IE 590ADVANCED DATA ANALYTICS - FINAL PROJECT REPORT

Analysis of Electricity used by Commercial Buildings in the U.S.

:DataSet:

Data Set: Commercial Buildings Energy Consumption Survey (CBECS) – 2003 by Energy Information Administration(EIA)

SWAPNIL RAI

Abstract

The following study aims at building predictive models for Commercial Buildings Energy Consumption for cooling in the United States using the 2003 EIA data set. This project illustrates an attempt at using logarithmic transformations on the electricity consumed, area and number of cooling degree days, as explained in detail later; selecting a model that will fit and predict the CBECS 2003 data set best. Our findings will be validated against the actual data published by EIA (Expected to be released in June 2016). This paper will also illustrate the important parameters that affect the cooling load of a building, and their degree of influence.

Introduction

A very common sight in the United States of America is that of the common office-goer wearing a sweater or a pullover while going to work on a warm summer day. This stems from the fact that the temperature set at office buildings is lower than what the average human being would term as "comfortable". Overcooling buildings during warmer months would seem counterproductive given the fact that the cost of setting the temperature lower during warmer days would linearly increase with the temperature difference. The intention of this paper is to highlight this fact using predictive analytical models.

This paper describes our efforts in initially building a model to fit the CBECS 2003 data and then using that model to validate the energy consumption due to cooling from the CBECS 2012 data. We will compare these results with actual data supplied by EIA and use a two sample t-test to validate the statistical significance of our hypothesis that Commercial buildings in the united states are overcooled. We expect that in spite of technological advances in the field of cooling methods and implementation of policies that encourage energy conservation this trend is prevalent amongst the office community.

We will build our model by careful selection and transformation of our data and using several methods like Random Forest, MARS and GAMs, GLM and select the best one based on R^2 values and also based on the MSE obtained from k-Fold cross validation.

Literature Review

One of our initial inspirations for writing this paper was (Derrible & Reeder, 2015) which was a short communication that used the eQuest v3.65 package to simulate the energy consumption in the United States from the CBECS 2012 data. (Contreras, Smith, M, & Jr., 2013) has used GLM to fit the overall energy consumption from the CBECS 2012 data but makes no attempt at validating predictive accuracy.

Artificial Neural Networks have been used in the past to make similar benchmarking models to predict energy consumption. (Melek Yalcintas1, 2007) uses ANN to build 9 separate models for each census division by recognizing that the energy consumption in each of the 9 divisions would depend on different factors. However, the range of R^2 values varied from 0.46 at the worst to 0.72 at the best. (Azadeh, Ghaderi, Tarverdian, & Saberi, 2007) uses a genetic algorithm and ANN to predict the energy consumption in Iran.

There have also been several papers detailing out the energy conservation strategies based on the CBECS data. (Griffith & Crawley, 2006) uses the 1999 data and reconstructs it as if all the buildings were built using energy use parameters in 2005 and analyses the technical potential of the commercial buildings to conserve energy. (Payne, 2006) attempts to deconstruct the commercial energy user.

Variable Imputation

The original CBECS 2003 data had a lot of values that needed to be imputed. Of the null values in the data we imputed the data based on the imputation flags provided by EIA for the specific variables.

We also made some specific changes to certain variables to make them more "R-friendly". Those details are listed below:

1) All null values were cross checked with the imputation flags for the same variable. The imputation flags indicated that there were two types of values which were marked NA. These were either "Data not Provided" or "Not Applicable to that Building". The NA values were filled in with these as factor variables.
2) Percentages in many places in this data set were used in conjunction with another variable having Yes/No data. For example, there were several percentage variables like PKGCP8 (Percent cooled by Package A/C) and its corresponding Yes/No variable PKGCL8 (Package A/C Present or not). In whichever PKGCL8 was set to No PKGCP8 was set to NA. All such values were set to 0%.
3) NFLOORS8 was used to indicate the number of floors of the building. This variable was filled with the correct number of floors for values from 0-20. For 20 – 30 floors a factor variable was used (998) and 30+ the factor 999 was used. These were replaced by 25 for 998 and 40 for 999.
4) The malls in particular had a large amount of missing data. All the missing data in malls was given a factor 23 so that we could account for malls separately in our analysis.

Data Description

Response Variable:

Electricity Consumed (Cooling) – ELCLBTU8

Major Predictors:

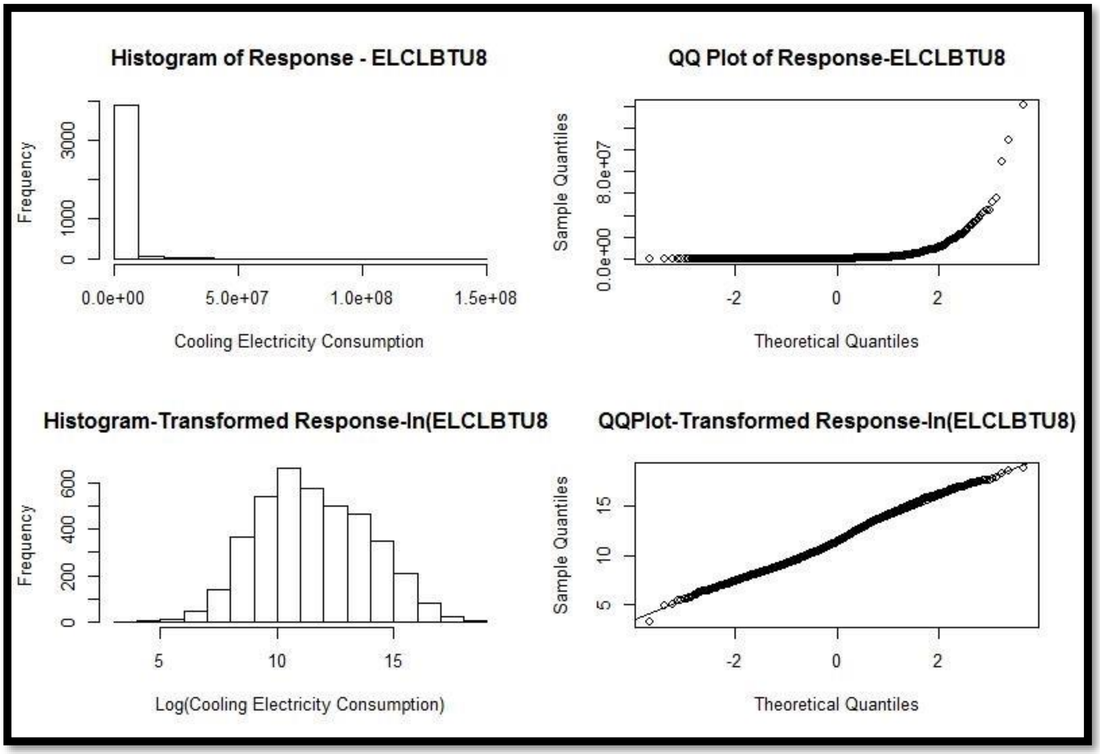
Our original dataset consisted of 151 predictors. Since the CBECS dataset contains energy consumption statistics from various sources there are a large number of predictors that we can intuitively eliminate based on relevance to our study. In order to give a brief description of our dataset we have outlined some of the major predictors are as follows:

