IE 590 Final Take Home Exam

Manan Shah

April 24, 2018

Contents

1 Introduction

Residential Energy Consumption Survey (RECS 2009) dataset is used to predict the Electricity used in space conditioning i.e heating, cooling and water heating. The prediction for Electricity usage is done for the state of Florida. The RECS data contains 948 observations for the state of Florida. Now,performing some elementary analysis, we find that out of 113.6 million households in United States, 6.15 percent of the households are in Florida state.

Now, to predict Electricity usage for space conditioning in FL, some important predictors are selected out of the 940 variables present in the RECS dataset. The EIA RECS dataset is collected through detailed surveys of the households in US. The algorithms used to estimated electricity usage for space conditioning are a function of building's average thermal conductance, heating and cooling degree days and also take into account the number of water heater tanks in a household and the type of roof material used. The predictors selected to conduct the statistical analysis are:

Table 1: Predictors Selected

Climatic and Population Variables
HDD65 - Heating Degree Day
CDD65
NWEIGHT
UR - urban/rural
Building Features
HIGHCEIL
DOOR1SUM
WINDOWS
TYPEHUQ - Type of Housing Unit
NUMFLRS
NUMAPTS
ROOFTYPE
TOTROOMS
CELLAR
ATTIC
Heating Features
MAINTHT - Maintenance of main space heating
EQUIPAGE - Age of Space heating equipments
EQMAMT - Portion of space heating provided by main heater
BASEHEAT
HEATROOM
ATTICHEAT
GARGHEAT
Water Heating Features
STEAMR
NUMH2ONOTNK - No of tanks
NUMH2OHTRS
H20TYPE1
WHEATOTH - Water heater used by other units
WHEATSIZ
WHEATAGE
H20TYPE2
WHEATSIZ2
WHEATAGE2
Cooling Features
AIRCOND
ACOTHERS
CDOLTYPE
AGECENAC - Age of AC
ACROOMS
USEWWAC - Most used window/wall AC in summer
BASECOOL
ATTICCOOL
GARGCOOL
Miscellaneous Features
TOTSQFT - Total heated square footage
TOTCSQFT
DOLLAREL - Total electricity cost

2 Explanatory Data Analysis

2.1 Correlation Plot

The response variable is the sum of the predictors 'KWHSPH', 'KWHCOL' and 'KWHWTH' which are the individual variables representing the electricity usage for heating, cooling and water heating respectively. The correlation matrix of the response variable (electricity usage for space conditioning) and the predictors is developed below in Figure 1. The blue color here represents positive correlation and the red color represents negative correlation. The degree of correlation is represented by the size of the circle, with larger circles indicating larger correlation.

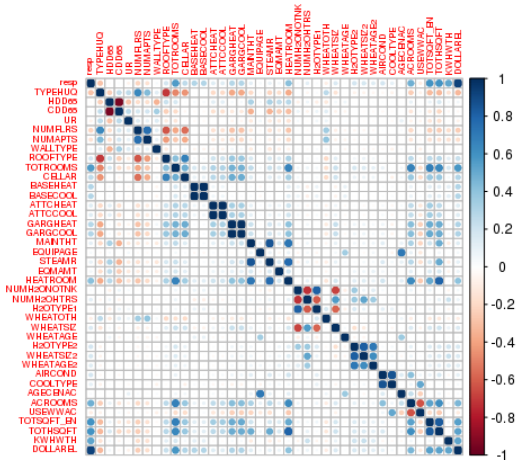

Figure 1: Correlation Plot

Inspecting the Figure 1, we can see that the response variable has significant correlation with 'DOLLAREL', 'TOTHSQFT', 'TOTCSQFT' and 'TOTROOMS'. It has a correlation greater than 0.6 with 'ACROOM' also. Also, we can see that 'CDD65' and 'HDD65' have high negative correlation between them. This is expected as both are somewhat correlated predictors. We have similar case for 'BASECOOL' and 'BASEHEAT' with high positive correlation. Lastly, the Type of Housing Unit has a strong negative correlation with the Rooftype.

