Report

IE590 Predictive Modeling

Yidan Nie

Objective: develop a predictive model of the total electricity used for cooling in the "South" region of the United and identify the main predictors.

Part 1: Modeling Fitting

1.1 Data Cleaning and Preprocessing

The original CBECS data set is highly granular and contains too much information. Hence, the first step of cleaning is deleting factors unrelate to cooling. Based on the common sense, I deleted features about heating and lighting and remained totally 70 features relate closely to the electricity consumption for cooling. With 2592 observations, I further grouped 39 predictors into six categories including electricity, climate, cooling, building construction, building occupation and other features.

For the data cleaning part, I first detected rows without label and columns with more than 20 percent of NAs. Those observations without response variable would be deleted since there is no ground truth value to be checked for our prediction. And for those features lost more than 20 percent of values, it is also dropped since the recoding would increase too much bias. Also, it should be noted that I dropped 140 observations which didn't use electricity for cooling (values in ELCOOL as 2).

For the imputation part, since several features with missing values provide detailed information about each building's construction, it is clear that simply imputation with mean is not a good method which would largely change those the distribution of those features. Hence, I also dropped all rows with NAs in building construction features and using the median value to recode other slot of NAs, typically for the numerical factors such as number of laptops.

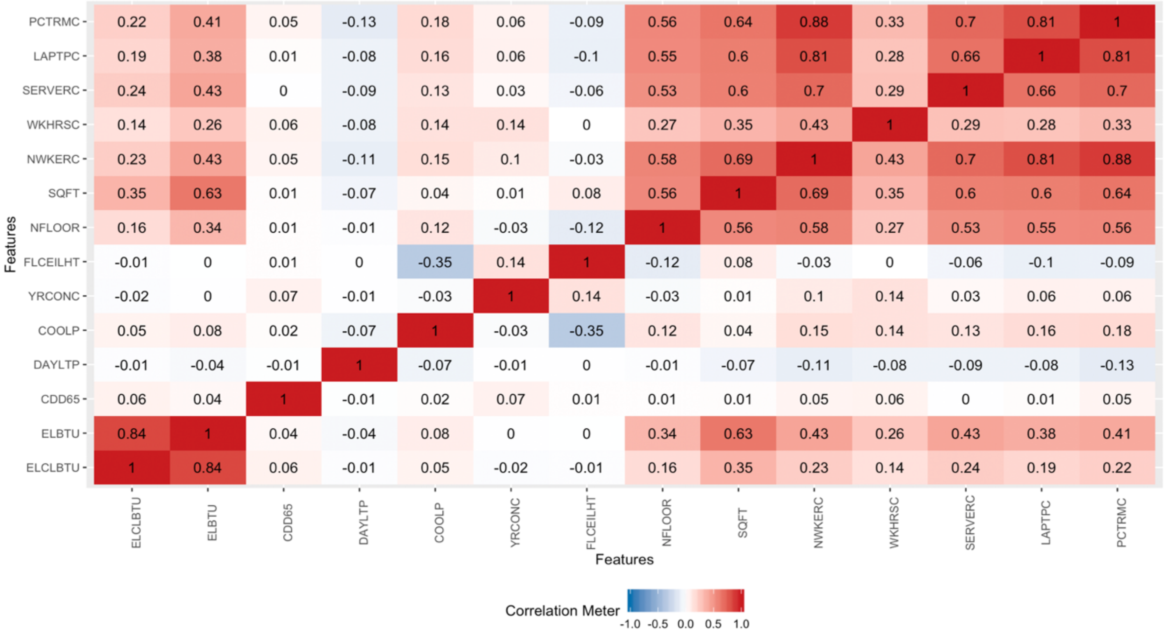
For the class of each feature, I consider both the physical meaning and convenience for manipulation parts. Since our data set is from census, majority of them are naturally in class of factor, such as feature "CENDIV" and "PUBCLIM". It can be seen in the table.1 that those features are in class of unordered factor. For some features about floor to ceiling's height I set them as numerical variables. Also, I treated some ordinal features as numerical features. For example, the categorial feature about the number of laptops describes the number in a bucketing setting. To make them as numerical features, the assumption that the numerical distance between each set of subsequent categories is equ.al was made.

