CUNY MSDS Capstone Project

COMMERCIAL BUILDING ENERGY CONSUMPTION

ANALYSIS AND PREDICTION

March 15, 2019

John Grando
john.grando@spsmail.cuny.edu

Contents

Abstract

Commercial Building Energy Consumption accounts for approximately 25%[1] of the United States energy production profile. Many economical and sociological factors are pushing owners of these buildings to reduce energy consumption and optimize performance. However, it is difficult to say whether a building is operating efficiently or not. Using publicly available data, models can be constructed to predict major fuel consumption. Keywords: building energy consumption, predicted energy consumption, baseline energy model.

Introduction

Building owners, local governments, and utility providers are all looking for ways to reduce energy consumption. Reasons for doing so can vary all the way from social responsibility to economic gain. Some people want to show off an efficient building, others want to identify properties that are in need of improvement. However, this concept of evaluating building performance typically requires two things; measuring the property in question's consumption, and comparing it to a standard practice equivalent. Additionally, in order for the final product to be useful, it is important that the final set of predictors be parsimonious and be realistically available to users.

Comparing summary statistics between buildings, such as energy use per square foot, is not as simple as it seems because there are a multitude of factors that affect a building's energy consumption profile. The building use type can cause energy use to vary by a large amount; such as office buildings and refrigerated warehouses. Also, some factors may have critical importance for some buildings and not others, which can lead to significant differences in consumption. This complexity of making similar comparisons creates a situation where it is difficult to determine whether a building is performing consistent with, or better than, other standard practice buildings.

Commercially, using the most popular example, ENERGY STAR [2] has implemented a benchmarking algorithm that scores buildings on a scale from 1 – 100 using market-available data. The output of this benchmarking algorithm is a unit-less score, as well as a reference 'baseline' building. However, the methodology is not released and it is unclear what factors are important to influence the energy consumption of the building. These barriers make it difficult to provide custom comparisons and nearly impossible to make batch predictions from a set of buildings, or variations of buildings.

Every few years, the U.S. Energy Information Administration (EIA) conducts a survey attempting to record pertinent features of these buildings, known officially as the Commercial Buildings Energy Consumption Survey (CBECS)[3]. While the survey is expansive (i.e. more than 600 tracked features), it is essential to create a model that is usable and only requires predictors that can easily be attained by building operators. Therefore, in this study a series of models will be evaluated in order to determine the most important predictors which will then be used to train a final, more complex, model. After completion of the model, the predictors will be evaluated on how easy they are to attain, and will possibly be exchanged with simpler variables that are highly correlated.

[1] EIA - `https://www.eia.gov/energyexplained/index.php?page=us_energy_commercial`

[2] ENERGY STAR - `https://www.energystar.gov/`

[3] EIA Microdata

Research

Related Work

The idea of determing building energy efficiency is not a novel concept in itself. As previously mentioned, ENERGY STAR has a building benchmarking tool[4]. Additionally, the United States Green Building Council has created the Arc Platform[5] which provides benchmarking and active monitoring features. While these platforms provide building comparisons in the form of an overall score, it is difficult to explore the space around the building attribute inputs themselves as well as compare consumption of a specific building to its equivalent standard practice building. With this functionality, a more direct comparison can be made and relative environmental impact can be measured.

Literature Review

There are a variety of texts that are dedicated to the analysis of building energy consumption, and determining operating efficiency, such as ASHRAE Guildine 14. Also, there are guidelines that must be followed for buildings undergoing new construction or major renovation, which have energy compliance sections (ASHRAE Guideline 90.1, 189.1, and International Energy Conservation Code). Particularly, there is a well thought out process for auditing commercial buildings, known as ASHRAE Audits, which start at the lowest level (I) and progress to the highest level (III) as the opportunity for energy and cost savings becomes more apparent [6].

Theory and Hypothesis

Commercial buildings are complex and encompass a wide variety of purposes. However, they all must be powered, and require a considerable amount of energy to operate properly. The most direct example of this complex issue is the ASHRAE energy audit process. As part of the initial audit process, an assessment of the building's overall operational efficiency is gauged. Typically, an auditor will walk through the buidling, analyze utility bills, and make the closest energy consumption comparisons they can. This takes years of experience and sometimes requires highly tuned spreadsheets that have been developed over a long period of time. It can take a surprising amount of effort just to determine if a building is operating efficiently or not.

The CBECS data set provide some insight as to what building attributes most greatly affect building energy consumption. Over 400 survey questions are recorded and coupled with major fuel consumption. These fuel sources are Electricity, Natural Gas, District Heat, and Fuel Oil. However, it would not be useful to construct a model with a large number of predictors, as it would require a large amount of time and effort to compile the necessary information in order to provide a prediction. Therefore, one of the main focuses for this study will be to extract the fewest amount of predictors necessary in order to make accurate predictions.

[4] https://www.energystar.gov/buildings/about-us/how-can-we-help-you/benchmark-energy-use/benchmarking

[5] https://arcskoru.com/

[6] http://aea.us.org/3143-2.html

Given the complex nature, it is unlikely a linear regression will provide the best prediction accuracy. This point is especially highlighted by the fact the the goal of this study is produce a parsimonious set of predictors, which means a small subset must be selected. Therefore, an investigation into more complex, nonlinear, algoriithms will be performed in order to keep the number of necessary predictors as low as possible while still capturing complex interactions.

Data and Methods

General Process

Due to the large number of features in the survey responses, it is not possible to analyze each one individually. Therefore, the first steps in the process will be centered around selecting a smaller subset. A few algorithms will be used in order to try and reduce bias. First, a principle component analysis will be performed. Second, a partial least squares model will be fit to the response. Third, a random forest regression will be used in order to try and extract any nonlinear relationships. Fourth, an attempt to construct a lasso regression model will be made. Fifth, a forward selection linear model will also be fit in order to see if an automated approach can be taken. Finally, a simple neural network model will be trained to gauge the possible effectiveness of using this model type. The magnitude and contribution percentage of each variable will be considered in selecting features from this model. Also, the various error rates from each preliminary model will be used as a benchmark for the final model performance.

After the prelminary set of models have been run and summarized, the extracted variables will be analyzed in order to verify their importance, gauge their potential predictive power, and to check whether they are easily attainable for a building operator/owner. This step is very important because it is essential worthwhile variables are used to predict the outcome. Selecting a variable that, for one reason or another, is erroneous may lead to reduced predictive power in the final model. If a variable did pass our initial analysis but doesn't actually have much predictive power (i.e. it only changes values by a slight amount) then it may not be worthwhile to select it at all. All selected variables increase the complexity of the model; therefore, we wish to only select those that will matter. Finally, the predictor must be usable, and 'knowable'. Ultimately, this tool will not be usable if a very difficult and hard to understand, and/or attain, variable value is used. These three concepts will be used in the analyzation of the candidate variables from the the preliminary analysis.

Finally, a neural network model will be built to take the verified subset of features and make predictions for the selected major fuel use. A variety of hyperparameters will be tested, using cross-validation, and compared on a common error metric. This step will reveal the optimal hyperparameter combination to use for the model. The prospective model will then be retrained on the entired entire training and validation data. This model's selected error metrics will then be compared to the preliminary models, which should be considered a floor for performance. Once this is done, the model can then be re-assessed for feature selection as well as analyzed for the value/tradeoff of adding/removing certain features.

Data Pre-Processing

The raw data set consists of 6,720 samples and 1,119 features. However, multiple steps of pre-processing were required in order to prepare the data. Note, there are many columns which are being used as imputation flags and statistical weights which, when removed, reduced the number of features down to more than 400. While these columns are useful to indicate where values have been imputed into the dataset by the source's own methodologies, rather than try to change back the data to the original records it has been determined that the imputed values were sufficiently applied and the dataset will not be blindly imputed any further.

After evaluating the reduced data set, some feature engineering efforts were taken. First, very specific cases which resulted in many null responses to follow-up survey questions (e.g. buildings less than 1,000 gross square feet), buildings open for less than a year, and features with a large amount of nulls were removed. Second, some null entries were converted to zero when logically appropriate. For example, if a building was indicated to not be cooled, then a follow up question asking what percentage of the building is cooled was not asked, resulting in an null. In this instance, the null value was replaced with a zero. Third, some values were removed as they simply did not apply to the study (e.g. expenditure for energy sources in USD). Fourth, nominal categorical values that had null responses were encoded to a special value. The thinking for this approach is that if, in fact, an null value for a feature ends up being a significant predictor, then it can be analzyed what factors make this situation occur. Fifth, the categorical features were then one-hot encoded to separate columns. The preprocessed data set was transformed to 6661 rows and 456 features (before one-hot encoding).

Electricity

General

The preprocessed data was passed to the following process in order to determine the best possible set of candidate predictors with one additional filter. Only buildings that indicated electricity being used `ELUSED` were included in the samples for this major fuel use. Then, one of each pair of predictors with correlations above 0.75 were removed, to avoid model selection issues. Additionally, the other major fuel consumption values were removed from the set of possible predictors since separate models will be made to predict these values as well. Also, the numeric predictors were transformed via BoxCox methodology as well as centered and scaled due to the varying scales and skewness.

Two potential outlier was found in the analysis. A public assembly space reported an energy consumption of 1E09 BTUs whereas the next highest consumption for this budiding type, with similar area was 3E08 BTU (less than one third of the value), and the 3rd quartile value of this subset is 5.5E06. While it is noted that there were significantly higher indications of refrigeration use than other comparables, the inclusion of this data point still vastly skews most models due to its high leverage. Similarly, an 'Other' space type has a reported energy consumption of 7E08 BTU whereas the next highest value is 2E08 BTU and the third quartile value is 2E06. While this building is large (1.4 mmSF), and has a lot of server equipment (¿500), it is still greatly beyond the next closest category and seems to be causing instability in the models due to lack of similar data

points. Therefore, these points have been removed and the caveat of instabilitity past a maximum limit will be instituted (¿5E08 BTU), due to lack of additional information.

Response Analysis

The response data appear to be unimodal and have a heavy right skew. After filtering for this model's end-use, there are 6499 samples in the data set. The energy use was convert to units mmBTU (1e6 BTU) and the log was taken in an attempt to maintain homoscedacity as the variance of the energy used also scales with the magnitude. *Appendix*

Variable Selection - PCA

RMSE: 0.0, Rsquared: NA
Top 5: `LAUNDR[NA]`, `LCOOK[NO]`, `EDSEAT`, `PBA.14[EDUCATION]`, `PRINTRN`

The principle component analysis indicates that only 4.6% of the variance in the data can be explained in the first principle component, which then drops to 2% for the second principle component. These results reveal that there does not appear to be a clear set of axes that can explain the variance of the data very well, which indicates there may be some very complex interactions taking place in the predictors. *Appendix*

Variable Selection - PLS

RMSE: 11784, Rsquared: 0.776
Top 5: `PRNTRN`, `COPEIERN`, `RFGVNN`, `RFGICN`, `RFGWIN`

This model returned a promising result; however, it must be noted that all predictors were used in this process. Looking at the output, it is obvious that the use of refrigeration equipment is dominating the variable importance plot (RFG prefix). *Appendix*

Variable Selection - Random Forest

RMSE: 12650, Rsquared: 0.883
Top 5: `PUBCLIM.7[WITHELD]`, `SQFTC.9[500K SF - 1mm SF]`, `SQFTC.8[200K SF - 500K SF]`, `SQFTC.7[100K SF - 200K SF]`, `SQFTC.6[50K SF - 100K SF]`

As with the PLS, model, the resulting error metrics were promising, with slightly better RMSE and Rsquared values. The selected variables are ver similar with a few exceptions. This model has placed higher importance on a yes/no response to the presence of walk in refrigerators as well as whether or not a building is a fast food establishment. *Appendix*

Variable Selection - Lasso

RMSE: NA, Rsquared: NA
Top 5: NA

This resulted in a very poor fit, which is not unexpected. Lasso models typically work when

a few variables can be used to predict the response, which does not appear to be the case in this instance. Due to the lack of fit, this model will not be used in the variable selection process. Additionally, the actual model was poor enough that predictions could not be made on the data, which is the reason for the lack of reported metrics. *Appendix*

Variable Selection - Forward Selection

RMSE: 26029, Rsquared: 0.267
Top 5: `PRNTRN`, `RFGVNN`, `COPIERN`, `RFGVEN[YES]`, `RFGICN`

This model was building using the leaps package which iteratively selected the best predictor variable up to a limit of 100. Unsurprisingly, the best model turned out to be the maximum setting; however, encouraging model metrics were still not returned. *Appendix*

Variable Selection - Recursive Feature Elimination

RMSE: 10548, Rsquared: 0.824
Top 5: `RFGICN`, `MRI[YES]`, `PRNTRN`, `HCBED_bin.6[>250]`

A more direct approach was taken with this model, which is specifically used to extract useful features from data sets. As can be seen, there are many of the same predictors chosen in this model as in previous models. *Appendix*

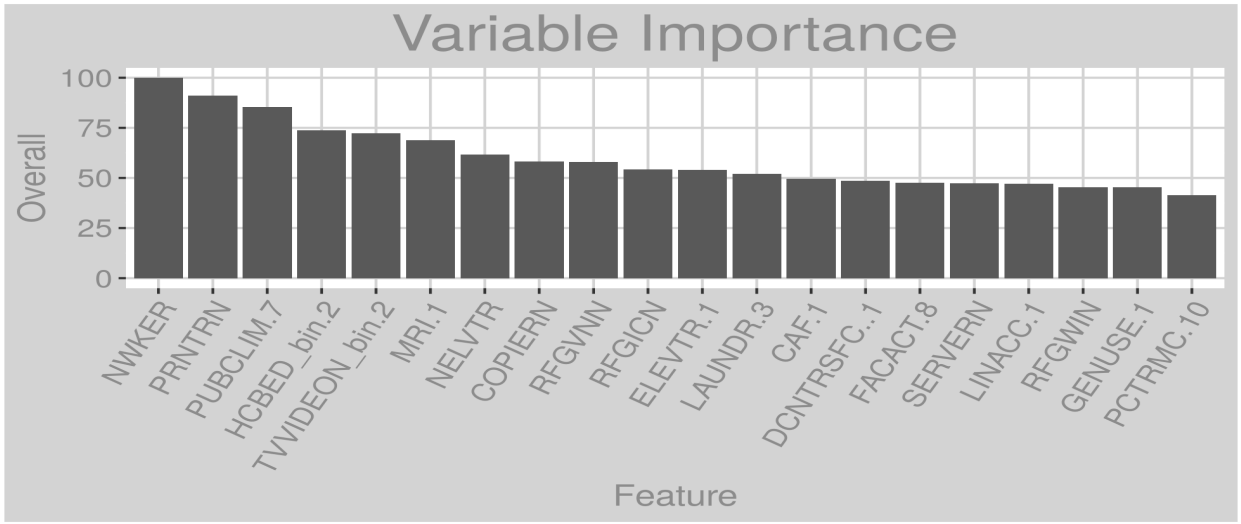
Variable Selection - Simple Neural Network

RMSE: 13080, Rsquared: 0.725
Top 5: `PUBCLIM[WITHELD]`, `HCBED_bin[>250]`, `TVVIDEON_bin[>200]`, `PRNTRN`, `MRI[YES]`

Given that the final model will be a neural network, it made sense to try a simple out-of-the-box training model to see if any particular features worked better with this specific process. However, low metrics were returned but it was decided to keep the predictor set and include it in the final selection summary. *Appendix*

Variable Selection - Selected Variable Analysis

Model	RMSE	R2	MAE
partialLeastSquares	1.169408e+04	0.7780371	3.237700e+03
leaps	2.327804e+04	0.3559985	5.652596e+03
randomForest	1.182305e+06	0.2788200	8.100512e+05
recursiveFeatureExtraction	2.221181e+07	0.8247289	8.762935e+06
neuralNetwork	4.453689e+07	0.1406134	1.061397e+07
lasso	8.242048e+13	0.0000032	2.936412e+13



In order to rank the most important features, the variable importance metrics from the selected models were all set to the same scale, the summed up. The number of selected features were then used as hyperparameters in the final model training. As a preliminary check, the top 20 predictors are plotted in the appendix and are generally discussed here. It seems the attempts to create stratified random samples, by building type, may have been beneficial in this case since there are some very building type specific end-uses that are highly ranked. It does cause some concern that there are some odd selections, such as the 'witheld' column of the climate encoded variables being selected; however, it seems as though whatever reasoning to classify these types of buidlings this way is a direct indicator of electrical consumption. When evaluating multiple selected atypical variables (e.g. LAUNDR.3[laundry provided by offsite services]) it seems that just the applicability of the question influences the consumption (building types - lodgining, healthcare, nursing). In an attempt to truly follow the imporant predictors, no variables have been removed from this set and the order of importance remains unchanged. *Appendix*

Natural Gas

General

As previously noted, only buildings that indicated natural gas being used NGUSED were included in the samples for this major fuel use. Then, one of each pair of predictors with correlations above 0.75 were removed, to avoid model selection issues. Numeric predictors were transformed via BoxCox methodology as well as centered and scaled due to the varying scales and skewness. Note, no further commentary will be made in the following sections unless it differs from previous sections.