- Building Capacity – CAPACITY8
- Cooling Degree Days – CDD658
- Census Division – CENDIV8
- Sq. Ft area – SQFT8
- Building Activity – PBAPLUS8
- Main Cooling Equipment – MAINCL8
- Temperature Adjustment Techniques – HWRDCL8
- Owner – OWNR8
- No. of Workers – NWKER8
- No. of PCs – PCNUM8
- Glass Percentage – GLSSPC8

Data Visualization

We aim at getting reasonable evidence about the specific traits (normality, variance, independence, etc.), and see effects of the key predictors on the response variable and corelations between predictors so as to help us build predictive models.

The following plots show comparison between frequency distribution and QQ plots of: Original response variable – Cooling Electricity Consumption – ELCLBTU8 Vs The log transformed response- log(ELCLBTU8)



Histogram of Response - ELCLBTU8

QQ Plot of Response-ELCLBTU8

Histogram-Transformed Response-ln(ELCLBTU8

QQPlot-Transformed Response-ln(ELCLBTU8)

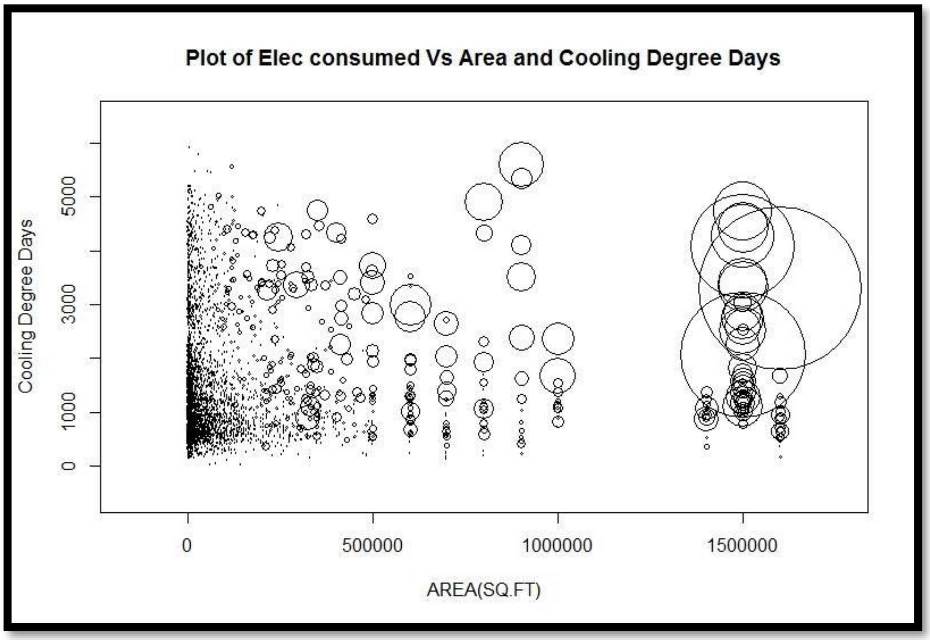
The QQ-Plot of the original response variable shows non-normal behavior. The log plots clearly indicate that transferring the response variable-ELCLBTU8 to the log scale significantly improves the frequency distribution, also the QQ plot indicates normality.

Summary statistics of Response in the Original scale (ELCLBTU8) and Log transformed scale(lELCLBTU8).

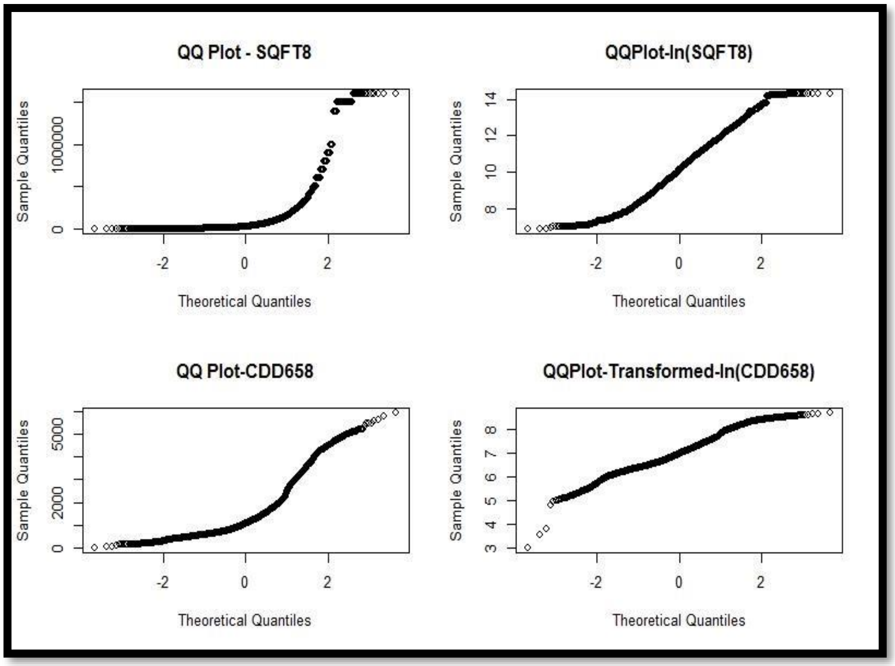
```
summary(ELCLBTU8)
      Min.   1st Qu.    Median       Mean     3rd Qu.       Max.
28       17880       83560      1179000      599700    141100000
summary(lELCLBTU8)
      Min. 1st Qu.  Median    Mean 3rd Qu.     Max.
    3.332   9.787  11.320  11.540  13.280  18.760
```

Log transformation is a good idea for the response variable, considering the original values are very widespread.