The clean data set for modeling contains 2269 observations with 26 predictors. According to the objective, the response variable is "BLCLBTU" which is electricity cooling use (thous Btu).

Table.1 The predictors variables for modeling

Electricity	ELBTU	Num	Annual electricity consumption (thous Btu)
	ELCOOL	Char	Electricity used for cooling
Climate	CENDIV	Char	Census division
	PUBCLIM	Char	Building America climate region
	CDD65	Num	Cooling degree days (base 65)
Cooling	MAINCL	Char	Main cooling equipment
	COOLP	Num	Percent cooled
Building Construction	AWN	Char	External overhangs or awnings
	REFL	Char	Reflective window glass
	TINT	Char	Tinted window glass
	WINTYP	Char	Window glass type
	YRCONC	Char	Year of construction category
	FLCEILHT	Num	Floor to ceiling height
	NFLOOR	Num	Number of floors
	RFCOOL	Char	Cool roof materials
	RFCNS	Char	Roof construction material
	WLCNS	Char	Wall construction material
	SQFT	Num	Square footage
Building Occupation	NWKERC	Num	Number of employees category
	WKHRSC	Num	Weekly hours category
	GOVOWN	Char	Government owned
	PBA	Char	Principal building activity
Others	SERVERC	Num	Number of servers category
	LAPTPN	Num	Number of laptops
	PCTRMC	Num	Number of computers category