Response Analysis

After filtering for this model's end-use, there are 4355 samples in the data set. The same transformations were applied to this response variable as electricity. *Appendix*

Variable Selection - PCA

RMSE: 0.0, Rsquared: NA
Top 5: NWKER, EDSEAT, PBA.14[EDUCATION], STRLZR..1[NA], PRINTRN

No further commentary. *Appendix*

Variable Selection - PLS

RMSE: 11784, Rsquared: 0.776
Top 5: NWKER, PRINTRN, NELVTR, RFGVNN, RFGWIN

No further commentary. *Appendix*

District Heat

General

Test

Fuel Oil

General

Test

Neural Network Models

General

The choice to use neural networks for the final model was multi-faceted. First, these types of models are very good at capturing complex non-linear interactions. This appears to be the case with the data set given the failure of lasso models as well as the low percentage of variance capture for the first few dimensions of the principal component and partial least squares analyses. Secondly, neural networks have the ability to use customized loss functions. This is beneficial because it is important to highlight practicality of the results returned. As the estimated energy consumption grows, it is somewhat acceptable for the error rate to grow proportionally if it results in the low estimates to have better error rates using a homoscedastic loss function. As an example, a large datacenter may use a lot of energy so a slightly higher relative error rate may not be a big issue since it could be a small portion of the overall consumption; however, if a non-heated warehouse with a moderate error rate, comparative to the rest of the data set, would be wildly innacurate. Therefore, the loss function chosen for this set of models was chosen to be the mean squared logarithmic error in an effort to reflect this reasoning.

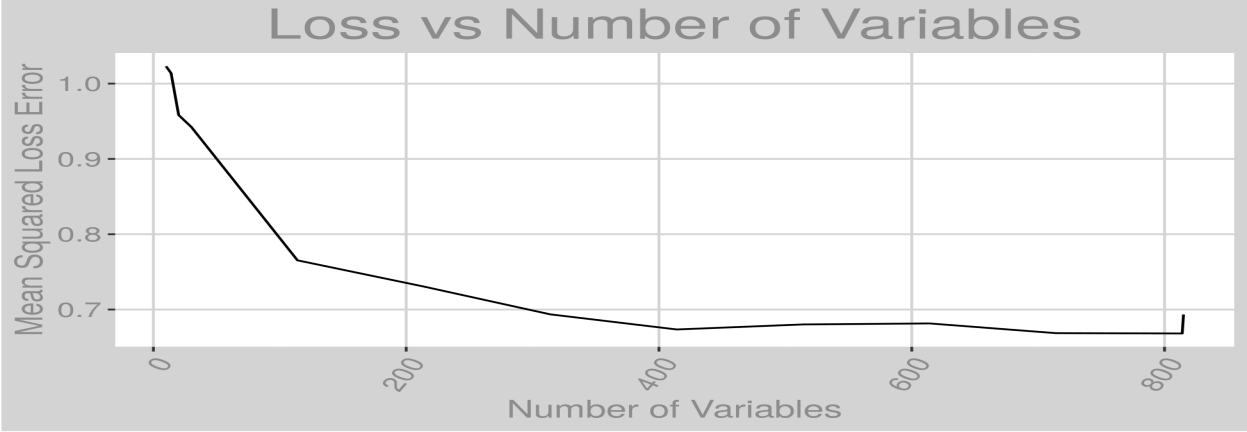
Hyperparameter Training

In order to select the most optimized set of parameters, some hyperparameter training was performed. Some standard searches were made, such as varying the dropout rate, regularization, learning rate, and batch size; however, one additional training set was incorporated to highlight the goals of this study. A series of models were tested which had an incrementally decreasing number of variables, by least importance, in order to test the loss of accuracy.

Electricity

Summary

The final selected model consisted of a 3 hidden layers of 300 nodes, using a dropout rate of 0.6, no regularization, batch sizes of 50 and XXX variables for selection. The number of variables is greatly reduced while maintaining relativively similar performance



The final selected model has a msle of XXX and mse of XXX (XX rmse). Comparing this to the previous feature extraction models, which used many more variables, the performance is ....

The resdiuals indicate that the variance scales with the response variable; however, since neural network models do not operate on a principle of homoscedacitity, only underlying patterns are of concern. Additionally, the noted error pattern is by design so that higher error in higher consumption projects are acceptable. *Appendix*

Future Work

While it was determined that some heteroscedacitity would be acceptable, there does appear to be area for improvement. Additionally, as mentioned at the beginning of this report, the sampling of this data set was stratified to reflect the building population. However, it is noted that there are some classes that have greater variance than others. Therefore, it may be useful to use this stratification as a weighted method, based on `PBAPLUS`, in order to try and emphasize accuracy on the most prevalent budiding types.