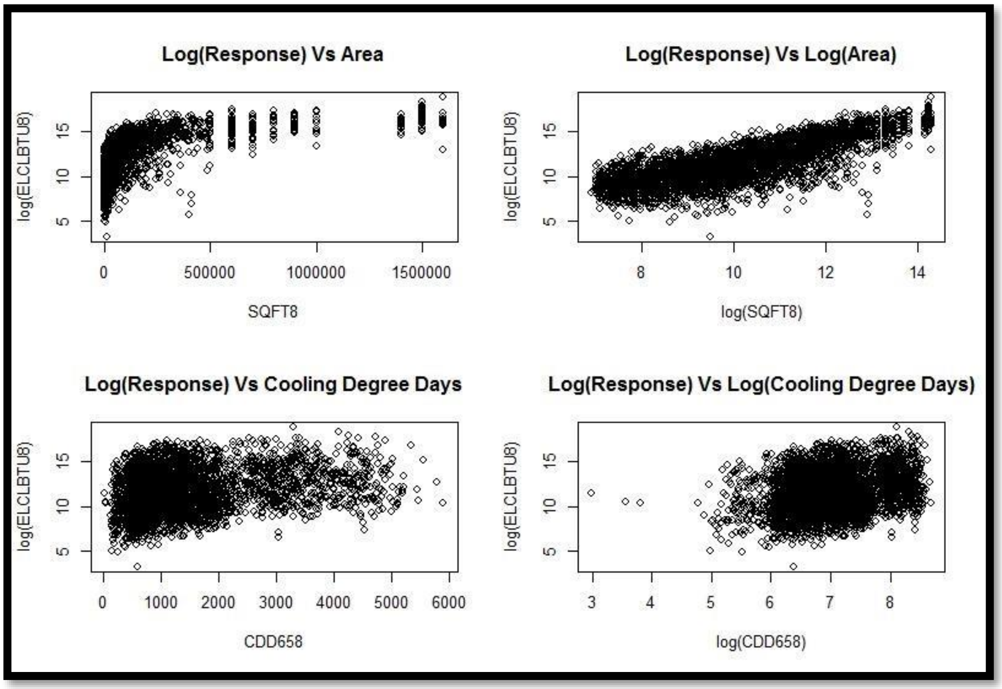
To check the effect of significance of CDDs and Area(SQFT) on the response, we used the bubble plot of CDD Vs Area, size of the bubble indicates electricity consumption- larger the diameter, more is the electricity consumption.

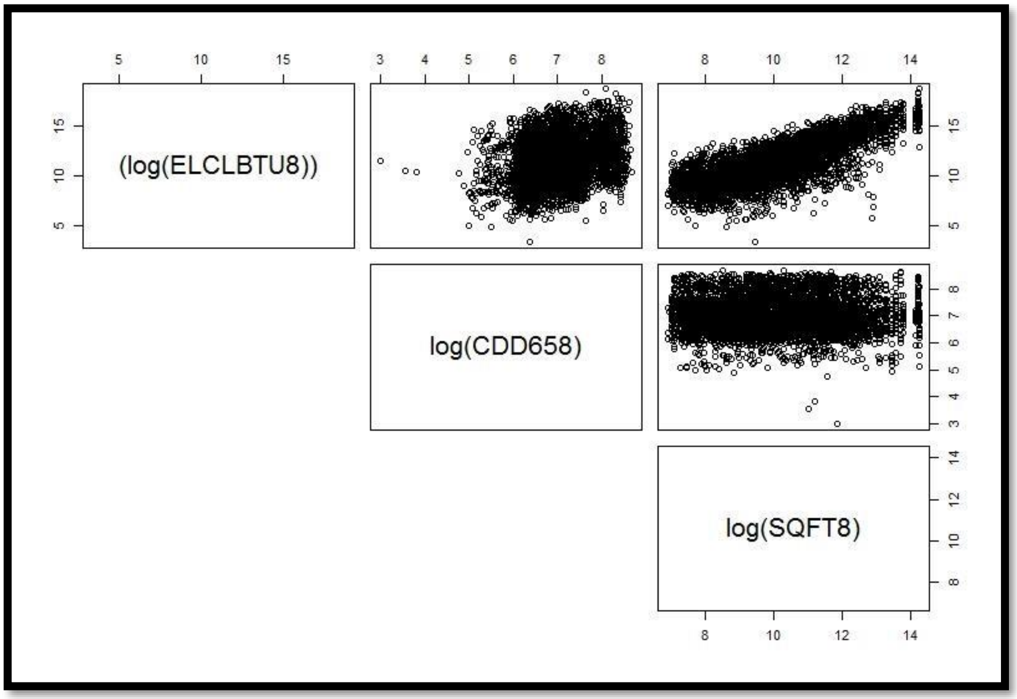


As the sq. feet area and CDD increases, as a general trend the size of the bubble increases indicating more electricity consumption due to cooling. We can also observe a few smaller bubbles in the large area region of the plot indicating the presence of other factors that affect our response variable.



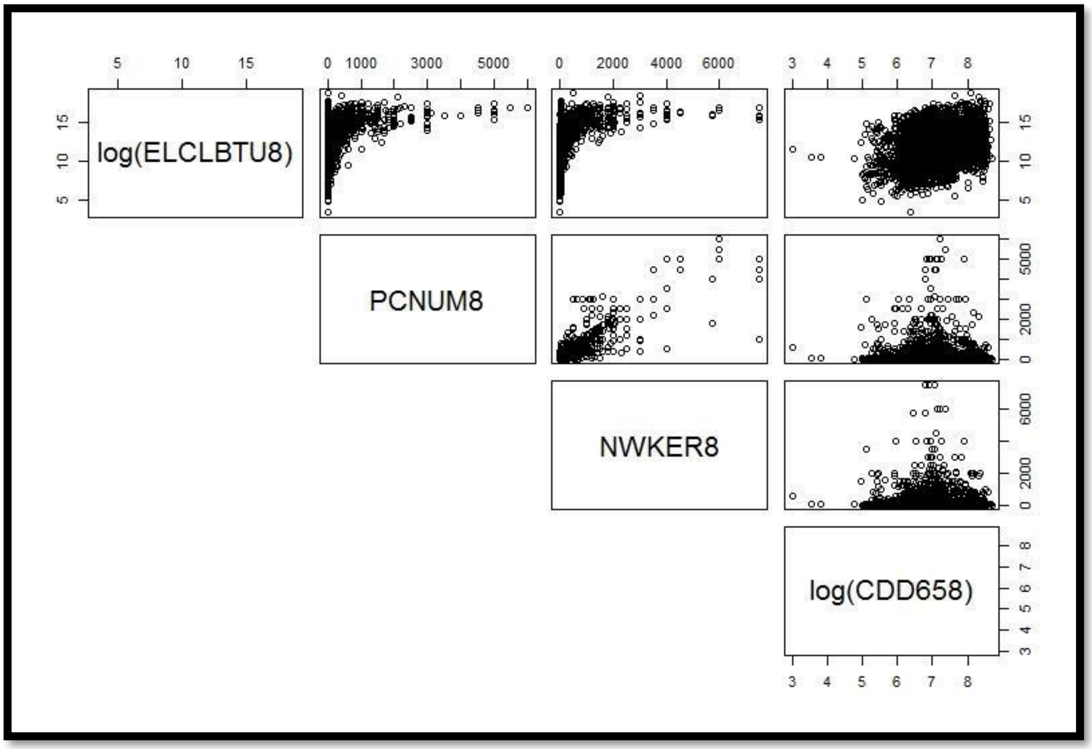
The comparison of Scatter plots of Log(Response) Vs Area, CDD and those of Log (Response) vs Log transformed Area and CDD are shown below. These plots clearly indicate that log transformation on Area and CDD improves their relationship with the response.





