CUNY MSDS Capstone Project

COMMERCIAL BUILDING ENERGY CONSUMPTION

ANALYSIS AND PREDICTION

May 13, 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

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)[2]. While the survey is expansive (i.e. more than 600 tracked features), it is useful to identify predictors that significantly affect consumptiona and that can be attained by building operators. This study will focus on creating a series of models to extract the most important survey questions and then use these values as predictors to train a final model that predicts fuel use consumption for standard practice buildings.

Research

Related Work

The idea of determing operational building energy efficiency is not a novel concept in itself. ENERGY STAR has a building benchmarking tool[3]. Additionally, the United States Green Building Council has created the Arc Platform[4] which provides benchmarking and active monitoring features. While these platforms provide building comparisons in the form of an overall score, it is beneficial to explore the space around various building attribute inputs themselves as well as compare

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

[2] EIA Microdata

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

[4] https://arcskoru.com/

consumption of a specific building to its equivalent standard practice building in a programmatic way.

Literature Review

There are a variety of texts that are dedicated to the analysis of building energy consumption, and determining operating efficiency. For example, ASHRAE Guildine 14[5] provides a standardized set of energy, demand, and water savings calculation procedures. 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)[6]. 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 increases [7].

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 smaller subsets from various feature extraction algorithms. The magnitude and contribution percentage of each variable will be considered in selecting features from this model. In order to try and normalize the data, the response variable was divided by the gross floor area of the building and reported in in units of BTU per square foot (e.g. ELBTU/PerSF, NGBTU/PerSF). A neural network model will be built to take the subset of extracted 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.

[5]-https://www.techstreet.com/standards/guideline-14-2014-measurement-of-energy-demand-and-water-savings?product_id=1888937

[6]https://www.energycodes.gov/status-state-energy-code-adoption

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

Data Pre-Processing

The raw data set consists of 6,720 records and 1,119 features based on a stratified sample of commercial buildings. Multiple steps of preprocessing were required in order to prepare the data.

Electricity

General

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. Numeric predictors were transformed via BoxCox methodology as well as centered and scaled. Two potential outliers were found in the analysis. Both data points had unusually high energy consumption for the building type noted, and were highly atypical in many other respsects.

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 BTU/SF 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: NA, Rsquared: NA

Top 5: `COOK.2[NO]`, `LAUNDR..1[NA]`, `ELCPLT..1[NA]`, `PBA.14[EDUCATION]`, `BLDPLT.2[NO]`

The principle component analysis indicates that only About 4.2% of the variance in the data can be explained in the first principle component, which then drops to about 1.7% 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: 49439, Rsquared: 0.512

Top 5: `NWKERPerSf`, `RGSTRNPerSf`, `FDSEATPerSf`, `RFGWINPerSf`, `PCTERMNPerSf`

This model returned a promising result; however, it must be noted that all predictors were used in this process. Looking at the output thus far, it is appears that the number of workers, receptical equipment, and refrigeration equipment, influence electrical consumption. *Appendix*

Variable Selection - Random Forest

RMSE: 165663, Rsquared: 0.138

Top 5: `RFGWINPerSf`, `RGSTRNPerSf`, `NWKERPerSf`, `RFGICNPerSf`, `PCTERMNPerSf`

The resulting error metrics were much less promising. However, similarly selected variables are are picked for this model when compared to the PLS. *Appendix*

Variable Selection - Forward Selection

RMSE: 91257, Rsquared: 0.316

Top 5: `NWKERPerSf`, `RFGWINPerSf`, `RFGWI.1[YES]`, `RFGICNPerSf`, `PCTERMNPerSf`

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. Large refrigeration equipment load and typical office space attributes dominated this analysis, as appears to be the case in previous models. It seems that in order to capture energy use for all buidling types, much more than 5 variables will be necessary. *Appendix*

Variable Selection - Recursive Feature Elimination

RMSE: 49714, Rsquared: 0.502

Top 5: `RGSTRNPerSf`, `RFGICNPerSf`, `NWKERPerSf`, `PBAPLUS.32[FAST FOOD]`, `RFGWINPerSf`

A more direct approach was taken with this model, which is specifically used to extract useful features from data sets. *Appendix*

Variable Selection - Simple Neural Network

RMSE: 51378, Rsquared: 0.463

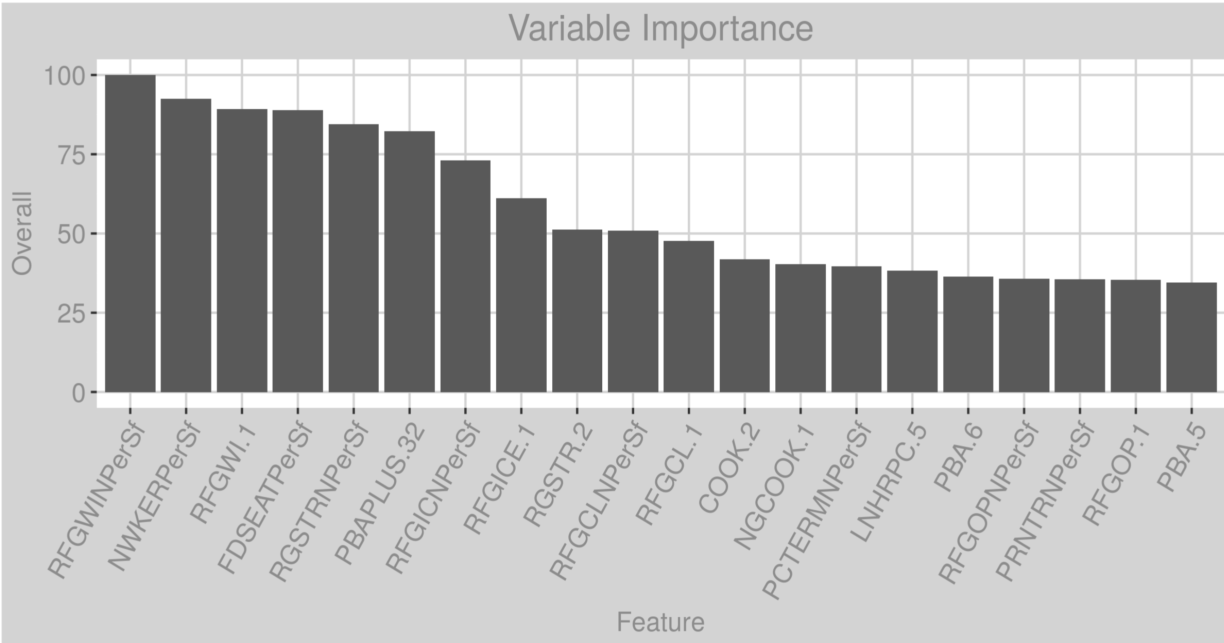
Top 5: `FDSeatPerSf`, `PBAPLUS.32[FAST FOOD]`, `RFGWINPerSf`, `RFGWI.1[YES]`, `RGSTRNPerSf`

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 work better with this process. As can be seen, there are some new attributes that surface which were not indicated to be of high importance previously. *Appendix*

Variable Selection - Selected Variable Analysis

Feature Extraction Model Results

Model	RMSE	R2	MAE	MAPE	MSLE
partialLeastSquares	49439.16	0.5119599	26622.19	85.70943	0.5680023
neuralNetwork	51378.59	0.4622766	29989.65	133.42257	0.8016741
recursiveFeatureExtraction	49713.91	0.5017194	31111.93	239.40187	1.0286980
randomForest	165662.98	0.1388473	144344.47	881.85161	3.3488969
leaps	91256.75	0.3157366	58679.30	99.92406	66.6999846



In order to rank the most impactful features, the variable importance metrics from the selected models were all set to the same scale then summed. 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 may have been beneficial in this case since there are some building type specific end-uses that are highly ranked. As previously noted, there are many attributes associated with refrigeration, office, and food sales equipment. Also, the attribute identifying one of the more atypical building types, speaking in an energy intensity sense, has made it into the top 20 (`PBA.5[NON-REFRIGERATED WAREHOUSE]`). Additionally, some occupancy features (`NWKERPerSf`, `FDSEATPerSf`) have been included which is expected given that they impact interior space cooling and ventilation loads. 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

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 6662 samples in the data set. The same transformations were applied to this response variable as electricity. *Appendix*

Variable Selection - PCA

RMSE: NA, Rsquared: NA

Top 5: `EDSEATPerSf`, `PBA.14[EDUCATION]`, `STRLZR.1[YES]`, `MCHEQP[NA]`, `ACT2PCT` *Appendix*

Variable Selection - PLS

RMSE: 67978, Rsquared: 0.232

Top 5: `FDSEATPerSf`, `HEATP`, `RFGWINPerSf`, `NWKERPerSf`, `RGSTRNPerSf`

Appendix

Variable Selection - Random Forest

RMSE: 174127, Rsquared: 0.122

Top 5: `DRYCL.1[YES]`, `FDSEATPerSf`, `RFGWINPerSf`, `LAUNDR.3[OFF-SITE]`, `STRLSZR.1[YES]` *Appendix*

Variable Selection - Forward Selection

RMSE: 69079, Rsquared: 0.184

Top 5: `FDSEATPerSf`, `TVVIDEONPerSf`, `NGWATR.2[NO]`, `RGSTRNPerSf`, `RFGWINPerSf` *Appendix*

Variable Selection - Recursive Feature Elimination

RMSE: 61465, Rsquared: 0.258

Top 5: `NWKERPerSf`, `RFGICNPerSf`, `RFGWINPerSf`, `FDSEATPerSf`, `RGSTRNPerSf` *Appendix*

Variable Selection - Simple Neural Network

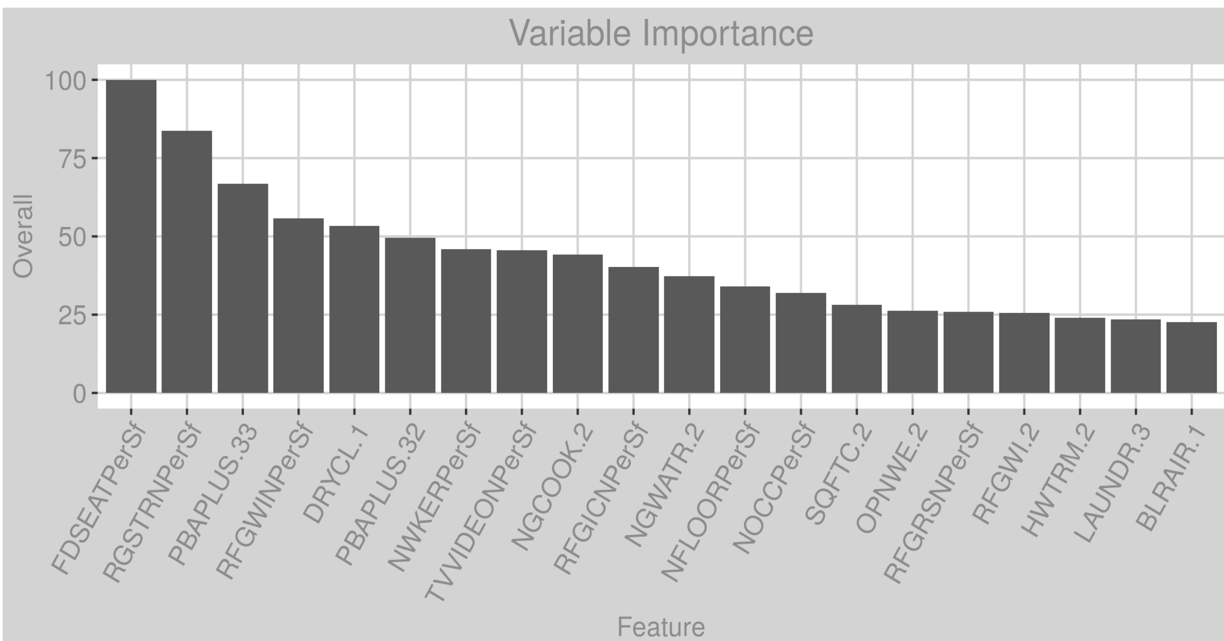
RMSE: 57929, Rsquared: 0.319

Top 5: `FDSEATPerSf`, `RGSTRNPerSf`, `DRYCL.1[YES]`, `PBAPLUS.33[RESTAURANT/CAFETERIA]`, `PBAPLUS.32[FAST FOOD]` *Appendix*

Variable Selection - Selected Variable Analysis

Feature Extraction Model Results

Model	RMSE	R2	MAE	MAPE	MSLE
partialLeastSquares	67978.44	0.2321160	31769.30	469.5262	1.215225
leaps	69079.81	0.1837522	34964.80	263.2334	1.367507
neuralNetwork	57929.72	0.3193293	33328.70	1660.1715	1.731125
recursiveFeatureExtraction	61464.86	0.2589020	35070.92	1887.8960	2.018073
randomForest	174127.85	0.1215693	147326.47	8693.7861	5.379562



As with the electricity model, attributes related to occupancy seem to have made a large impact, possibly due to the need to heat ventilation air, especially given some of these occupancy types are associated with 24/7 operation. Specifically, buildings which report a high density of food service seating appear to be directly correlated with high gas usage, which is not surprising. Also as expected, cooking and large heating equipment attributes are high on the list. Surprisingly, the number of floors per gross floor area has shown some importance, perhaps due to building shape and its relationship with heating needs (i.e. volume to area ratio). *Appendix*

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. Note, lasso models were not reported in the previous section due to these issues. Secondly, neural networks have the ability to select different 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 better fits for the low estimates. To reflect this reasoning, additional metrics, and loss function, for this set of models were chosen to be the mean absolute percentage error (MAPE) and mean square logarithmic error (MSLE). These metrics will be evaluated for all models to determine effectiveness.

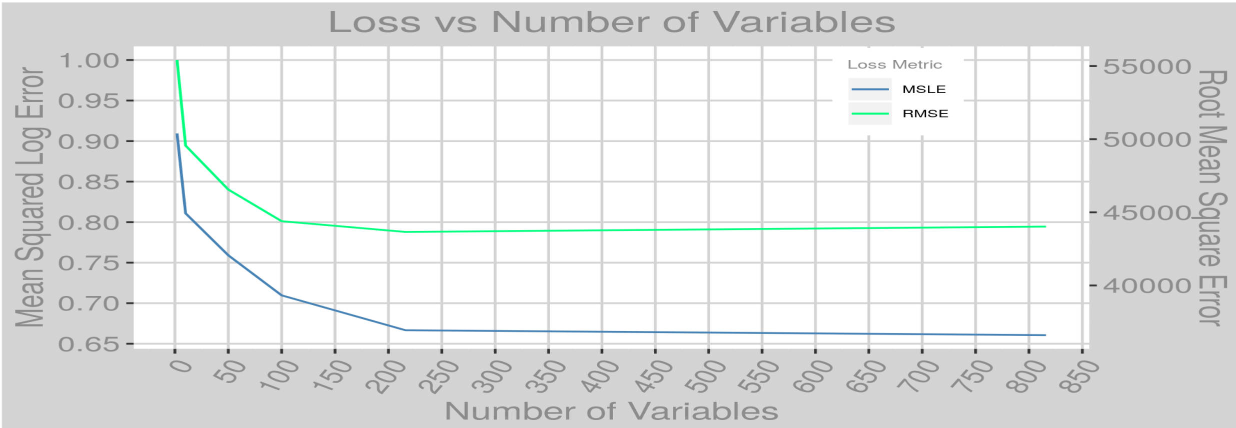
Hyperparameter Training

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 5 hidden layers, 1000 hidden layer nodes, a dropout rate of 0.3, no regularization, batch sizes of 150, using the rmsprop() algorithm with a learning rate of 0.0005, and 100 predictors. As can be seen in the graph below, the number of variables needed to obtain near-peak performance, is much less than the full set.



The final selected model, after re-training, has a MSLE of 0.92 and RMSE of 54696. Comparing this model ('Full Neural Network') to the previous feature extraction models, which used many more variables, the performance is competitive. Additionally, the results were then multiplied by their respect gross floor area and then compared to the set of feature extraction models, with the same transformation, in order to evaluate the total consumption prediction error. Again, it can be seen that this neural network model has shown to be competetive in this manner and, in fact, has a better Rsquared value.

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 since the loss function (MSLE) allows for higher error in higher consumption projects. *Appendix*

Per SF Model Comparison [BTU/SF]

Model	RMSE	R2	MAE	MAPE	MSLE
partialLeastSquares	49439	0.51	26622	86	0.57
neuralNetwork	51379	0.46	29990	133	0.80
Full Neural Network	54696	0.49	28463	67	0.92
recursiveFeatureExtraction	49714	0.50	31112	239	1.03
randomForest	165663	0.14	144344	882	3.35
leaps	91257	0.32	58679	100	66.70

Total Consumption Prediction vs. Feature Extraction Models [mmBTU]

Model	RMSE	R2	MAE	MAPE	MSLE
partialLeastSquares	16180	0.62	3813	86	0.57
neuralNetwork	14267	0.68	3616	133	0.80
Full Neural Network	13542	0.72	3514	67	0.92
recursiveFeatureExtraction	13308	0.71	3535	239	1.03
randomForest	52723	0.68	21239	882	3.35
leaps	26036	0.65	8831	100	66.70

Variable Selection Summary

The final model chosen uses 100 variables. Many of the features within the set have to do with the amount of receptacle equipment within the building as well as major electrical devices (e.g. MRI machines) and essential equipment (e.g. data center servers, refrigeration). Also, building type identifiers have been included for various categories. *Appendix*

The automated selection process does seem to have included some highly correlated pairs, such as the number of workers per square foot as well as the categorical bin of workers. This does reduce the number of necessary questions, but it is unclear if both are necessary and/or if they are possibly detrimental. Also, there are a number of questions that may be automatically known just based on the usage type as some questions do not apply to all buidlings. It is possible take the steps used in asking questions in the survey in order to build a live form that can automatically parse out the meaningful questions based on building type, which could reduce the need to enter a value for all the selected variables.

Natural Gas

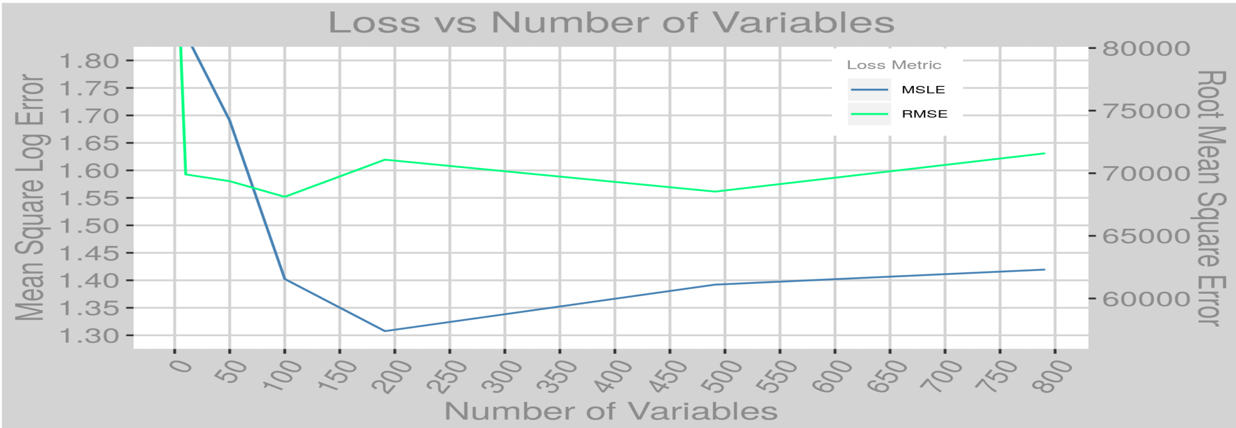
Summary

The final selected model consisted of a 5 hidden layers, 800 hidden layer nodes, a dropout rate of 0.3, no regularization, batch sizes of 150, using the rmsprop() algorithm with a learning rate of 0.0005, and 100 predictors. As can be seen in the graph below, the number of variables needed to obtain near-peak performance, is much less than the full set. *Appendix*

The final selected model, after re-training, has a MSLE of 1.42 and RMSE of 61341. Comparing this model ('Full Neural Network') to the previous feature extraction models, which used many more variables, the performance is actually better (when using RMSE). *Appendix*

Variable Selection Summary

This fuel source prediction appears to highly depend on predictor variables that indicate occupancy and length of building operation, such as food service seating capacity, number of cash registers,



Per SF Model Comparison [BTU/SF]

Model	RMSE	R2	MAE	MAPE	MSLE
partialLeastSquares	67978	0.23	31769	470	1.22
leaps	69080	0.18	34965	263	1.37
Full Neural Network	61341	0.29	32608	713	1.42
neuralNetwork	57930	0.32	33329	1660	1.73
recursiveFeatureExtraction	61465	0.26	35071	1888	2.02
randomForest	174128	0.12	147326	8694	5.38

Total Consumption Prediction vs. Feature Extraction Models [mmBTU]

Model	RMSE	R2	MAE	MAPE	MSLE
partialLeastSquares	15567	0.45	3671	293	1.19
leaps	15769	0.43	3950	213	1.35
Full Neural Network	13191	0.69	4267	713	1.42
neuralNetwork	13365	0.59	3641	955	1.70
recursiveFeatureExtraction	14798	0.50	4016	1172	2.00
randomForest	43661	0.43	16812	5332	5.36

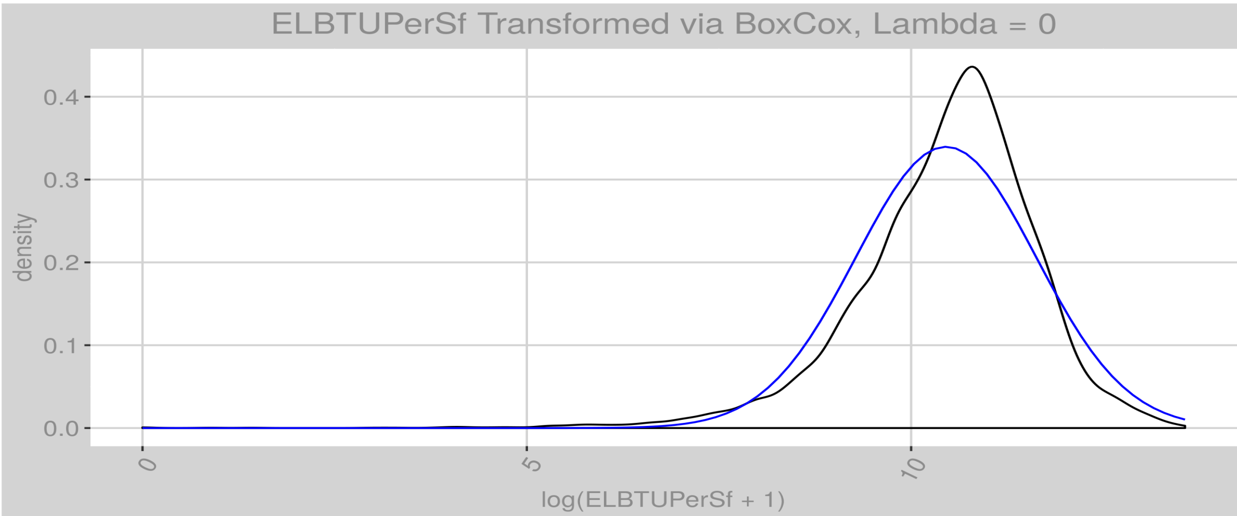
weekend operation, etc. Additionally, indicators of commercial cooking are included. All these selections are unsurprising as natural gas, in the U.S.A, is mainly used for heating energy and food preparation.