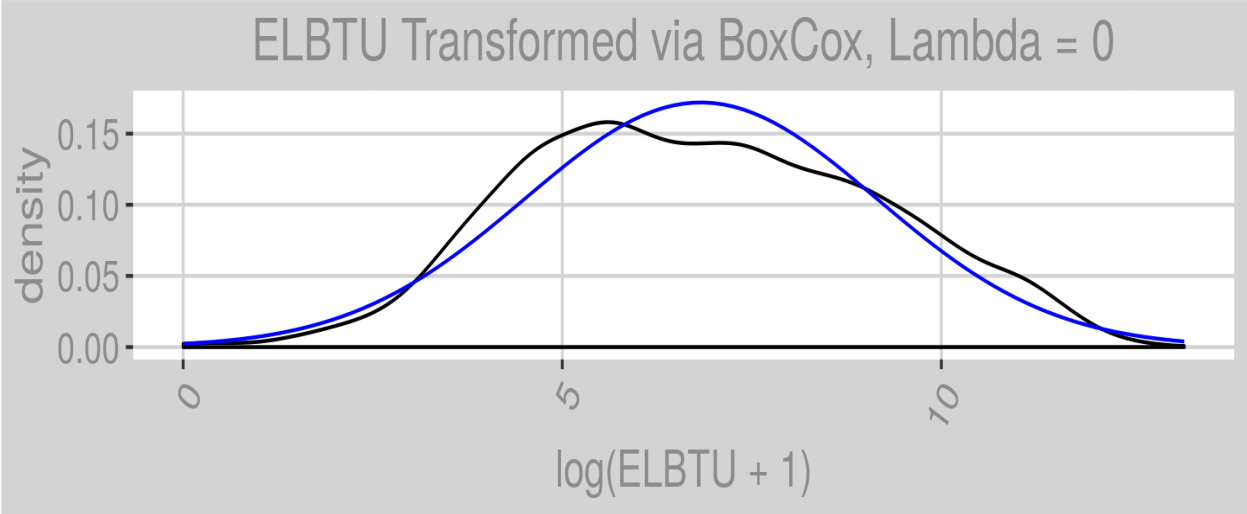
Natural Gas

District Heat

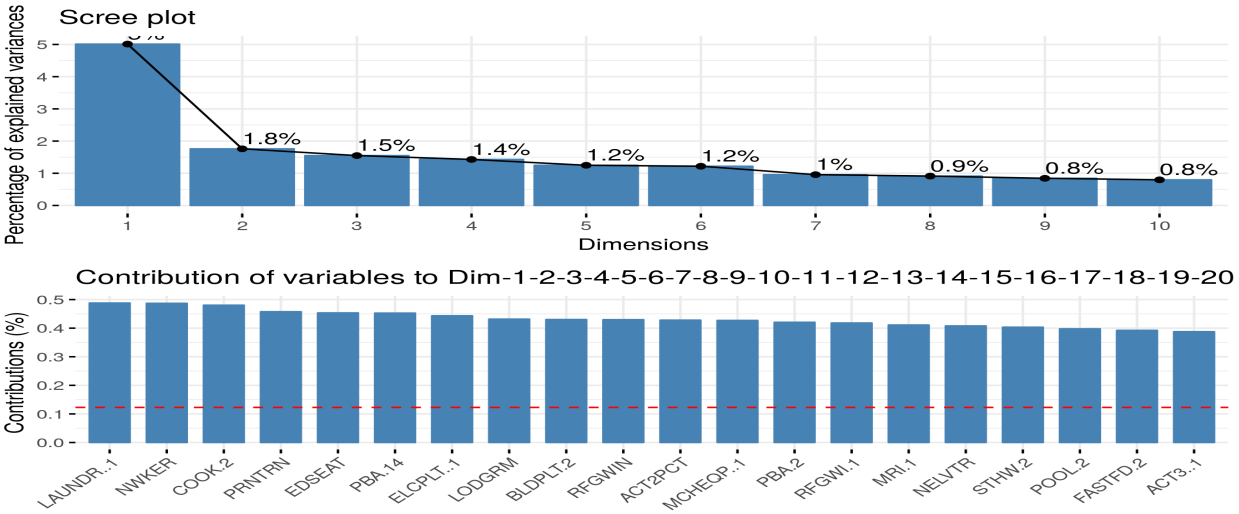
Fuel Oil

Appendix - Electricity

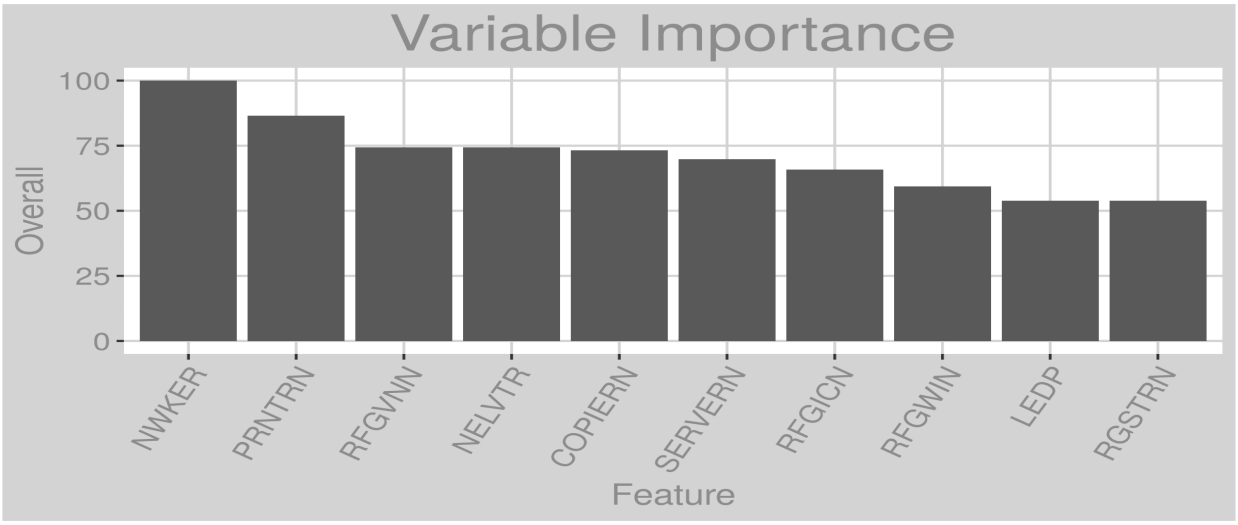
Response

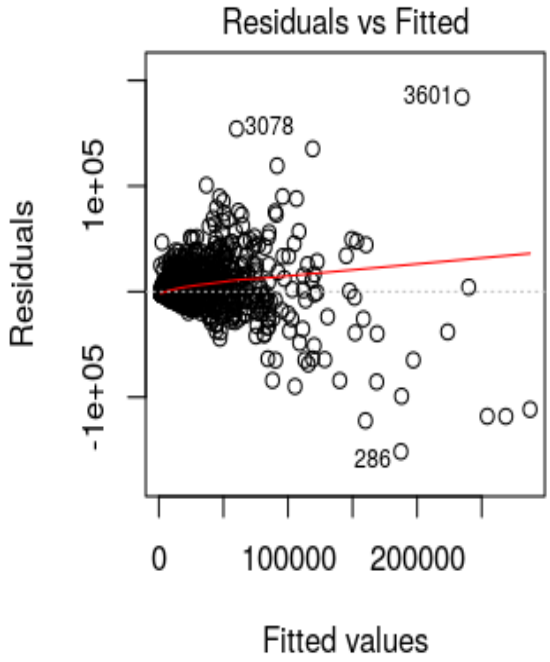


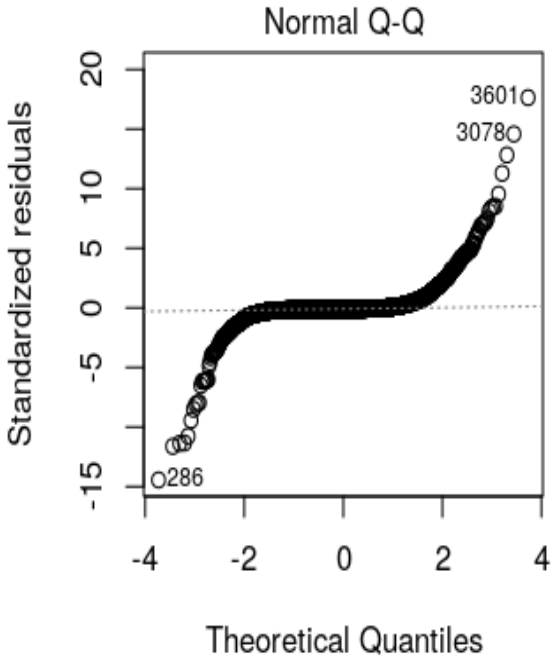
PCA

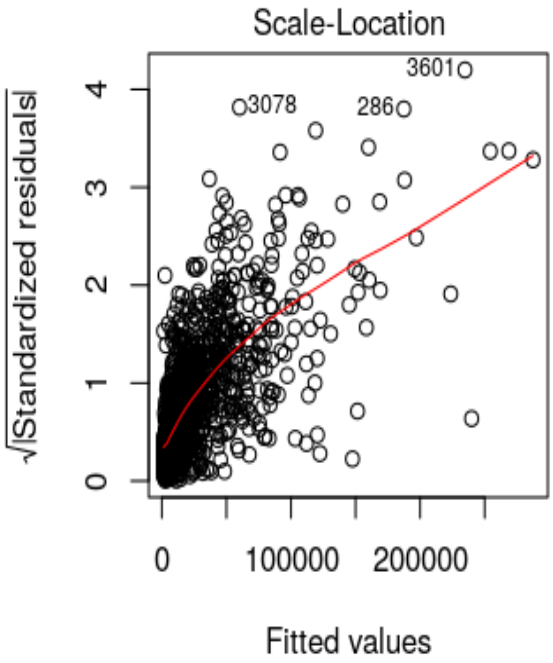


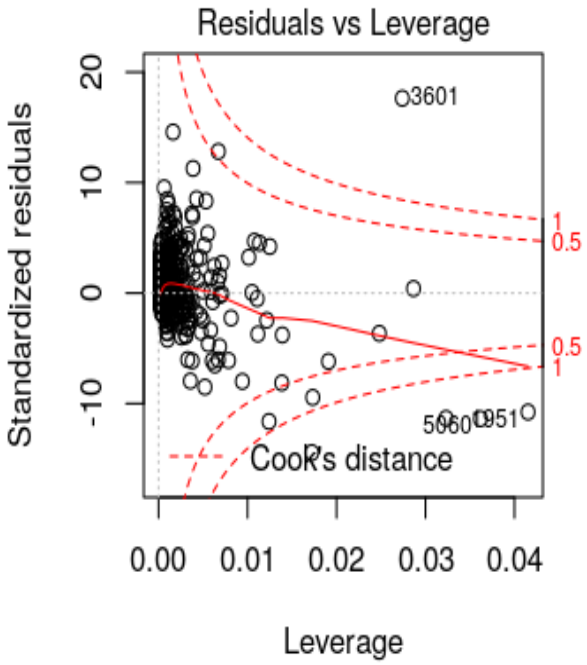
PLS

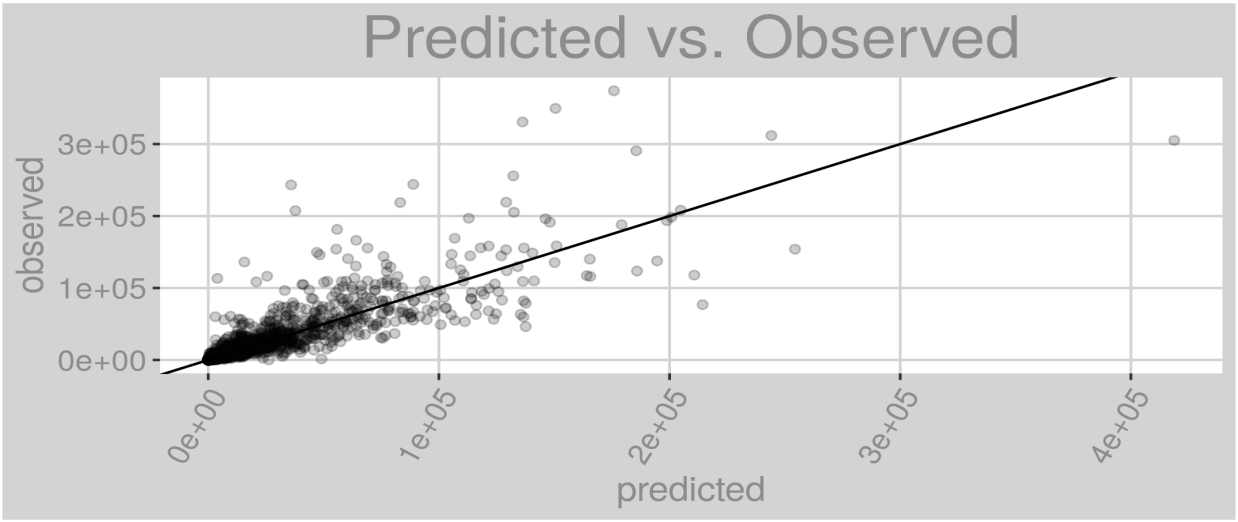

Variable Importance
Overall
100
75
50
25
0
NWKER
PRNTRN
RFGVNN
NELVTR
COPIERN
SERVERN
RFGICN
RFGWIN
LEDP
RGSTRN
Feature


Residuals vs Fitted
Residuals
1e+05
-1e+05
3601
3078
286
0
100000
200000
Fitted values

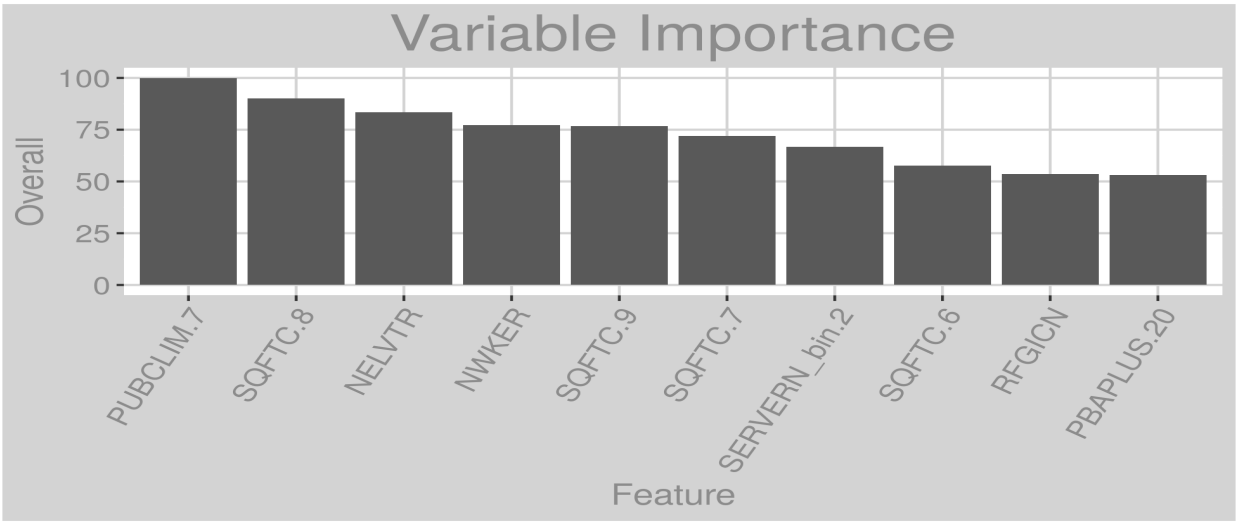

Normal Q-Q
Standardized residuals
20
10
5
0
-5
-15
3601
3078
286
-4
-2
0
2
4
Theoretical Quantiles

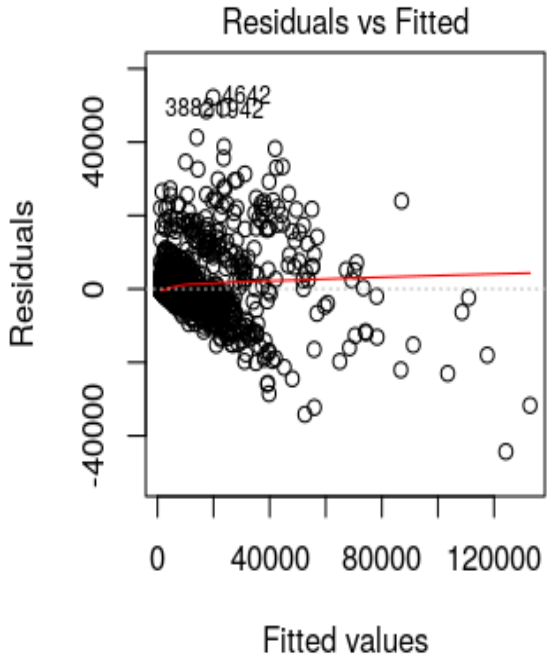

Scale-Location
√|Standardized residuals|
3601
3078
286
0
1
2
3
4
0
100000
200000
Fitted values

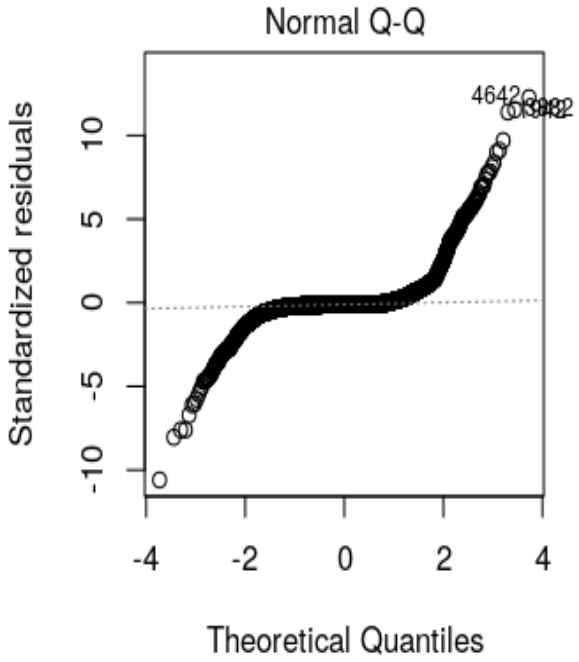

Residuals vs Leverage
Standardized residuals
3601
5060
951
Cook's distance
1
0.5
0.5
1
20
10
0
-10
0.00
0.01
0.02
0.03
0.04
Leverage

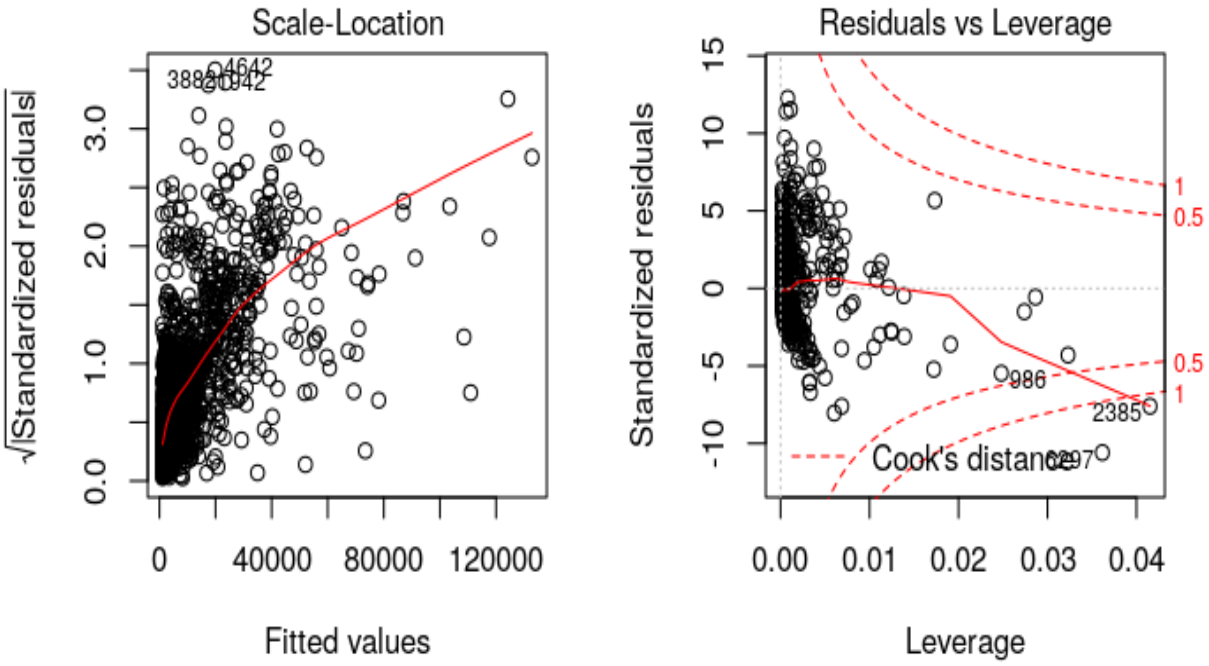

Predicted vs. Observed
observed
3e+05
2e+05
1e+05
0e+00
0e+00
1e+05
2e+05
3e+05
4e+05
predicted

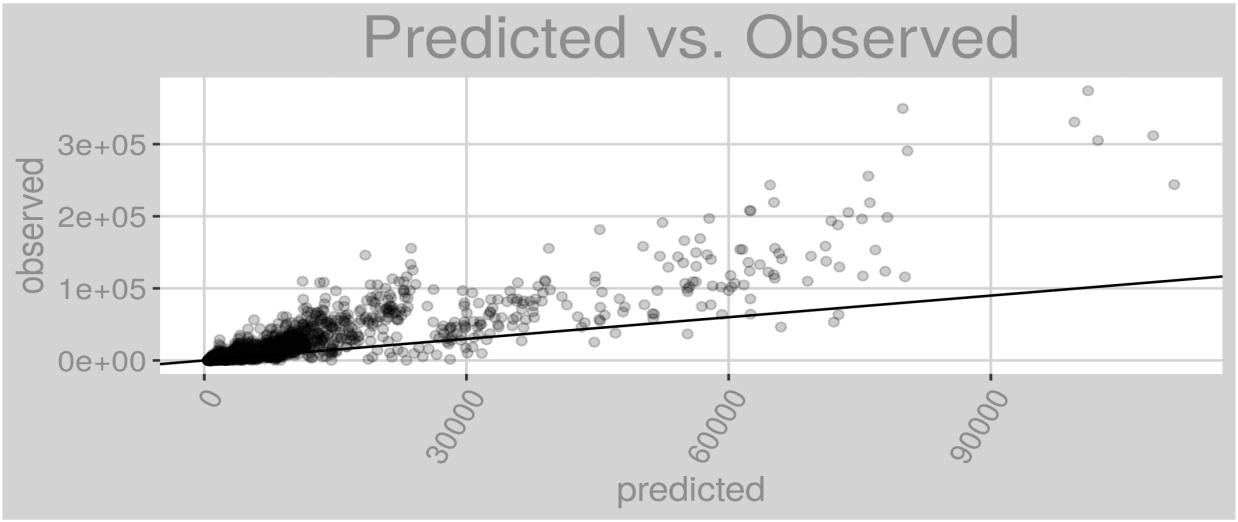
Random Forest



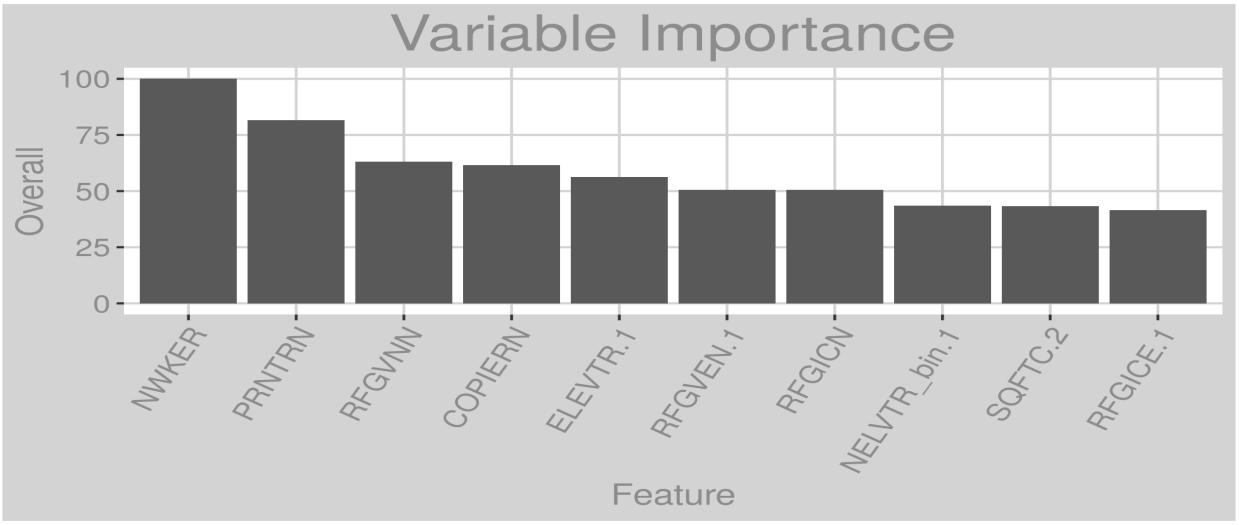


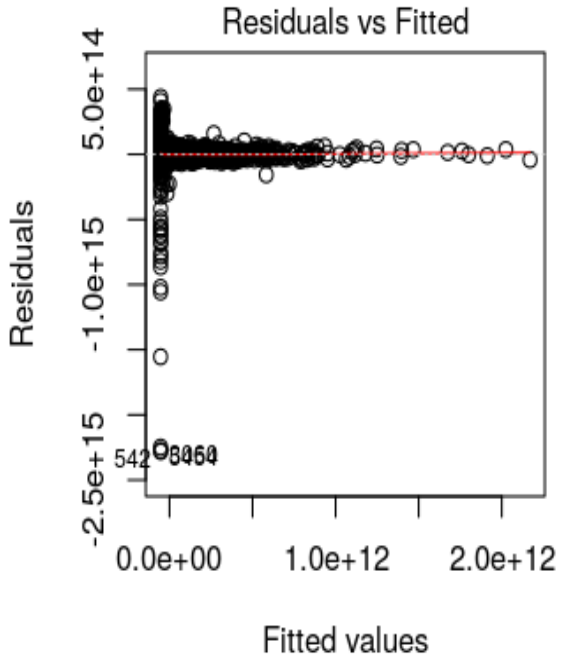


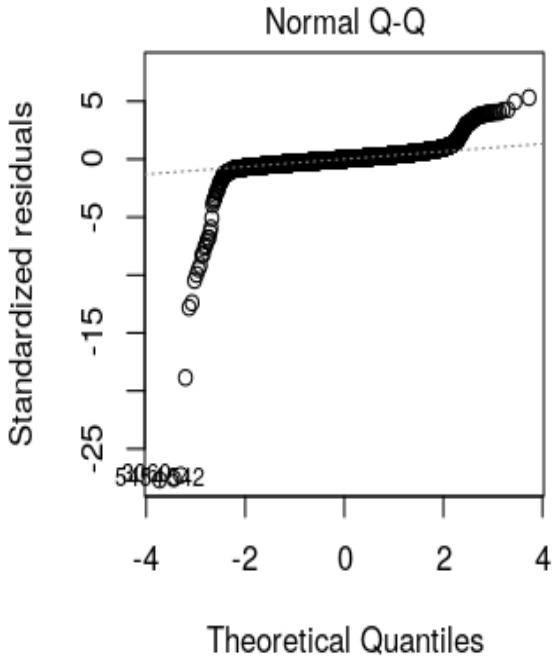

Scale-Location
√|Standardized residuals|
Fitted values
0
40000
80000
120000
0.0
1.0
2.0
3.0
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.00
0.01
0.02
0.03
0.04
-10
-5
0
5
10
15
0.5
1
986
2385