From the above scatter plot between our 3 major predictors (which have been transformed to the log scale we see no apparent co-relation.

Co – Relations



- No. of PCs (Computers) is co-related with No. of Workers
- For models like LM and GLM which are affected by multi-collinearity, we used a ratio of No. of PCs/No. of Workers to remove the collinearity.

Methodology

We implemented the following models to fit our data set:

1. **Linear Model (LM):**(James, Witten, Hastie, & Tibshirani, n.d.) Assumptions:

- Linear relationship between response and predictors
- Normal errors
- Independent errors
- Constant Variance

In general, suppose that we have *p* distinct predictors. Then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \ldots \ldots . + \beta_p X_p + \epsilon$$

2. **LM step** (James et al., n.d.)

To perform *best subset selection*, we fit a separate least squares regression for each possible combination of the 'p' predictors. That is, we fit all *p* models that contain exactly one predictor, all

${}^{P}C_2 = p(p-1)/2$ models that contain exactly two predictors, and so forth.

We then look at all of the resulting models, with the goal of identifying the one that is best.

3. **Generalized Linear Model (GLM)** (James et al., n.d.)

GLM is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

The GLM consists of three elements:

1. A probability distribution from the exponential family.
2. A linear predictor $\eta = \boldsymbol{X\beta}$.
3. A link function g such that $E(\boldsymbol{Y}) = \boldsymbol{\mu} = g^{-1}(\eta)$

4. **Generalized Additive Models (GAMs)** (James et al., n.d.)

In order to allow for non-linear relationships between each feature and the response, we replace each linear component βjxij with a (smooth) nonlinear function fj(xij). We would then write the model as

$$y_i = \beta_0 + \sum f_j(x_{ij}) + \epsilon_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

In GAMS, we calculate a separate f_j for each X_j , and then add together all of their contributions.

We noticed that two of our predictors No of Workers and PC Number did not have a linear relationship with the response variable. Hence we applied a cubic smoother to these variables with 1 knot.

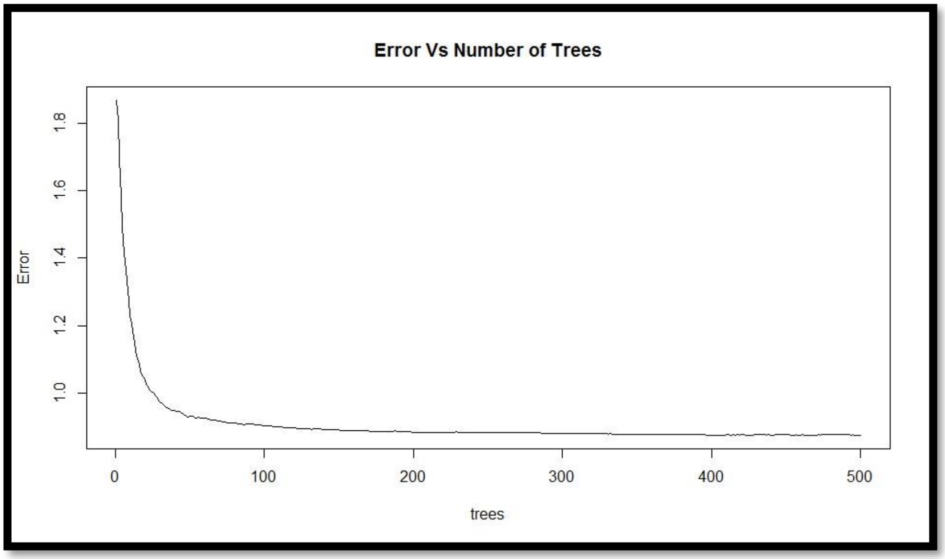
5. **Random Forest (RF)** (Hastie, Tibshirani, & Jerome, 2009)

Random Forest is an ensemble learning- tree based method. It is a modification of applying Boosting to Classification and regression trees by averaging out a large number of un-correlated trees. The algorithm that Random Forest follows is mentioned below: 1) Set B as number of trees to grow 2) For b = 1 to B:

a. Draw a bootstrap sample Z* of size N form the training data
b. Grow a tree T_b to the bootstrap sample similar to CART by following the below steps
    i. Select m variables at random from the predictors
    ii. Pick the best variables and their split points among the m variables
    iii. Split the tree at that node into two child nodes

3) Output the ensemble of trees $\{T_b\}_1^B$

We initially fitted a large number of predictors to our Random Forest model. We then observed the variable importance plot and reduced the number of variables to prevent overfitting of the data.



Optimum number of trees used, ntree = 190

6. **Multivariate Adaptive Regression Splines (MARS)** (Hastie et al., 2009)

Multivariate Adaptive Regression Splines is an adaptive procedure for regression. MARS uses expansion of piecewise linear basis functions with knots at certain values 't'. We represent the collection of these linear basis function as below.

$$C = \{(X_j - t)_+, (t - X_j)_+\}_{\substack{t \epsilon \{x_{1j}, x_{2j}, x_{3j} \ldots\ldots, x_{Nj}\} \\ j=1,2,3,\ldots p}}$$

The model is built on a forward stepwise linear regression but we use functions from C to build our linear model. Thus the final model is represented as

$$f^{(x)} = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$$

Where each $h_m(X)$is a function in C or a product of two or more functions.

We fitted MARS by observing that our data set had no variable whose relation with the response had multiple changes in curvature. From this observation we chose to fit MARS with degree 2 and penalty 1 to the data set.

7. <u>Support Vector Machines (SVM)</u> (Hastie et al., 2009)

We have used SVM implemented for linear regression. We take the concept of Support Vector Machines used for classification and use it to modify the linear regression model. The linear model is described by

$f(x) = x^T\beta + \beta_0$

We then handle non-linear generalizations. We consider the minimization of the below function to estimate β

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} ||\beta||^2$$

$$where\ V_\epsilon(r) = \begin{cases} 0 & if\ |r| < \epsilon \\ |r| - \epsilon & otherwise \end{cases}$$

For any choice of $V(r)$ the solution of $f(x) = \sum_{i=1}^{N} \hat{a}_i K(x, x_i)$

Where $K(x, x_i) = \sum_{i=1}^{M} h_m(x) h_m(y)$ this has the same form as a radial kernel function. Hence we have used a radial kernel to fit SVM on the data set.

For fitting SVM we tuned the parameters (gamma and cost) by setting different values of the same from $gamma = \{0.001, 0.01, 0.1, 1, 10, 100\}$ and $cost = \{0.1, 1, 5, 10, 100\}$ and observed the MSE of all the fitted SM models.

