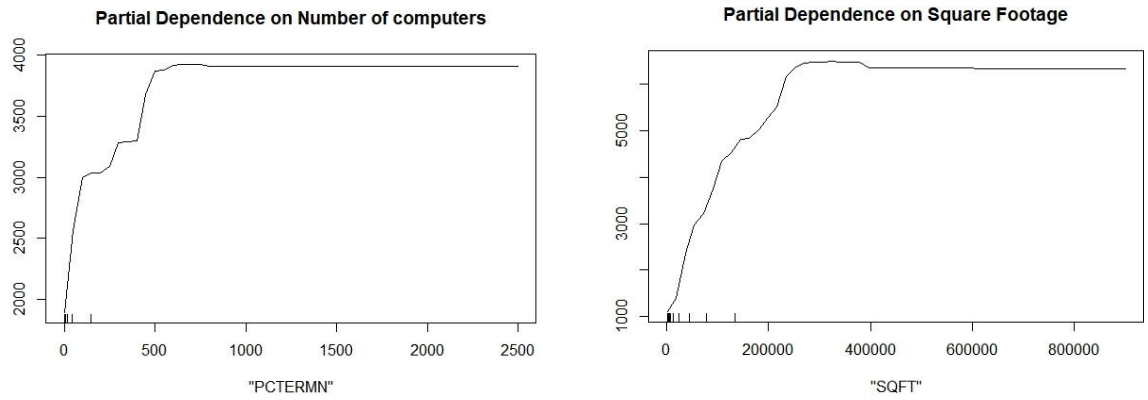


Figure-15: Partial dependence plots on Number of Computers and Square Footage Area

As you can see from Figure-15, the total energy increases in both the plots when the number of computers or the square footage area is increased. This is also logical because, more computers need more cooling and hence need more energy.

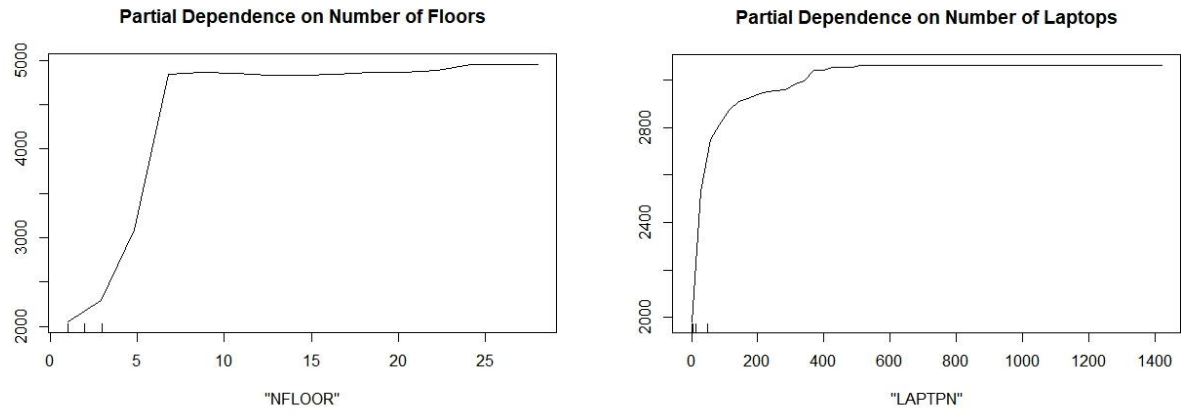


Figure-16: Partial dependence plots on Number of floors and Number of Laptops

As you can see from Figure-16, the total energy increases in both the plots when the number of laptops or the number of floors is increased. This is also logical because, more computers need more cooling and hence need more energy.