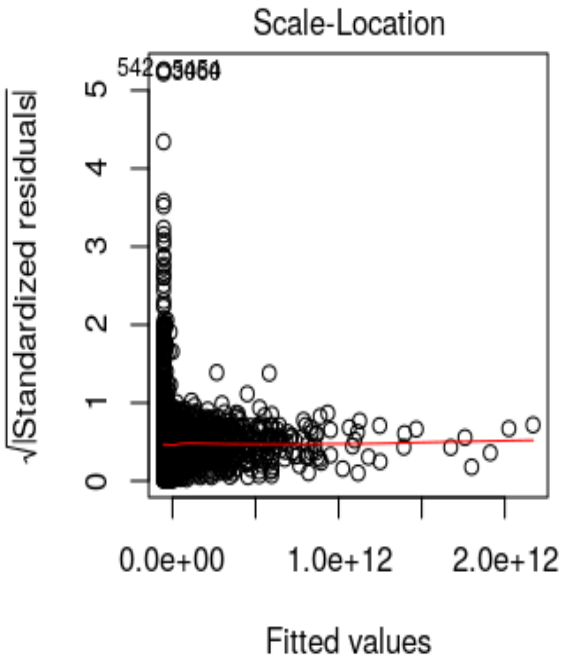

Predicted vs. Observed
observed
predicted
0e+00
1e+05
2e+05
3e+05
0
30000
60000
90000

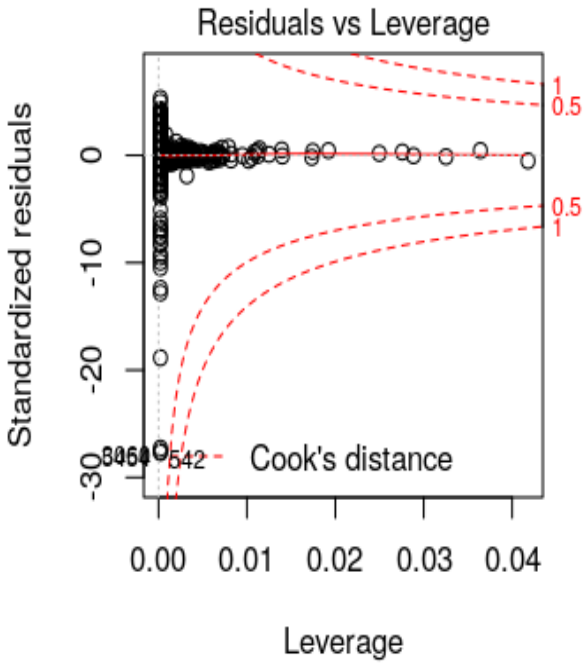
Lasso

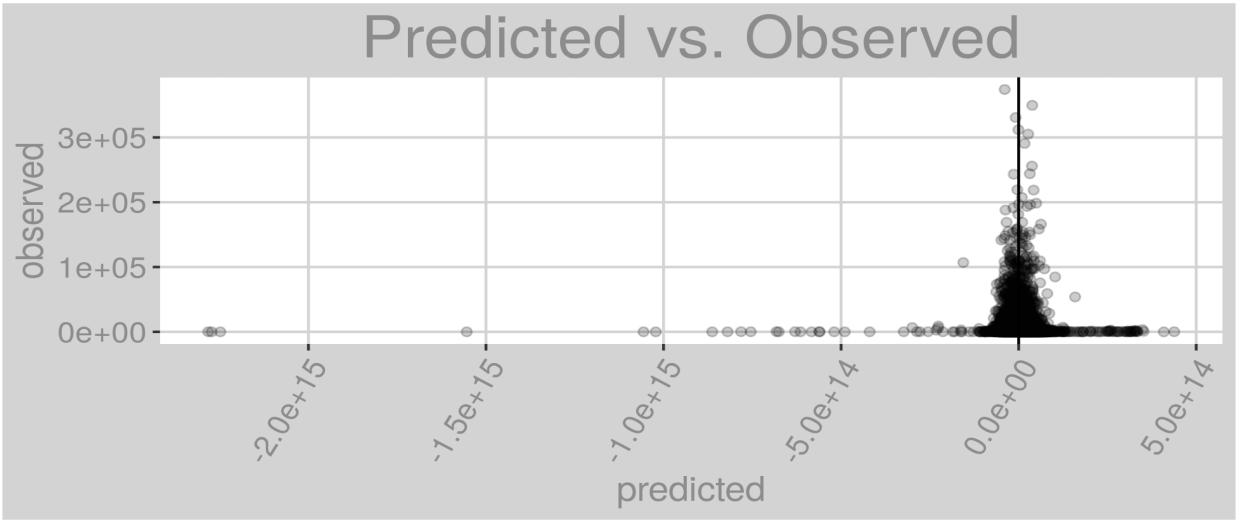

Variable Importance
Overall
100
75
50
25
0
NWKER
PRNTRN
RFGVNN
COPIERN
ELEVTR.1
RFGVEN.1
RFGICN
NELVTR_bin.1
SQFTC.2
RFGICE.1
Feature


Residuals vs Fitted
Residuals
5.0e+14
-1.0e+15
-2.5e+15
0.0e+00
1.0e+12
2.0e+12
Fitted values

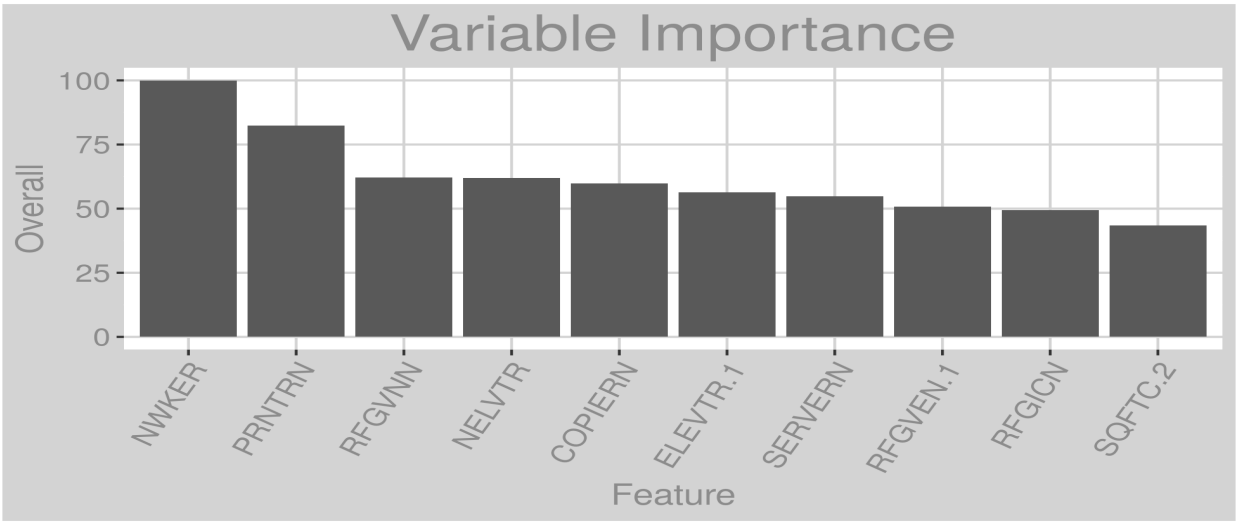

Normal Q-Q
Standardized residuals
5
0
-5
-15
-25
-4
-2
0
2
4
Theoretical Quantiles

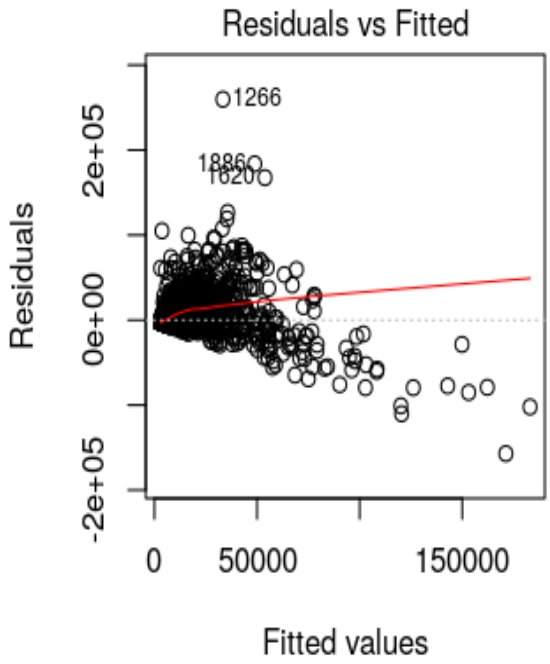

Scale-Location
√|Standardized residuals|
0
1
2
3
4
5
0.0e+00
1.0e+12
2.0e+12
Fitted values

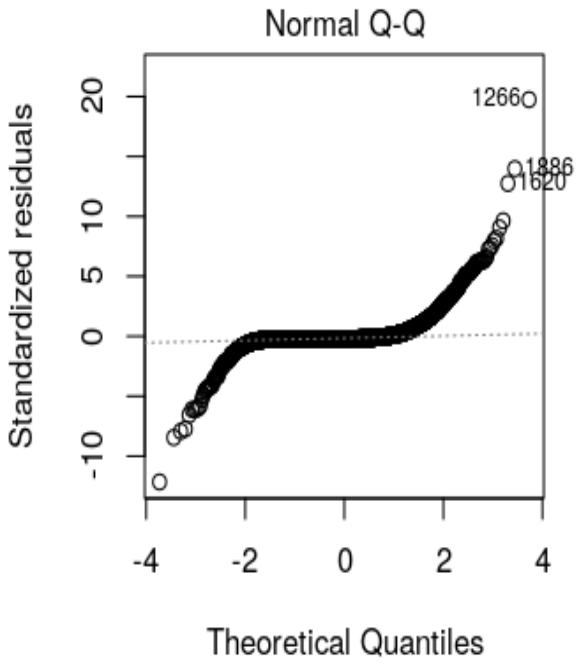

Residuals vs Leverage
Standardized residuals
0
-10
-20
-30
0.00
0.01
0.02
0.03
0.04
Leverage
Cook's distance
1
0.5
0.5
1

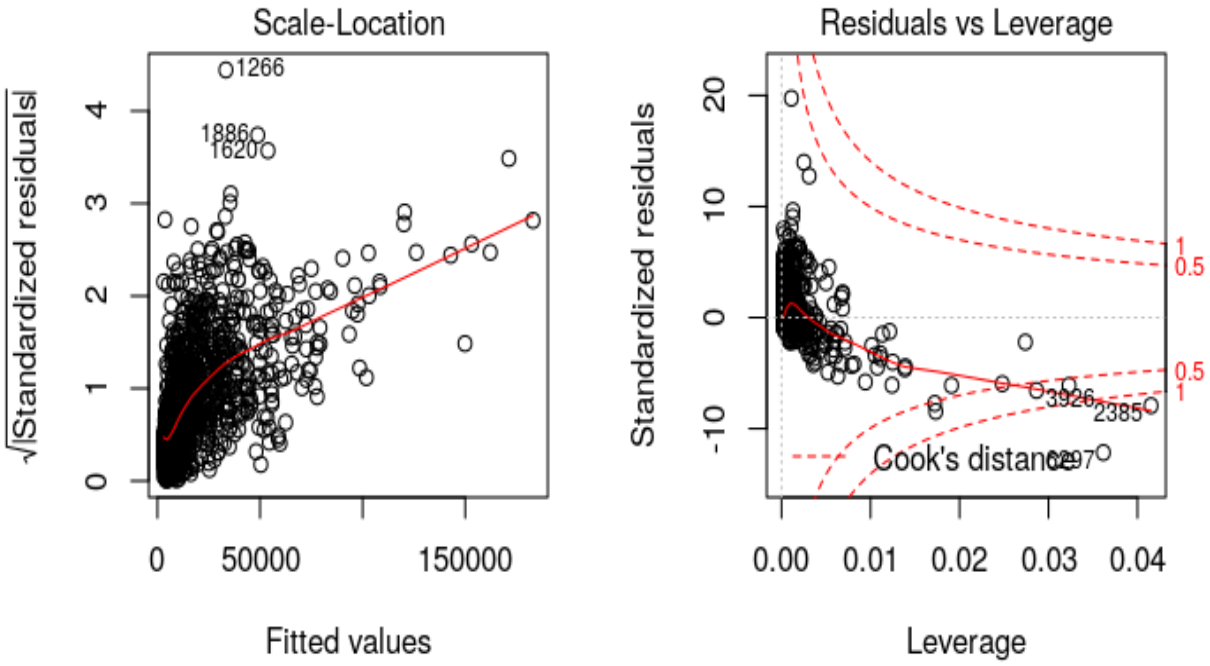

Predicted vs. Observed
observed
3e+05
2e+05
1e+05
0e+00
-2.0e+15
-1.5e+15
-1.0e+15
-5.0e+14
0.0e+00
5.0e+14
predicted