Future Work

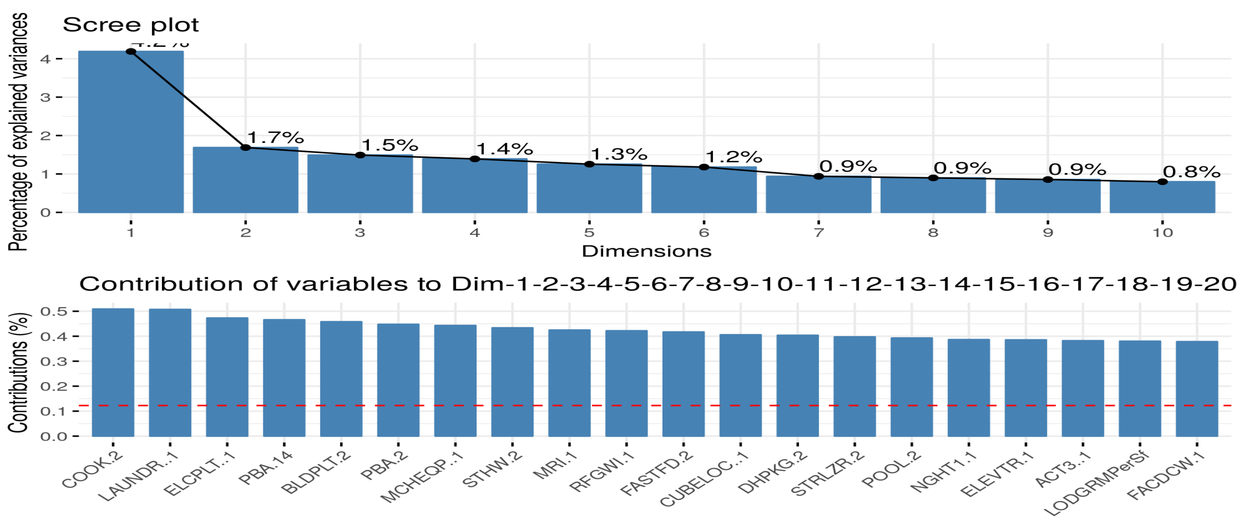
While it was determined that some heteroscedacitity would be acceptable, there does appear to be area for improvement. Additionally, as mentioned at the beginning of this report, the sampling of this data set was stratified to reflect the building population. However, it is noted that there are some building 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. Also, there was not a lot of attention paid to the actual transformations of the predictors given the large quantity of them. It is possible a better fit can be obtained with more intelligent transformations applied to the features after further analysis.

Appendix - Electricity

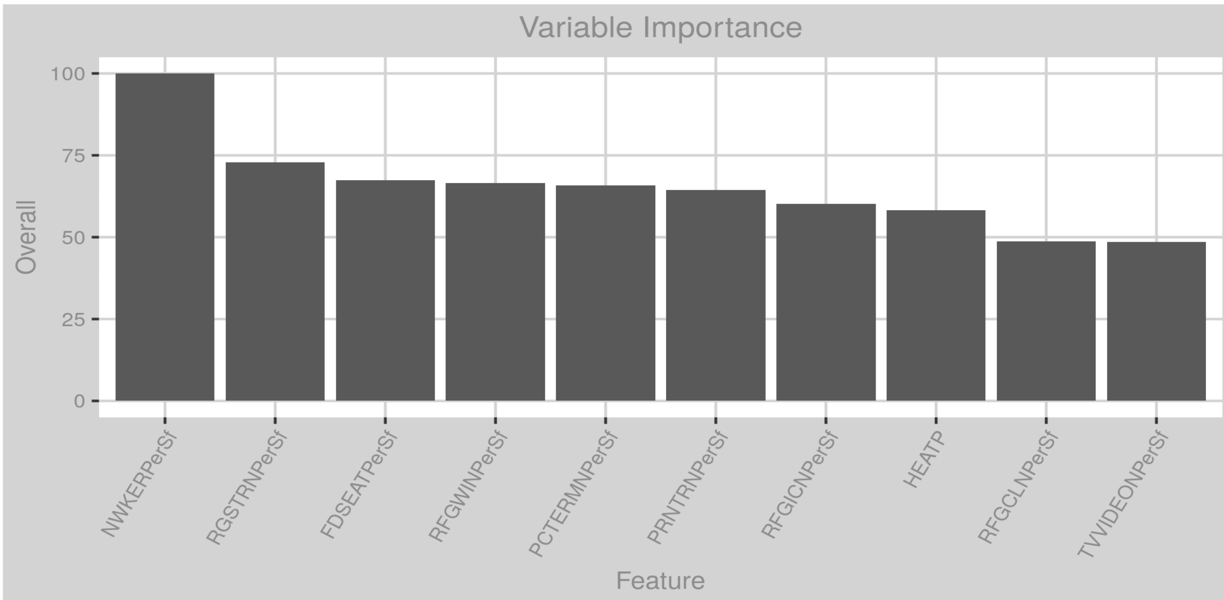
Response



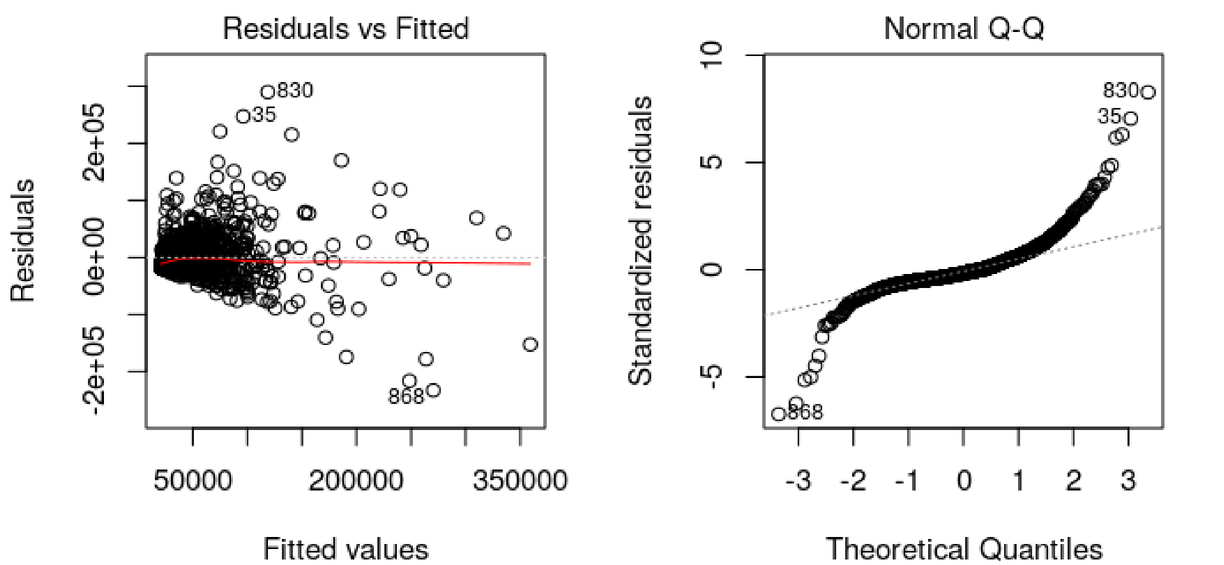
PCA



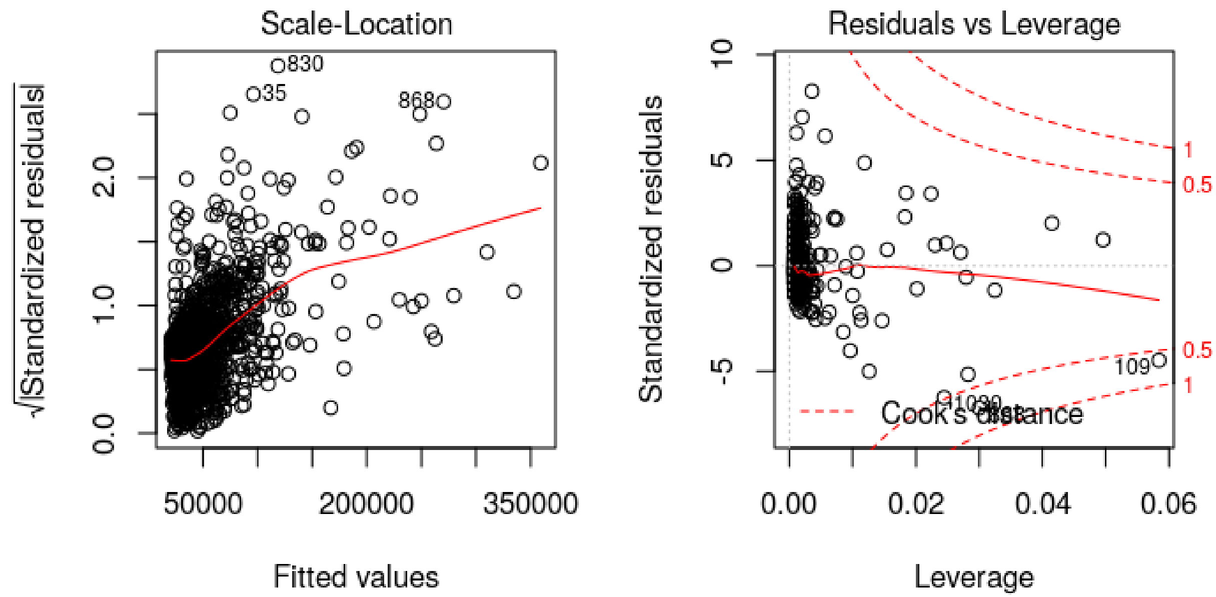
PLS

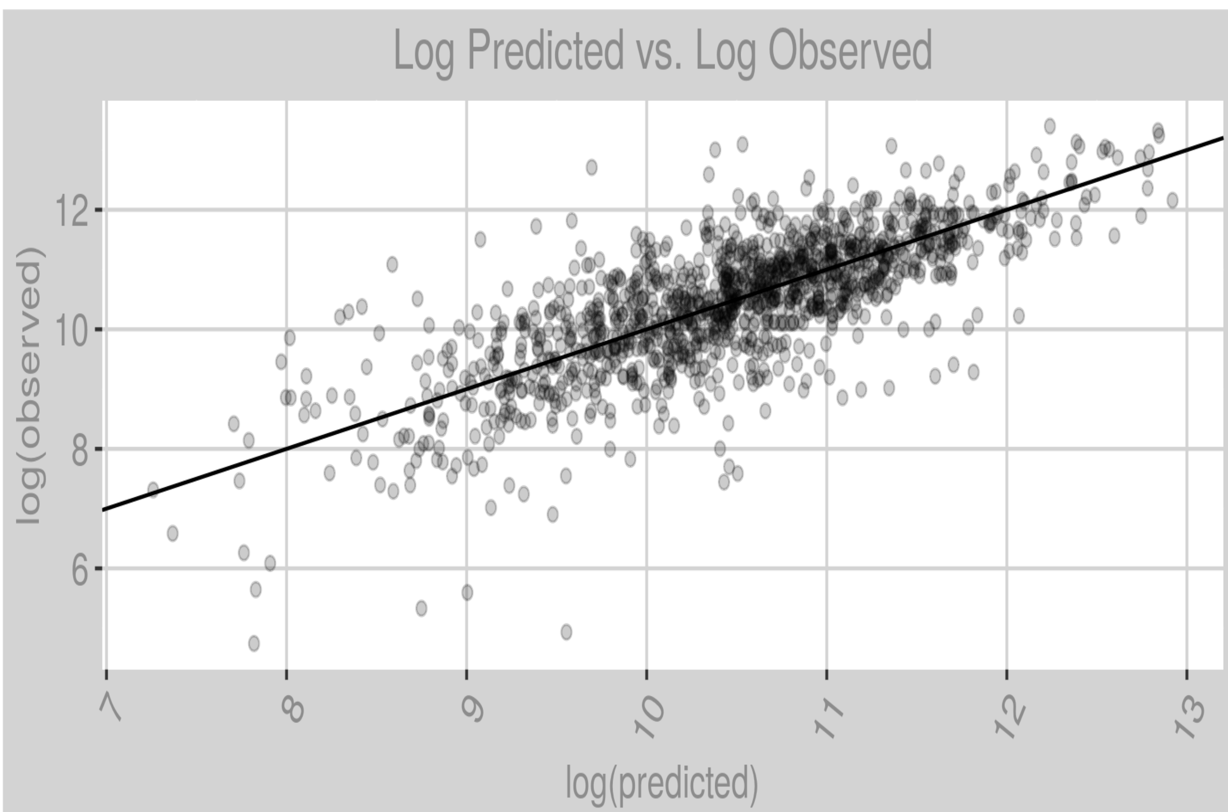


Variable Importance
Overall
100
75
50
25
0
NWKERPerSf
RGSTRNPerSf
FDSEATPerSf
RFGWINPerSf
PCTERMNPerSf
PRNTRNPerSf
RFGICNPerSf
HEATP
RFGCLNPerSf
TVVIDEONPerSf
Feature

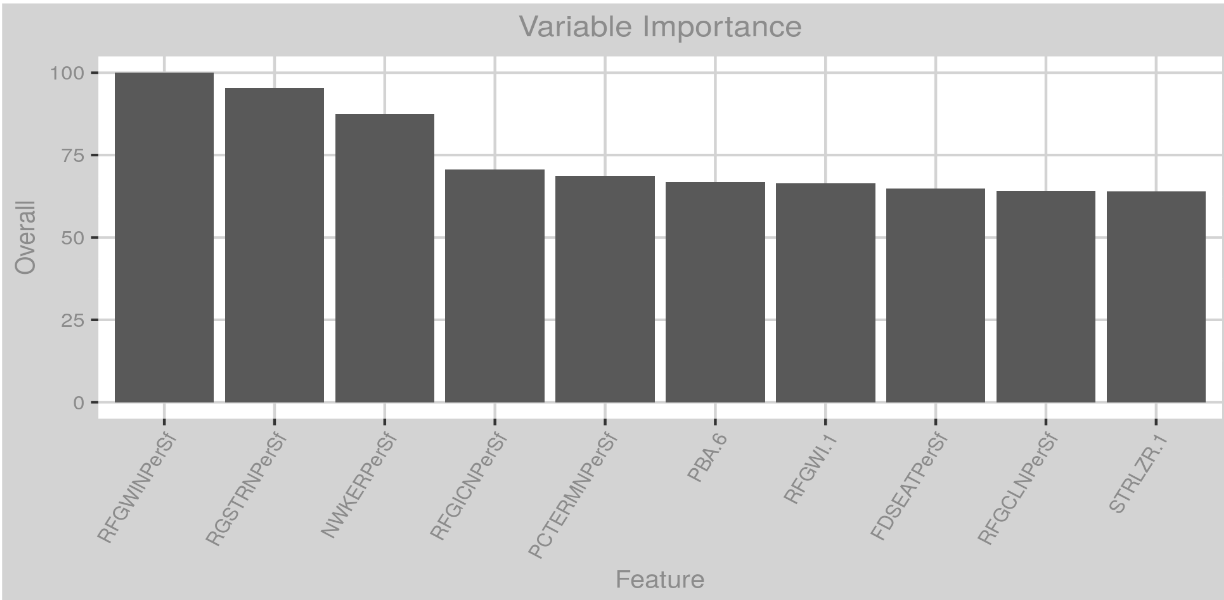


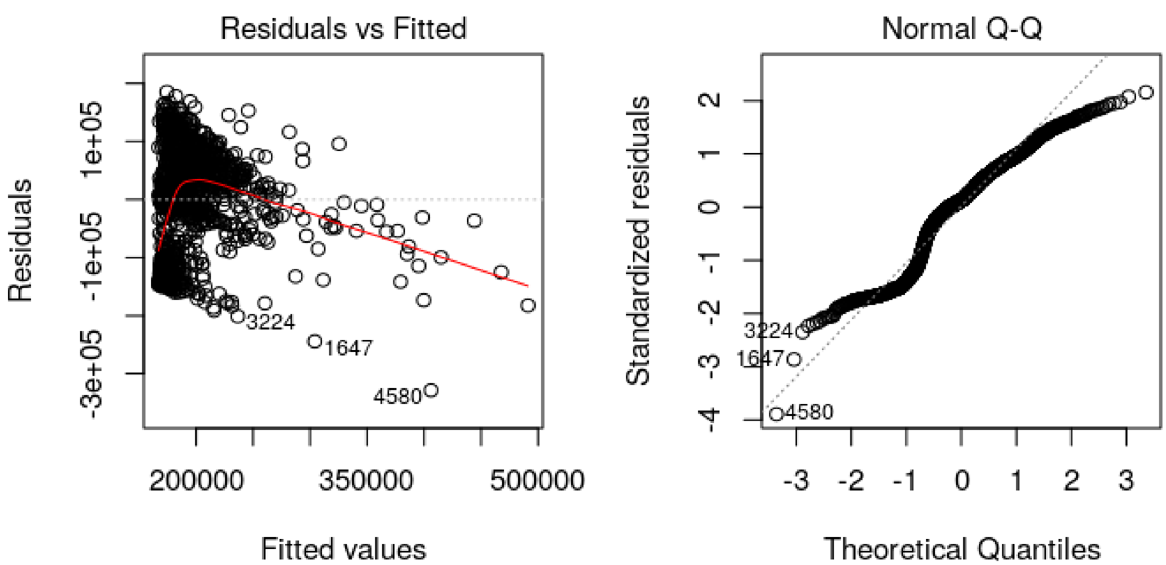
Residuals vs Fitted
Residuals
Fitted values
2e+05
0e+00
-2e+05
50000
200000
350000
830
35
868
Normal Q-Q
Standardized residuals
Theoretical Quantiles
10
5
0
-5
-3
-2
-1
0
1
2
3
830
35
868

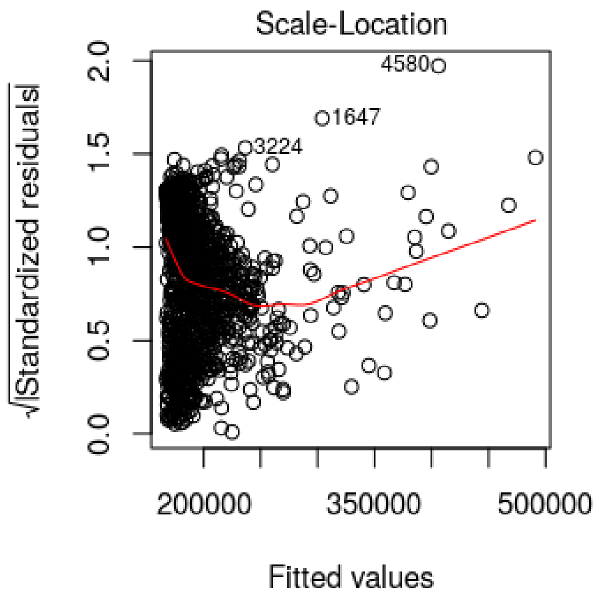

Scale-Location
√|Standardized residuals|
Fitted values
0.0
1.0
2.0
50000
200000
350000
830
35
868
Residuals vs Leverage
Standardized residuals
Leverage
10
5
0
-5
0.00
0.02
0.04
0.06
1
0.5
0.5
1
109
Cook's distance

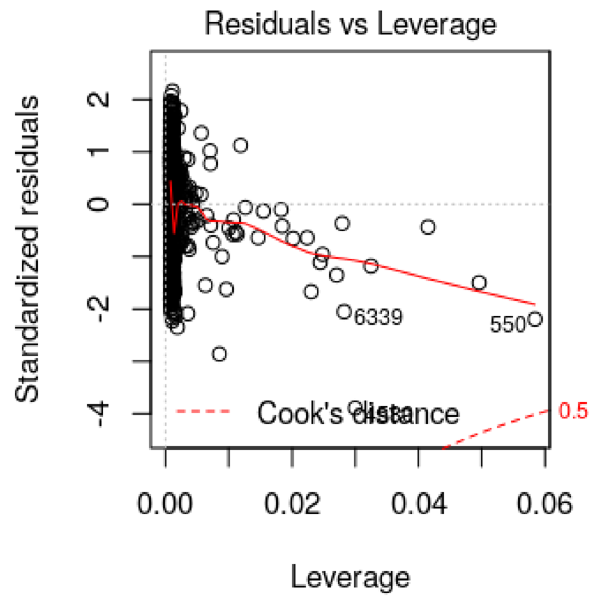

Log Predicted vs. Log Observed
log(observed)
log(predicted)
12
10
8
6
7
8
9
10
11
12
13