<u>Models Not Used</u>

Reason for not using CART – A set of Ys will give the same average Y_{pred} using cart, so we will have very high MSEs. Also we have used Random Forest which eliminates this problem, hence there is no need to use CART.

To test which model has the highest predictive power, the Mean Squared Errors and Mean Absolute Errors for each model were compared by using K-fold Cross validation (k=10) to find mean values after 10 number of holdouts.

Model Selection

Model	MSE (Log scale)	R^2	Best Variable Subset
Mean Only	5.201	-	-
LM	0.926	0.76	CDD, Area, MAINCL8, Construction Material, Temp Control Technique, Mainframe, Server-farm, Glass %, Building Capacity, Walk in Refrigeration Units, Census Division, PCs/ No. of workers, No. of workers
LM Step	1.043	0.76	
GAMS	0.926	0.767	CDD, Area, Cooling Equipment, Construction Material, Temp Control Technique, Mainframe, Server-farm, Glass %, Building Capacity, Walk in Refrigeration Units, Census Division, (Splines: PCs, No. of workers)
Random Forest	**0.764**	**0.82**	CDD, Area, Cooling Equipment, Construction Material, Temp Control Technique, Building Capacity, Walk in Refrigeration Units, Census Division
MARS	0.80	0.82	CDD, Area, MAINCL8, Construction Material, Temp Control Technique, Mainframe, Server-farm, Glass %, Building Capacity, Walk in Refrigeration Units, Census Division, PCs/ No. of workers, No. of workers
SVM	0.80	0.81	

- From the above table, we can see that Random forest and MARS give low prediction errors.
- To select the best predictive model out of the models above, we checked whether the difference in MSE and MAE values of Random Forest, LM, GAMS and Mars are statistically significant. We performed an ANOVA test and a Pairwise t-test with Bonferroni correction for multiple testing on the residuals of each model. The results are shown below:

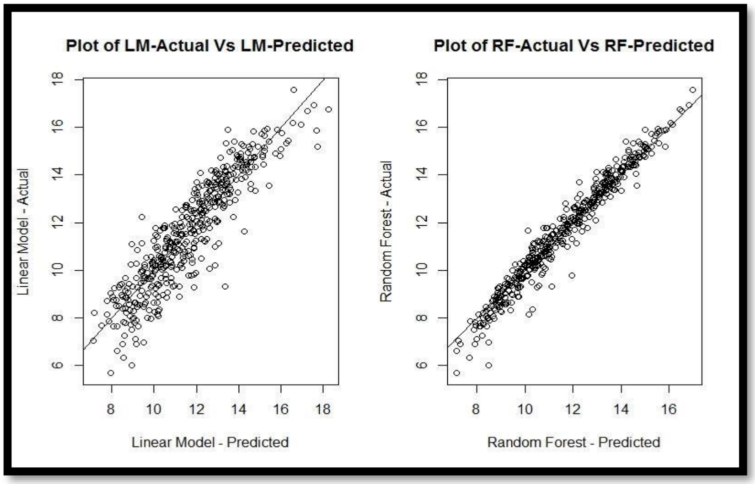
```
   summary(anova)
                Df Sum Sq Mean Sq F value Pr(>F) models
   3      1   0.360   0.099  0.961 Residuals   1784
   6488   3.637
       pairwise.t.test(Residuals, models ,
p.adjust="bonferroni")  Pairwise comparisons using t tests
with pooled SD
   data:  Residuals and models
   a b c b 1 - - c 1 1 - d 1 1
   1
   P value adjustment method: bonferroni
```

We can see that the values for RF, LM, GAMS and MARS are not significantly different.

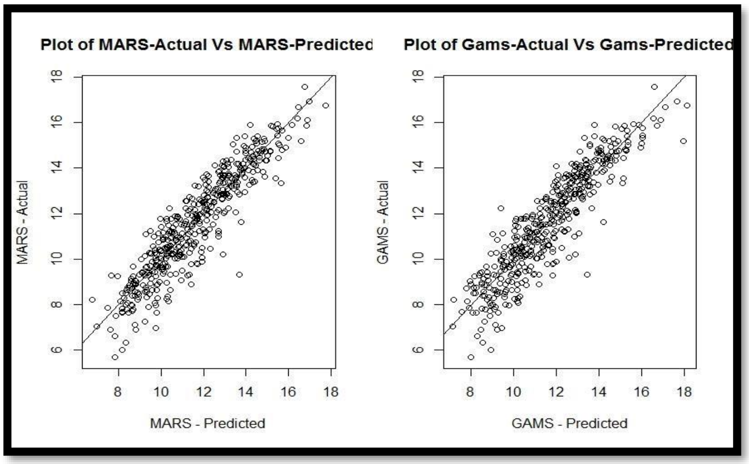
However, we chose random Forest as our best model since it had

- Superior fit
- Best MSE
- Least number of predictors

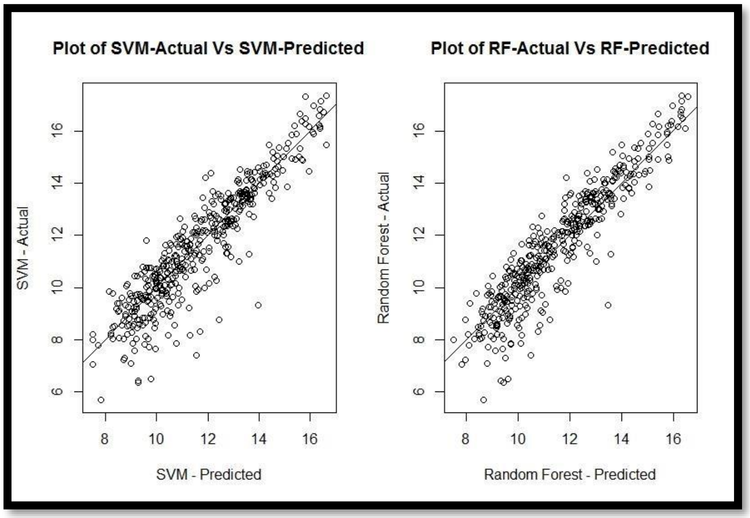
Plots of Actual Vs Fitted Values for RF, GAMS, MARS, SVM and LM