1.2 Data Exploration

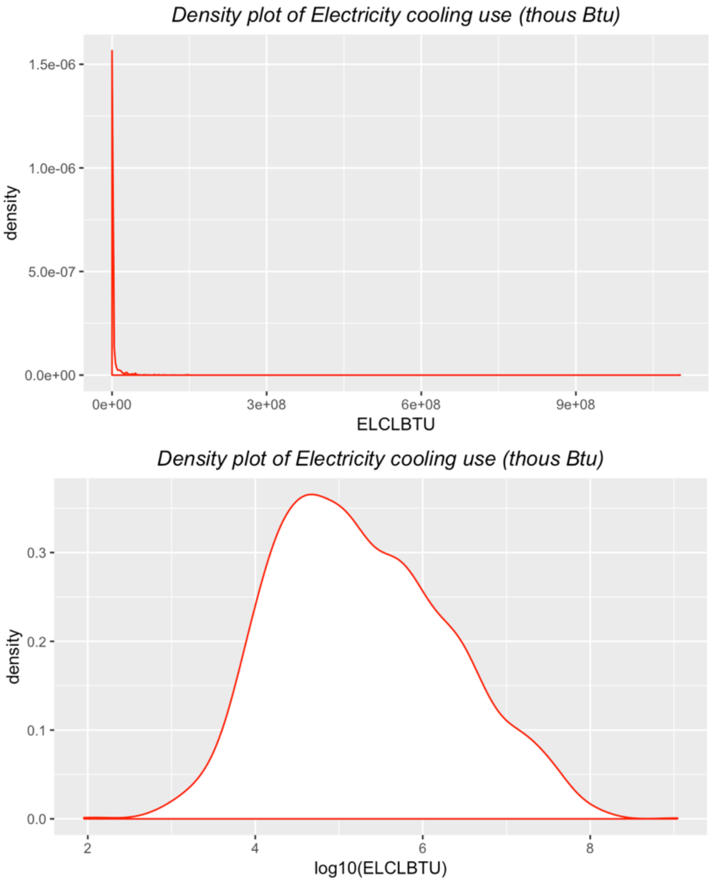


	ELCLBTU	ELBTU	CDD65	DAYLTP	COOLP	YRCONC	FLCEILHT	NFLOOR	SQFT	NWKERC	WKHRSC	SERVERC	LAPTPC	PCTRMC
PCTRMC	0.22	0.41	0.05	-0.13	0.18	0.06	-0.09	0.56	0.64	0.88	0.33	0.7	0.81	1
LAPTPC	0.19	0.38	0.01	-0.08	0.16	0.06	-0.1	0.55	0.6	0.81	0.28	0.66	1	0.81
SERVERC	0.24	0.43	0	-0.09	0.13	0.03	-0.06	0.53	0.6	0.7	0.29	1	0.66	0.7
WKHRSC	0.14	0.26	0.06	-0.08	0.14	0.14	0	0.27	0.35	0.43	1	0.29	0.28	0.33
NWKERC	0.23	0.43	0.05	-0.11	0.15	0.1	-0.03	0.58	0.69	1	0.43	0.7	0.81	0.88
SQFT	0.35	0.63	0.01	-0.07	0.04	0.01	0.08	0.56	1	0.69	0.35	0.6	0.6	0.64
NFLOOR	0.16	0.34	0.01	-0.01	0.12	-0.03	-0.12	1	0.56	0.58	0.27	0.53	0.55	0.56
FLCEILHT	-0.01	0	0.01	0	-0.35	0.14	1	-0.12	0.08	-0.03	0	-0.06	-0.1	-0.09
YRCONC	-0.02	0	0.07	-0.01	-0.03	1	0.14	-0.03	0.01	0.1	0.14	0.03	0.06	0.06
COOLP	0.05	0.08	0.02	-0.07	1	-0.03	-0.35	0.12	0.04	0.15	0.14	0.13	0.16	0.18
DAYLTP	-0.01	-0.04	-0.01	1	-0.07	-0.01	0	-0.01	-0.07	-0.11	-0.08	-0.09	-0.08	-0.13
CDD65	0.06	0.04	1	-0.01	0.02	0.07	0.01	0.01	0.01	0.05	0.06	0	0.01	0.05
ELBTU	0.84	1	0.04	-0.04	0.08	0	0	0.34	0.63	0.43	0.26	0.43	0.38	0.41
ELCLBTU	1	0.84	0.06	-0.01	0.05	-0.02	-0.01	0.16	0.35	0.23	0.14	0.24	0.19	0.22

Figure. 1 correlation matrix

The correlation matrix of the response variable and continuous predictors are visualized in Figure. 1. It can be seen there is a high correlation among building occupation features and other features. For example, number of workers (NWKERC) and number of computers (PCTRMC) has a Pearson correlation of 0.88. Also, the linear correlation between electricity for cooling and annual electricity consumption is 0.84.

The distribution of the response variable is showing below in Fig. 2. It is clear that its distribution is heavily positive-skewed with a long tail leaning to the right side. After the log transformation, the distribution looks better.


Density plot of Electricity cooling use (thous Btu)

Density plot of Electricity cooling use (thous Btu)