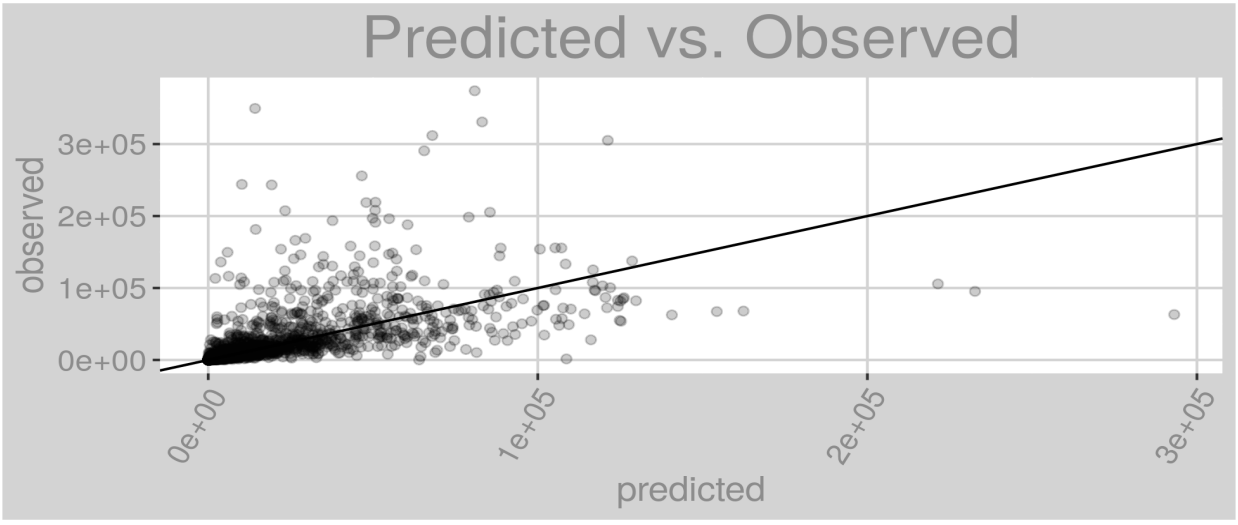
Forward Selection



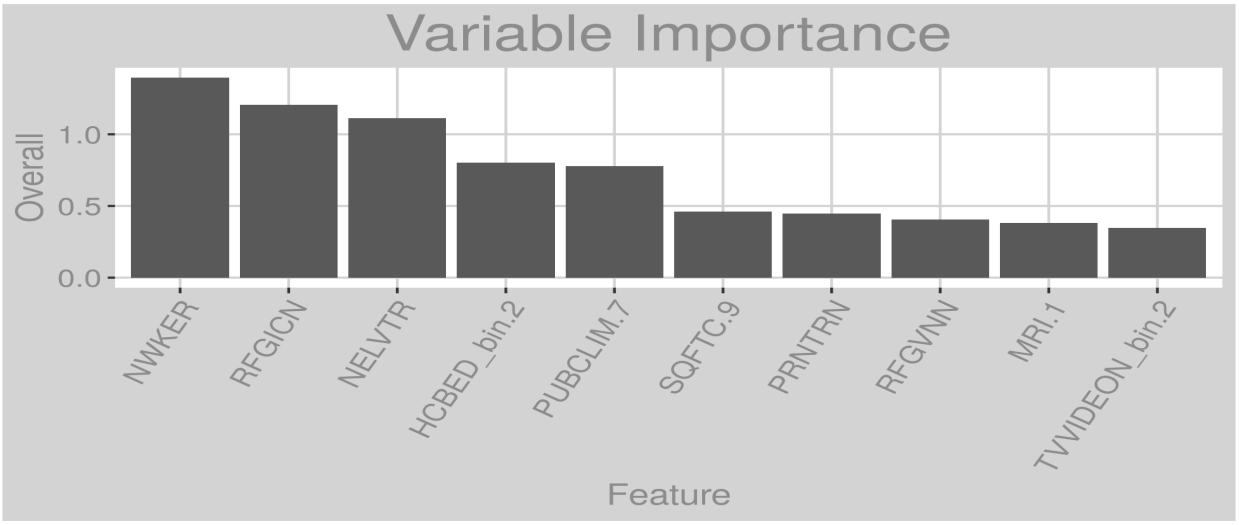


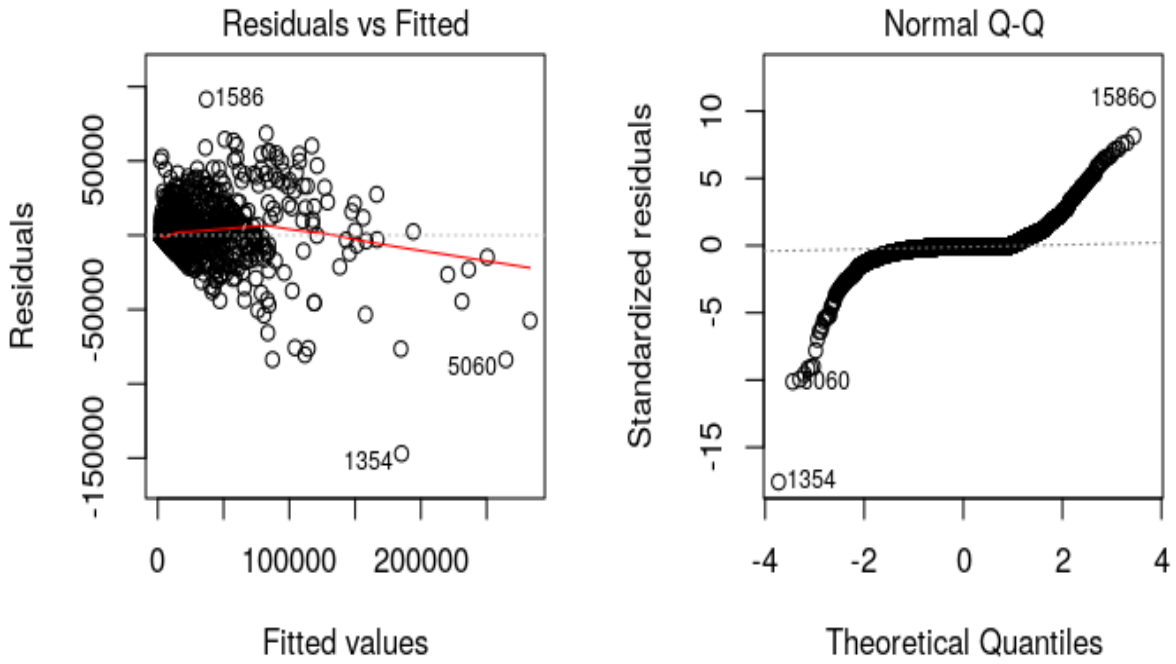


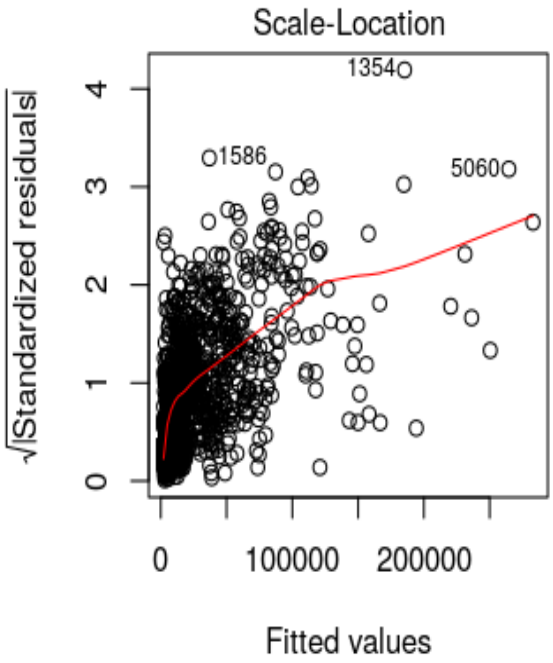

Scale-Location
√|Standardized residuals|
1266
1886
1620
0
1
2
3
4
0
50000
150000
Fitted values
Residuals vs Leverage
Standardized residuals
20
10
0
-10
1
0.5
0.5
1
3926
2385
Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage

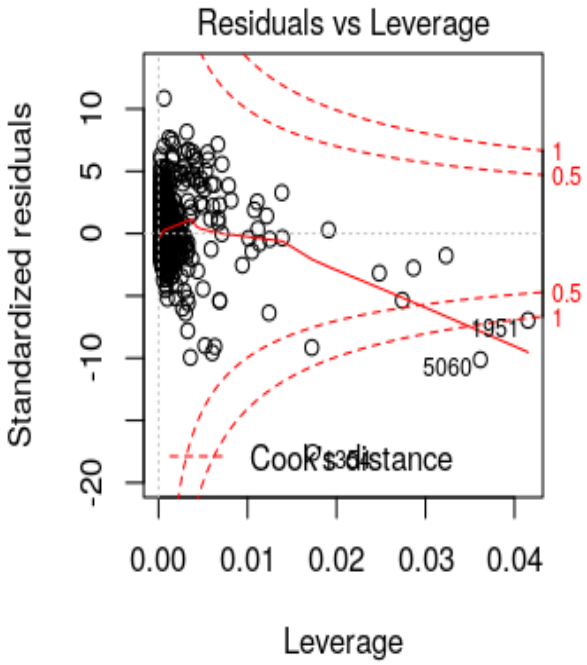

Predicted vs. Observed
observed
3e+05
2e+05
1e+05
0e+00
0e+00
1e+05
2e+05
3e+05
predicted

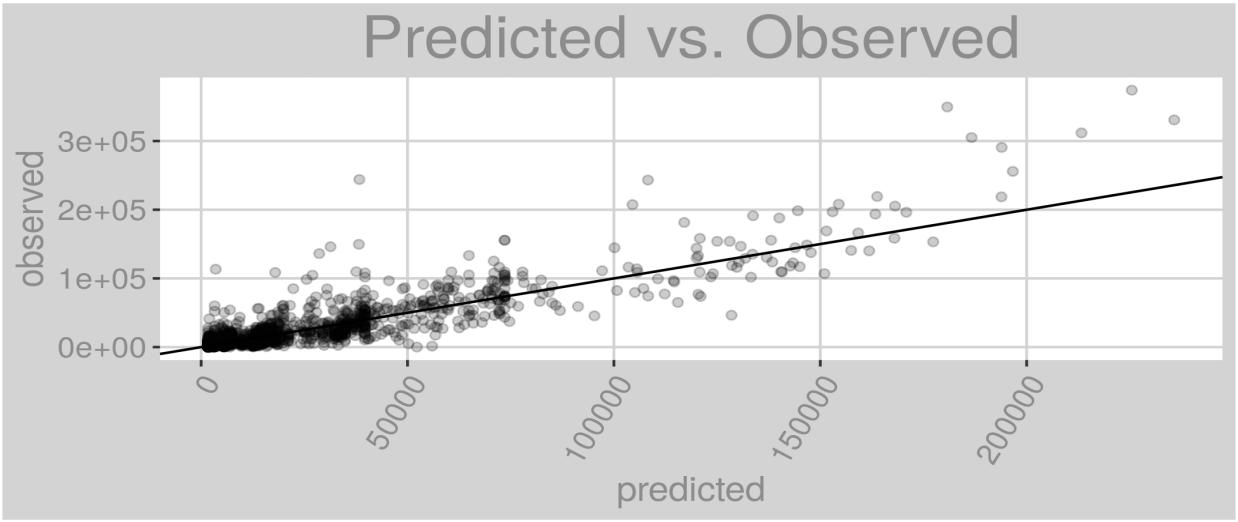
Recursive Feature Extraction



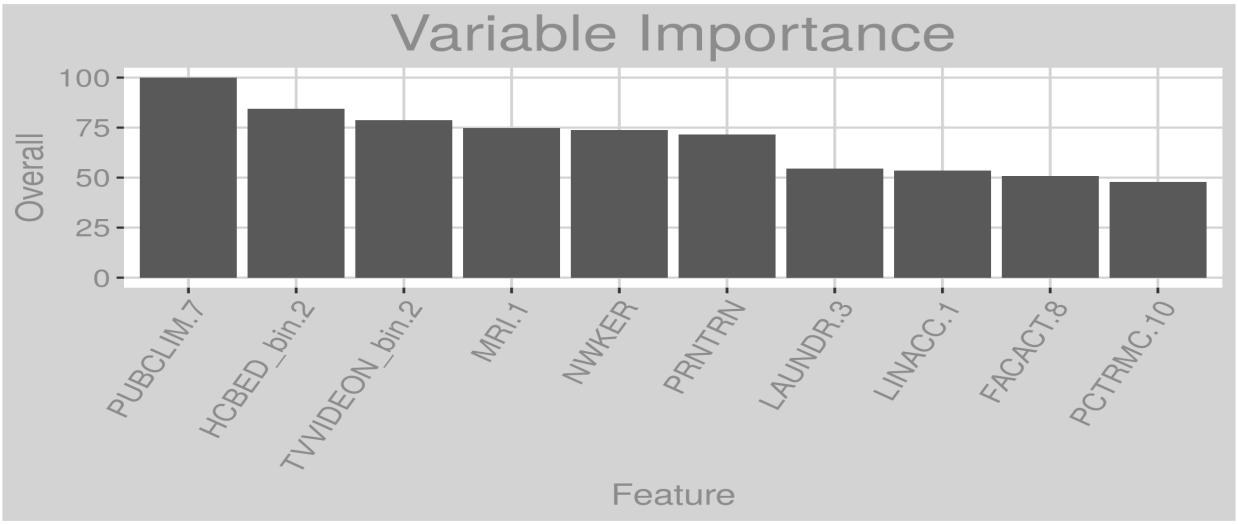


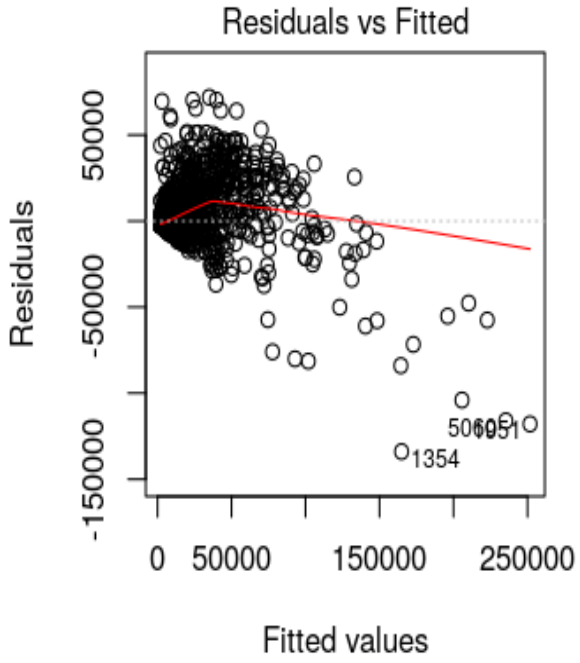

Scale-Location
√|Standardized residuals|
1354
1586
5060
0
1
2
3
4
0
100000
200000
Fitted values

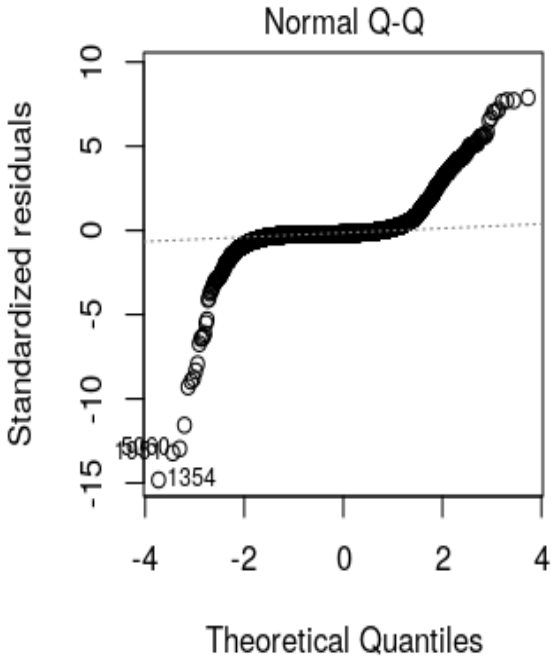

Residuals vs Leverage
Standardized residuals
1
0.5
0.5
1
1951
5060
Cook's distance
-20
-10
0
5
10
0.00
0.01
0.02
0.03
0.04
Leverage

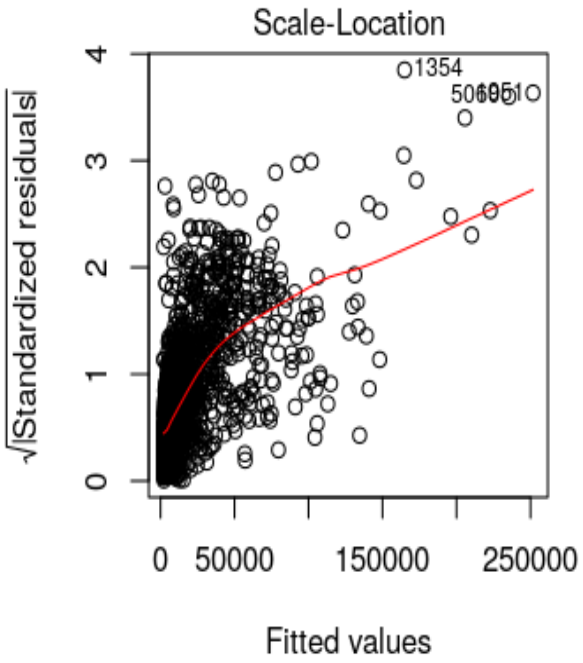

Predicted vs. Observed
observed
3e+05
2e+05
1e+05
0e+00
0
50000
100000
150000
200000
predicted