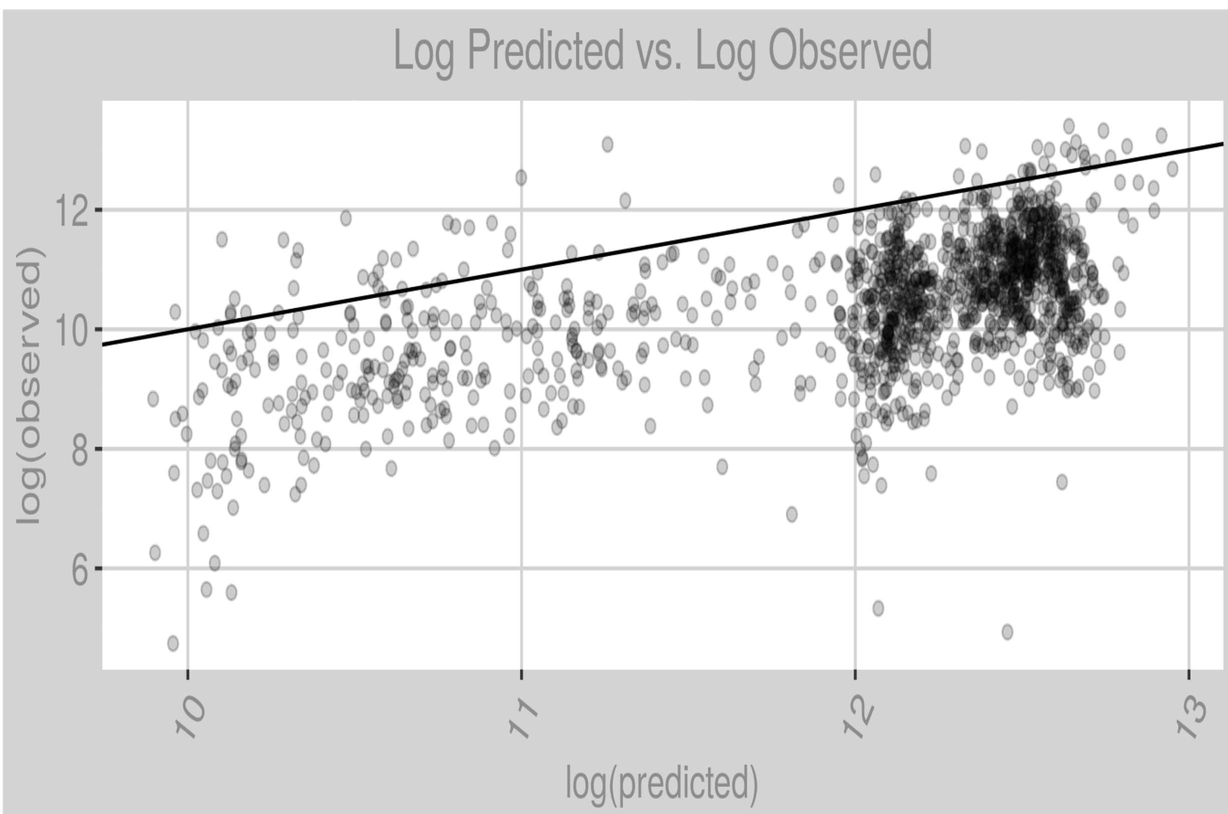
Random Forest



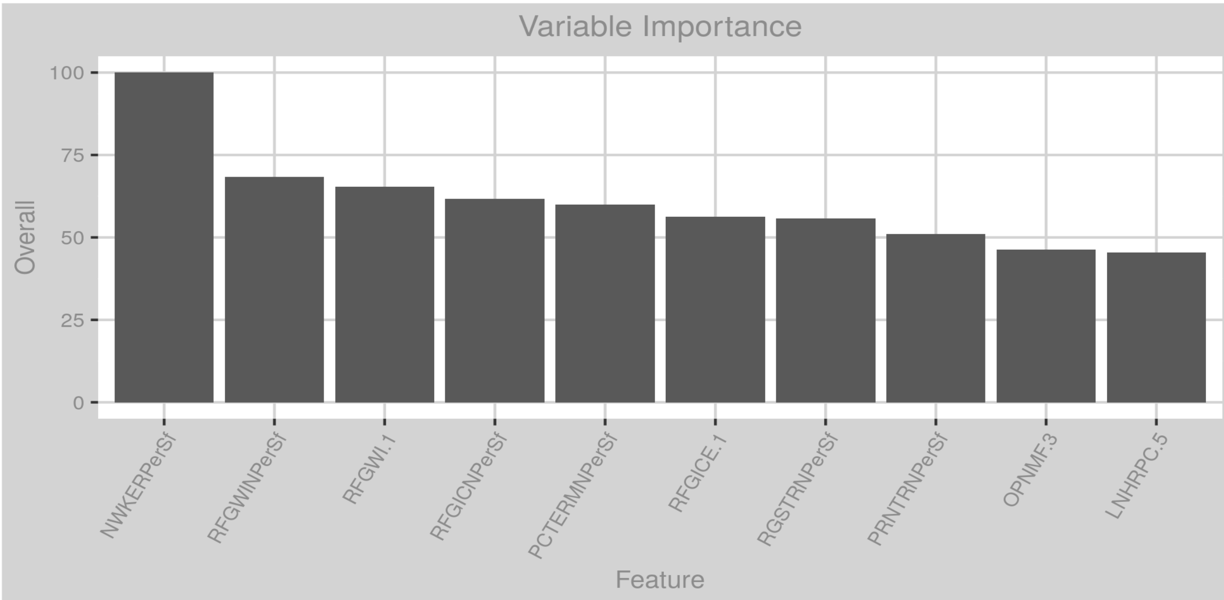


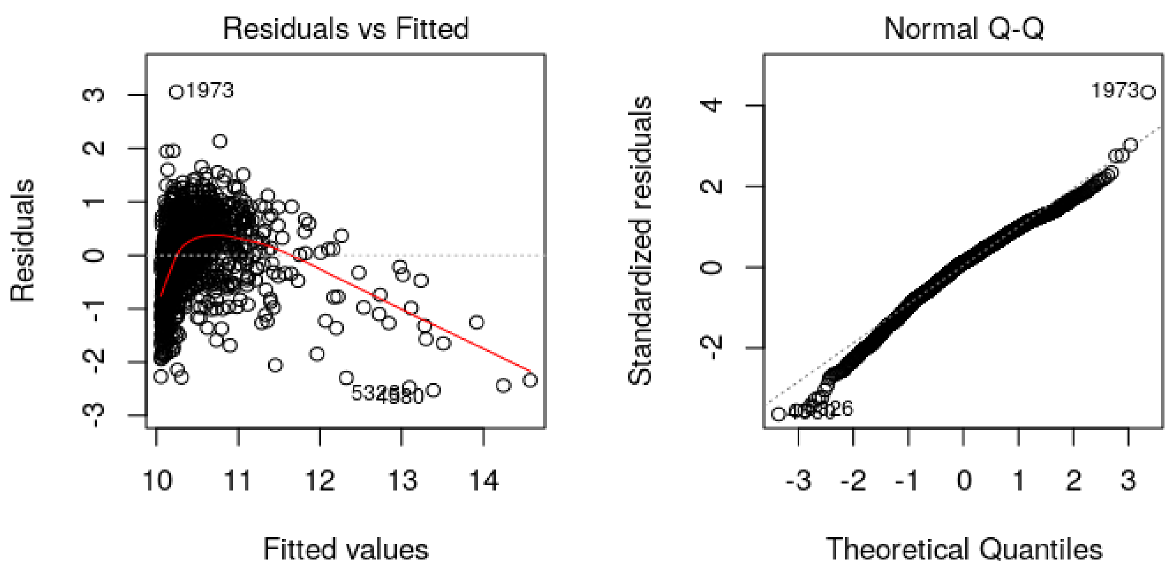

Scale-Location
√|Standardized residuals|
Fitted values
4580
1647
3224
0.0
0.5
1.0
1.5
2.0
200000
350000
500000

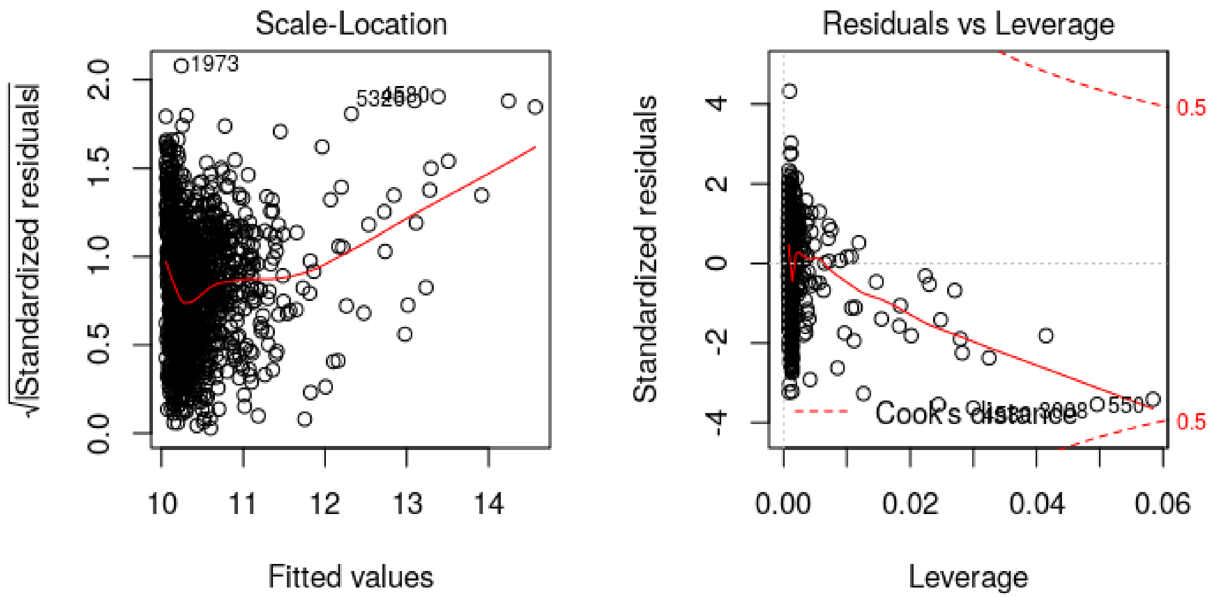

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
6339
550
-4
-2
0
1
2
0.00
0.02
0.04
0.06

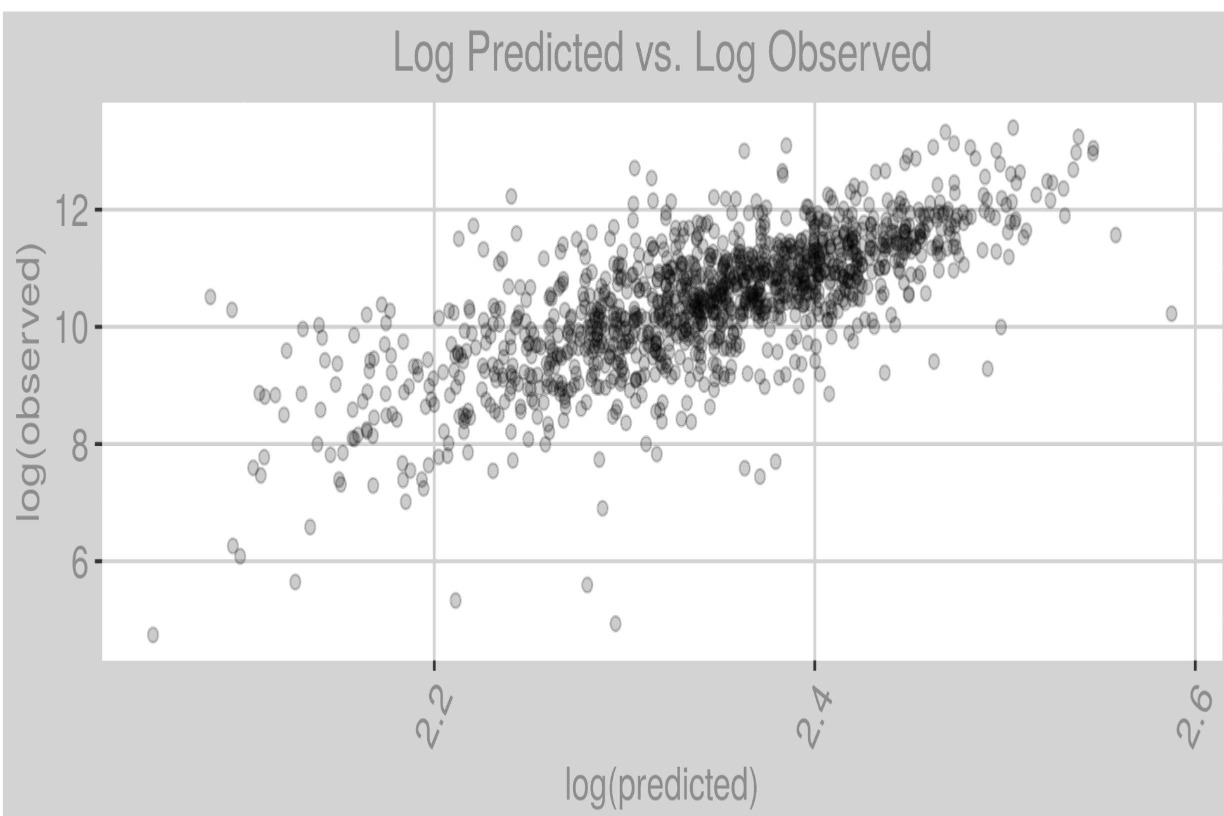

Log Predicted vs. Log Observed
log(observed)
log(predicted)
6
8
10
12
10
11
12
13

Forward Selection

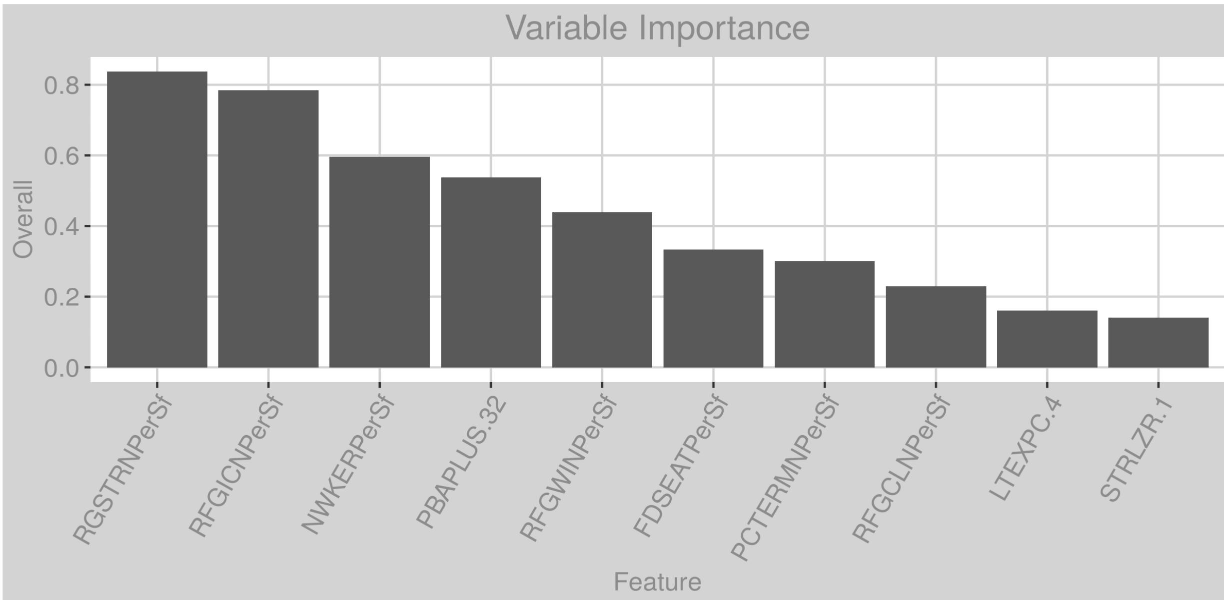


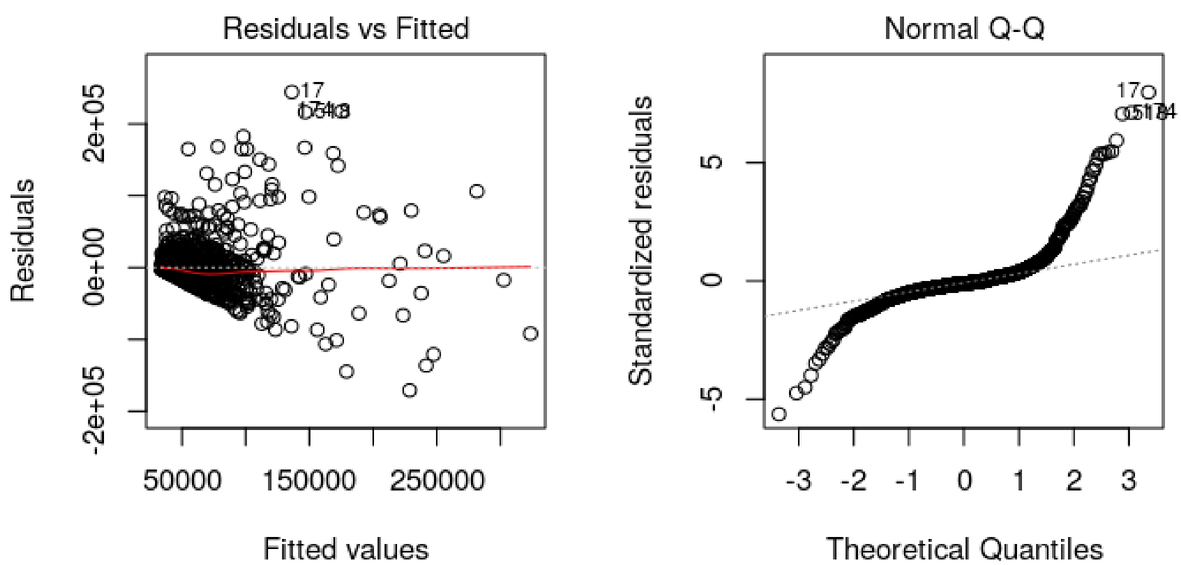


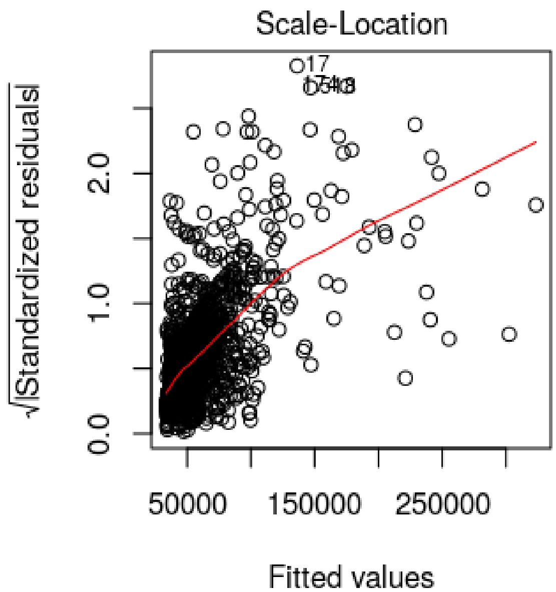

Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
0.5

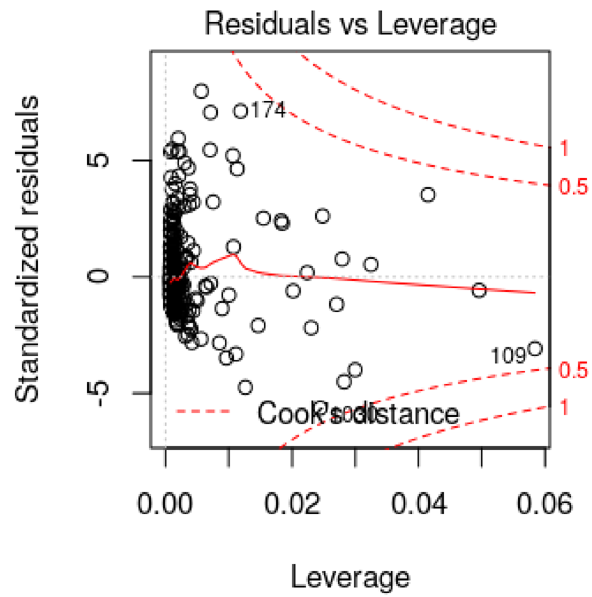

Log Predicted vs. Log Observed
log(observed)
log(predicted)

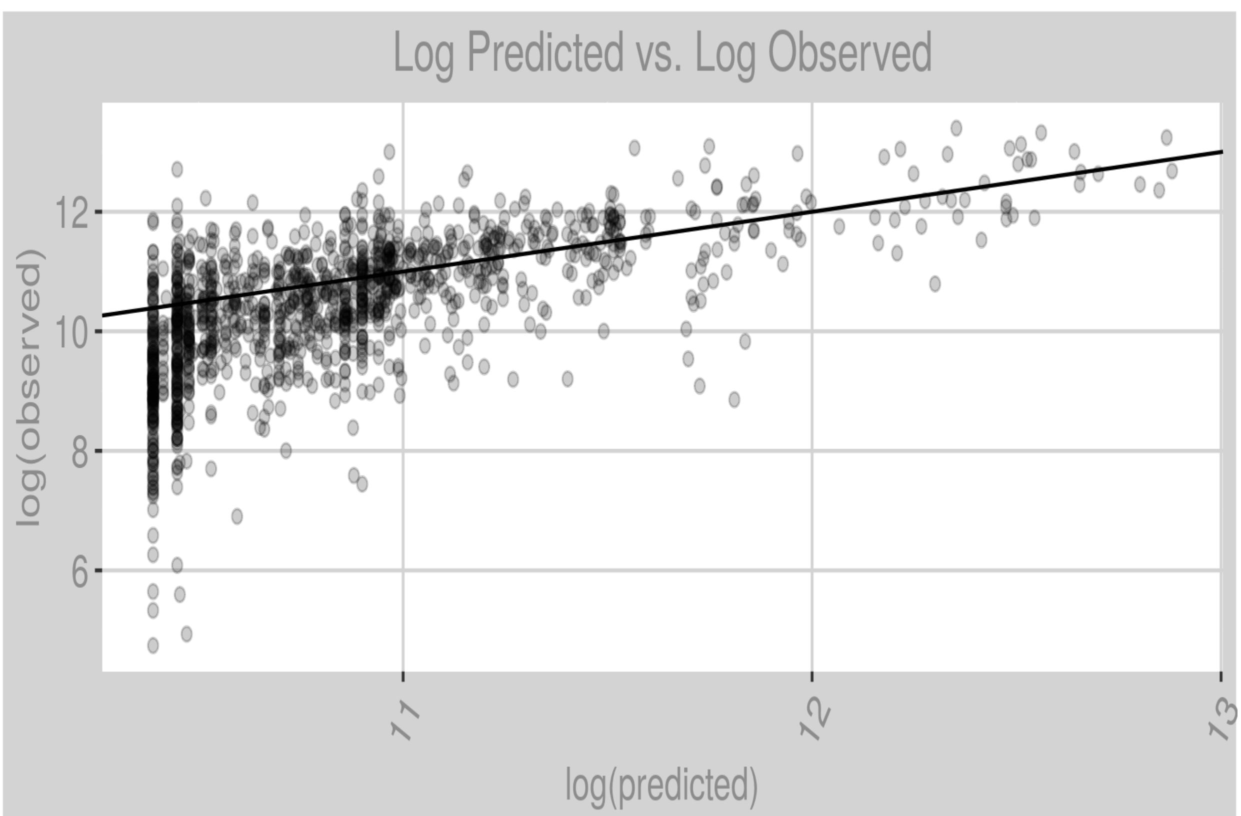
Recursive Feature Extraction



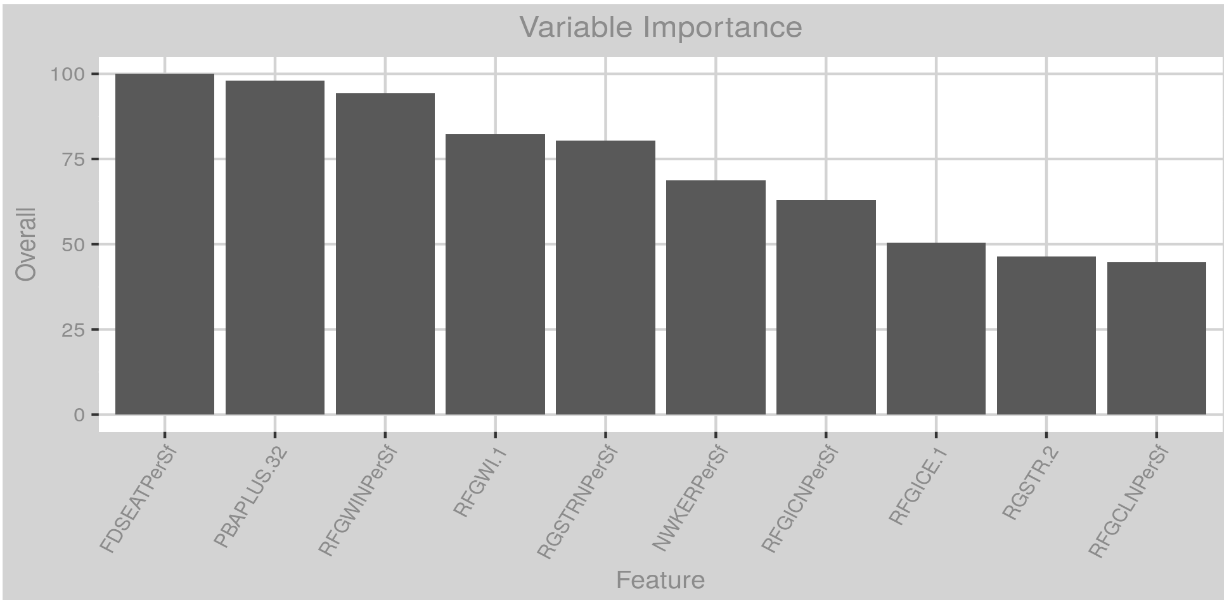


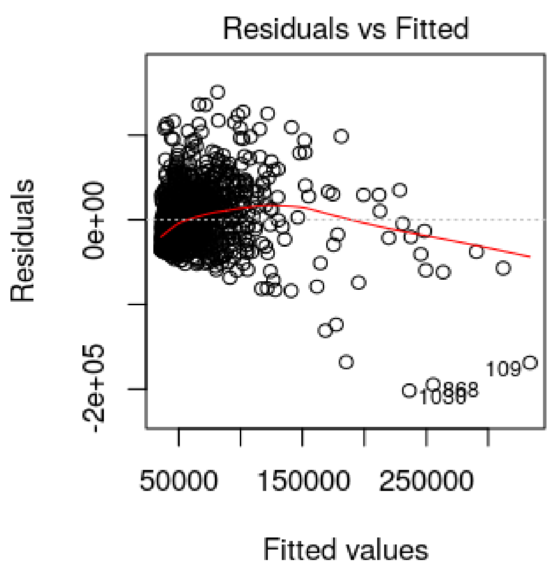

Scale-Location
√|Standardized residuals|
Fitted values
17
0.0
1.0
2.0
50000
150000
250000