2.2 Descriptive Statistics

To understand the response variable better, a histogram of the response variable is plotted. Figure 2 depicts the empirical cumulative distribution of the electricity usage for space conditioning in Florida. The blue tail indicates the normal distribution fit to the tail for the empirical electricity usage data. Fig. 2 indicates that the mean electricity usage is around 5000 KWh. The histogram is skewed a little to the left, but the figure indicates that the response variable follows a normal distribution.

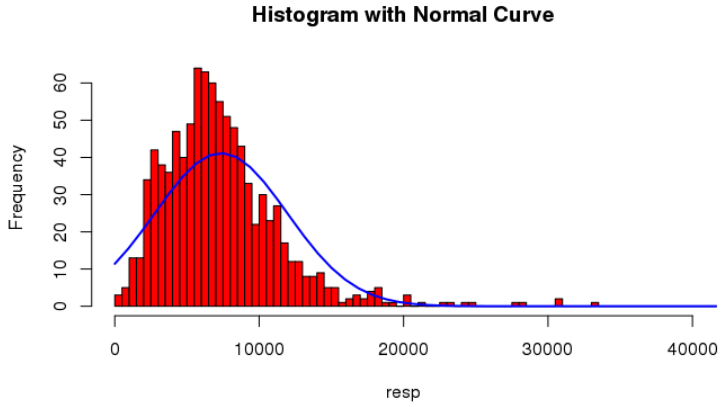

Figure 2: Histogram of response variable

Figure 3 depicts the violin plot of electricity usage for space conditioning for different types of housing units. A violin plot combines a box-plot and a kernel density plot. Fig. 3 indicates that the Single Family detached housing unit has a lot of variation in electricity used for space conditioning. The fatter violin plot due to skewness of electricity usage for Single Family detached housing unit.

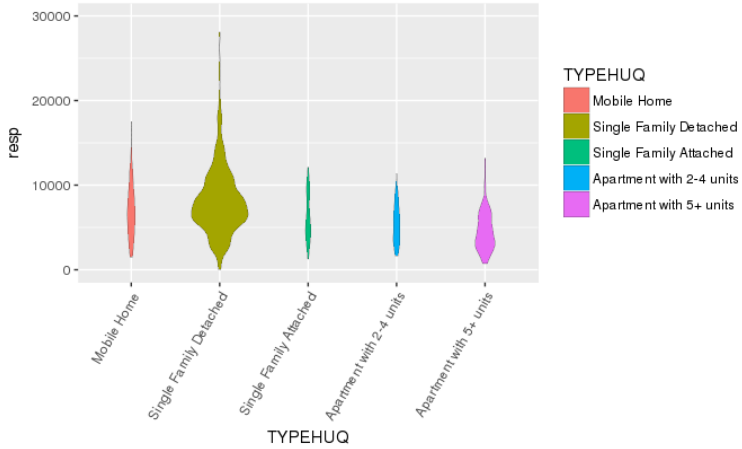

Figure 3: Histogram of response variable

Figure 4 depicts a violin plot of electricity usage for space conditioning for urban and rural areas. The plot indicates that the electricity usage is skewed a lot in the urban areas with higher usage than in rural areas.