Figure. 2 Distribution of response variable

1.3 Data Modeling

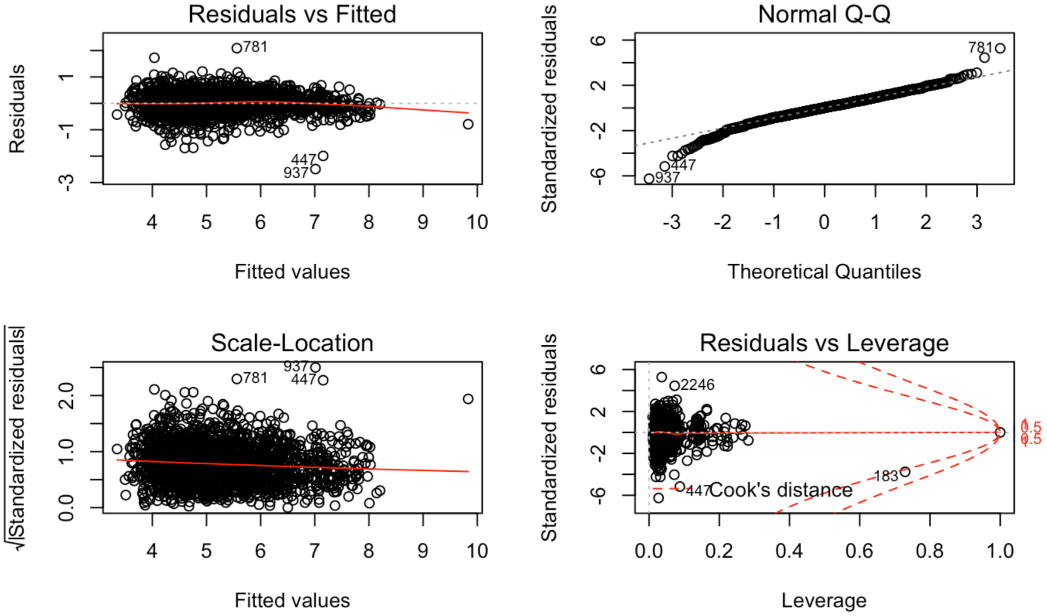
I first standardized all the numerical features for better fitting of linear models. Then I set the seed as 590 and split my data into 80% training and 20% testing. In the modeling part, from relatively simple but interpretable models like linear regression to highly complex models like support vector machine are considered into the method set. When the model is training, grid search method is used to determine the optimal hyperparameter's setting. To utilize the most of dataset, a 5-fold cross validation is used to provide the mean error metrics to evaluate model performance. After model fitting, RMSE, MAE and percentage improvement over the null model are used to assess the predictive performance of each model.

1.3.1 Null Model

The null model is only using the mean of response variable to predict cooling electricity usage. It is used to be a baseline and compare with metrics of other models to check if there is any improvement.

1.3.2 Generalized Linear Model

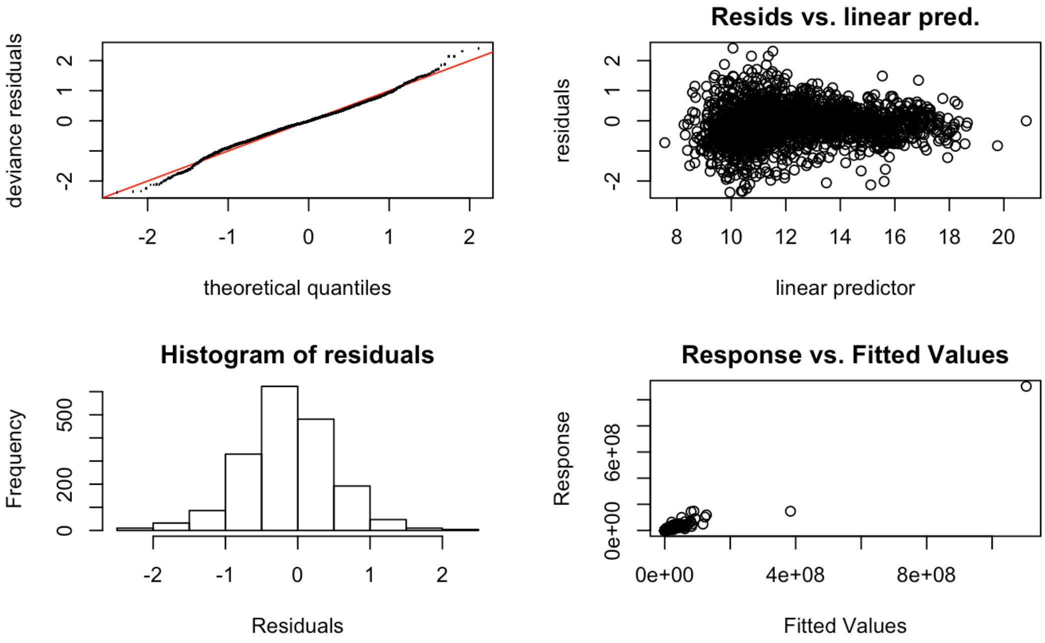
Through the diagnostic graphs, it can be seen that many assumptions of linearity are violated. The funnel shape of residuals plot gives the signal that the variance of residual is not constant and there may be some non-linear terms or interaction of terms should be added into the model. Since the distribution of response variable is not normally distributed, even it is transformed by log, the performance of prediction of linear model is still not good.