Corelation = 0.904 **Corelation= 0.980**



Correlation = 0.915 **Correlation= 0.906**

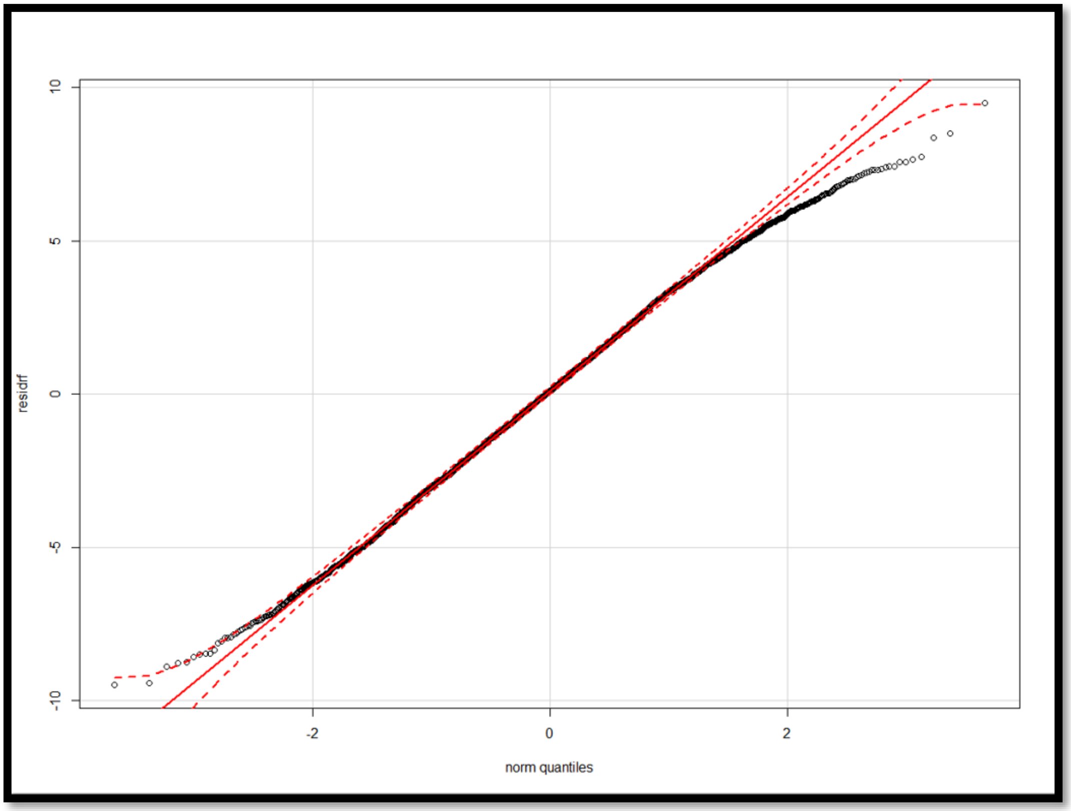


Correlation = 0.91 **Correlation = 0.92**

Results

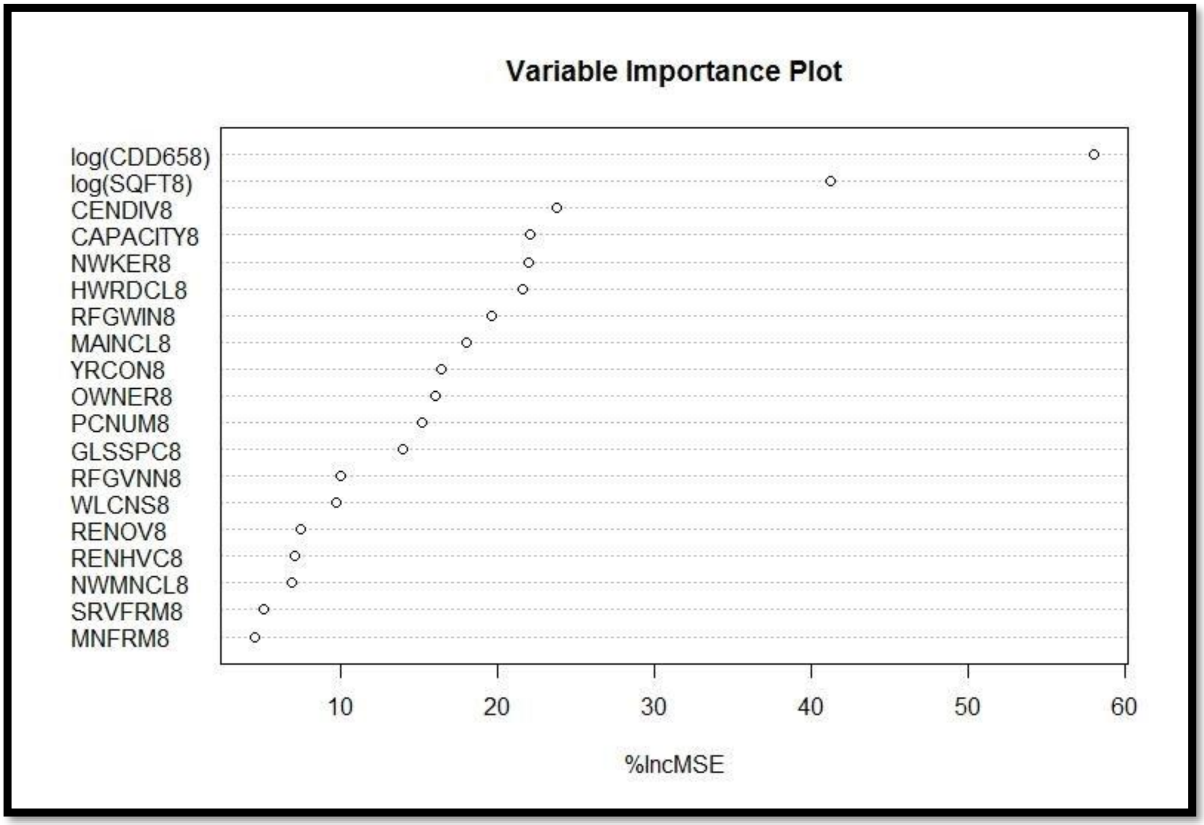
From R^2, MSE values, and the above plots, we can conclude that Random Forests fits the data best, and also gives a very high predictive accuracy. Also, it has useful features like Variable Importance plots, Partial Dependence plots. Hence we chose Random Forest as our final model.