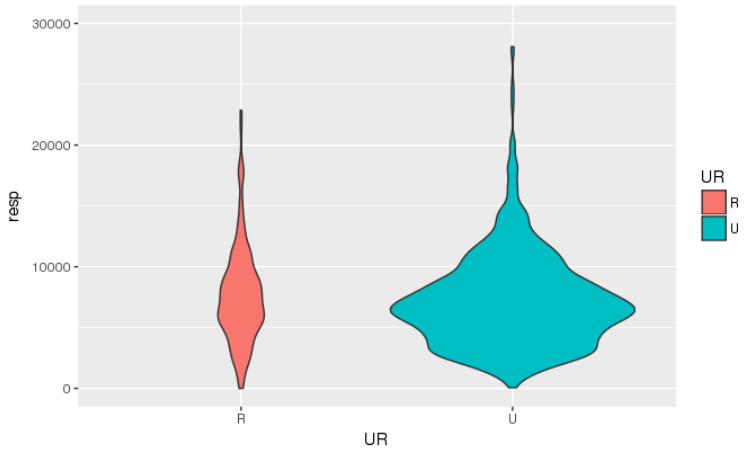

Figure 4: Histogram of response variable

It is important to note that a lot of predictors had high placeholder values like 41 instead of 4 which have been changed appropriately. Also, predictor like 'NUMFLRS' have values like -2 for Not Applicable which do not make analytical sense as there cannot be negative number of floors. So, these values have been changed to 0.

3 Models

3.1 GLM

The Y vs Y-hat and other residuals plots for GLM are given as:

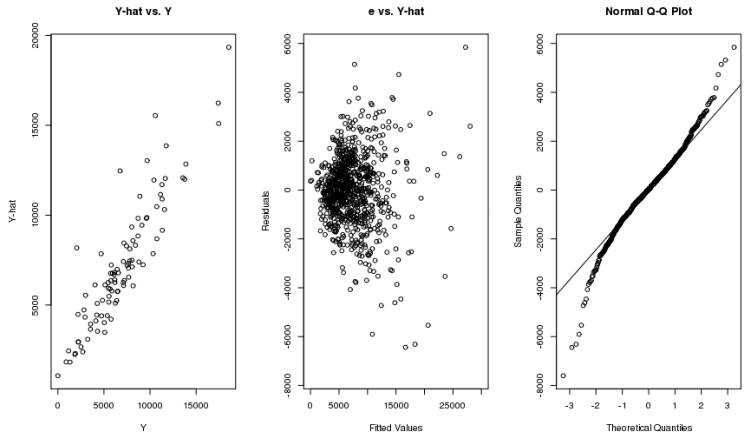

Figure 5: Inference plots

The Y vs Y-hat graph seems okay but the QQ-plot has residuals tailing off at the ends. This may mean the glm is not able to build a good model due to the presence of outliers.

3.2 CART

The boxplot comparing the rmseOS for unpruned and pruned CART models is given as:

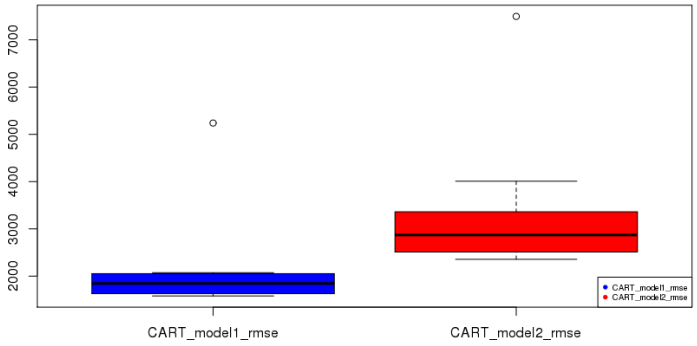

Figure 6: Comparing rmseOS of unPruned and Pruned CART

The Y vs Y-hat graph seems okay. The residuals also behave satisfactorily.

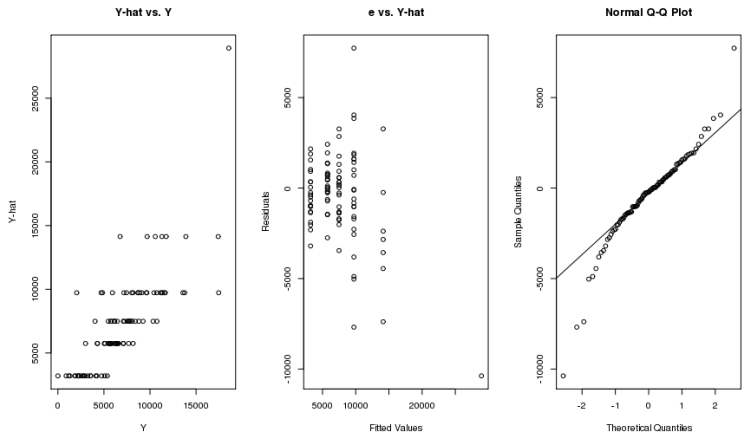

Figure 7: Residual Plots of CART

3.3 MARS

The boxplot comparing the rmseOS for unpruned and pruned MARS models is given as:

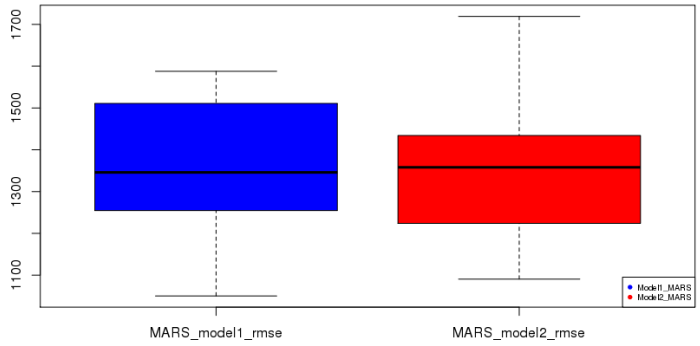

Figure 8: Comparing rmseOS of unPruned and Pruned MARS

The Y vs Y-hat graph seems okay. The residuals also behave satisfactorily.

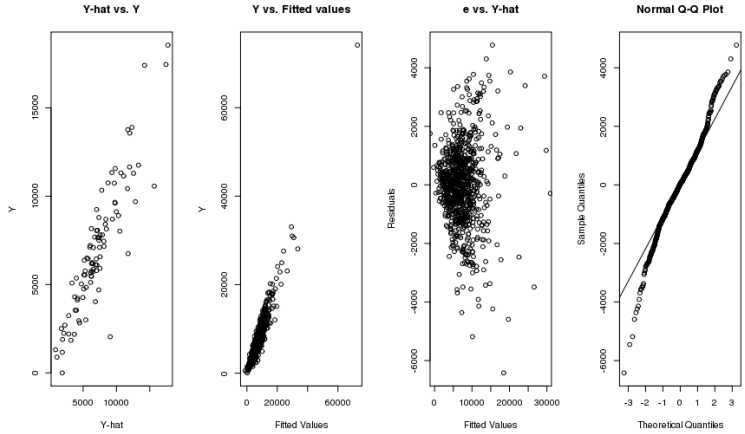

Figure 9: Residual Plots of MARS

3.4 Random Forest

The variable importance plot to select the important variables is:

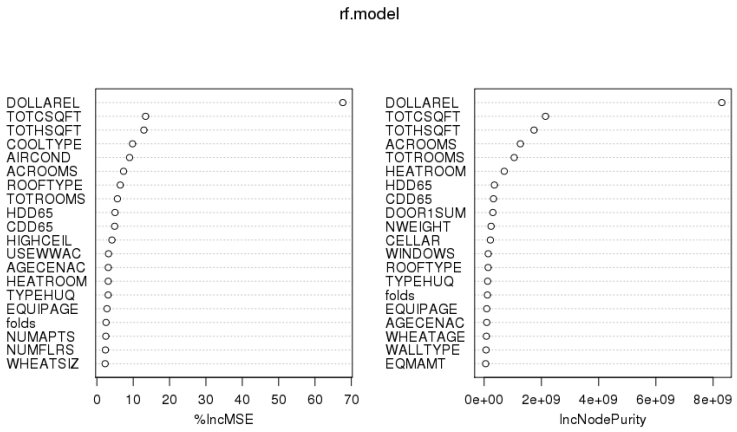

Figure 10: Variable Importance plot for Random Forest

The Y vs Y-hat graph seems okay. The residuals also behave satisfactorily.