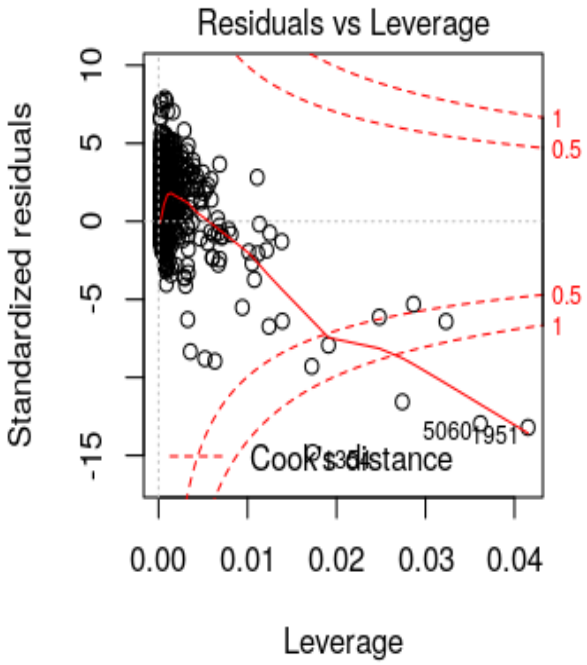
Simple Neural Network

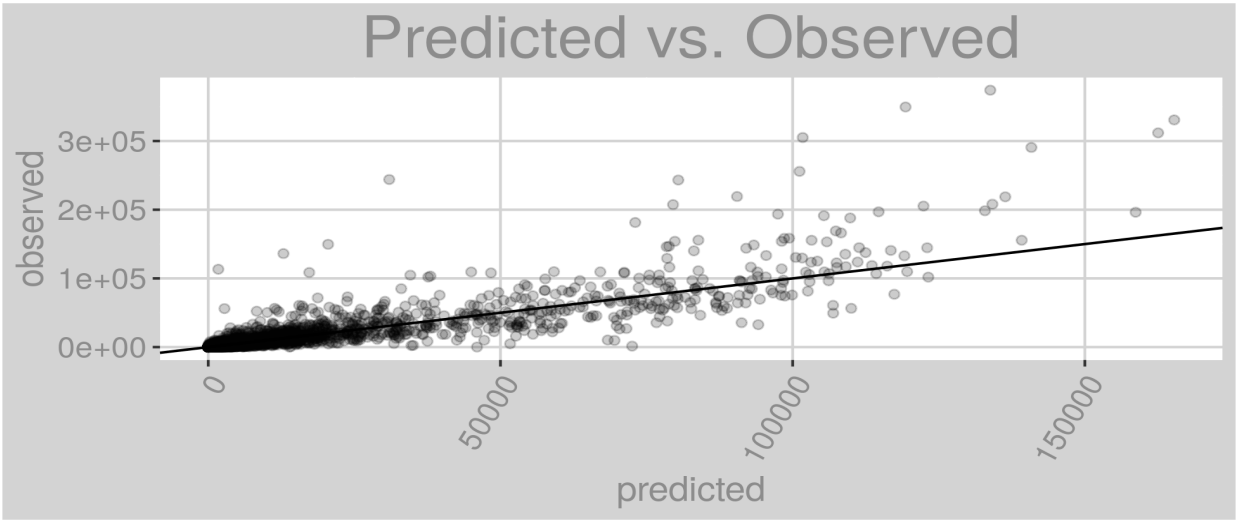






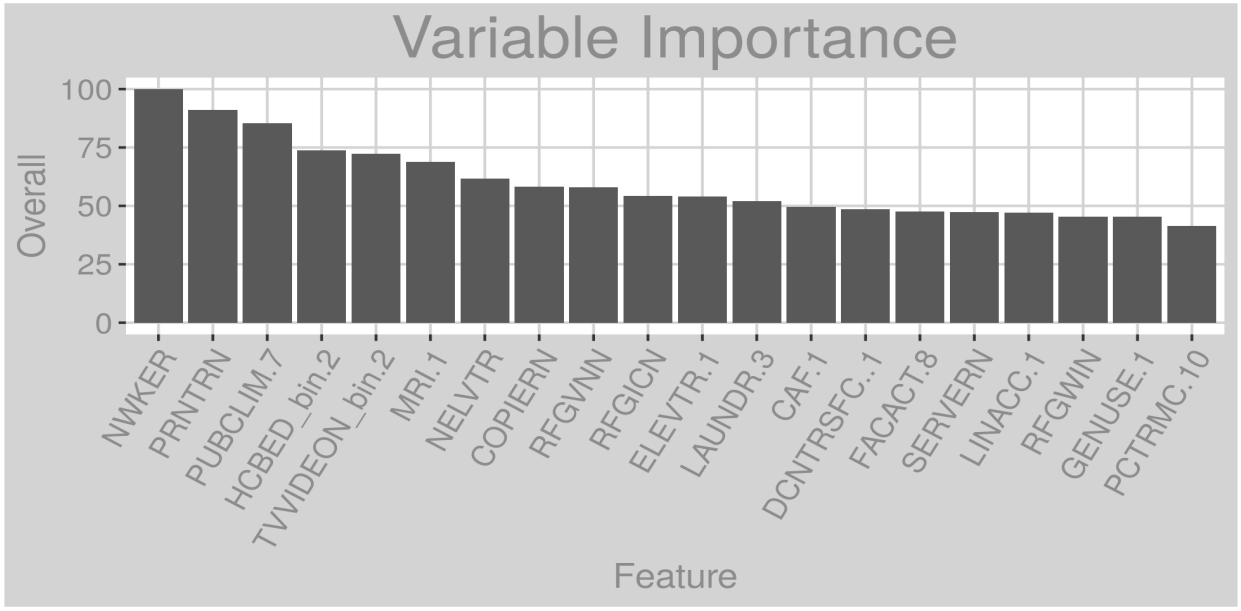

Scale-Location
√|Standardized residuals|
Fitted values
0
50000
150000
250000
1354

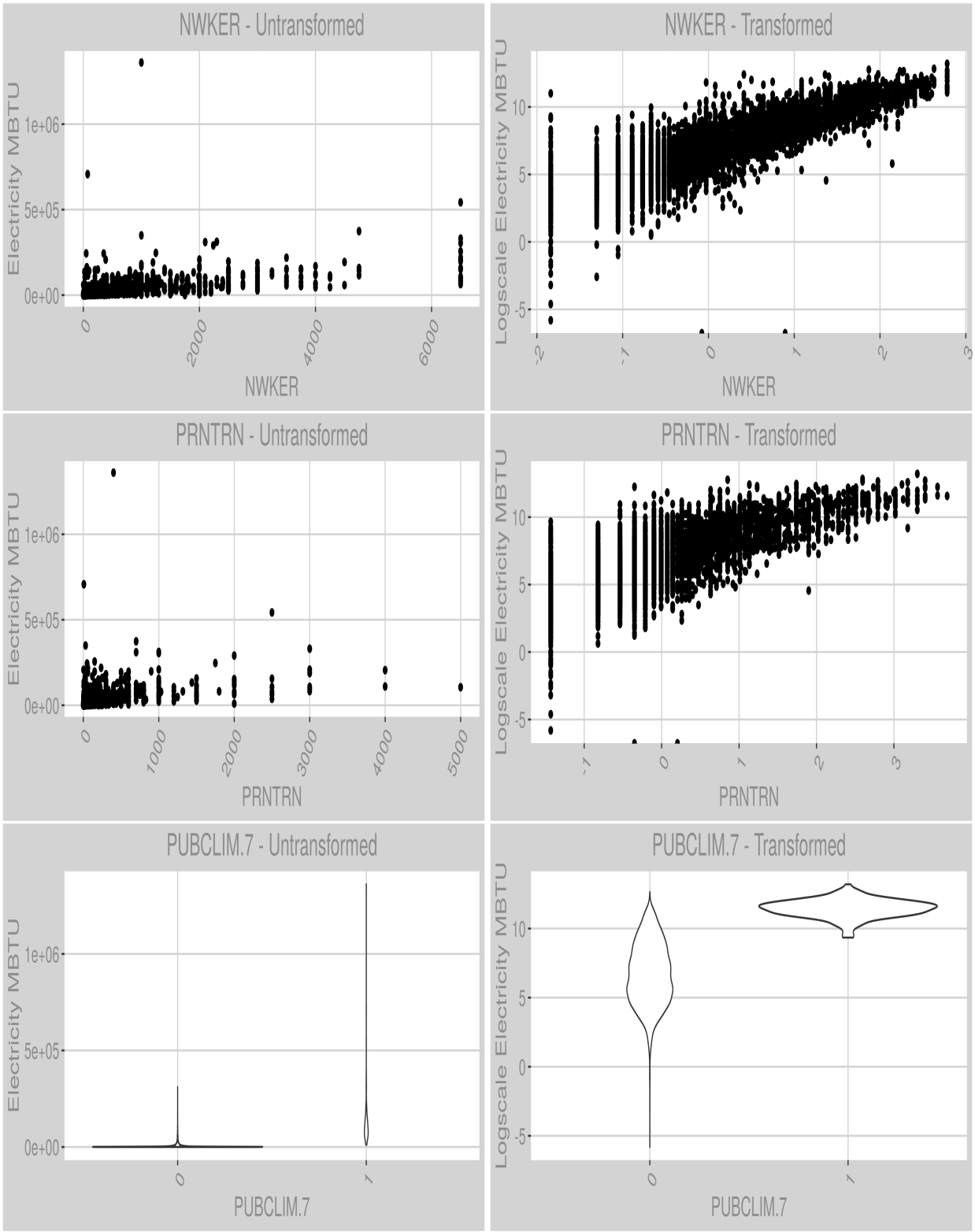

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.00
0.01
0.02
0.03
0.04
50601951

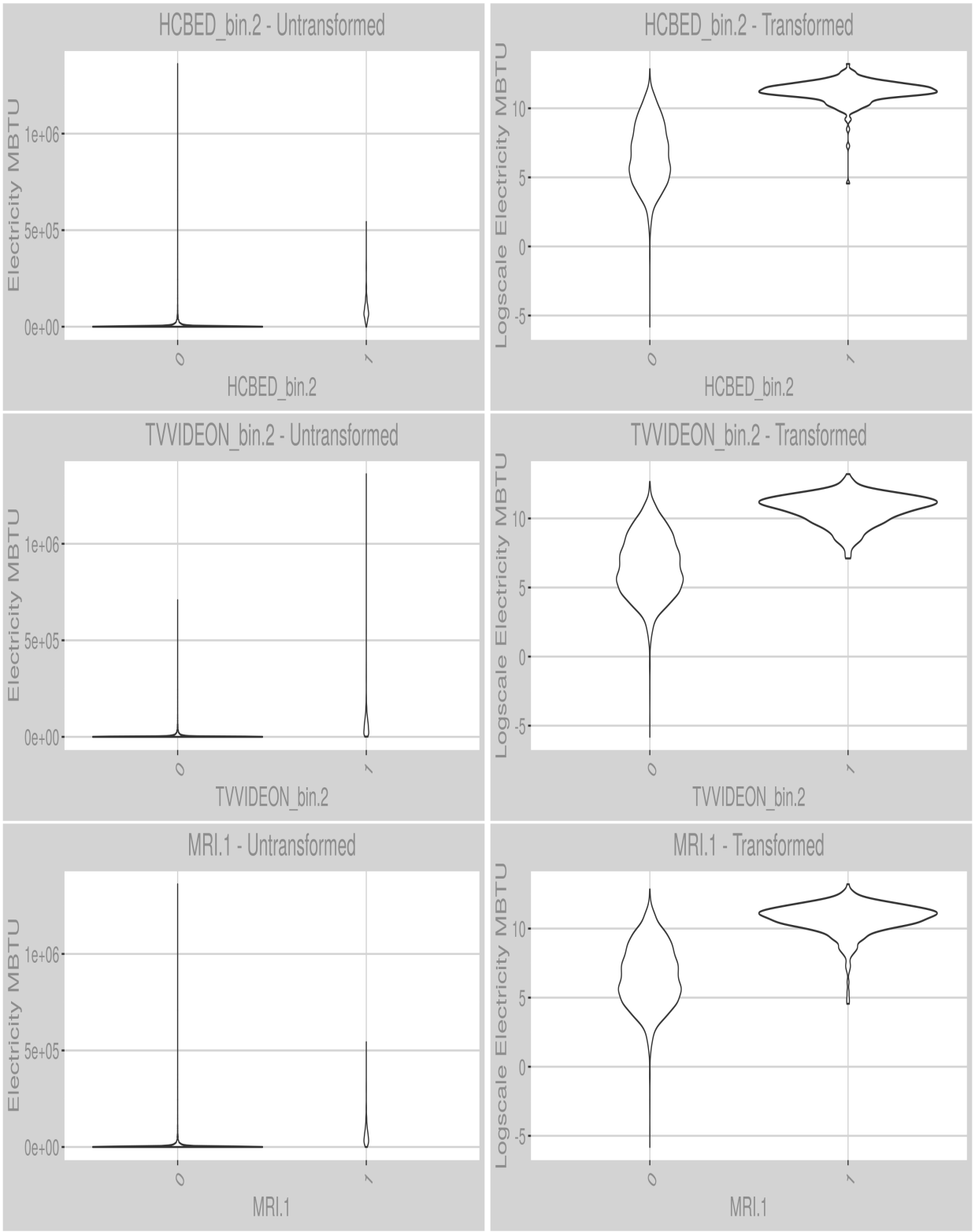

Predicted vs. Observed
observed
predicted
3e+05
2e+05
1e+05
0e+00
0
50000
100000
150000