1.3.3 GAM

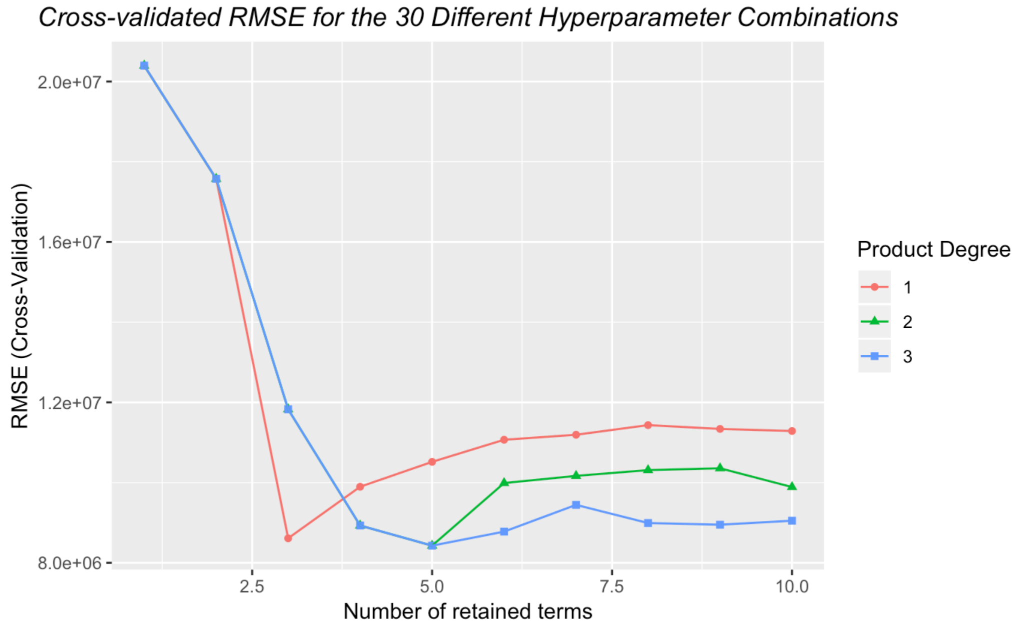
GAM allows non-linear and keeps the variables depend linearly on smooth functions. Since it is a generalized linear model, the standardization of numerical predictors and log transformation of response variable are required for this model.

The model is trained by only setting all the numerical variables as non-linear terms. Then the variables have weak effect for non-linearity are dropped out of the model. Through the diagnostic graphs, it can be seen that a little deviation in the bottom left-hand side and upper left-hand side which indicates the normality of residual is not perfect. And the plot of residual vs. fitted value shows a funnel shape which means the assumption of constant variance of residual is violated.



1.3.4 MARS

To figure out the optimal hyperparameter setting, I use a five-fold cross validation to do grid search based on degree of terms and number of terms retained. And the graph below shows that when number of retained term is 5, both green and blue lines reach the lowest RMSE point. To keep model simpler, I picked the optimal combination as nprune equals 6 and degree 2.