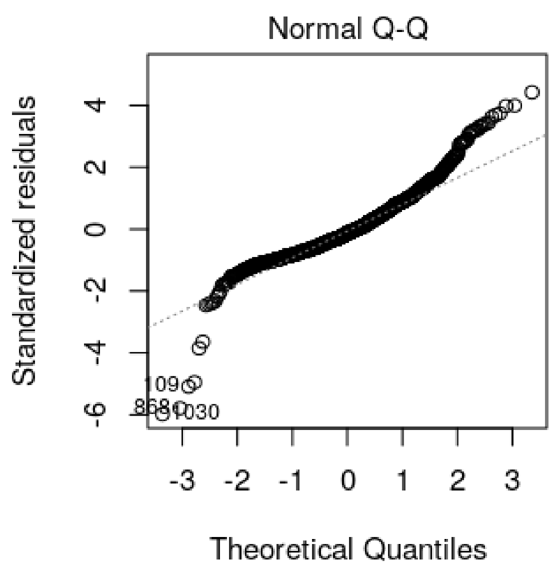

Residuals vs Leverage
Standardized residuals
Leverage
174
109
Cook's distance
1
0.5
0.5
1
-5
0
5
0.00
0.02
0.04
0.06

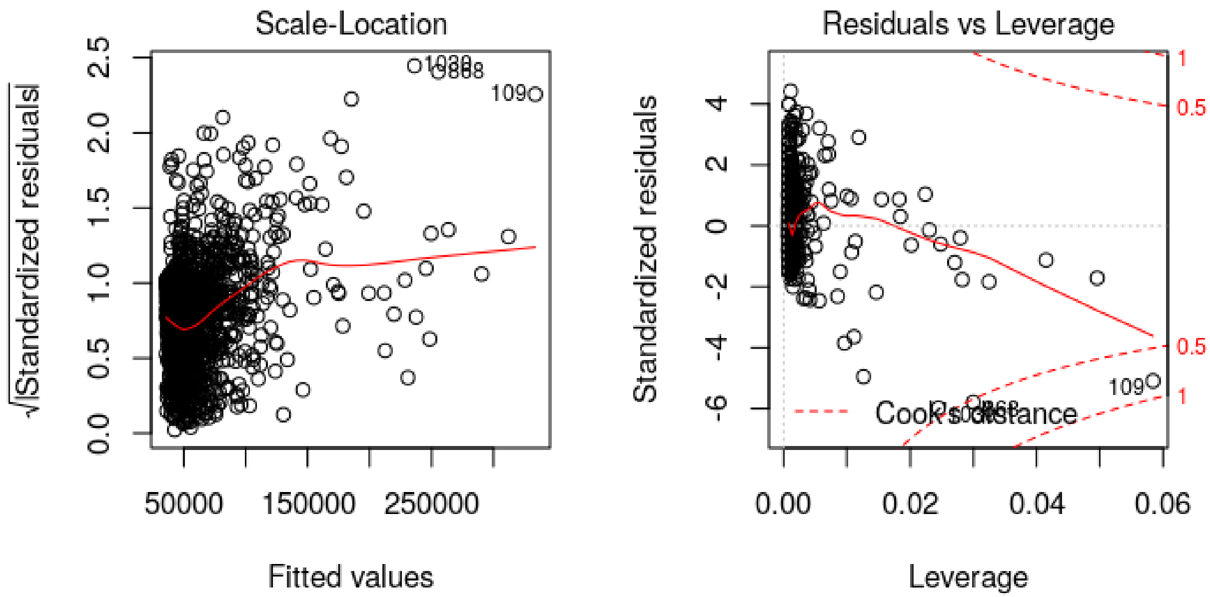

Log Predicted vs. Log Observed
log(observed)
log(predicted)
6
8
10
12
11
12
13

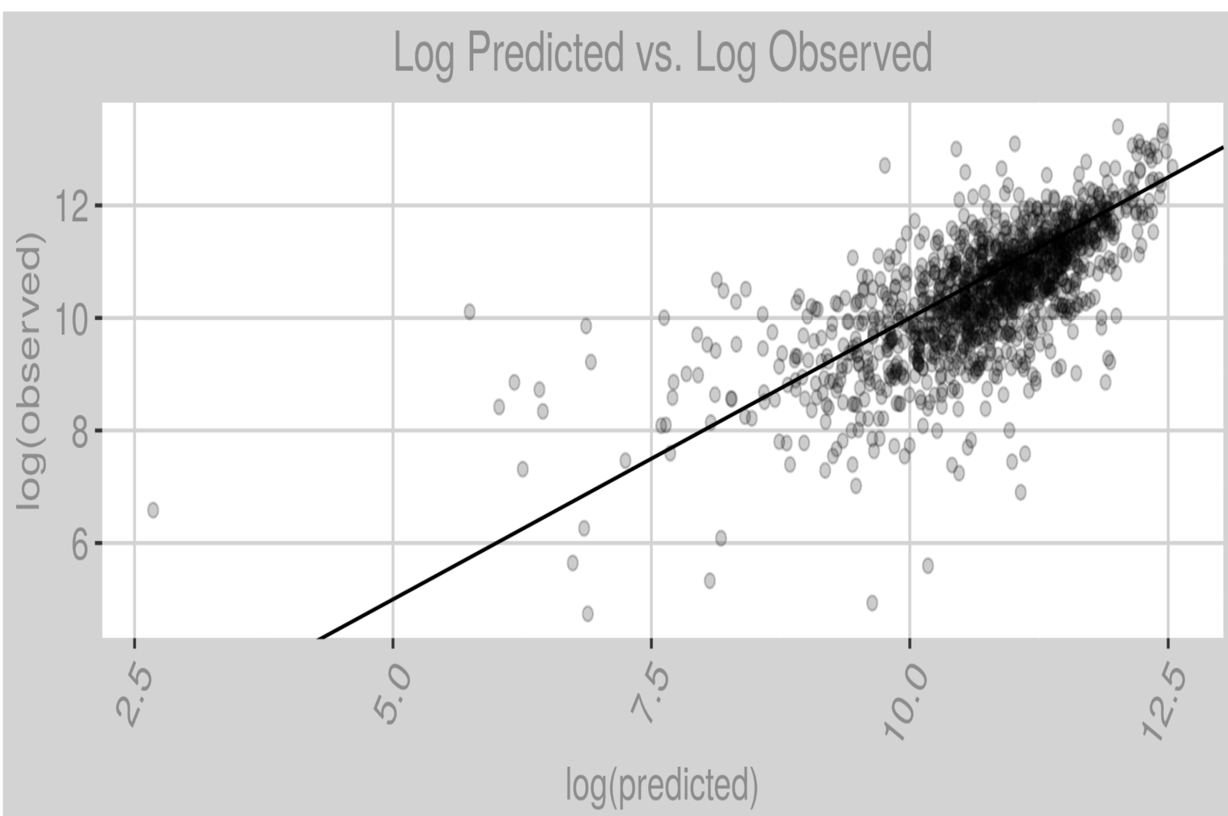
Simple Neural Network







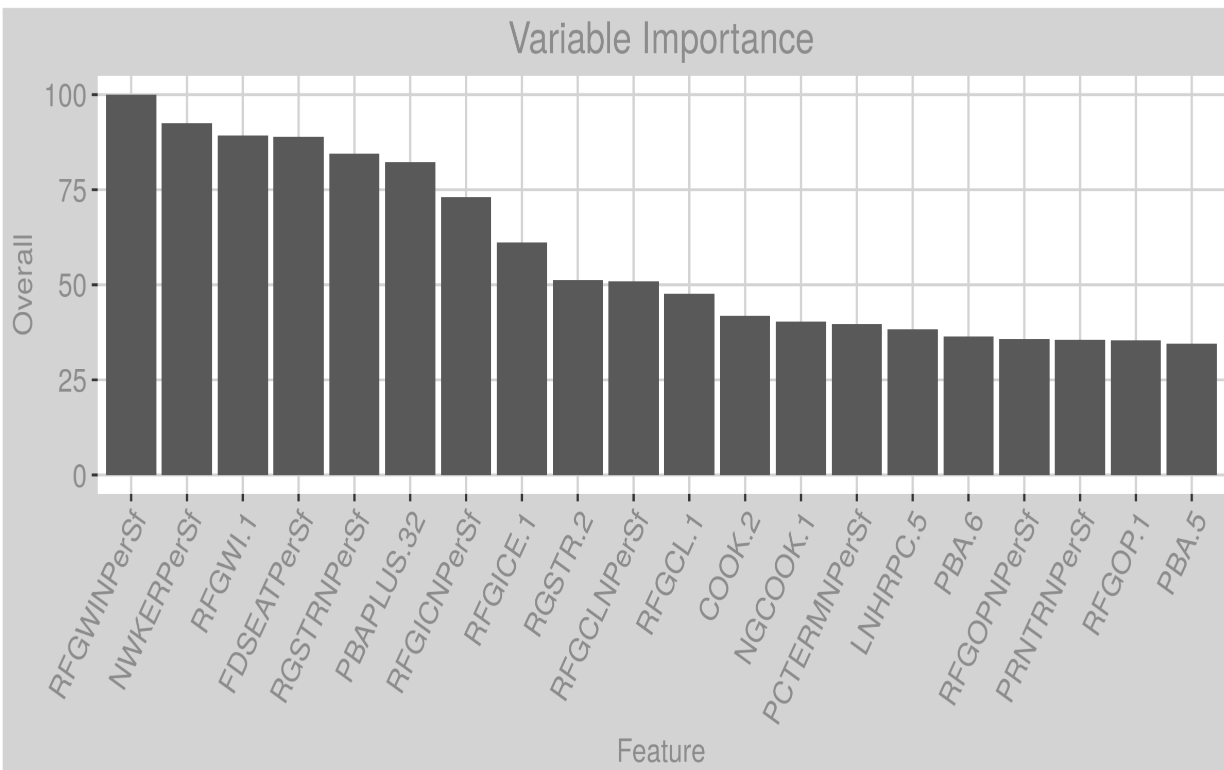

Scale-Location
√|Standardized residuals|
Fitted values
50000
150000
250000
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.00
0.02
0.04
0.06

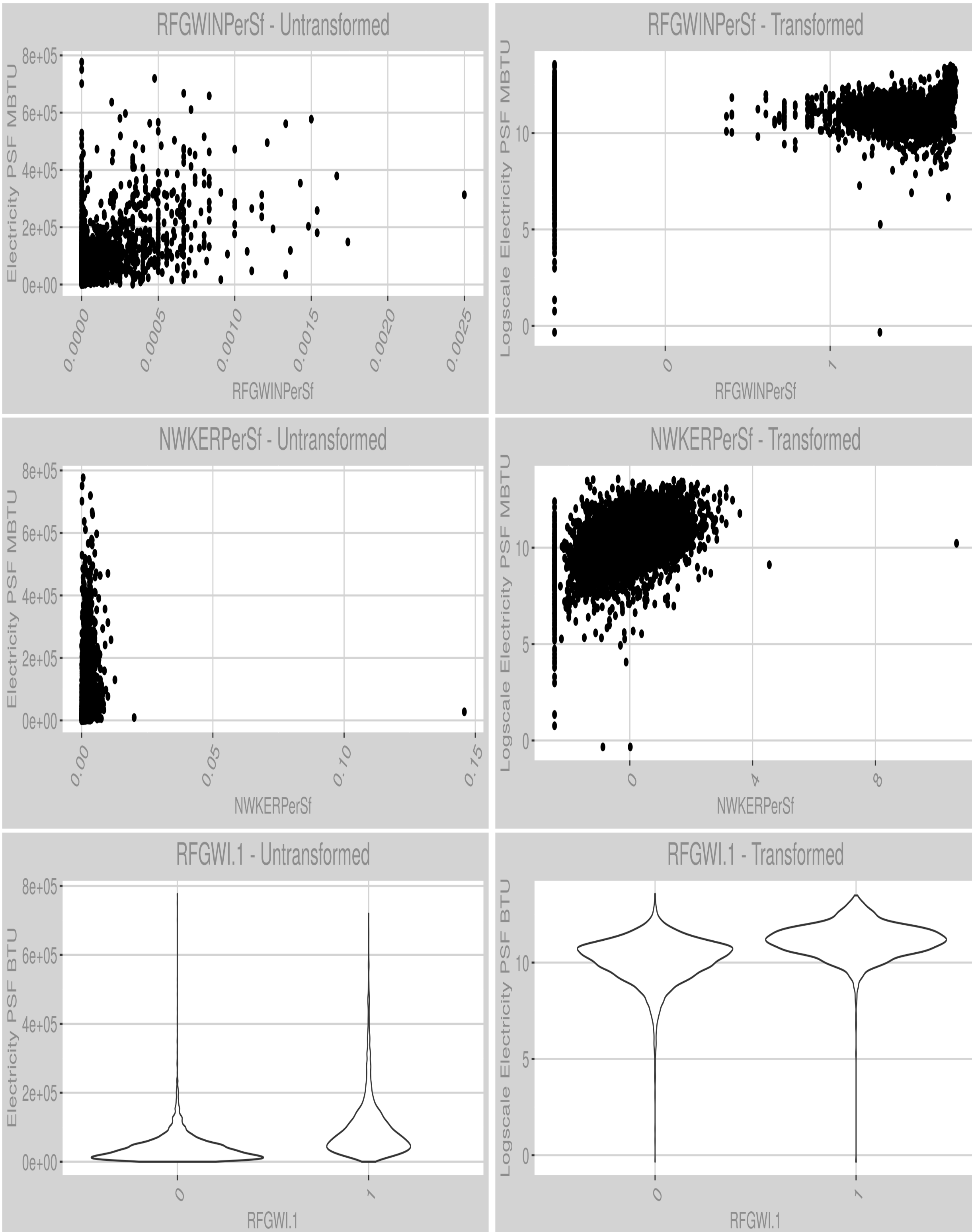

Log Predicted vs. Log Observed
log(observed)
log(predicted)
2.5
5.0
7.5
10.0
12.5

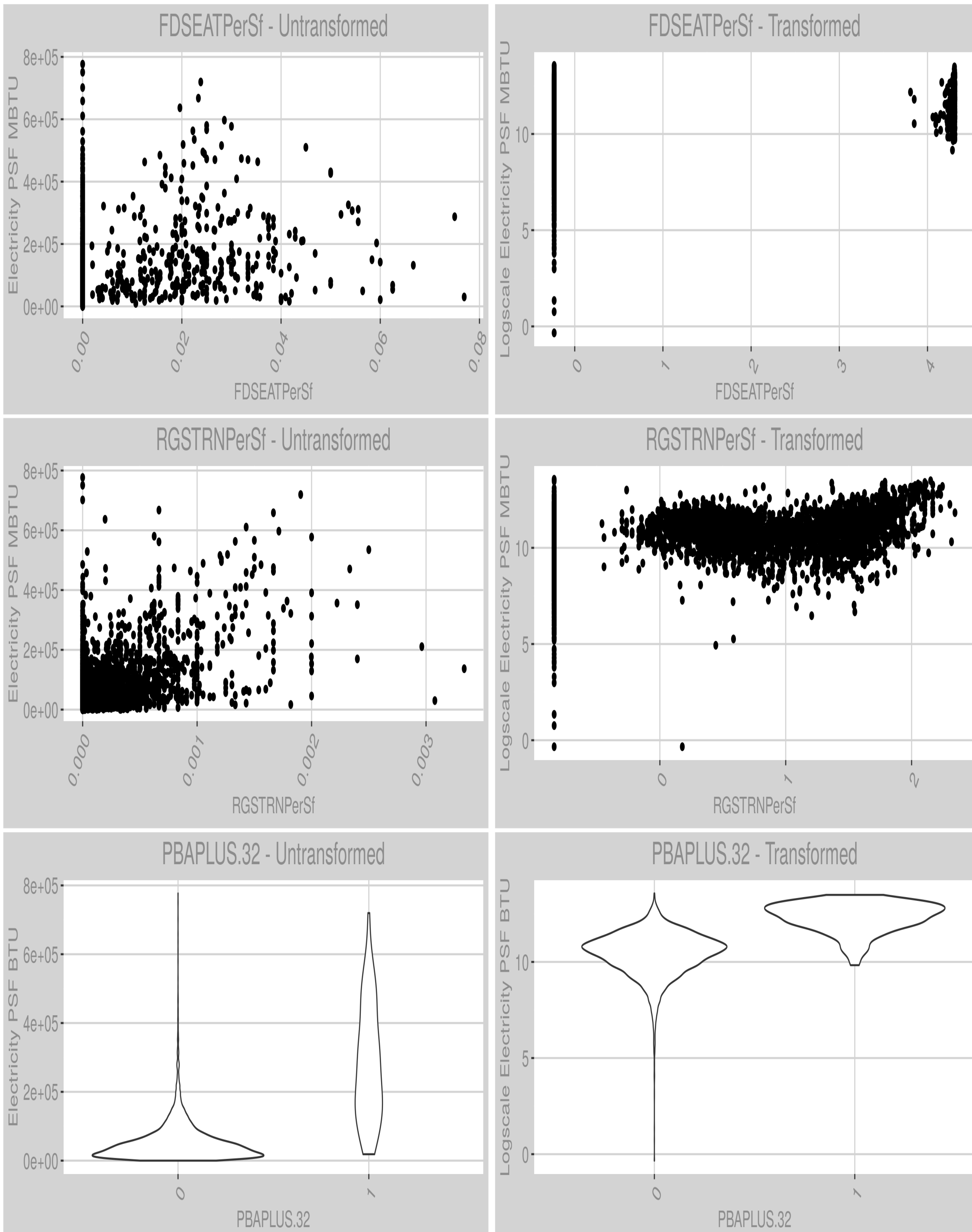
Select Variable Analysis

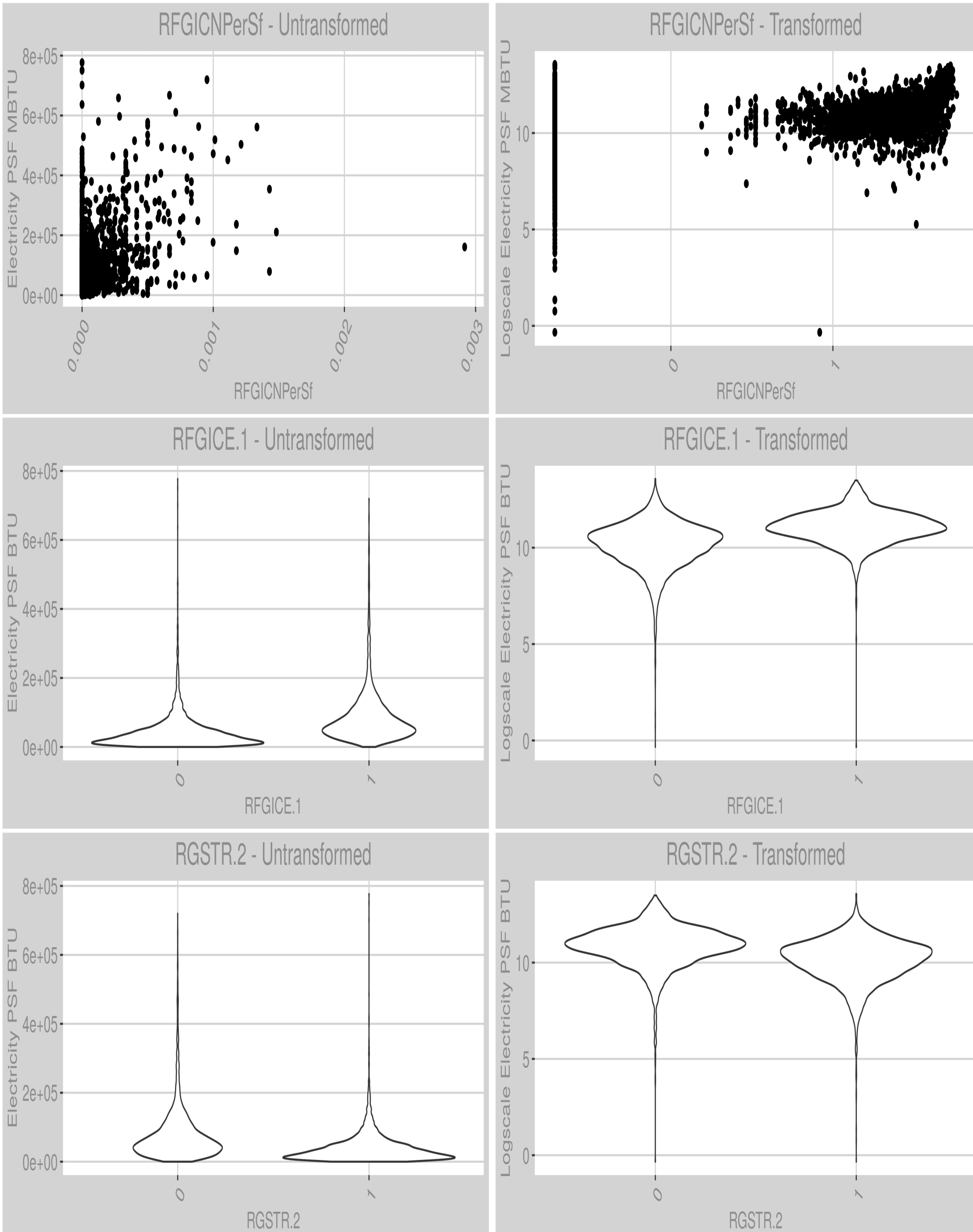
Feature Extraction Model Results

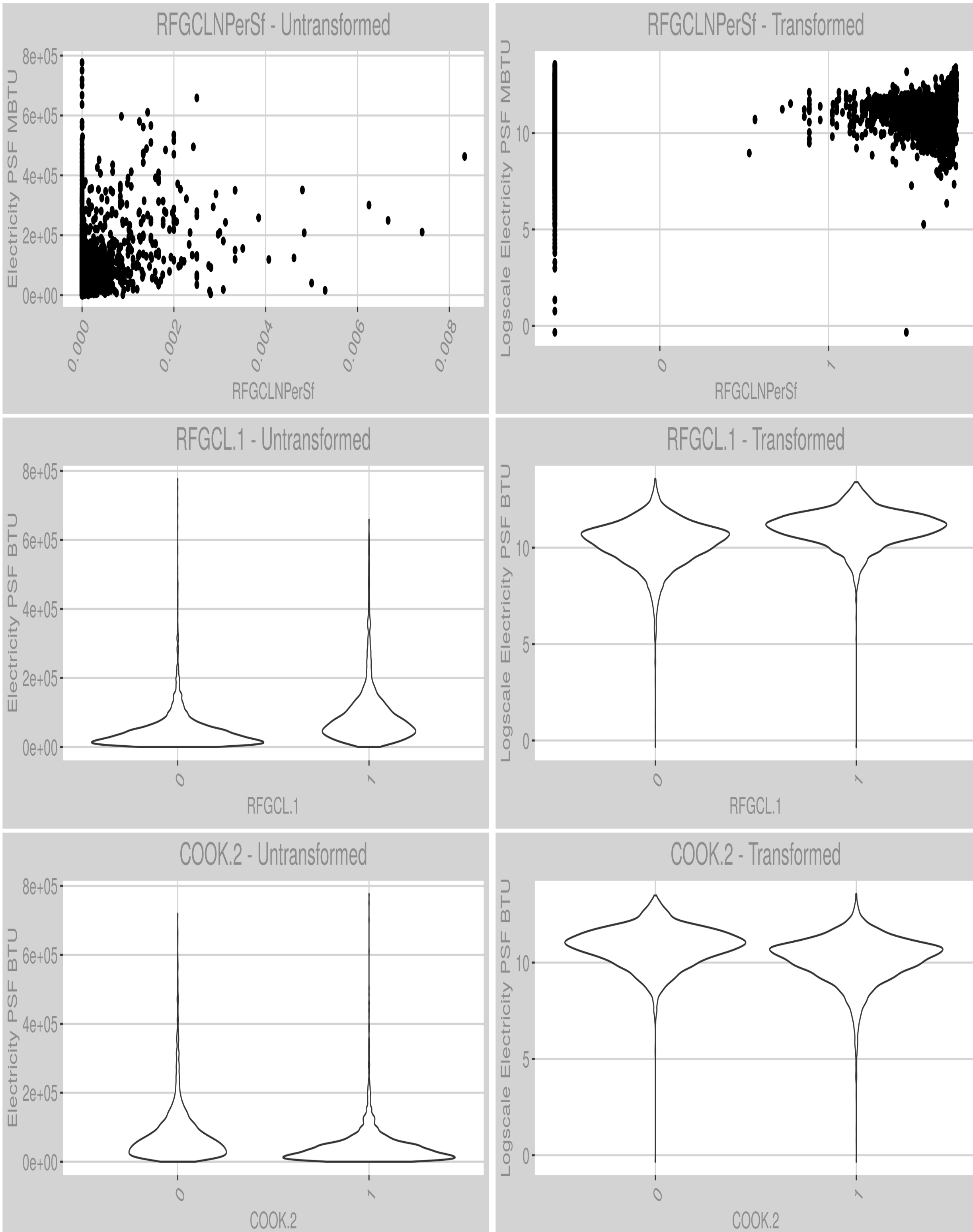
Model	RMSE	R2	MAE	MAPE	MSLE
partialLeastSquares	49439.16	0.5119599	26622.19	85.70943	0.5680023
neuralNetwork	51378.59	0.4622766	29989.65	133.42257	0.8016741
recursiveFeatureExtraction	49713.91	0.5017194	31111.93	239.40187	1.0286980
randomForest	165662.98	0.1388473	144344.47	881.85161	3.3488969
leaps	91256.75	0.3157366	58679.30	99.92406	66.6999846