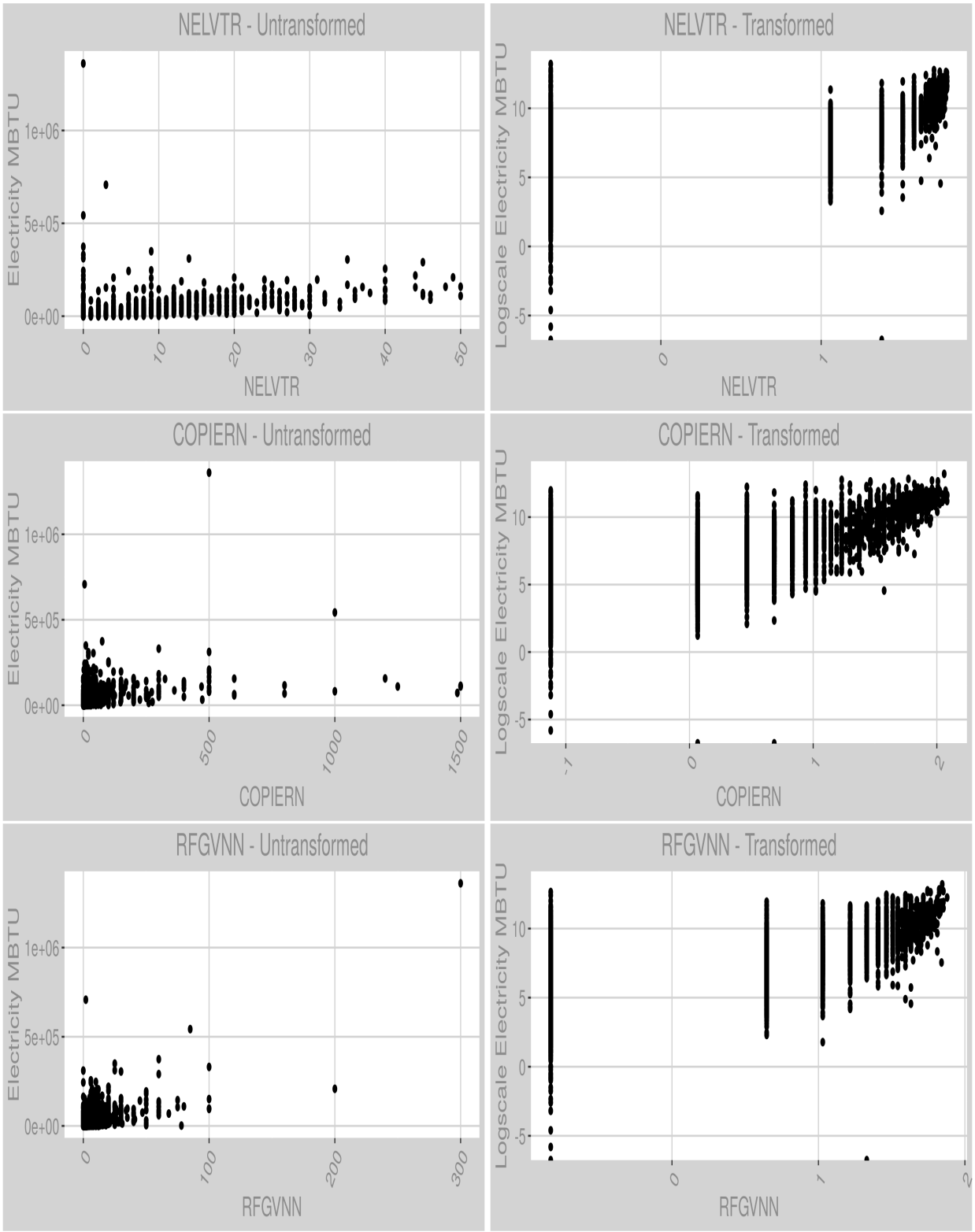
Select Variable Analysis

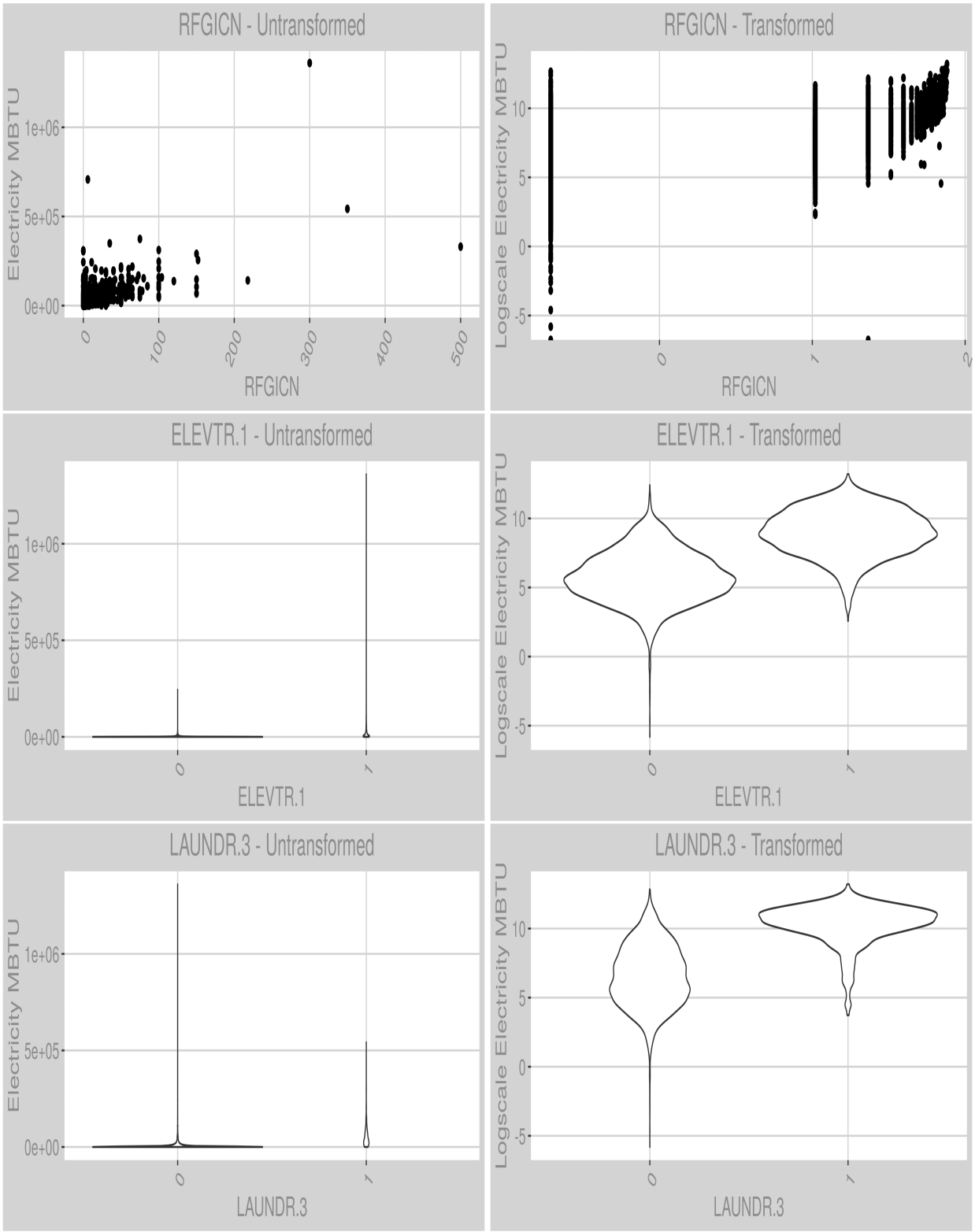
Model	RMSE	R2	MAE
partialLeastSquares	1.169408e+04	0.7780371	3.237700e+03
leaps	2.327804e+04	0.3559985	5.652596e+03
randomForest	1.182305e+06	0.2788200	8.100512e+05
recursiveFeatureExtraction	2.221181e+07	0.8247289	8.762935e+06
neuralNetwork	4.453689e+07	0.1406134	1.061397e+07
lasso	8.242048e+13	0.0000032	2.936412e+13

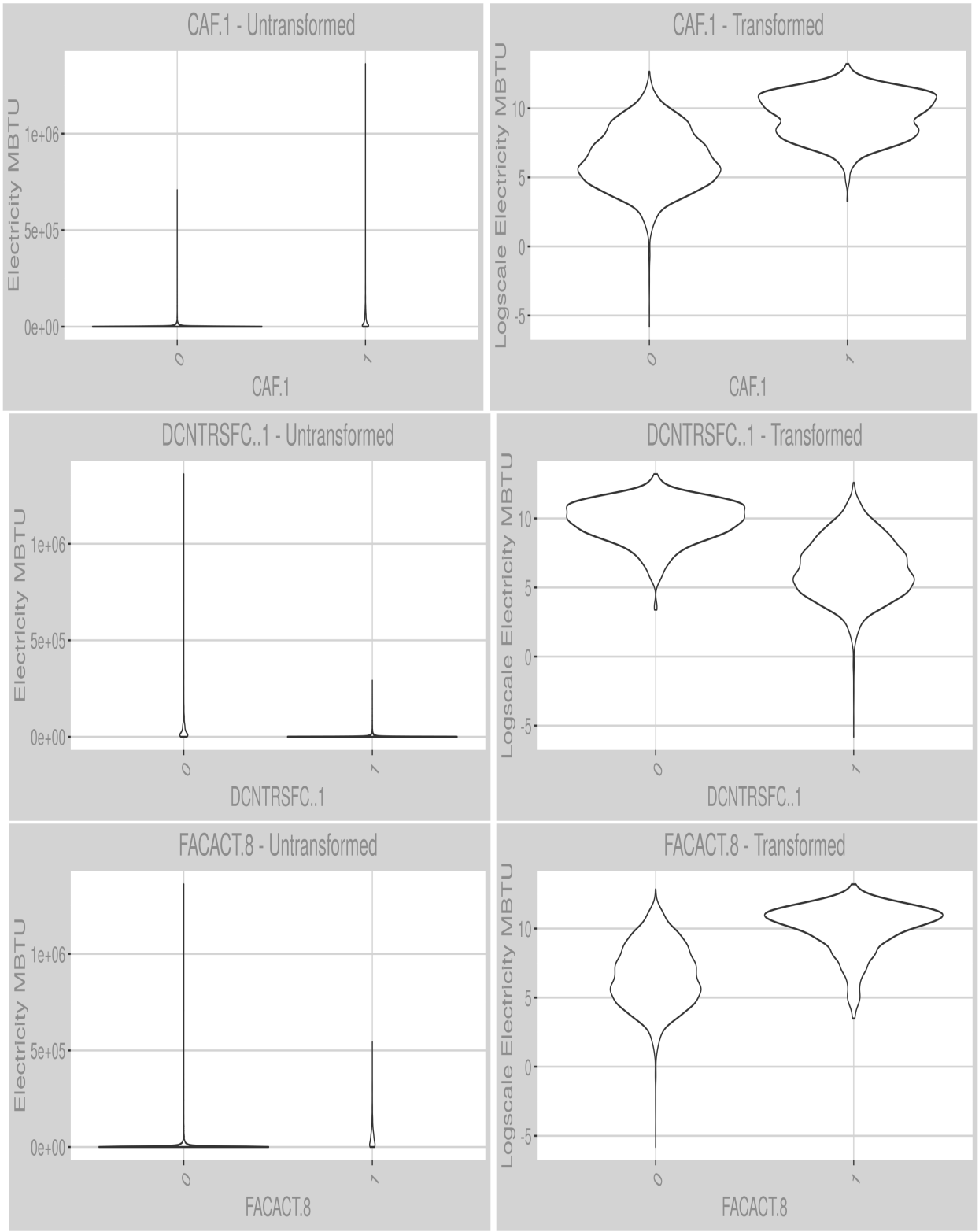


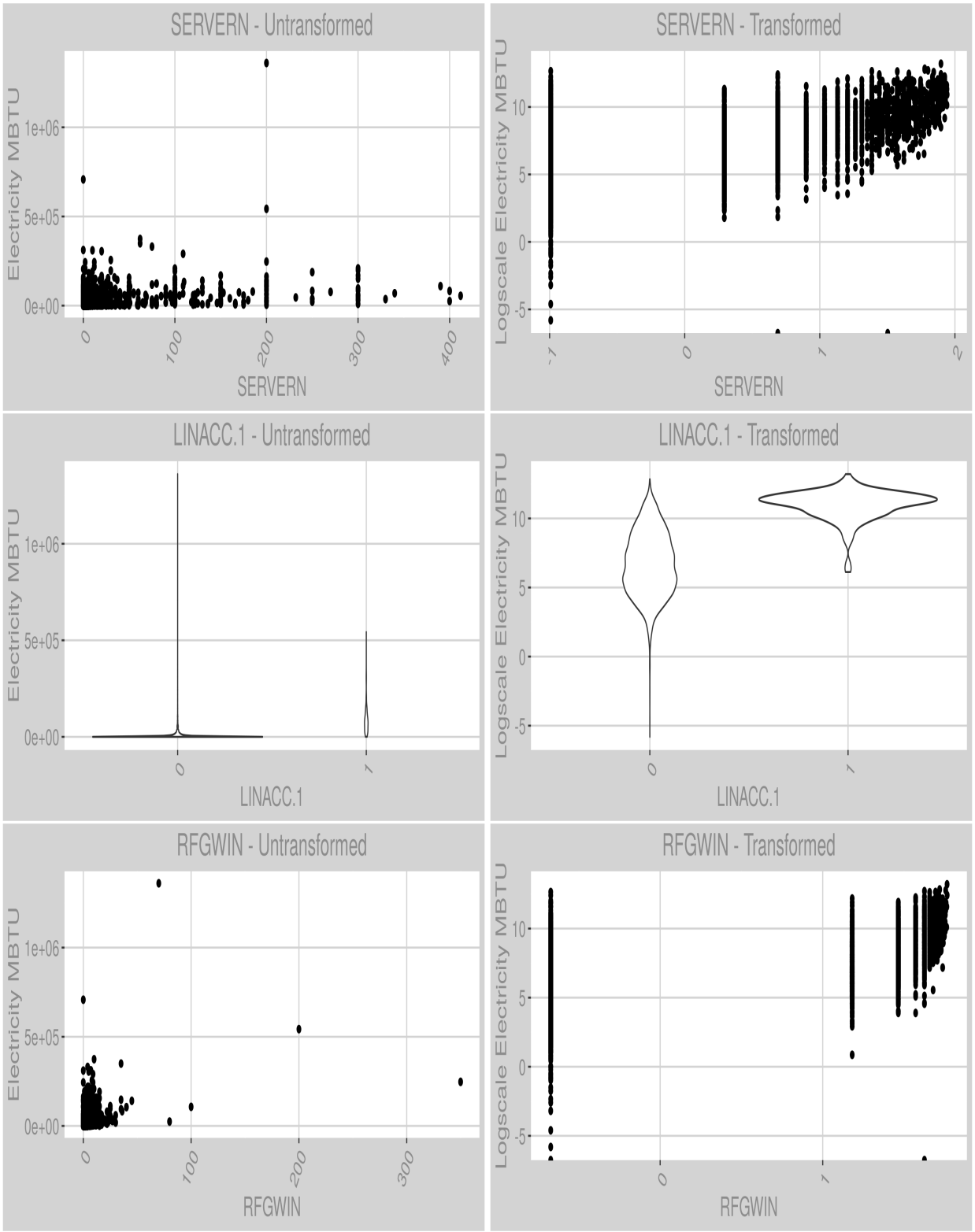

NWKER - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
2000
4000
6000
NWKER
NWKER - Transformed
Logscale Electricity MBTU
10
5
0
-5
-2
-1
0
1
2
3
NWKER
PRNTRN - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1000
2000
3000
4000
5000
PRNTRN
PRNTRN - Transformed
Logscale Electricity MBTU
10
5
0
-5
-1
0
1
2
3
PRNTRN
PUBCLIM.7 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
PUBCLIM.7
PUBCLIM.7 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
PUBCLIM.7

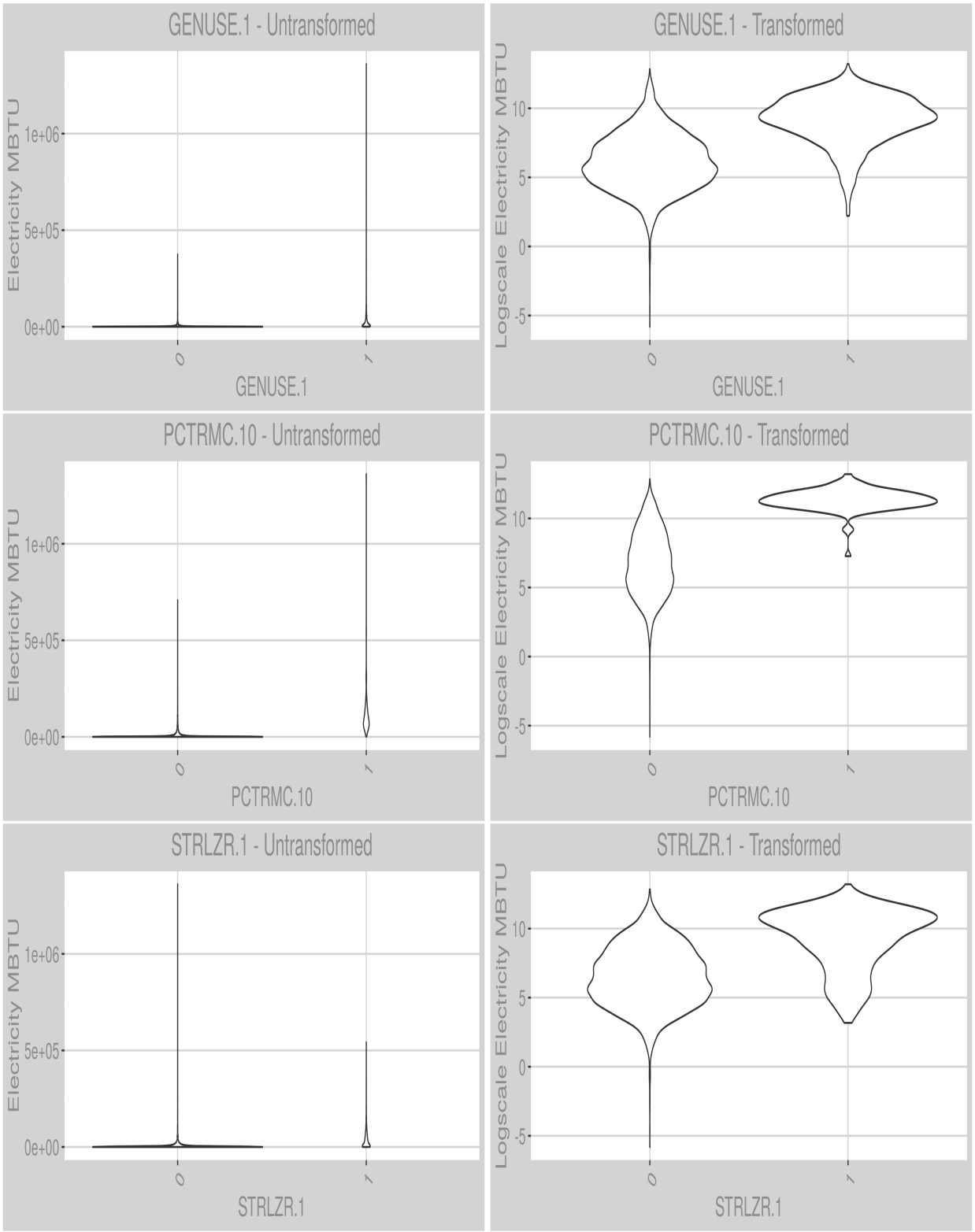

HCBED_bin.2 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
HCBED_bin.2
HCBED_bin.2 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
HCBED_bin.2
TVVIDEON_bin.2 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
TVVIDEON_bin.2
TVVIDEON_bin.2 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
TVVIDEON_bin.2
MRI.1 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
MRI.1
MRI.1 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
MRI.1