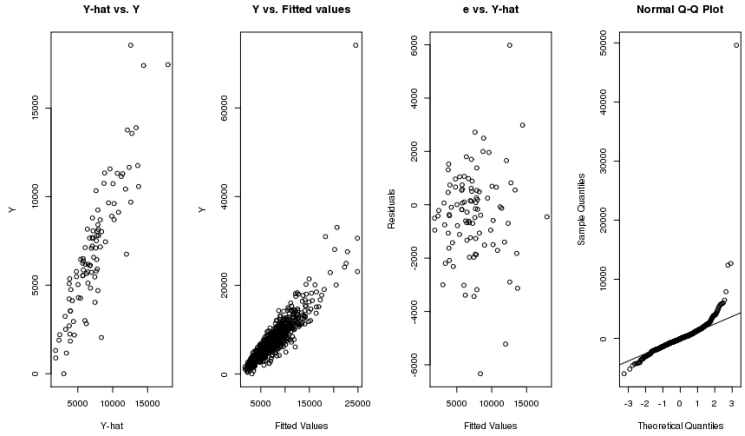

Figure 11: Residual Plots of Random Forest

3.5 BART

The variable importance plot to select the important variables is:

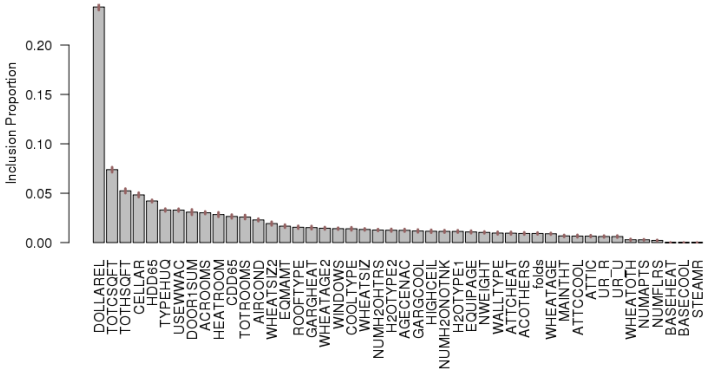

Figure 12: Variable Importance plot for BART

The Y vs Y-hat graph seems satisfactory.

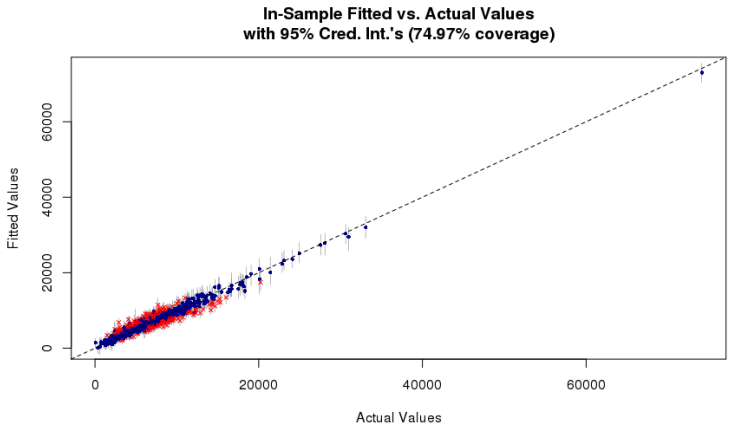

Figure 13: Fitted vs Actual BART

The residuals seem to be satisfactory.