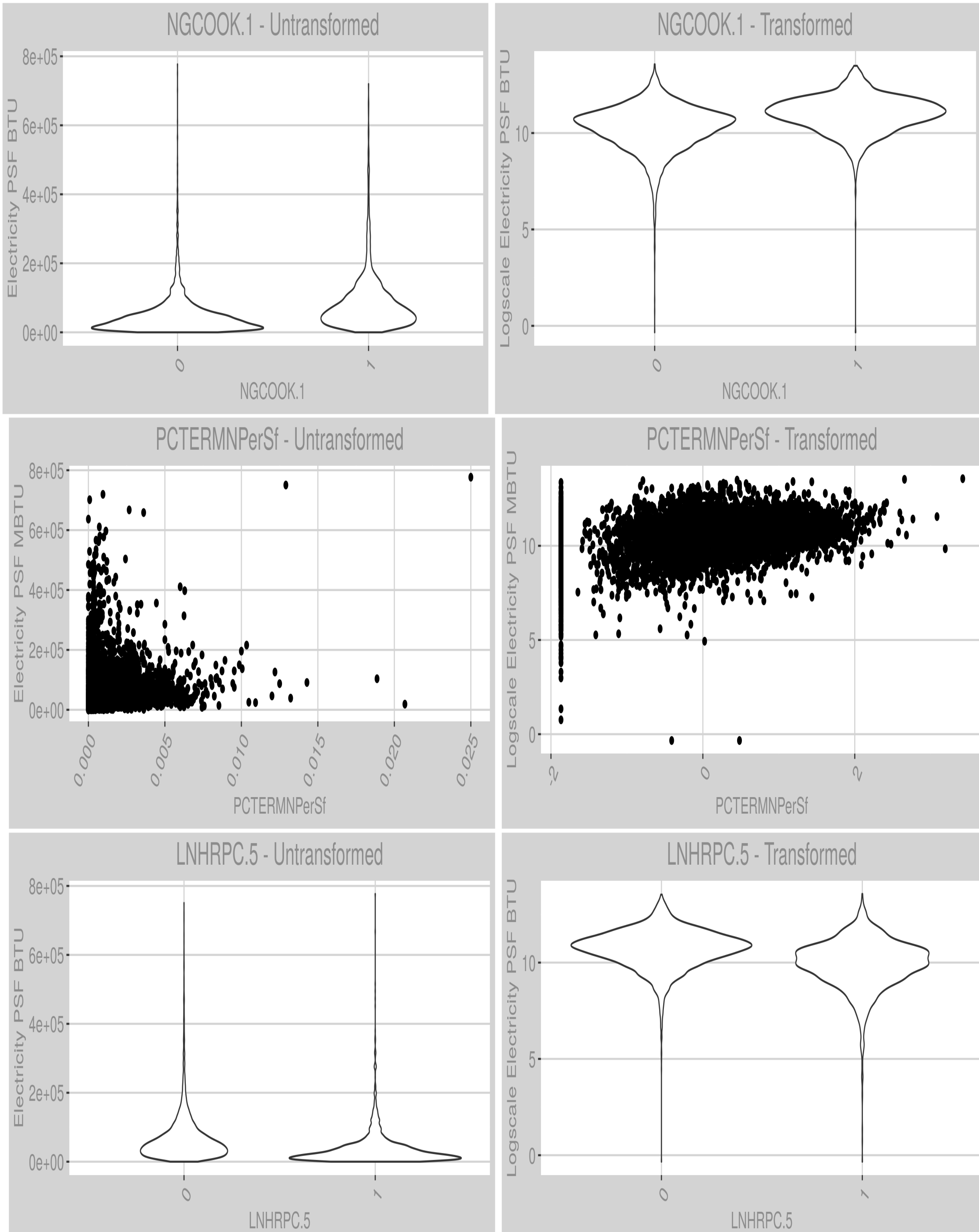


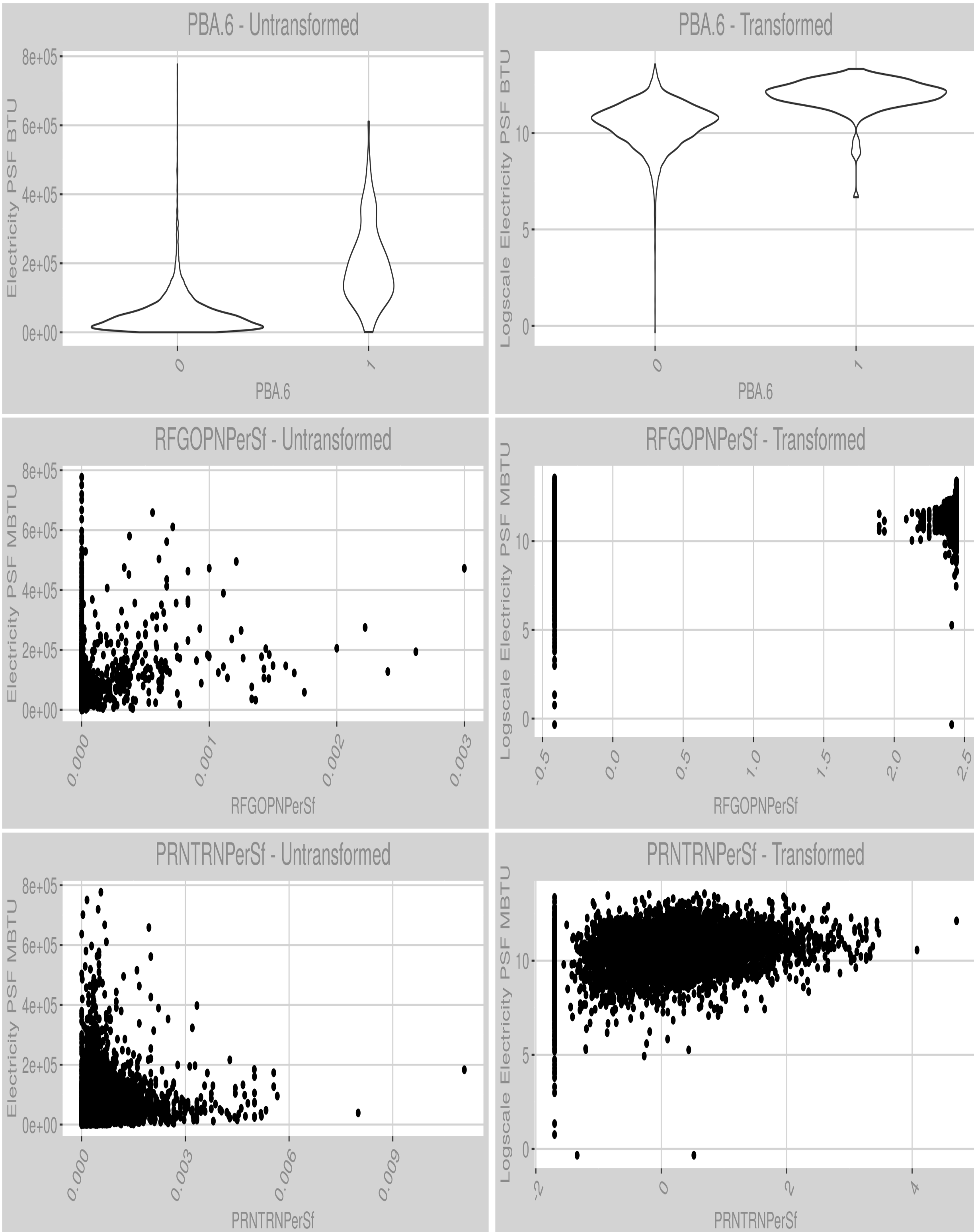

RFGWINPerSf - Untransformed
Electricity PSF MBTU
8e+05
6e+05
4e+05
2e+05
0e+00
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
RFGWINPerSf
RFGWINPerSf - Transformed
Logscale Electricity PSF MBTU
10
5
0
0
1
RFGWINPerSf
NWKERPerSf - Untransformed
Electricity PSF MBTU
8e+05
6e+05
4e+05
2e+05
0e+00
0.00
0.05
0.10
0.15
NWKERPerSf
NWKERPerSf - Transformed
Logscale Electricity PSF MBTU
10
5
0
0
4
8
NWKERPerSf
RFGWI.1 - Untransformed
Electricity PSF BTU
8e+05
6e+05
4e+05
2e+05
0e+00
0
1
RFGWI.1
RFGWI.1 - Transformed
Logscale Electricity PSF BTU
10
5
0
0
1
RFGWI.1

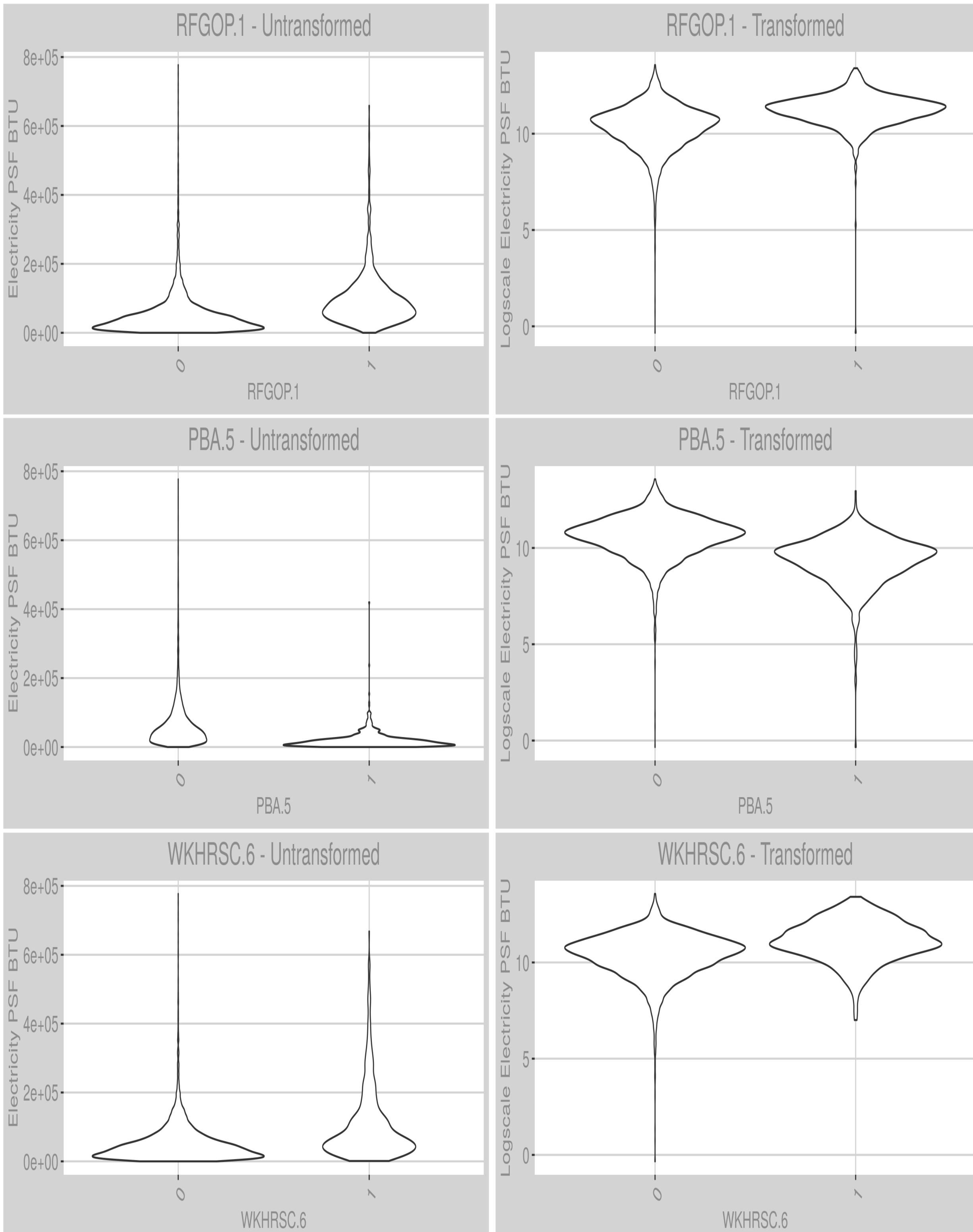

FDSEATPerSf - Untransformed
Electricity PSF MBTU
FDSEATPerSf
FDSEATPerSf - Transformed
Logscale Electricity PSF MBTU
FDSEATPerSf
RGSTRNPerSf - Untransformed
Electricity PSF MBTU
RGSTRNPerSf
RGSTRNPerSf - Transformed
Logscale Electricity PSF MBTU
RGSTRNPerSf
PBAPLUS.32 - Untransformed
Electricity PSF BTU
PBAPLUS.32
PBAPLUS.32 - Transformed
Logscale Electricity PSF BTU
PBAPLUS.32


RFGICNPerSf - Untransformed
Electricity PSF MBTU
8e+05
6e+05
4e+05
2e+05
0e+00
0.000
0.001
0.002
0.003
RFGICNPerSf
RFGICNPerSf - Transformed
Logscale Electricity PSF MBTU
10
5
0
0
1
RFGICNPerSf
RFGICE.1 - Untransformed
Electricity PSF BTU
0
1
RFGICE.1
RFGICE.1 - Transformed
Logscale Electricity PSF BTU
RFGICE.1
RGSTR.2 - Untransformed
Electricity PSF BTU
RGSTR.2
RGSTR.2 - Transformed
Logscale Electricity PSF BTU
RGSTR.2


RFGCLNPerSf - Untransformed
Electricity PSF MBTU
8e+05
6e+05
4e+05
2e+05
0e+00
0.000
0.002
0.004
0.006
0.008
RFGCLNPerSf
RFGCLNPerSf - Transformed
Logscale Electricity PSF MBTU
10
5
0
0
1
RFGCLNPerSf
RFGCL.1 - Untransformed
Electricity PSF BTU
0
1
RFGCL.1
RFGCL.1 - Transformed
Logscale Electricity PSF BTU
RFGCL.1
COOK.2 - Untransformed
Electricity PSF BTU
COOK.2
COOK.2 - Transformed
Logscale Electricity PSF BTU
COOK.2

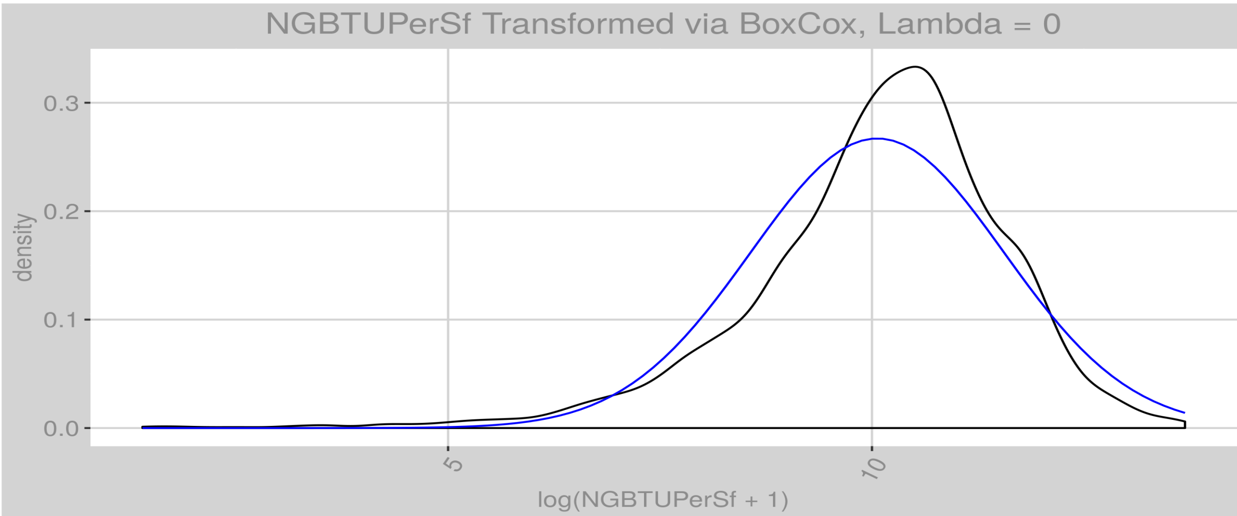

NGCOOK.1 - Untransformed
Electricity PSF BTU
8e+05
6e+05
4e+05
2e+05
0e+00
0
1
NGCOOK.1
NGCOOK.1 - Transformed
Logscale Electricity PSF BTU
10
5
0
0
1
NGCOOK.1
PCTERMNPerSf - Untransformed
Electricity PSF MBTU
8e+05
6e+05
4e+05
2e+05
0e+00
0.000
0.005
0.010
0.015
0.020
0.025
PCTERMNPerSf
PCTERMNPerSf - Transformed
Logscale Electricity PSF MBTU
10
5
0
-2
0
2
PCTERMNPerSf
LNHRPC.5 - Untransformed
Electricity PSF BTU
8e+05
6e+05
4e+05
2e+05
0e+00
0
1
LNHRPC.5
LNHRPC.5 - Transformed
Logscale Electricity PSF BTU
10
5
0
0
1
LNHRPC.5


PBA.6 - Untransformed
PBA.6 - Transformed
Electricity PSF BTU
Logscale Electricity PSF BTU
PBA.6
8e+05
6e+05
4e+05
2e+05
0e+00
10
5
0
RFGOPNPerSf - Untransformed
RFGOPNPerSf - Transformed
Electricity PSF MBTU
Logscale Electricity PSF MBTU
RFGOPNPerSf
0.000
0.001
0.002
0.003
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
PRNTRNPerSf - Untransformed
PRNTRNPerSf - Transformed
PRNTRNPerSf
0.003
0.006
0.009
-2
2
4

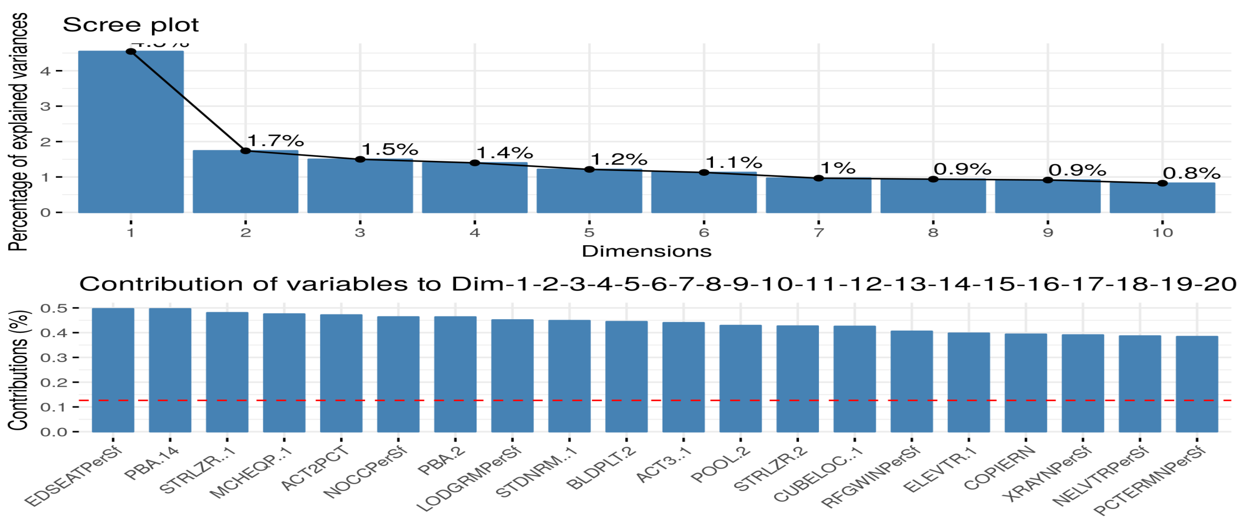

RFGOP.1 - Untransformed
Electricity PSF BTU
8e+05
6e+05
4e+05
2e+05
0e+00
0
1
RFGOP.1
RFGOP.1 - Transformed
Logscale Electricity PSF BTU
10
5
0
0
1
RFGOP.1
PBA.5 - Untransformed
Electricity PSF BTU
8e+05
6e+05
4e+05
2e+05
0e+00
0
1
PBA.5
PBA.5 - Transformed
Logscale Electricity PSF BTU
10
5
0
0
1
PBA.5
WKHRSC.6 - Untransformed
Electricity PSF BTU
8e+05
6e+05
4e+05
2e+05
0e+00
0
1
WKHRSC.6
WKHRSC.6 - Transformed
Logscale Electricity PSF BTU
10
5
0
0
1
WKHRSC.6