NELVTR - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
10
20
30
40
50
NELVTR
NELVTR - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
NELVTR
COPIERN - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
500
1000
1500
COPIERN
COPIERN - Transformed
Logscale Electricity MBTU
10
5
0
-5
-1
0
1
2
COPIERN
RFGVNN - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
100
200
300
RFGVNN
RFGVNN - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
2
RFGVNN


RFGICN - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
100
200
300
400
500
RFGICN
RFGICN - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
RFGICN
ELEVTR.1 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
ELEVTR.1
ELEVTR.1 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
ELEVTR.1
LAUNDR.3 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
LAUNDR.3
LAUNDR.3 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
LAUNDR.3

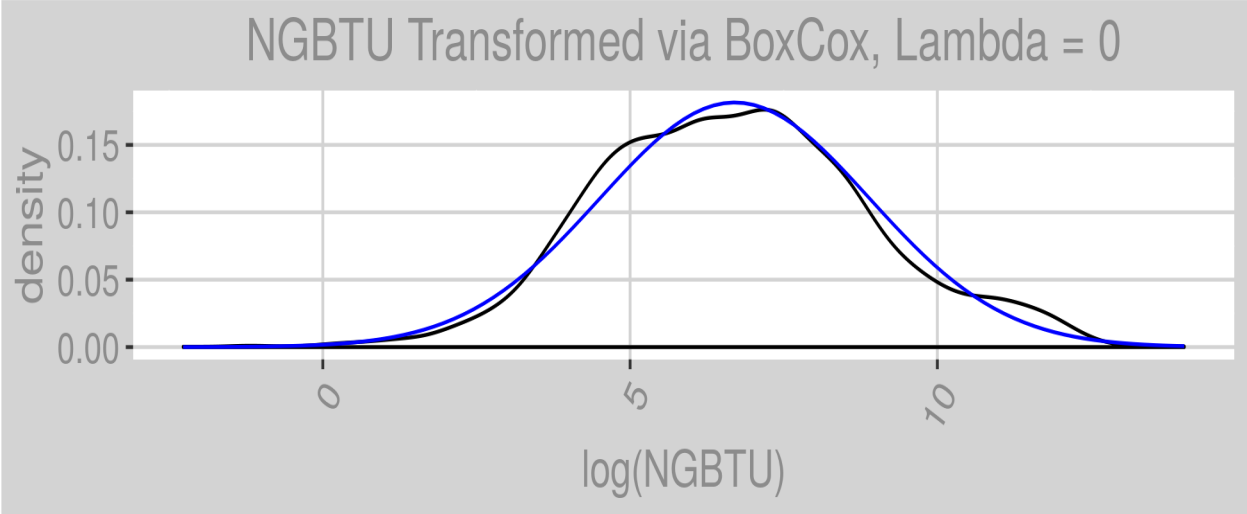

CAF.1 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
CAF.1
CAF.1 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
CAF.1
DCNTRSFC..1 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
DCNTRSFC..1
DCNTRSFC..1 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
DCNTRSFC..1
FACACT.8 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
FACACT.8
FACACT.8 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
FACACT.8


SERVERN - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
100
200
300
400
SERVERN
SERVERN - Transformed
Logscale Electricity MBTU
10
5
0
-5
-1
0
1
2
SERVERN
LINACC.1 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
LINACC.1
LINACC.1 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
LINACC.1
RFGWIN - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
100
200
300
RFGWIN
RFGWIN - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
RFGWIN

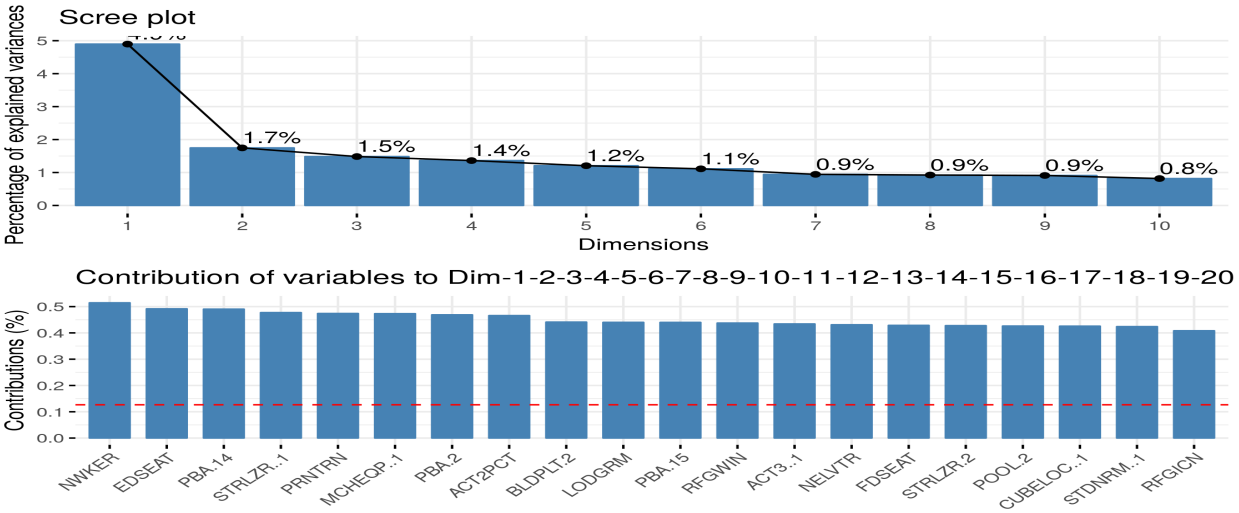

GENUSE.1 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
GENUSE.1
GENUSE.1 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
GENUSE.1
PCTRMC.10 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
PCTRMC.10
PCTRMC.10 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
PCTRMC.10
STRLZR.1 - Untransformed
Electricity MBTU
1e+06
5e+05
0e+00
0
1
STRLZR.1
STRLZR.1 - Transformed
Logscale Electricity MBTU
10
5
0
-5
0
1
STRLZR.1

Appendix - Natural Gas

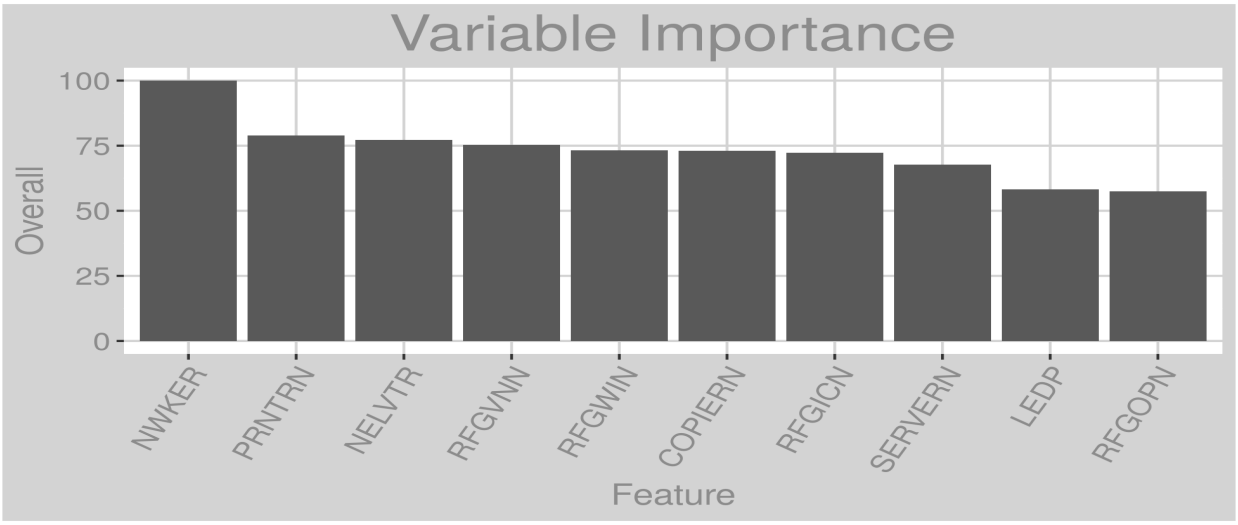
Response

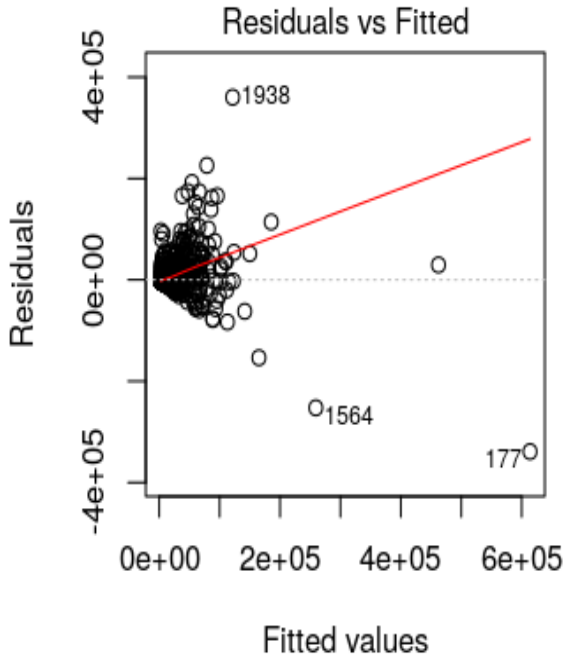


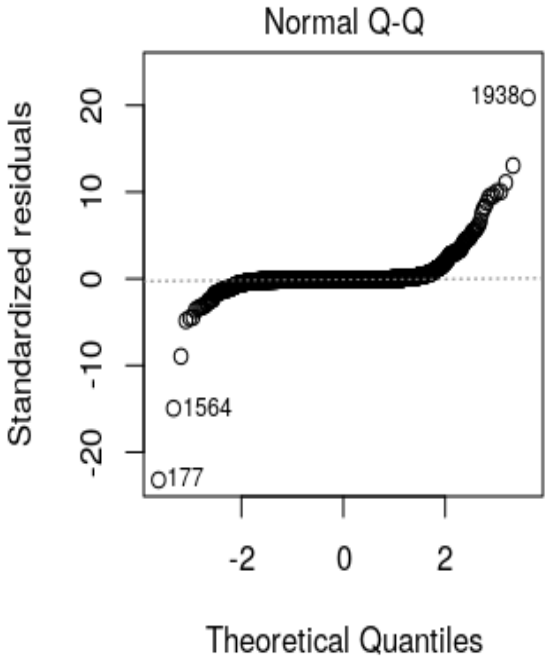
PCA

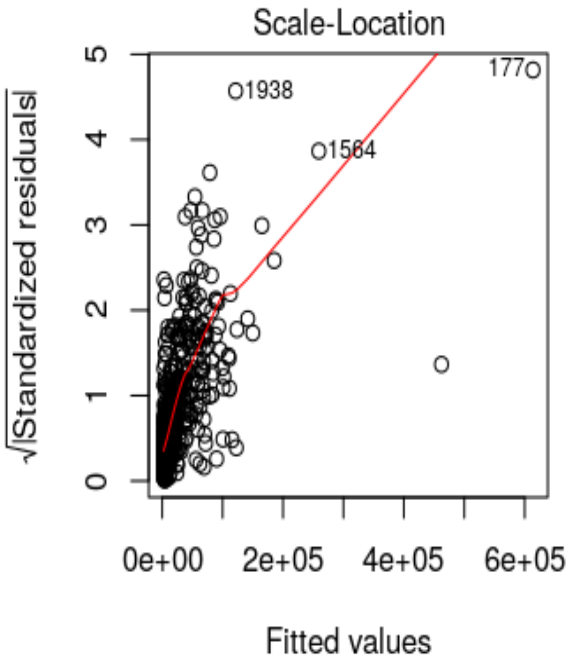


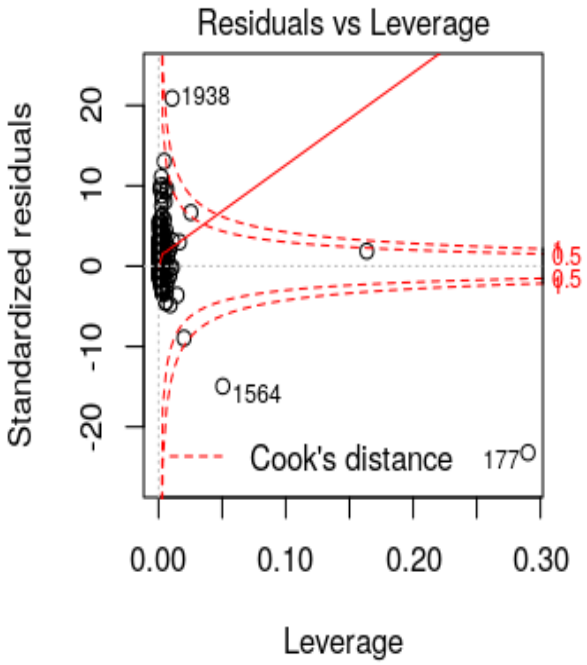
PLS

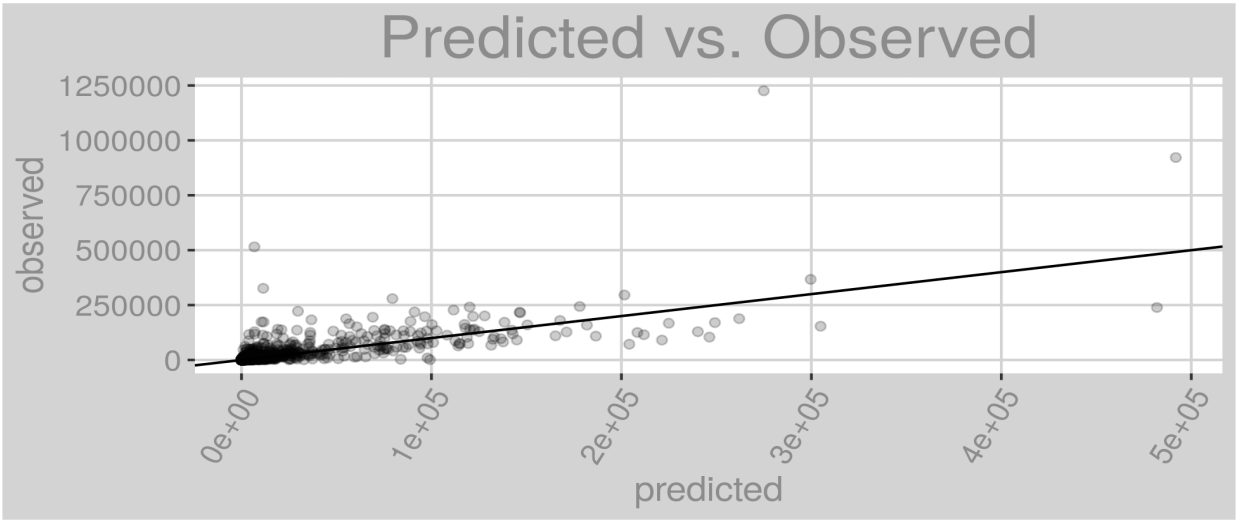

Variable Importance
Overall
100
75
50
25
0
NWKER
PRNTRN
NELVTR
RFGVNN
RFGWIN
COPIERN
RFGICN
SERVERN
LEDP
RFGOPN
Feature


Residuals vs Fitted
Residuals
4e+05
0e+00
-4e+05
1938
1564
177
0e+00
2e+05
4e+05
6e+05
Fitted values


Normal Q-Q
Standardized residuals
20
10
0
-10
-20
1938
1564
177
-2
0
2
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
0
1
2
3
4
5
1938
1564
177
0e+00
2e+05
4e+05
6e+05
Fitted values


Residuals vs Leverage
Standardized residuals
-20
-10
0
10
20
1938
1564
177
Cook's distance
0.5
0.5
0.00
0.10
0.20
0.30
Leverage


Predicted vs. Observed
observed
1250000
1000000
750000
500000
250000
0
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
predicted

Appendix - Neural Networks

Electricity