Following is the QQ Plot for residuals:



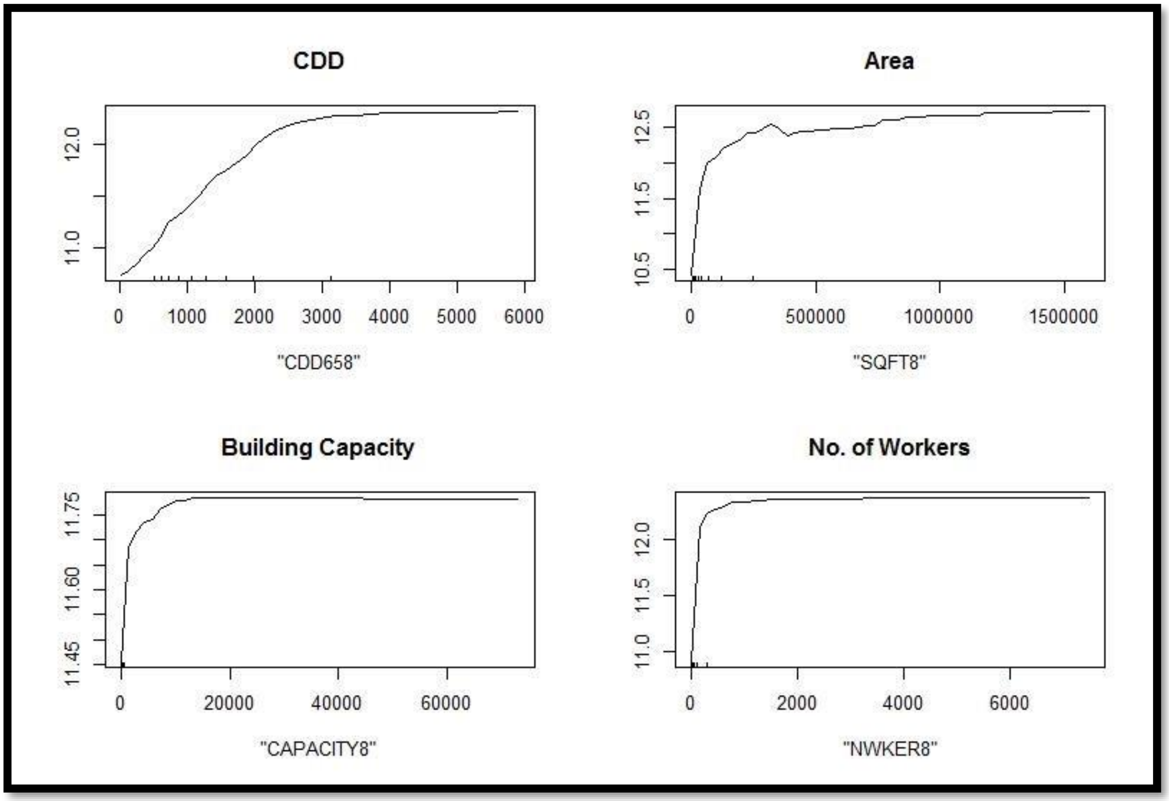
We can see that the plot is fairly normal in the middle, however it has heavy tails.

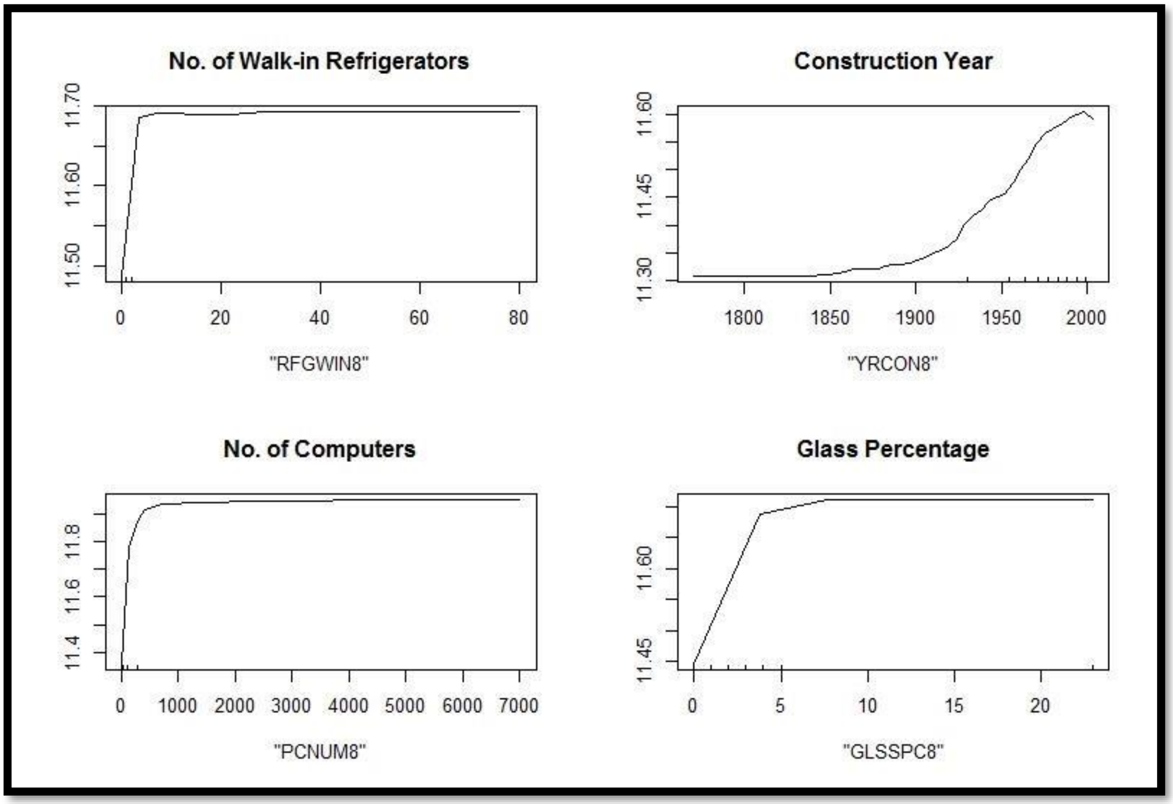
The Variable Importance plot is as shown below:

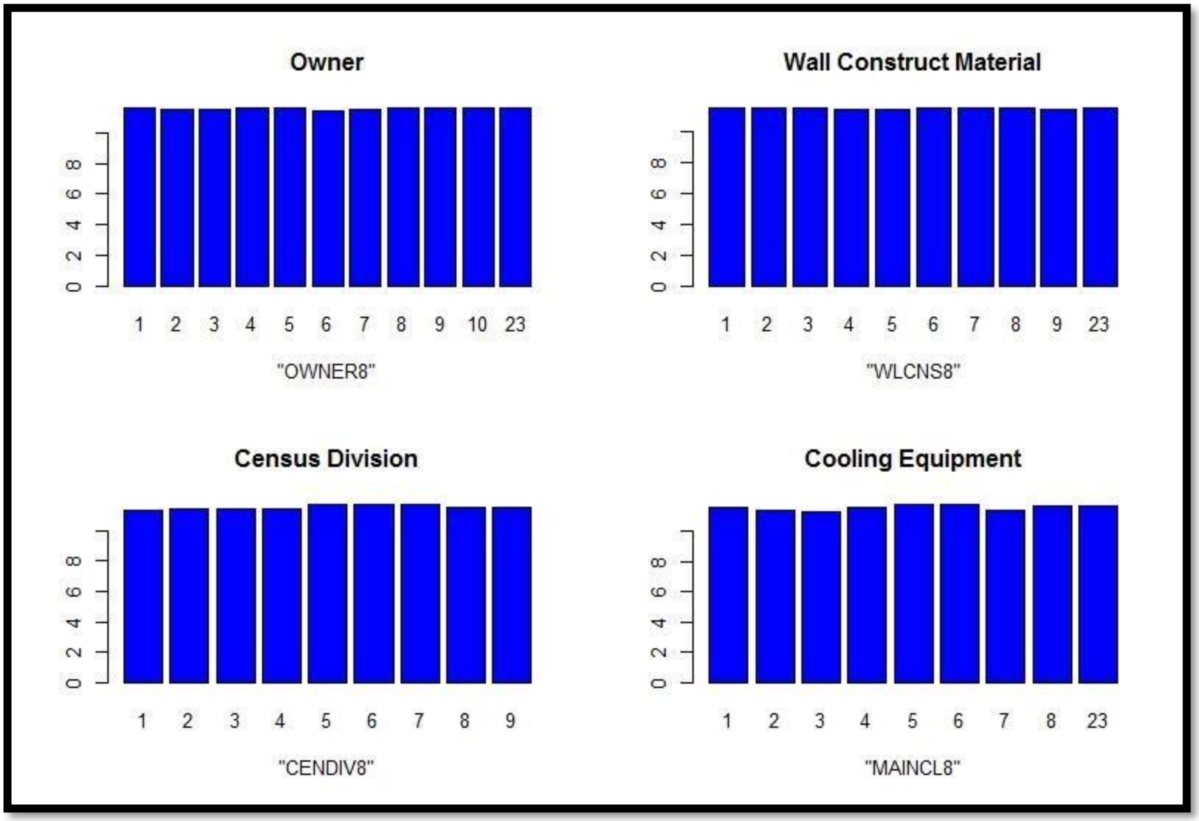


The 16 influential variables in descending order of influence are:

1. Cooling Degree Days – CDD658
2. Area – SQFT8
3. Census Division – CENDIV8
4. Building Capacity – CAPACITY8
5. No. of Workers – NWKER8
6. Cooling Reduction Technique – HWRDCL8
7. No. of Walk-In Refrigerators – RFGWIN8
8. Main Cooling Equipment - MAINCL8
9. Year of Construction – YRCON8
10. Building Owner – OWNER8
11. No. of PCs – PCNUM8
12. Glass Percentage – GLSSPC8
13. No. of Vending Machines – RFGVNN8
14. Wall Construction Material – WLCNS8
15. Renovation done since 1990 – (Yes/No) – RENOV8
16. HVAC Renovation done since 1990 – (Yes/No) – RENHVC8
17. Main Cooling Changed since 1990 - (Yes/No) – NWMNCL8 The Partial Dependence Plots of these 16 predictors have been shown below.


CDD
12.0
11.0
0
1000
2000
3000
4000
5000
6000
"CDD658"
Area
12.5
11.5
10.5
0
500000
1000000
1500000
"SQFT8"
Building Capacity
11.75
11.60
11.45
0
20000
40000
60000
"CAPACITY8"
No. of Workers
12.0
11.5
11.0
0
2000
4000
6000
"NWKER8"


No. of Walk-in Refrigerators
11.70
11.60
11.50
0
20
40
60
80
"RFGWIN8"
Construction Year
11.60
11.45
11.30
1800
1850
1900
1950
2000
"YRCON8"
No. of Computers
11.8
11.6
11.4
0
1000
2000
3000
4000
5000
6000
7000
"PCNUM8"
Glass Percentage
11.60
11.45
0
5
10
15
20
"GLSSPC8"



<u>Inferences</u>

- CDD: We can see that there is a lag in the response of electricity consumed for cooling as compared to other variables. This might be due to good insulation systems which mitigate the effect of sudden temperature changes.

- YRCON: Initially, the development of HVAC systems affected Electricity consumption more than it did after 2000. This might be due to the fact that the marginal contribution of other factors now affects the electricity consumption more than improvements in HVAC system.

- Glass Percentage: We can see a linear increase in partial dependence of Glass percentage in predicting cooling electricity consumption. We can see that cooling electricity consumed increases with increase in glass percentage ie. More glass percentage means more sunlight/heat, and hence needs more cooling, especially during the Summers.

- Census Divisions: Census divisions 5,6 and 7 (Southern divisions) show higher electricity consumption levels, which is pretty intuitive.

- Main Cooling Equipment: Packaged A/C (1), Chilled Water (5) and Central Chillers (6) affect cooling electricity consumption to a higher extent.

Impact

- These models and inferences can be used to study and inspect the reasons for higher electricity consumption (for cooling) by Commercial Buildings in the USA.
- Remedial and preventive measures can be designed in order to prevent over-usage of electricity (ie. Overcooling)
- Different policies / penalties and be formulated for different parts of the country by studying the particular regional electricity consumption patterns.

Future Work

- We intend to use this predictive model to predict electricity consumption for 2012 CBECS data and validate against the actual data published by EIA (Expected to be released in June 2016).
- We expect the predicted values to be different than actual values, and will draw comparisons between commercial buildings- ten years ago and now.
- We also note that in partial dependence plot for Random Forest we see that there is a dip in the marginal contribution of it to the energy consumed due to cooling. We would like to explore the reason for this dip and also see if this decreasing trend continues for the 2012 data.

References

Azadeh, A., Ghaderi, S. F., Tarverdian, S., & Saberi, M. (2007). Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption. *Applied Mathematics and Computation*, *186*(2), 1731–1741. http://doi.org/10.1016/j.amc.2006.08.093

Contreras, S., Smith, W. D., M, T., & Jr., F. (2013). U.S. commercial electricity consumption, (47061).

Derrible, S., & Reeder, M. (2015). The cost of over-cooling commercial buildings in the United States. *Energy and Buildings*, *108*, 304–306. http://doi.org/10.1016/j.enbuild.2015.09.022

Griffith, B., & Crawley, D. (2006). Methodology for analyzing the technical potential for energy performance in the U.S. commercial buildings sector with detailed energy modeling. *SimBuild Conference*. Retrieved from http://www.nrel.gov/docs/fy07osti/40124.pdf

Hastie, T., Tibshirani, R., & Jerome, F. (2009). *Elements of Statistical Learning, Data Mining, inference and Prediction* (2nd editio). http://doi.org/10.1007/b94608

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). *An Introduction to Statistical Learning : With Applications in R*. Springer. http://doi.org/10.1007/978-1-4614-7138-7

Melek Yalcintas1, U. A. O. (2007). An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database. *International Journal of Energy Research*, *31*(August 2007), 135–147. http://doi.org/10.1002/er

Payne, C. (2006). The Commercial Energy Consumer : About Whom Are We Speaking ? Commercial Sector Investment in Energy Efficiency. *ACEEE Summer Study on Energy Efficiency in Buildings*, 215–225.

Appendix

The Codebook and link to the data source is as shown below:

https://www.eia.gov/consumption/commercial/data/2003/index.cfm?view=microdata