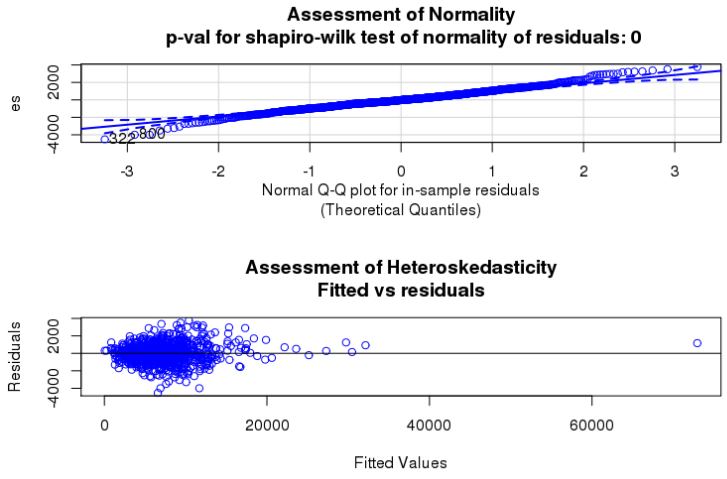

Figure 14: Residual Analysis BART

The partial dependency plots for some important variables are:

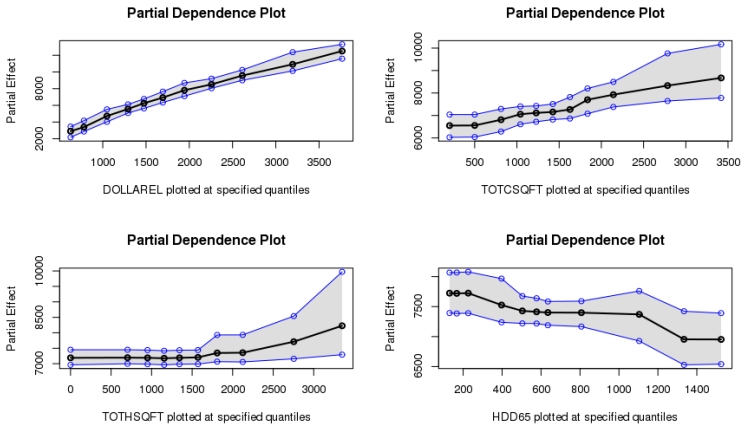

Figure 15: Partial Dependency plots BART

From the partial dependency plots, it can be observed that electricity usage increases with increase in predictors 'DOLLAREL', 'TOTHSQFT' and 'TOTCSQFT' while decreases with increase in 'HDD65'. This seems logical as electricity increases with more square footage of heated/cooled area as well as the total cost.

4 Final Model

Several models were used for predictions leveraging 7 different algorithms. Generalized Linear Model (GLM), Generalized Additive Model (GAM), Classification and Regression Trees (CART), Random Forest, Multi-Adaptive Regression Splines (MARS), Bayesian Additive Regression Trees (BART) and Support Vector Machines (SVM) were the algorithms used to build various models for prediction of electricity usage for space conditioning in Florida. These models were evaluated using Cross-Validation for the purpose of model selection and predictive accuracy.

The data was split into a training and testing dataset. A 10-fold cross-validation was used to train the data using the above mentioned algorithms. Variable selection for the model was done using the indigenous variable selection techniques available for each of the algorithm. Finally, the model was tested for predictive accuracy on the out-of-sample testing data. The Root Mean Squared Error for the out-of-sample test data was calculated and this metric was used for the model comparison. Lower the average out-of-sample RMSE value for a model, higher would be the predictive accuracy of the model. The average out-of-sample rmse values for all the models is given as:

Table 2: rmseOS Comparison of Models

GLM	GAM	CART	randomForest	MARS	BART	SVM
1498.5	1515	1838.6	1565	1615.5	2033.6	4211

Although the glm model has the least rsmeOS value, the final model selected is the gam model. This is because the dataset contains a lot of predictors with categorical values, for which the generalized linear model would not be useful. Also, it wouldn't be able to explain the non-linearities in the data.

The final model equation is given as:

finalmodel<-gam(resp s(TYPEHUQ, d = 4) + s(HDD65, d = 4) + s(TOTCSQFT, d = 4) +s(TOTCSQFT, d = 4) + s(DOLLAREL, d = 4) + NUMH2OHTRS + WHEATSIZ2 + ACOTHERS, data = df.train)

4.1 Final Model Diagnostics

The final gam model was build iteratively by initially giving all the variables a smoothing function. Using the ANOVA tests, the significant variables were selected iteratively by reducing the non-significant variables over each iteration. For the model with all the variables significant, **step.Gam()** function was used to determine the smoothing constant for the selected variables. The final model was selected as the one with the lowest AIC value.