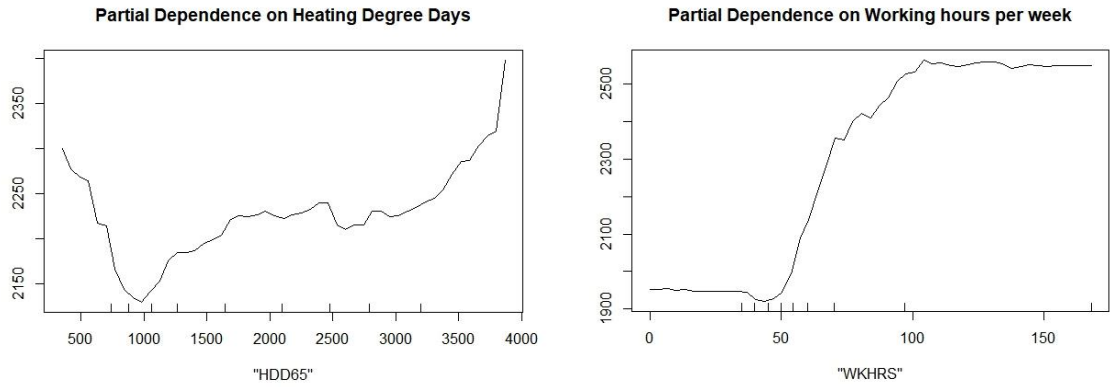


Figure-17: Partial dependence plots on Heating degree days and working hours per week

As you can see from Figure-17, the total energy has a general increasing trend. As working hours increases the buildings will be used for more time and more energy will be consumed.

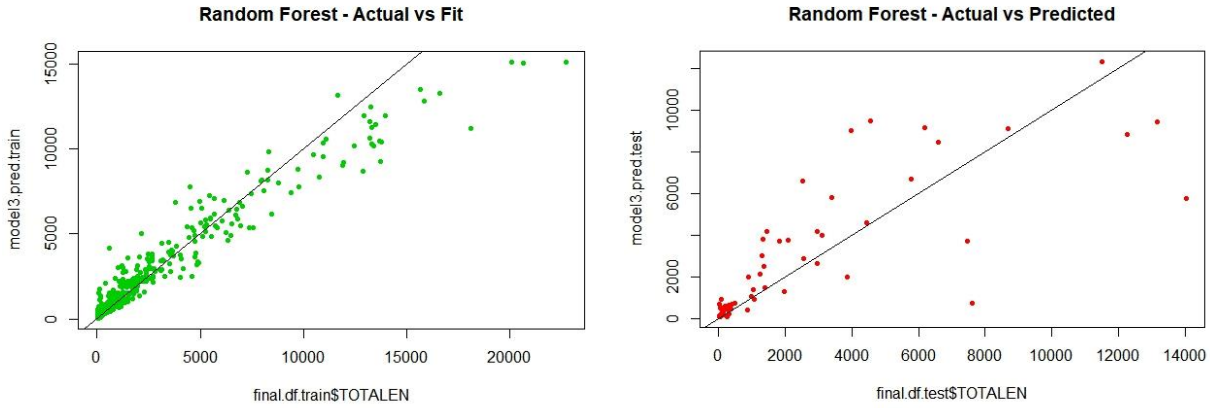


Figure-18: Comparison Plots – Actual VS Fitted values & Actual VS Predicted values

7. CONCLUSION

It can be observed that the Random Forest Model is the best model to describe the data set and also it can be seen that the 'SQFT' (square foot area) is one of the top predictor and by looking at the variable important plot it can be seen that the variables which was predicted by the model can also be explained intuitively.