Appendix - Natural Gas

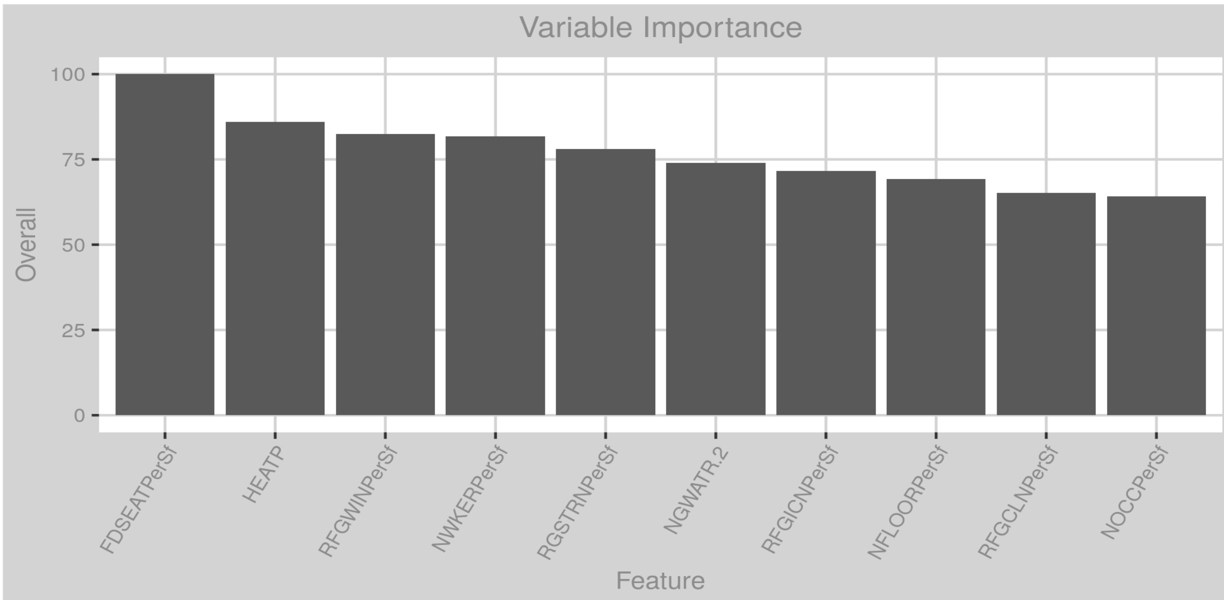
Response

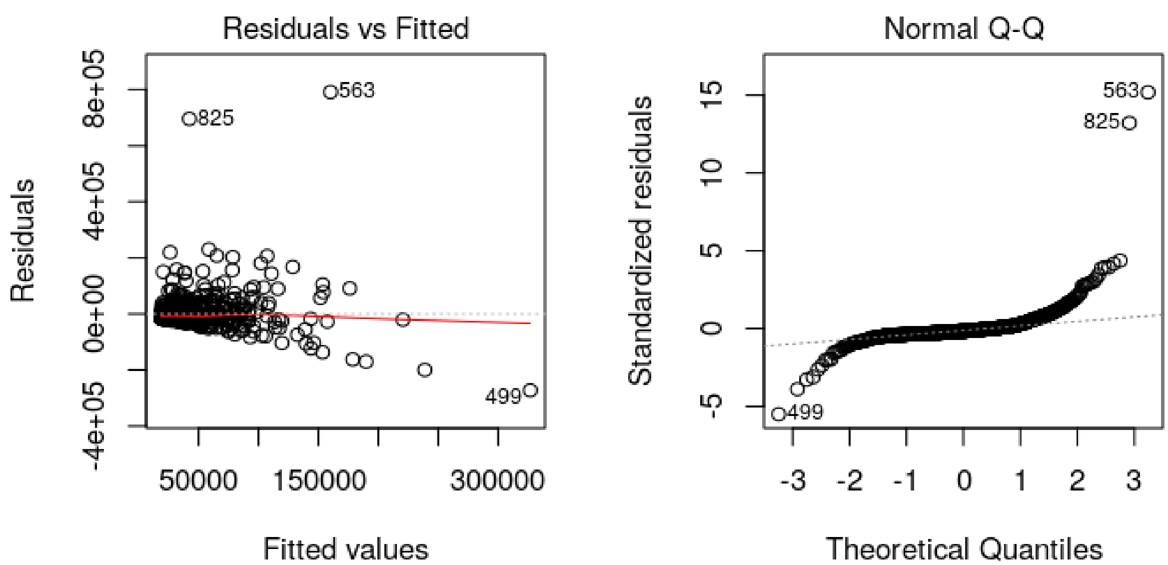


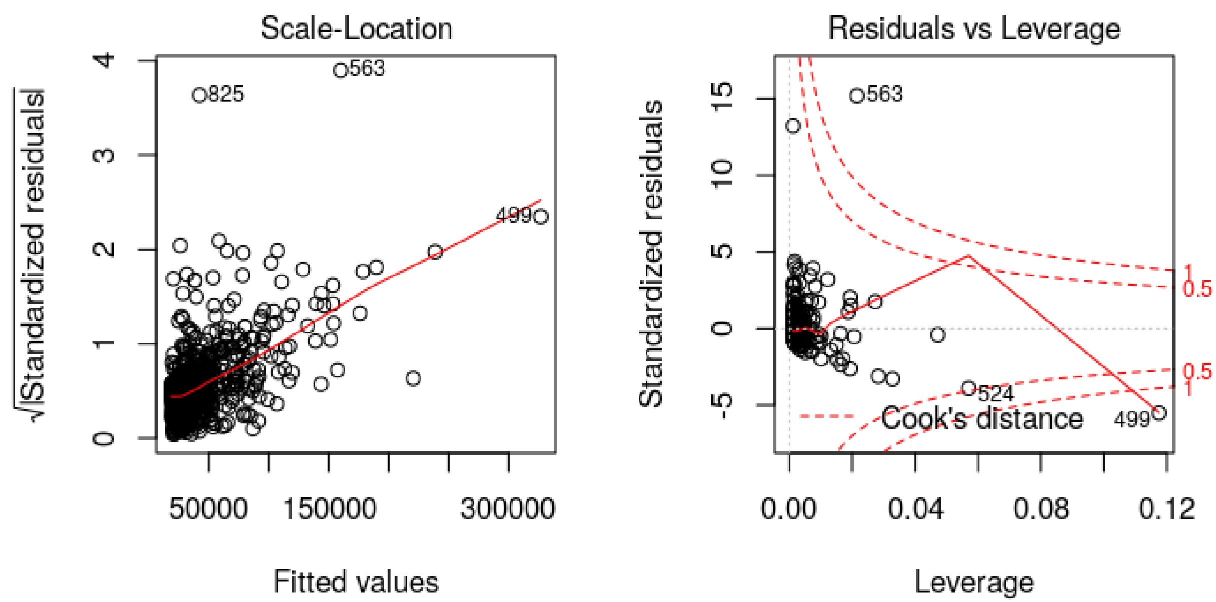
PCA

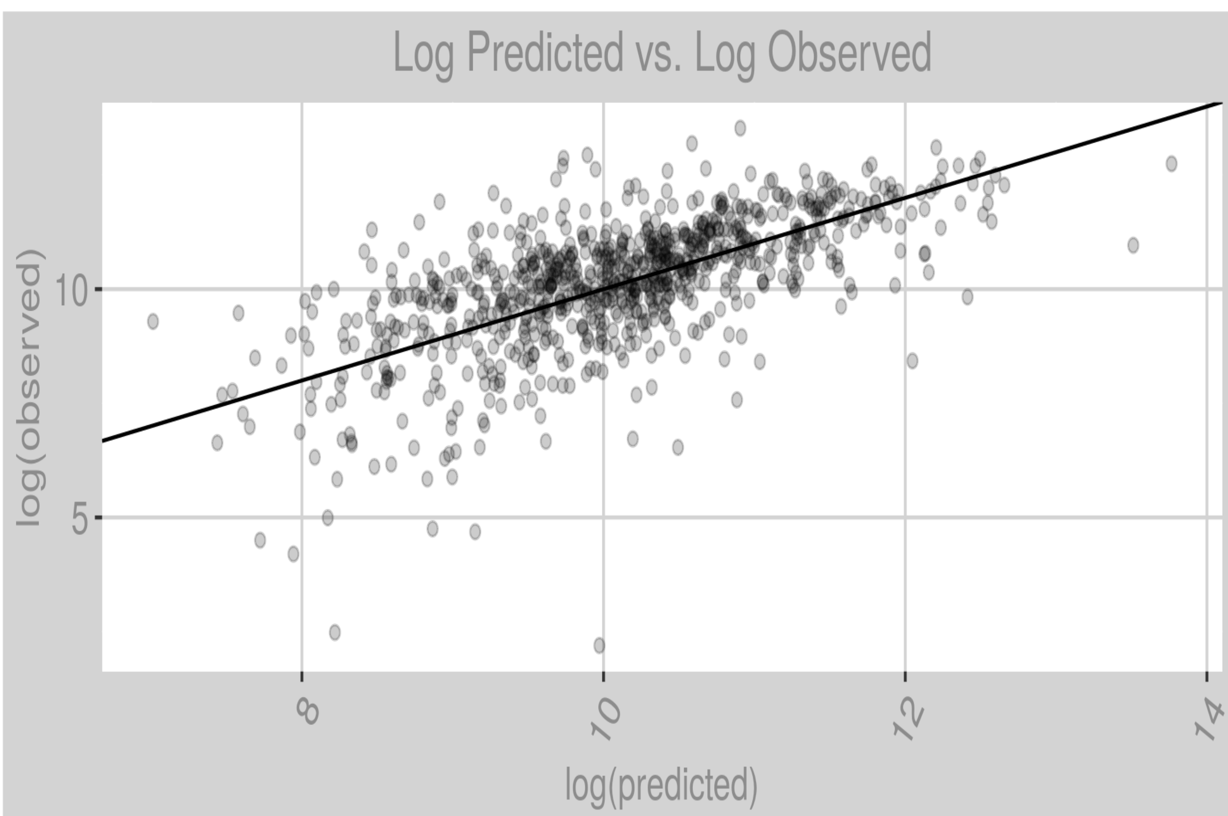


PLS

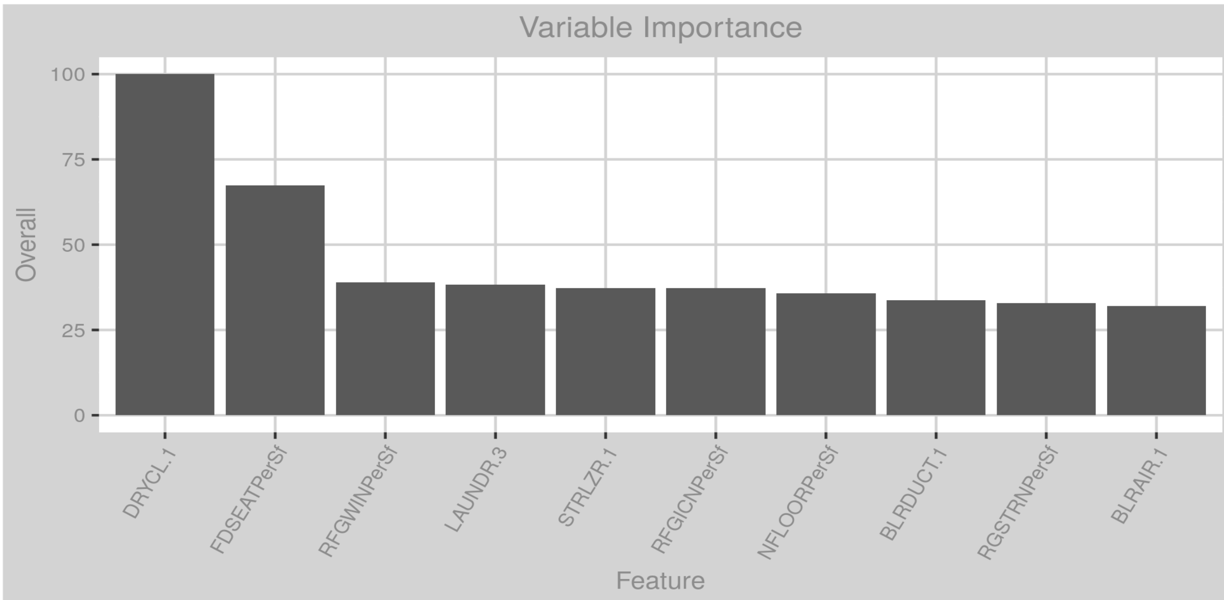

Variable Importance
Overall
100
75
50
25
0
FDSEATPerSf
HEATP
RFGWINPerSf
NWKERPerSf
RGSTRNPerSf
NGWATR.2
RFGICNPerSf
NFLOORPerSf
RFGCLNPerSf
NOCCPerSf
Feature

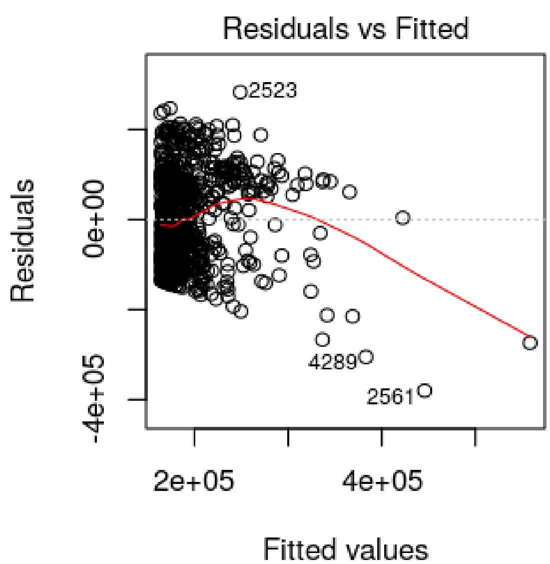

Residuals vs Fitted
Residuals
Fitted values
825
563
499
-4e+05
0e+00
4e+05
8e+05
50000
150000
300000
Normal Q-Q
Standardized residuals
Theoretical Quantiles
563
825
499
-5
0
5
10
15
-3
-2
-1
0
1
2
3

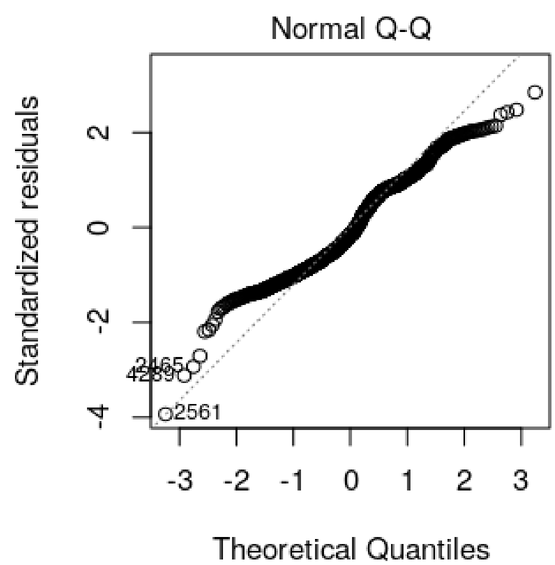

Scale-Location
√|Standardized residuals|
Fitted values
0
1
2
3
4
50000
150000
300000
825
563
499
Residuals vs Leverage
Standardized residuals
Leverage
-5
0
5
10
15
0.00
0.04
0.08
0.12
563
524
499
Cook's distance
1
0.5
0.5
1

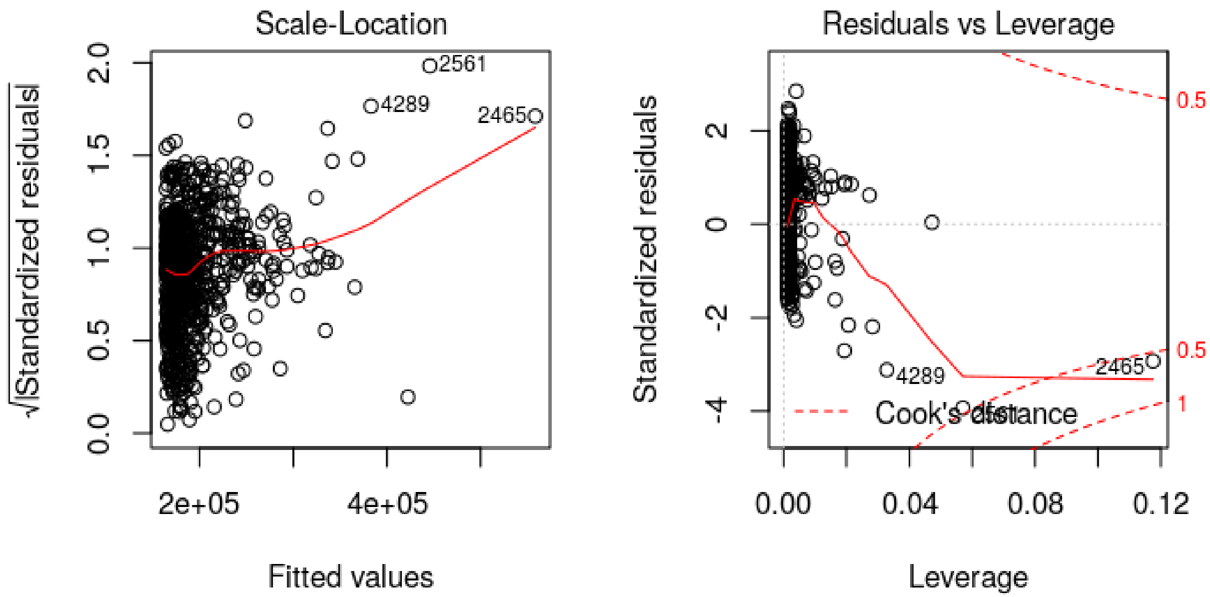

Log Predicted vs. Log Observed
log(observed)
log(predicted)
5
10
8
10
12
14

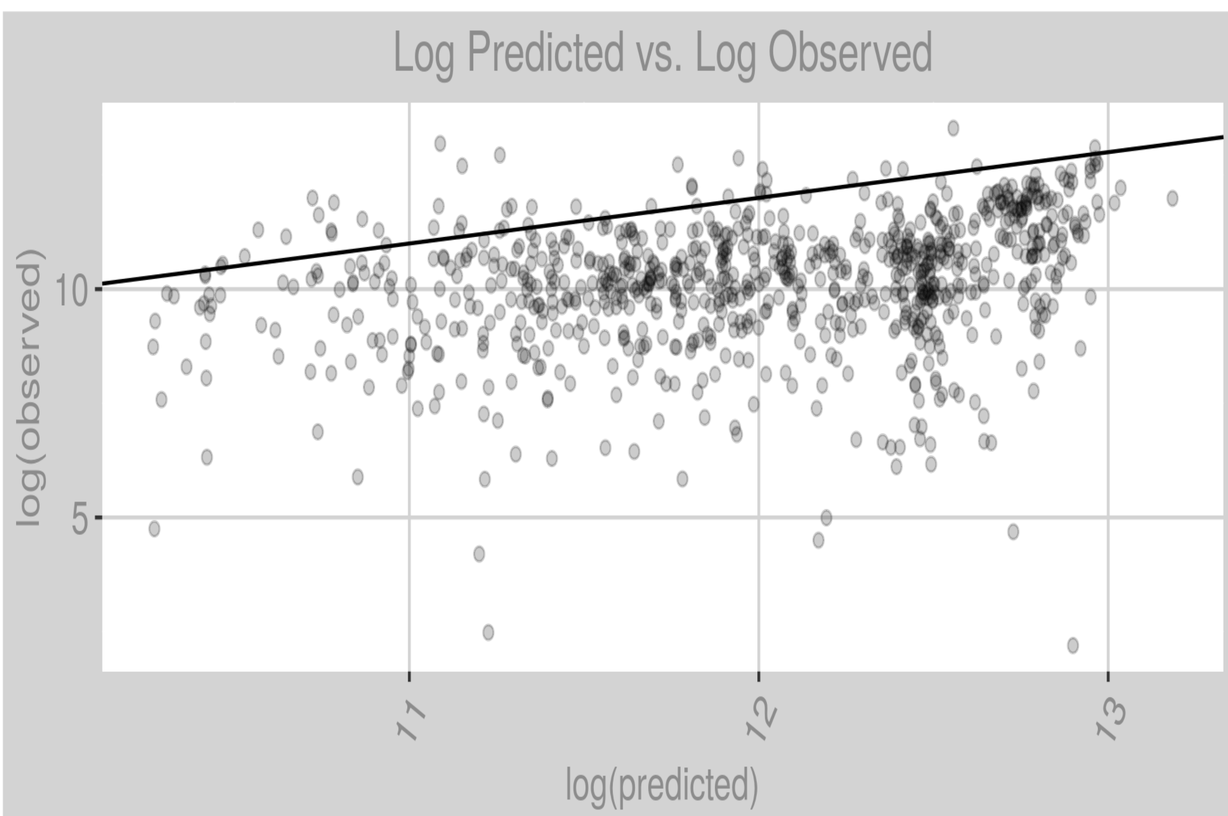
Random Forest



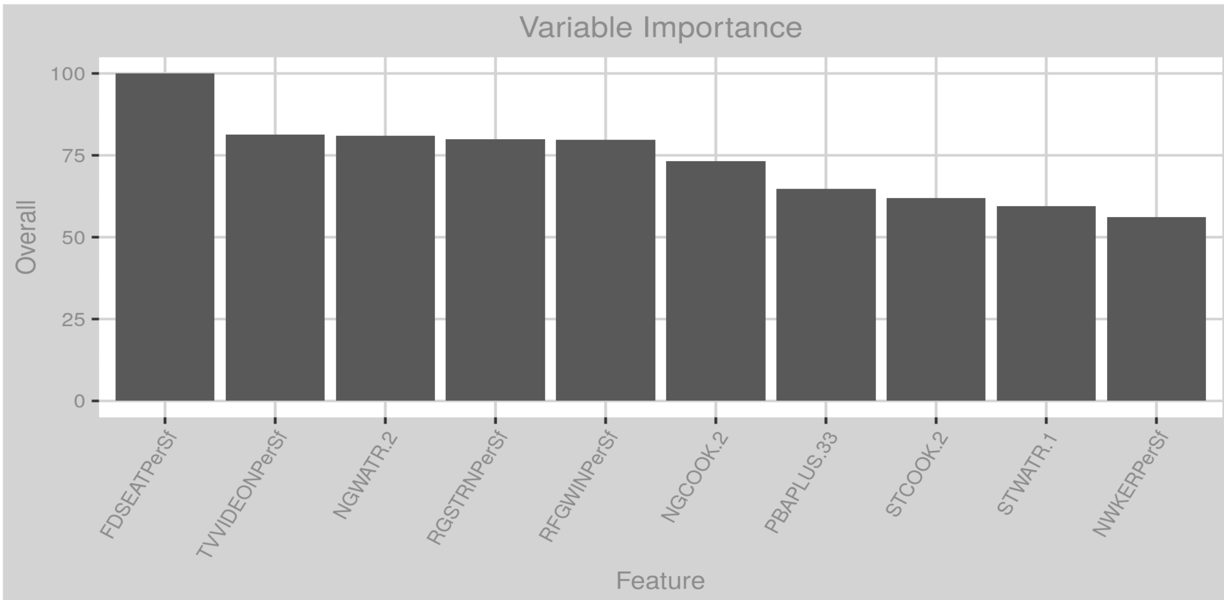


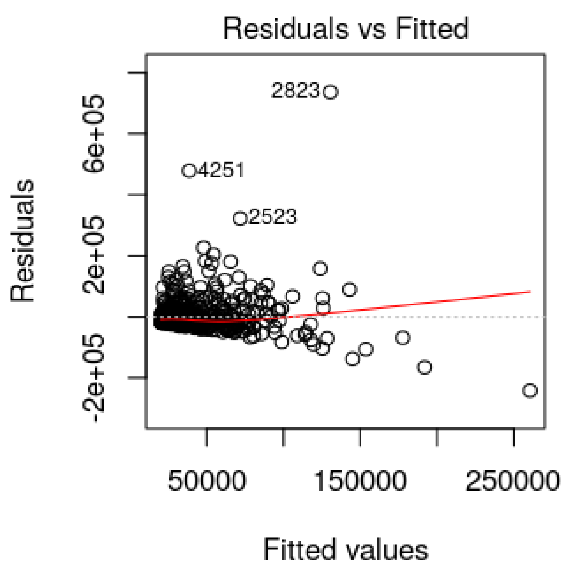


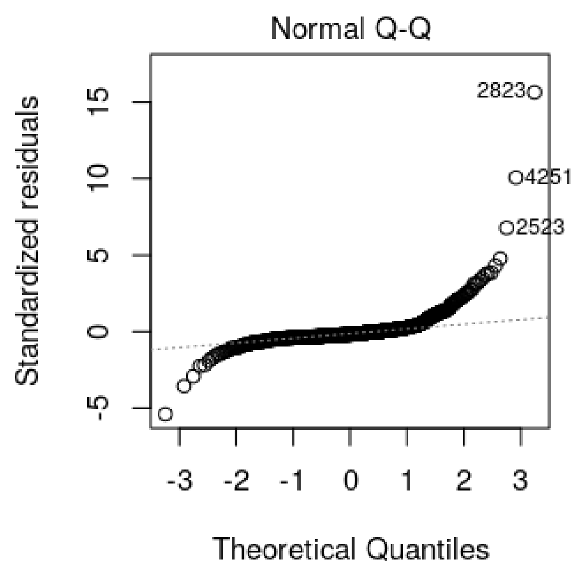

Scale-Location
√|Standardized residuals|
Fitted values
2e+05
4e+05
0.0
0.5
1.0
1.5
2.0
2561
4289
2465
Residuals vs Leverage
Standardized residuals
Leverage
0.00
0.04
0.08
0.12
-4
-2
0
2
0.5
0.5
1
4289
2465
Cook's distance