The figures below indicate the Observed vs Predicted and Residuals vs Predicted plots. Figure 16 indicates the accuracy of the predictions with respect to the actual values. If the points are away from the linear line, it means the predictive accuracy is not good enough. Since the predicted values and observed values are somewhat linearly related, the predictive accuracy is good for this model.We can observe that there are a couple of outliers in the data which may skew the results a little.

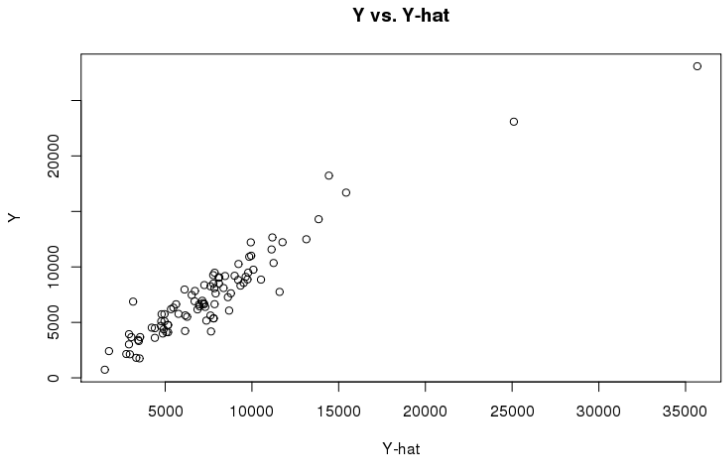

Figure 16: Y vs Y-hat

Figure 17 indicates whether the residuals hold the assumption of homoscedasticity. Although the points appear to be skewed, there doesn't seem to be a pattern which means the assumption of constant variance of the residuals is not violated.

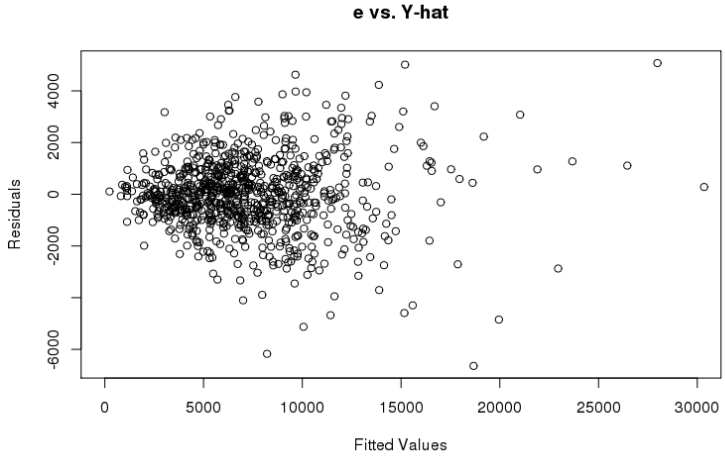

Figure 17: Residuals vs fitted values

Figure 18 indicates the qq-plot for the residuals. Figure depicts that the residuals seem to be on the line except at the tails which may be due to huge outliers in the data.