8. APPENDIX

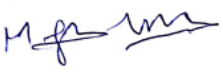
CENDIV	Census division	'1' = 'New England' '2' = 'Middle Atlantic' '3' = 'East North Central' '4' = 'West North Central' '5' = 'South Atlantic' '6' = 'East South Central' '7' = 'West South Central' '8' = 'Mountain' '9' = 'Pacific'
PBA	Principal building activity	'01' = 'Vacant' '02' = 'Office' '04' = 'Laboratory' '05' = 'Nonrefrigerated warehouse' '06' = 'Food sales' '07' = 'Public order and safety' '08' = 'Outpatient health care' '11' = 'Refrigerated warehouse' '12' = 'Religious worship' '13' = 'Public assembly' '14' = 'Education' '15' = 'Food service' '16' = 'Inpatient health care' '17' = 'Nursing' '18' = 'Lodging' '23' = 'Strip shopping mall' '24' = 'Enclosed mall' '25' = 'Retail other than mall' '26' = 'Service' '91' = 'Other'
SQFT	Square footage	1,001 - 1,500,000
WLCNS	Wall construction material	'1' = 'Brick, stone, or stucco' '2' = 'Pre-cast concrete panels' '3' = 'Concrete block or poured concrete (above grade)' '4' = 'Aluminum, asbestos, plastic, or wood materials (siding, shingles, tiles, or shakes)' '5' = 'Sheet metal panels' '6' = 'Window or vision glass (glass that can be seen through)' '7' = 'Decorative or construction glass' '8' = 'No one major type' '9' = 'Other'

RFCNS	Roof construction material	'1' = 'Built-up (tar, felts, or fiberglass and a ballast, such as stone)' '2' = 'Slate or tile shingles' '3' = 'Wood shingles, shakes, or other wooden materials' '4' = 'Asphalt, fiberglass, or other shingles' '5' = 'Metal surfacing' '6' = 'Plastic, rubber, or synthetic sheeting (single or multiple ply)' '7' = 'Concrete' '8' = 'No one major type' '9' = 'Other'
GLSSPC	Percent exterior glass	'1' = '1 percent or less' '2' = '2 to 10 percent' '3' = '11 to 25 percent' '4' = '26 to 50 percent' '5' = '51 to 75 percent' '6' = '76 to 100 percent' Missing = Not applicable
NFLOOR	Number of floors	1 - 14 994 = 15 to 25 995 = More than 25
FLCEILHT	Floor to ceiling height	6 - 50 995 = More than 50
YRCONC	Year of construction category	'01' = 'Before 1920' '02' = '1920 to 1945' '03' = '1946 to 1959' '04' = '1960 to 1969' '05' = '1970 to 1979' '06' = '1980 to 1989' '07' = '1990 to 1999' '08' = '2000 to 2003' '09' = '2004 to 2007' '10' = '2008 to 2012'
GOVOWN	Government owned	'1' = 'Yes' '2' = 'No'
OWNTYPE	Building owner	'01' = 'Real estate investment trust (REIT)' '02' = 'Other public or private corporation, partnership, LLC, or LLP' '03' = 'Individual owner(s)' '04' = 'Religious organization' '05' = 'Non-profit organization (other than religious or government)' '06' = 'Private academic institution' '07' = 'Other non-government' '08' = 'Federal government' '09' = 'State government' '10' = 'Local government' '97' = Withheld to protect confidentiality'

MONUSE	Months in use	0 - 12
WKHRS	Total hours open per week	0 - 168
ELUSED	Electricity used	'1' = 'Yes' '2' = 'No'
NGUSED	Natural gas used	'1' = 'Yes' '2' = 'No'
PRUSED	Bottled gas/LPG/propane used	'1' = 'Yes' '2' = 'No'
STUSED	District steam used	'1' = 'Yes' '2' = 'No'
HWUSED	District hot water used	'1' = 'Yes' '2' = 'No'
PCTERMN	Number of computers	0 - 4,195 Missing = Not applicable
LAPTPN	Number of laptops	0 - 1,420 Missing = Not applicable
HDD65	Heating degree days (base 65)	
CDD65	Cooling degree days (base 65)	
PUBCLIM	Building America climate region	1' = 'Very cold/Cold' '2 '= 'Mixed-humid' '3' = 'Hot-dry/Mixed-dry/Hot-humid' '5' = 'Marine' '7' = 'Withheld to protect confidentiality'

Declaration

"I have obeyed all rules for this exam and have not received any unauthorized aid or advice."



GOPI MANTHENA

Date : 03/26/18