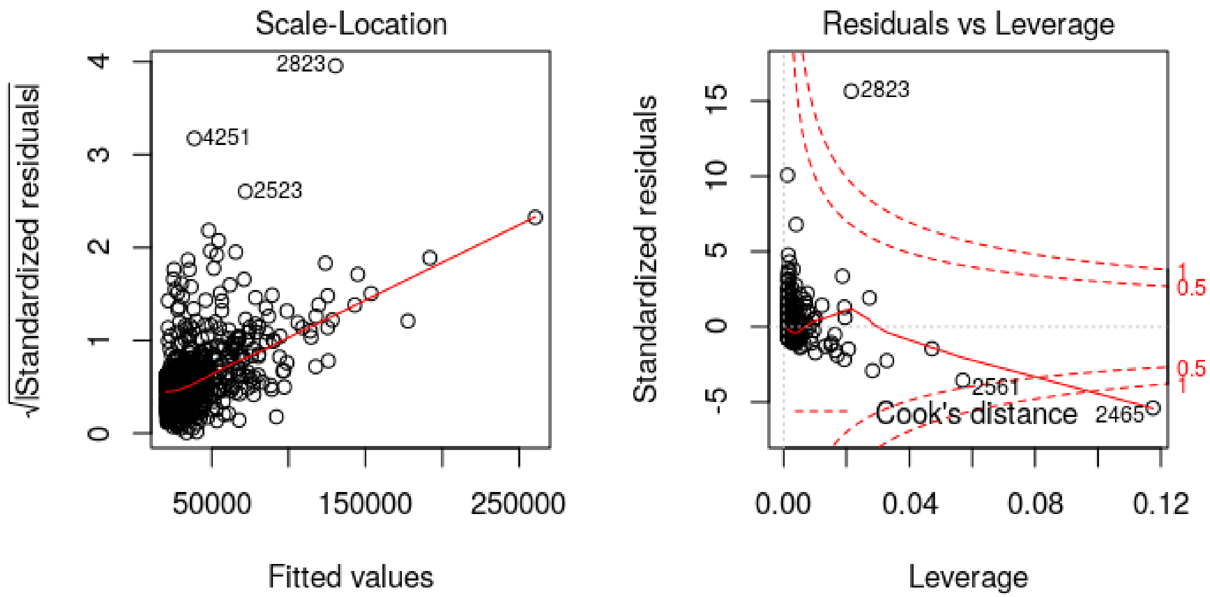

Log Predicted vs. Log Observed
log(observed)
log(predicted)
5
10
11
12
13

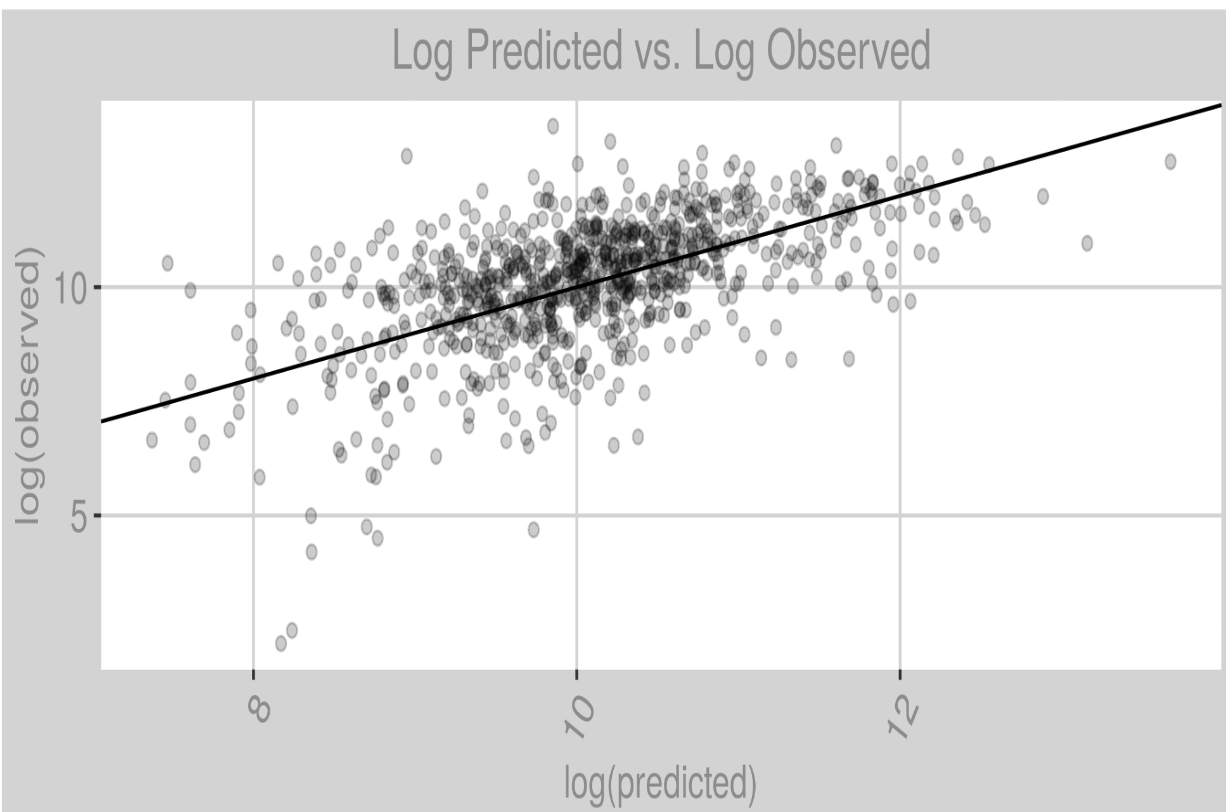
Forward Selection



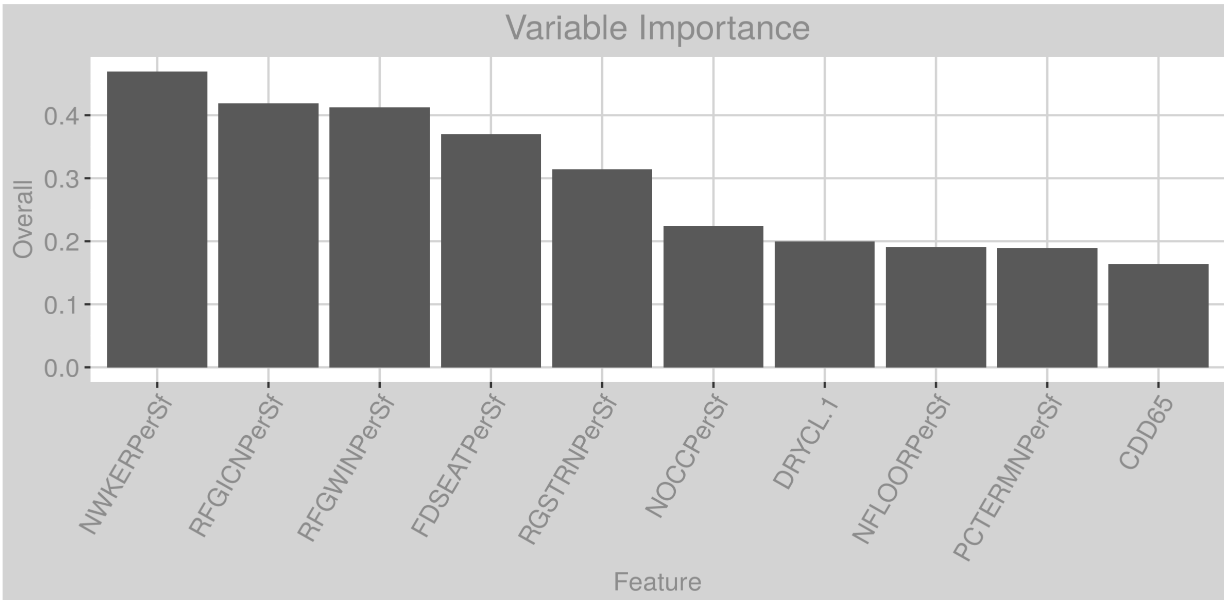


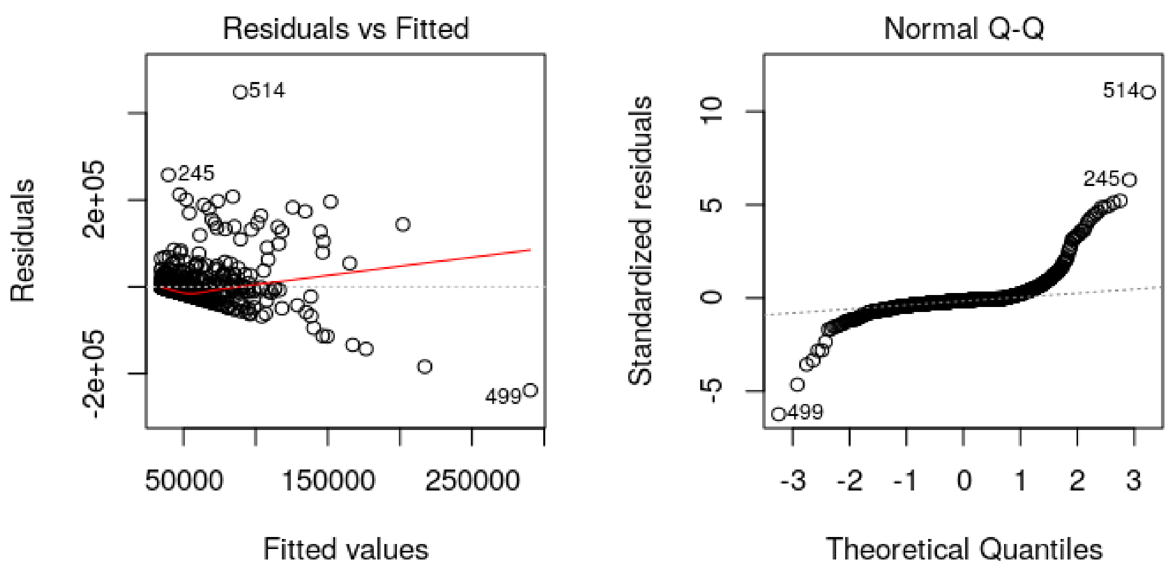


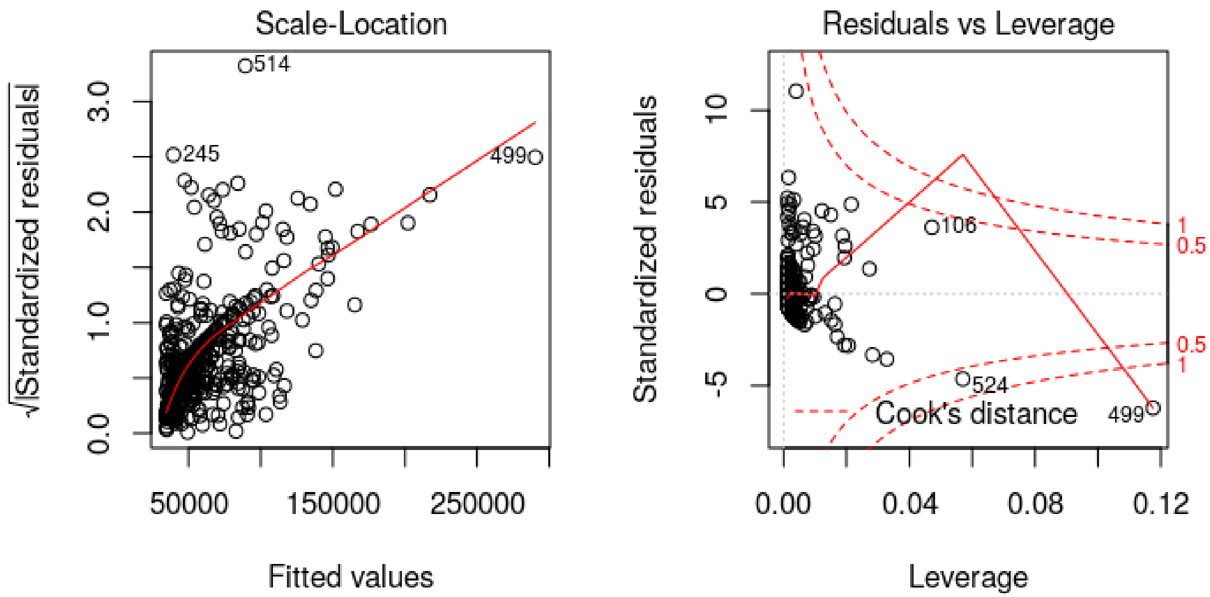

Scale-Location
$\sqrt{|\text{Standardized residuals}|}$
Fitted values
2823
4251
2523
Residuals vs Leverage
Standardized residuals
Leverage
2823
2561
2465
Cook's distance
1
0.5
0.5
1

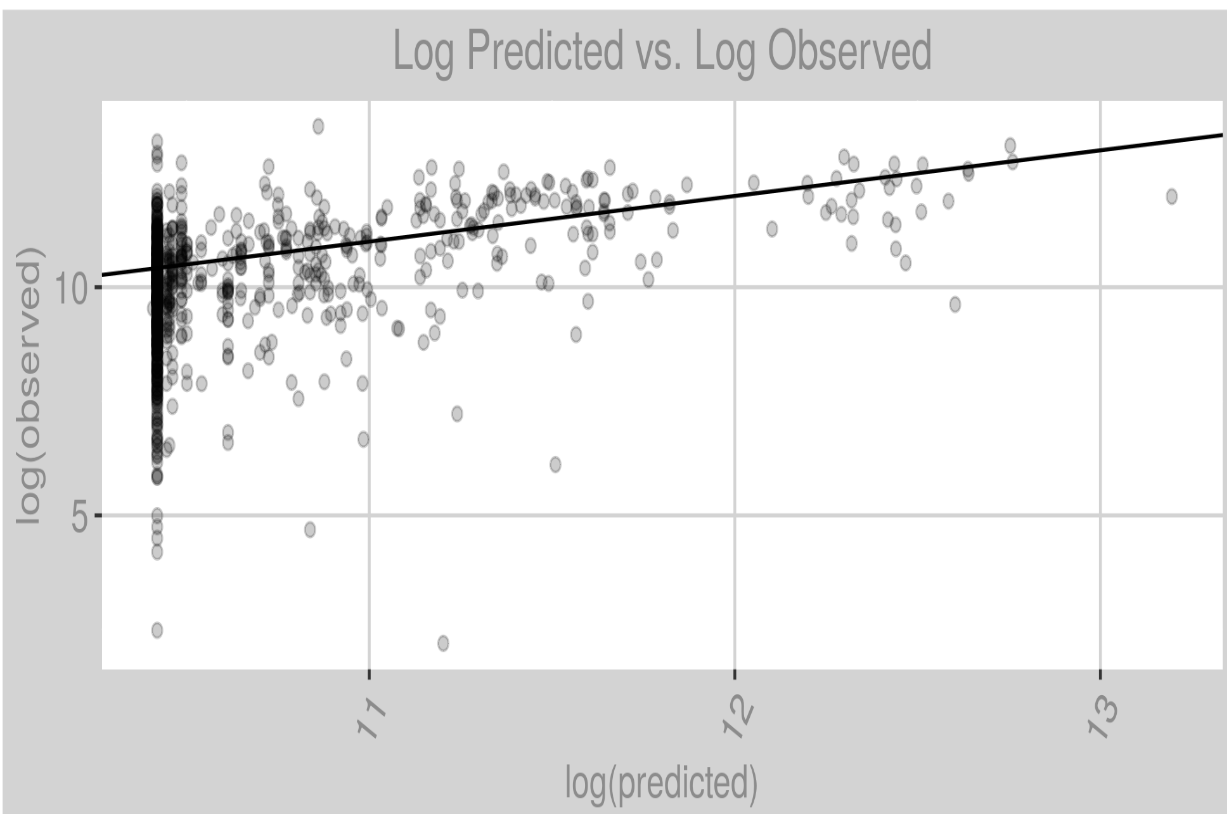

Log Predicted vs. Log Observed
log(observed)
log(predicted)

Recursive Feature Extraction

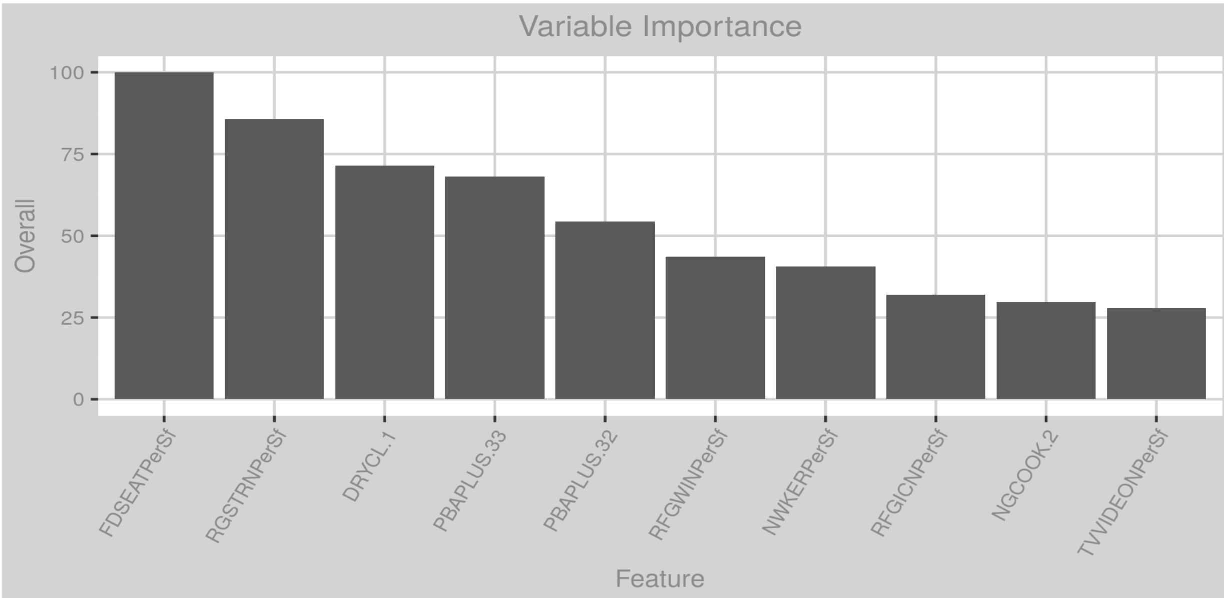


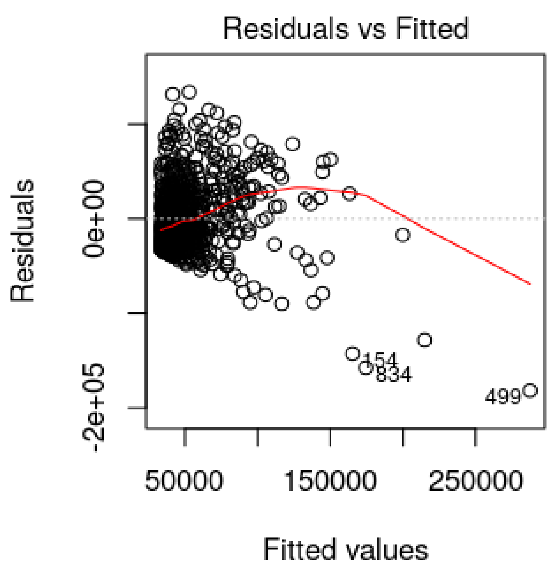


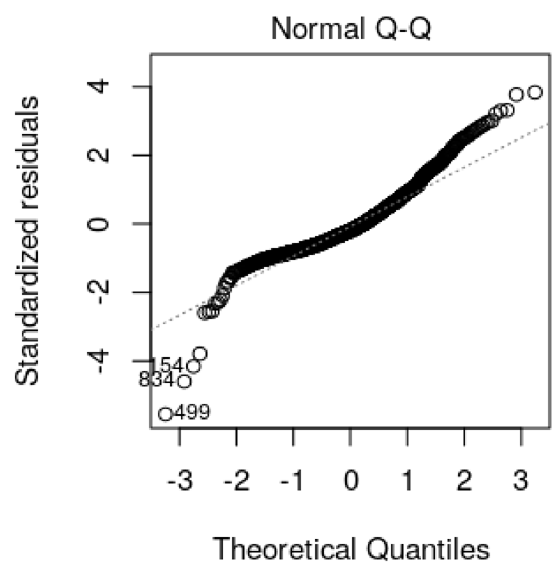

Scale-Location
√|Standardized residuals|
Fitted values
514
245
499
0.0
1.0
2.0
3.0
50000
150000
250000
Residuals vs Leverage
Standardized residuals
Leverage
106
524
499
Cook's distance
1
0.5
0.00
0.04
0.08
0.12
-5
0
5
10