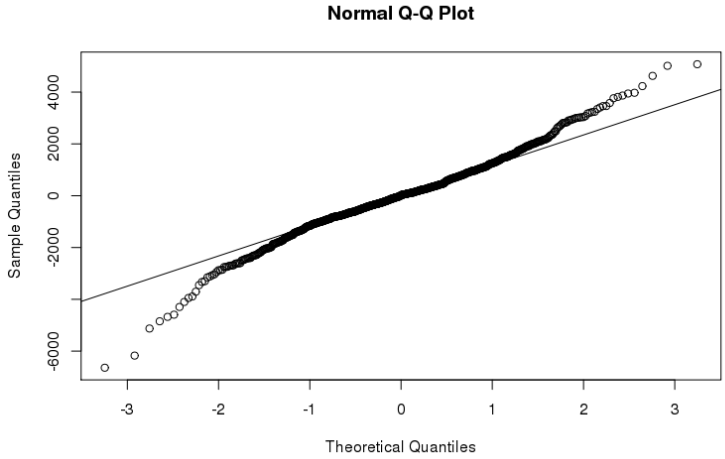

Figure 18: QQ-Plot

4.2 Final Model Inferences

From the individual predictor plots with the response variable, we can understand which predictors are a non-linear function of the response. Also, we can gauge the effect of each individual predictor.

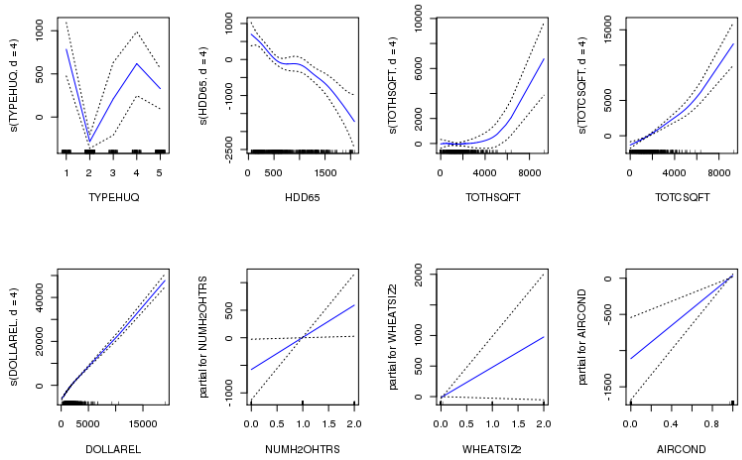

Figure 19: Individual Predictors Plot for Final Model

From Figure 19, it can be seen that predictors 'TYPEHUQ', 'HDD65', 'TOTCSQFT', 'TOTHSQFT' and 'DOLLAREL' have a non-linear relationship with the response variable. From the plots, we can observe that the electricity consumption for space conditioning is most for Mobile Homes and apartment buildings with 2-4 units than other type of housing units. Also, the electricity usage decreases as the number of heating degree days increase. This means on a hotter day, less electricity is consumed. This makes sense as heater and water heaters are not used during hot days although ACs may have been used. But no comment on the usage of AC can be made as the temperature should be much higher than 65 degrees (temperature considered for HDD). Also, the model seems to slightly over-fit the data here.

Electricity usage for space conditioning increases as 'TOTHSQFT' and 'TOTCSQFT' increases. This is logical as heated or cooled rooms with higher square footage would obviously consume more electricity. Higher total cost for electricity would mean higher electricity consumption. Electricity usage in Florida increases as 'NUMH2OHTRS' and 'WHEATSIZ2' increase. This means as the number of water heaters and size of the secondary water heaters increase the electricity usage for water heating goes up. Finally, as the number of Air Conditioners increase, the electricity usage for cooling increases.

Here, we get the predictor - number of storage water heaters as significant. This might have been to do with the Mobile homes being a significant factor for high electricity usage. Mobile Homes would have storage water heater tanks instead of main water tank. We can see that 'HDD65', 'TOTCSQFT', 'TOTHSQFT', 'NUMH2OHTRS', 'WHEATSIZ2' and 'DOLLAREL' are pretty significant estimators of electricity usage for heating, cooling and water heating.

The reason we get very bad rmseOS values for SVM is because the data is not scaled. The gam model is pretty good in capturing the non-linearities in the data. But a limitation of the GAM model is the propensity to over-fit the data, which can lead to high bias. But the issue of over-fitting is not really encountered in this case. Another limitation of gam model is that it loses predictability when the variables are outside the range of training set. But, here the data is within the range of training data which means the precision is not lost.