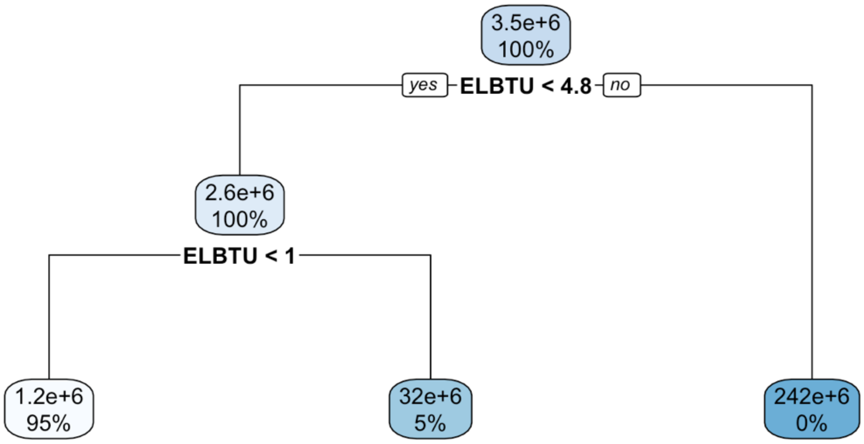
From the coefficient of terms, it can be seen that the annual electricity usage (ELBTU) is the dominated predictor, and main cooling equipment (MAINCL) as well as percentage of cooled space (COOLP) also plays an important role in prediction.

```
                               coefficients
(Intercept)                        51378404
h(3.45756-ELBTU)                  -13946211
h(ELBTU-3.45756)                  -37536127
h(ELBTU-3.45756) * MAINCL3         53992375
h(ELBTU-3.45756) * COOLP           34201531

Selected 5 of 8 terms, and 3 of 66 predictors
```

1.3.5 CART

To tune the hyperparameter of CART model, a 5-fold cross validation is used. With 2 split nodes, the optimal cp-value is 0.01. And we could see that only ELBTU is split since it’s the most dominant predictor which makes the model tends to over-fit and have a high variance.



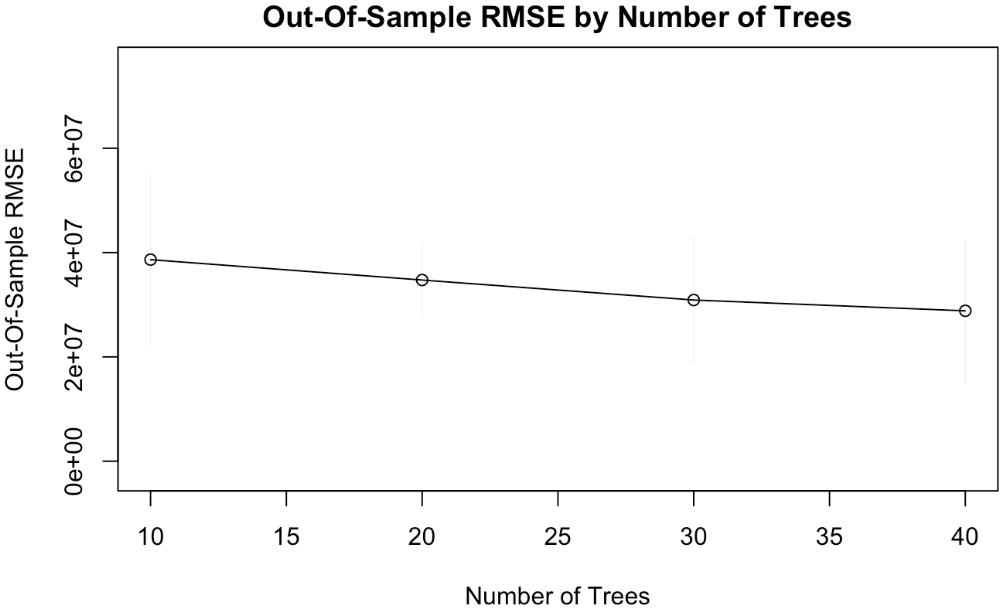
1.3.6 Random Forest

Random forest model could decrease the correlation among trees and improve the predictive power. For this random forest model, I tuned four hyperparameters and the result is showing below.

Name	Range	optimal
Number of variables randomly sampled at split	[2, 7 … 25]	22
minimum number of samples within the terminal nodes	[3, 5, 7]	3
number of samples to train on	[0.55, 0.632, 0.8]	0.8
Number of trees	[100, 200, 300, 400]	100

1.3.7 BART

The BART model uses a prior and likelihood to estimate the posterior distribution of the prediction. Since our data set is small and response variable has a really large range, by intuition I think BART model would not perform well on this trial. First a 5-fold cross validation is performed to figure out the optimal hyperparameter setting. Through the figure we could see that increasing number of trees gives a decreasing trend of out-of-sample RMSE.



1.3.8 SVM

SVM is a supervised model aiming to find a hyperplane which could maximized the number of observations which fall within the given margin region. The value of cost of a constraints violation and gamma was tuned by 5-fold cross-validation to determine the optimal hyperparameter setting. The result shows that the optimal value for cost and gamma are 10 and 0.5 respectively.

Part 2: Model Evaluation

The candidate models are compared to null model respectively by using out-of-sample Mean Absolute Error (MAE) and out-of-sample Root Mean Squared Error (RMSE). The percentage improvement is the percentage of improvement against null model and red colored number means a negative value. From the table below, it can be seen that only two model have both metrics are negative which means the value of its error metrics is smaller than the null model. And the best model is the random forest model which largely improve the predictive power among other competing models.

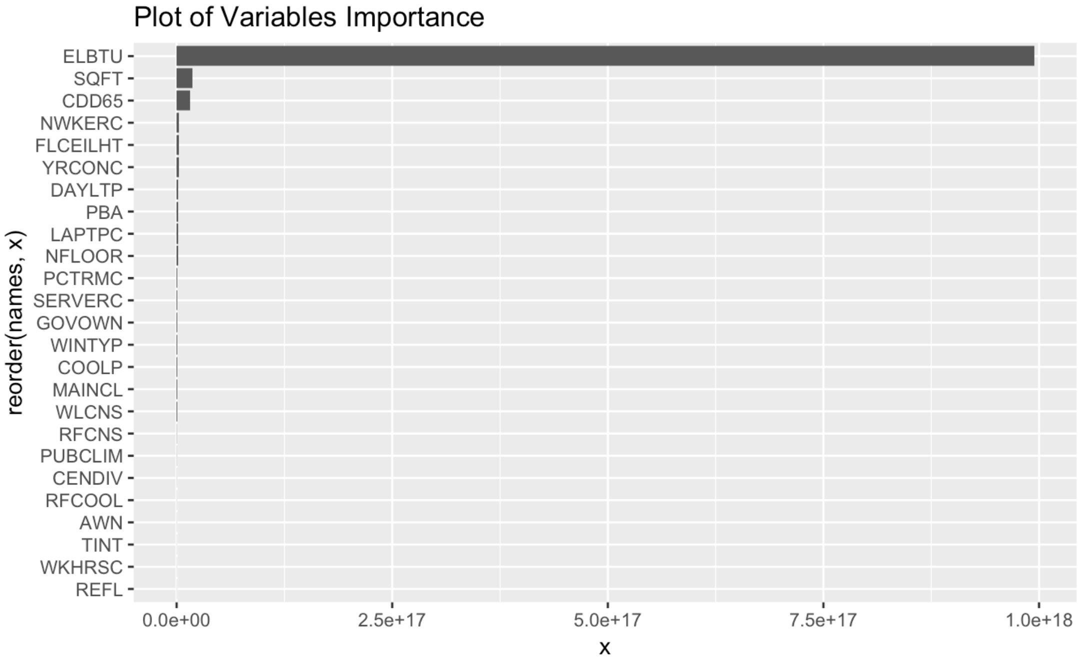
Model	Metric	Value	Percentage Improvement
Null Model	R2	-	-
	RMSE	9758669	-
	MAE	4525803	-
GLM	R2	0.85	-
	RMSE	9869598	0.01
	MAE	989919	(0.78)
GAM	R2	0.93	-
	RMSE	10075851	0.03
	MAE	2649631	(0.41)
MARS	R2	0.98	-
	RMSE	7392946	(0.24)
	MAE	1760455	(0.61)
CART	R2	0.38	-
	RMSE	14738392	0.51
	MAE	3252657	(0.28)
Random Forest	R2	0.44	-
	RMSE	6288001	(0.36)
	MAE	1388738	(0.69)
BART	R2	-	-
	RMSE	29887787	2.06
	MAE	5987049	0.32
SVM	R2	-	-
	RMSE	10038574	0.03
	MAE	2604797	(0.42)

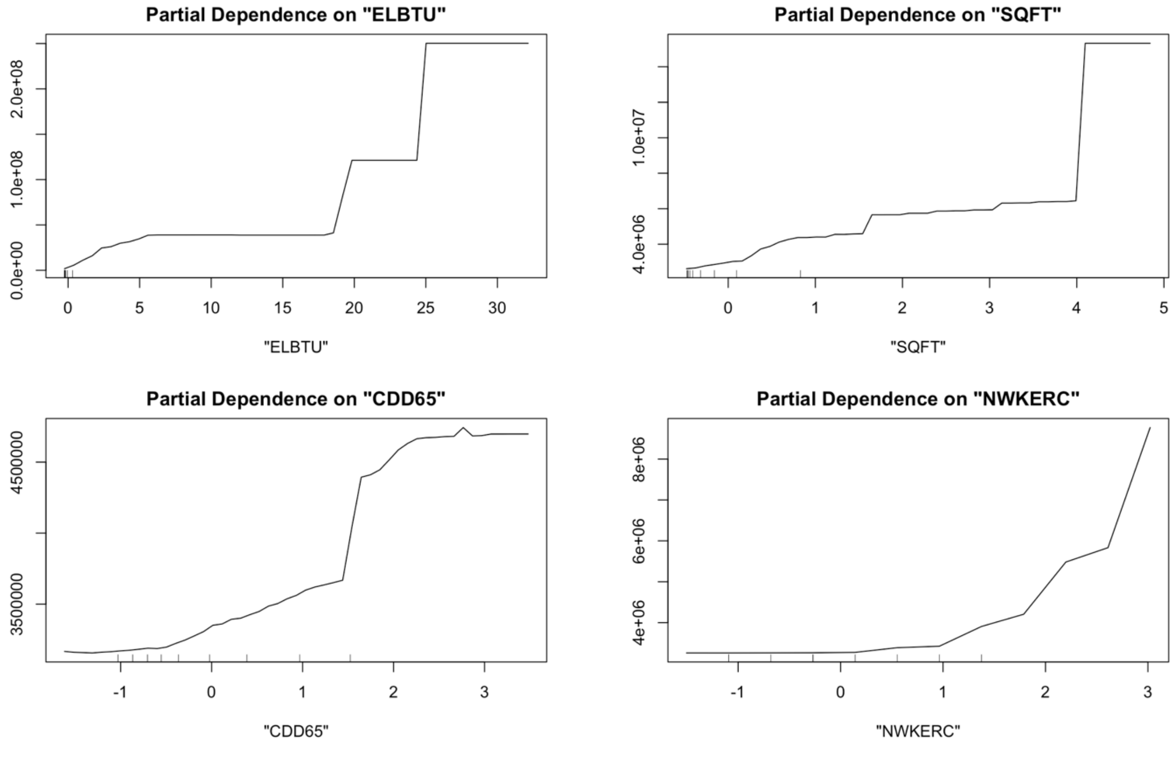
Besides the great performance on error metrics, Random Forest is a non-parametric model which could better capture the non-linearly of the response variable. Also, the Random Forest model could do feature selection in its nature development.

Part 3: Model Inference

The inference of Random Forest model is implemented by plots of variable importance and plots of partial dependence between predictors and response variable.

The Plot below shows variable importance which is measured by recording the decrease in MSE each time a variable used as a node split in a tree. It can be seen that the annual electricity usage, square footage, cooling degree days are the most important variables in year 2012. And the most of the top ten variables are from building occupation features and climate features.





Because my predictor is standardized, the x axis has the negative side. For the annual electricity usage, it is clear that cooling electricity usage is a portion of annual electricity usage, so there is a positive trend. From the CDD65 partial dependence plot, it can be seen that the higher CDD is associated with higher electricity usage for cooling. Also, the relationship between square footage and electricity usage for cooling is also positive, which indicates that the increasing number of buildings would increase the cooling space. The plot of partial dependence between number of employers and response variable shows a positive relationship and the growth rate of electricity for cooling increases when number of employers reach a specific category.