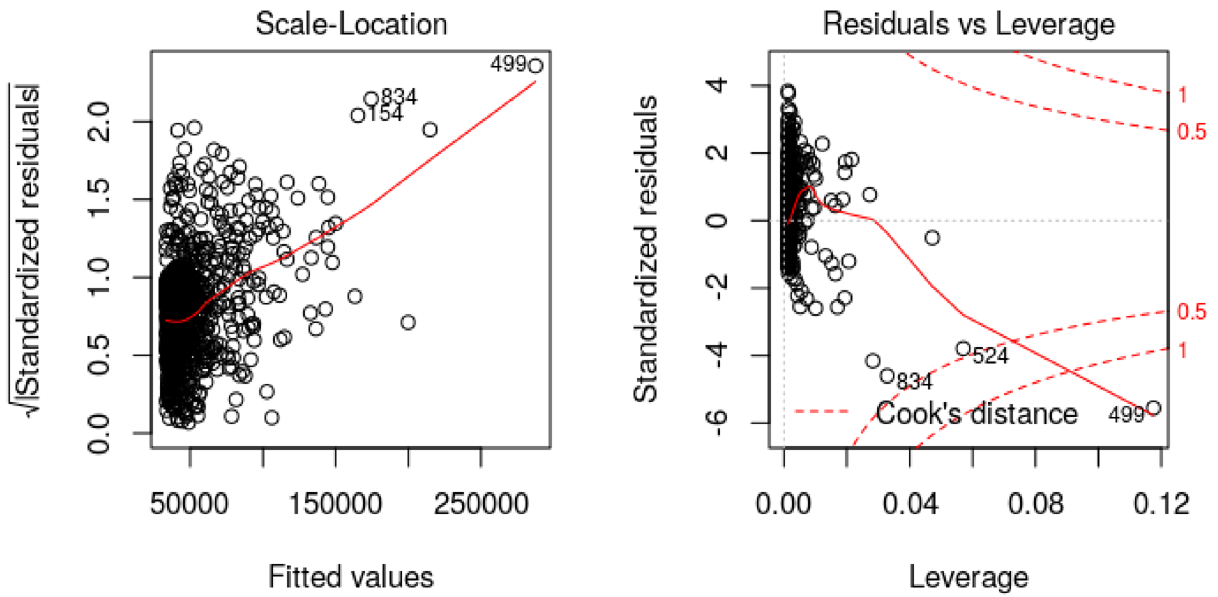

Log Predicted vs. Log Observed
log(observed)
log(predicted)
5
10
11
12
13

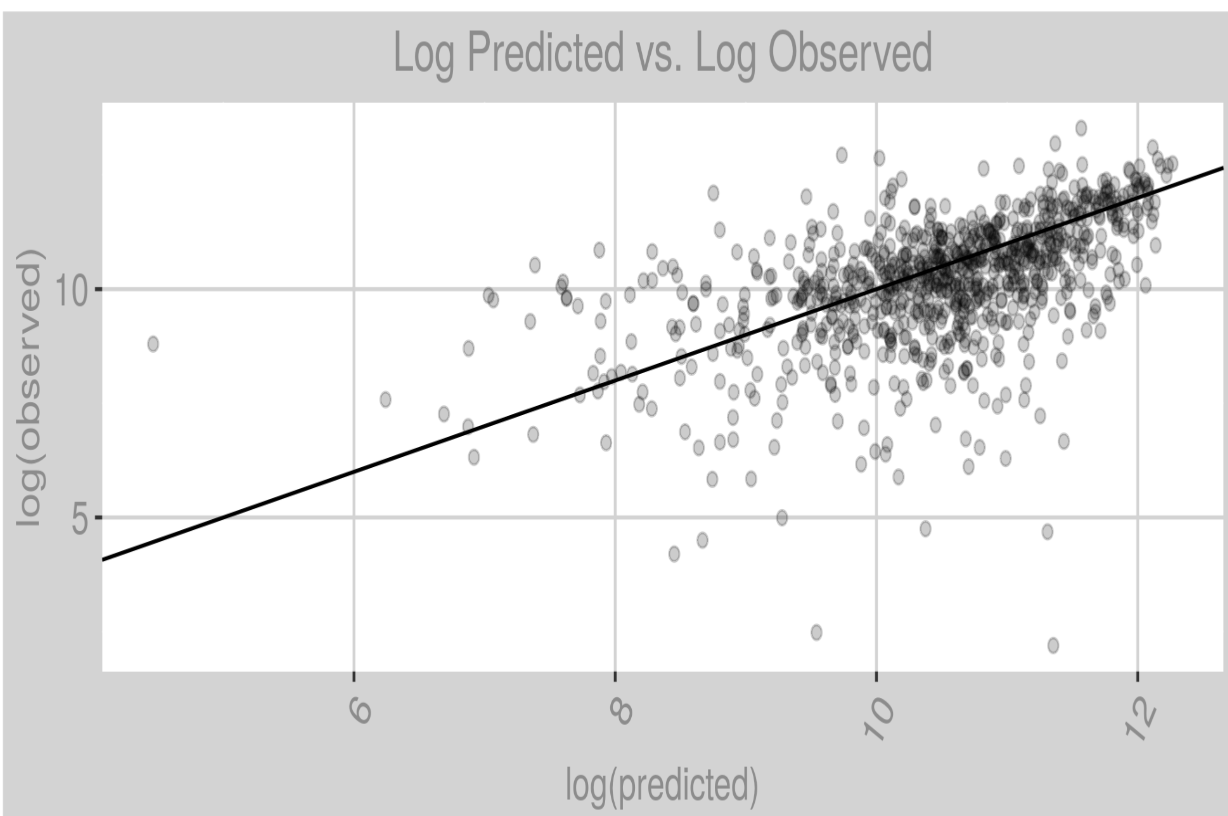
Simple Neural Network







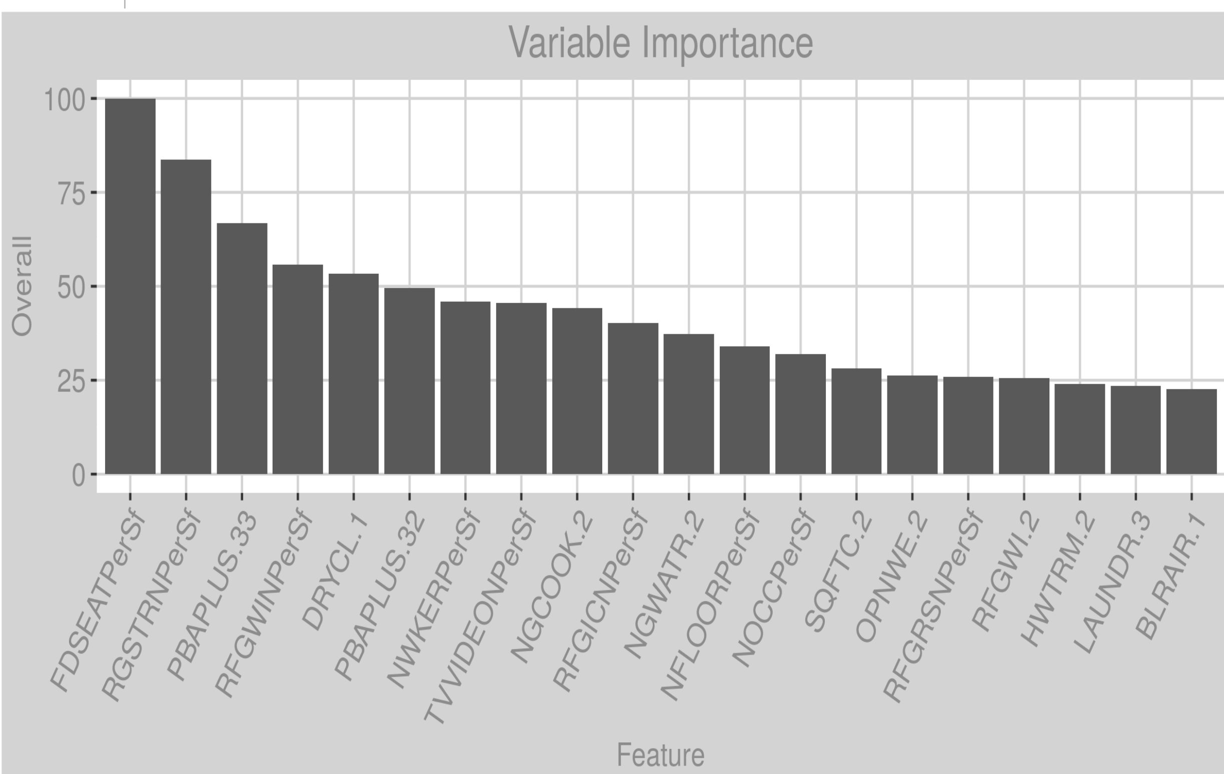

Scale-Location
√|Standardized residuals|
Fitted values
499
834
154
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
524
834
499
1
0.5
0.5
1

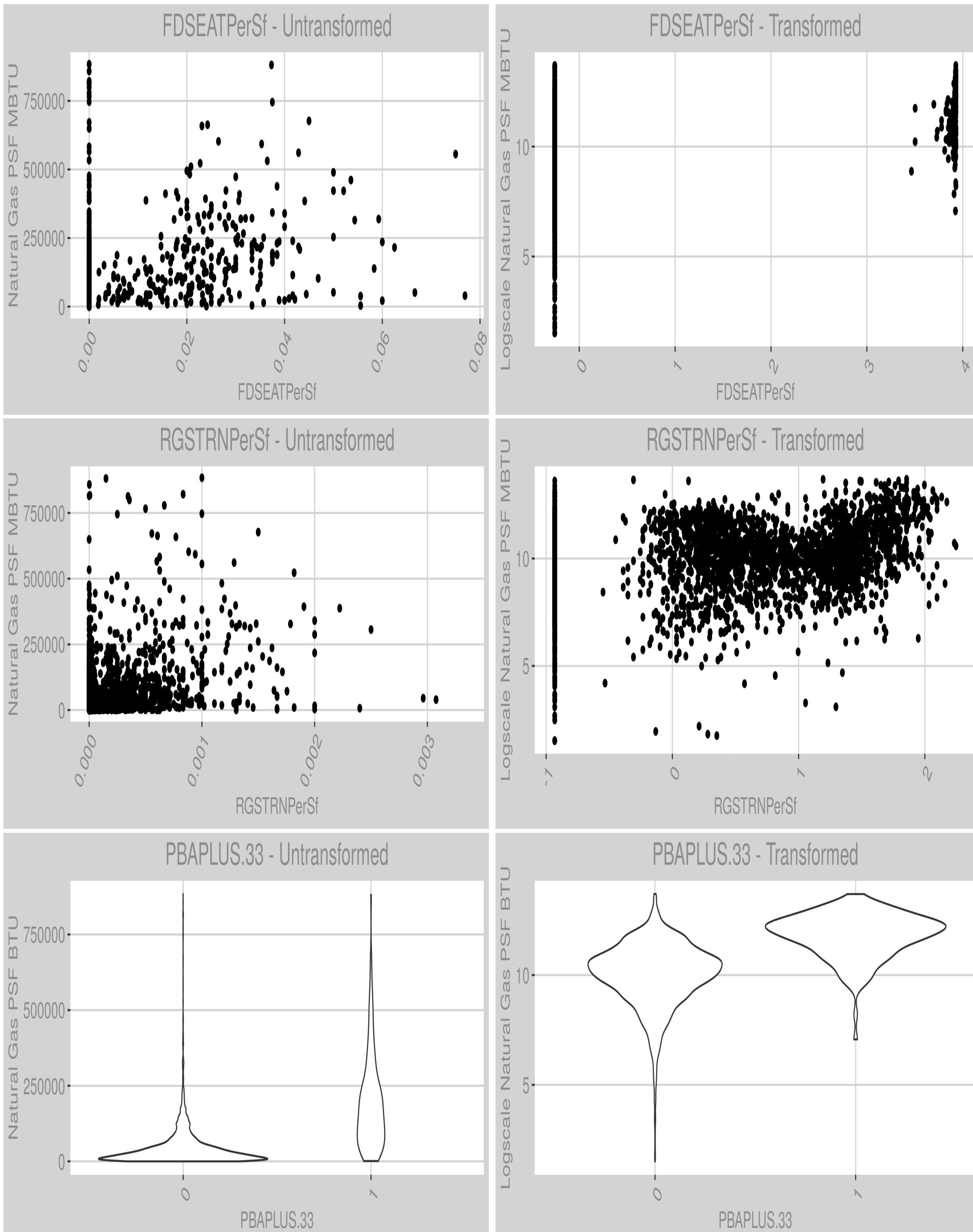

Log Predicted vs. Log Observed
log(observed)
log(predicted)

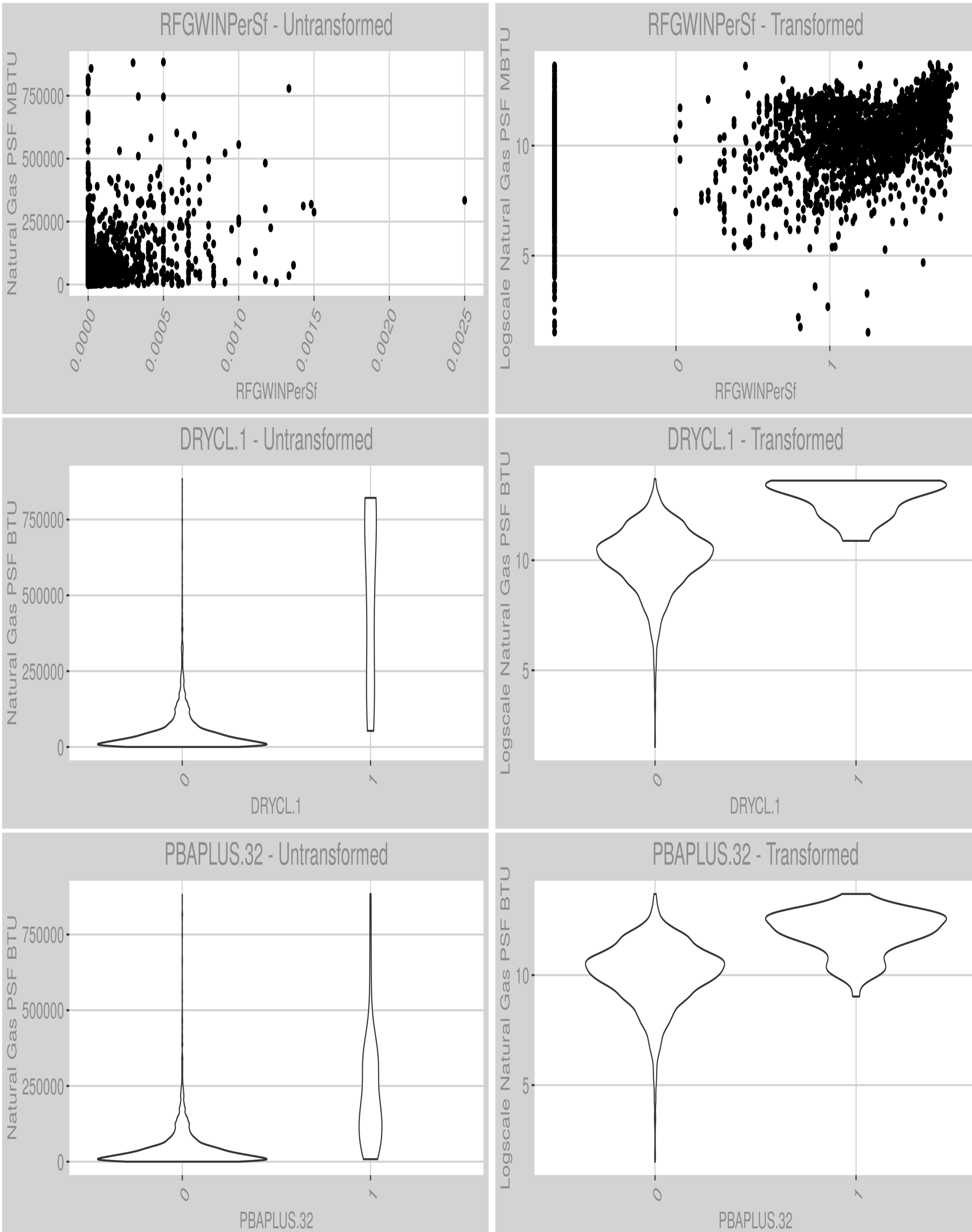
Select Variable Analysis

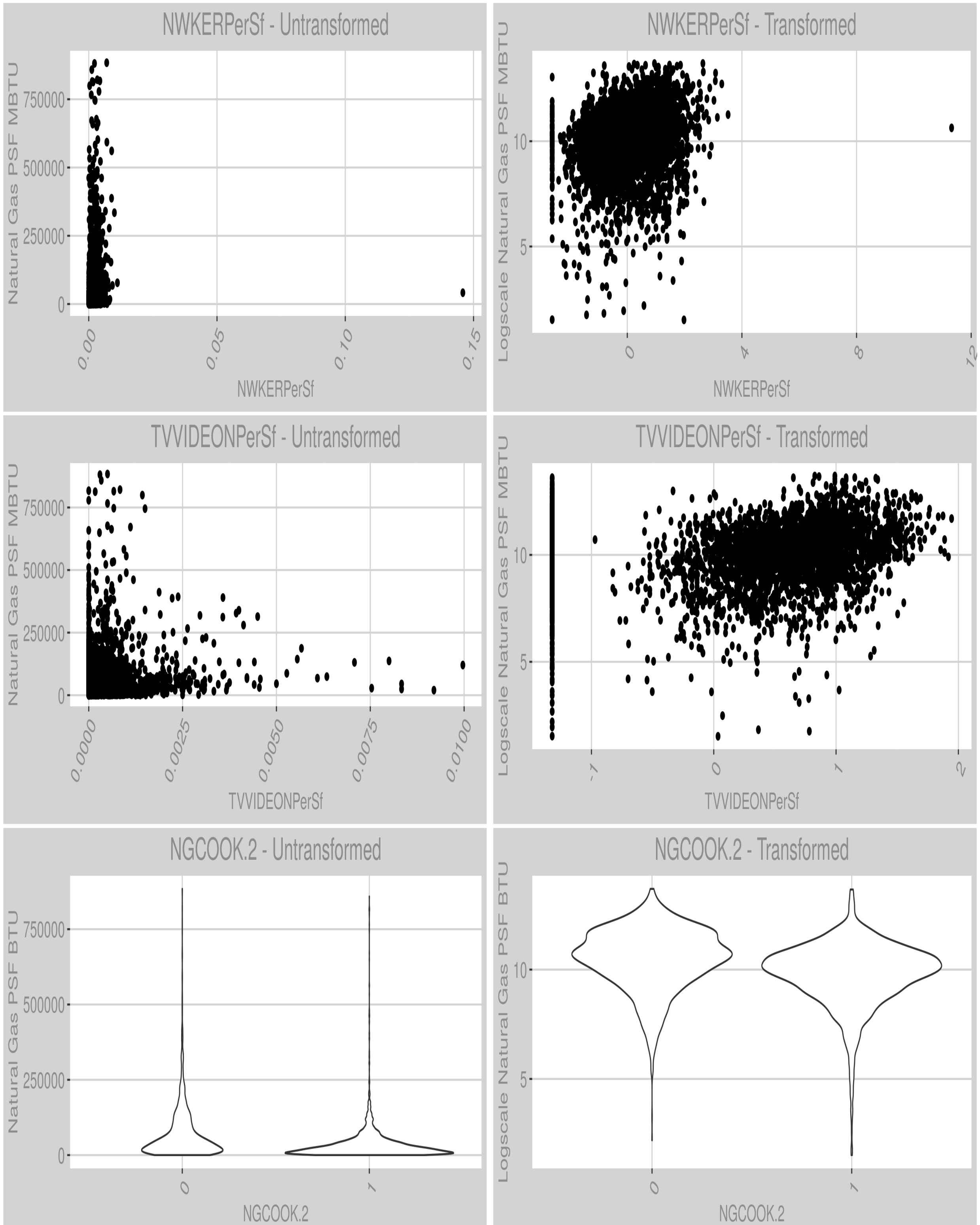
Feature Extraction Model Results

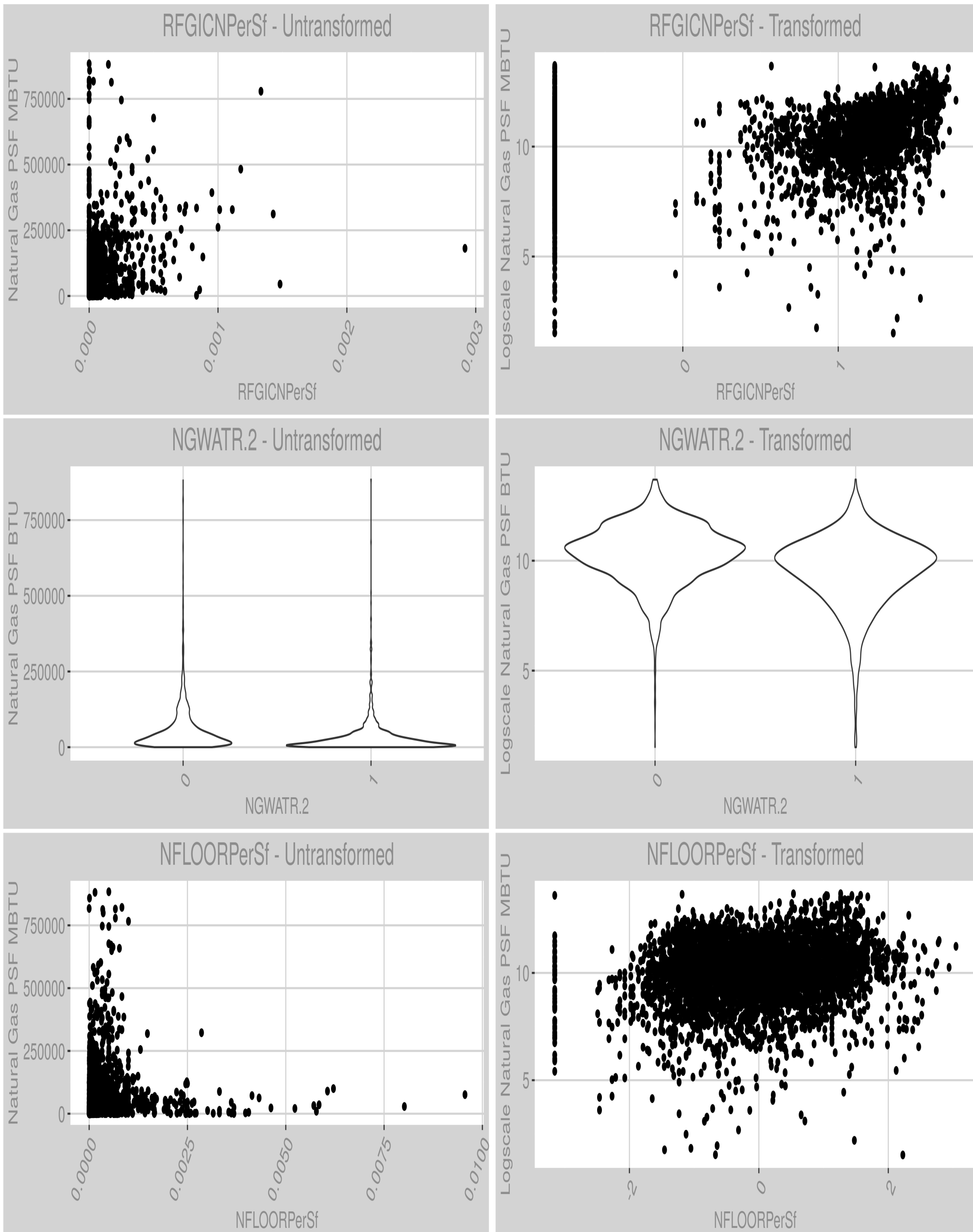
Model	RMSE	R2	MAE	MAPE	MSLE
partialLeastSquares	67978.44	0.2321160	31769.30	469.5262	1.215225
leaps	69079.81	0.1837522	34964.80	263.2334	1.367507
neuralNetwork	57929.72	0.3193293	33328.70	1660.1715	1.731125
recursiveFeatureExtraction	61464.86	0.2589020	35070.92	1887.8960	2.018073
randomForest	174127.85	0.1215693	147326.47	8693.7861	5.379562

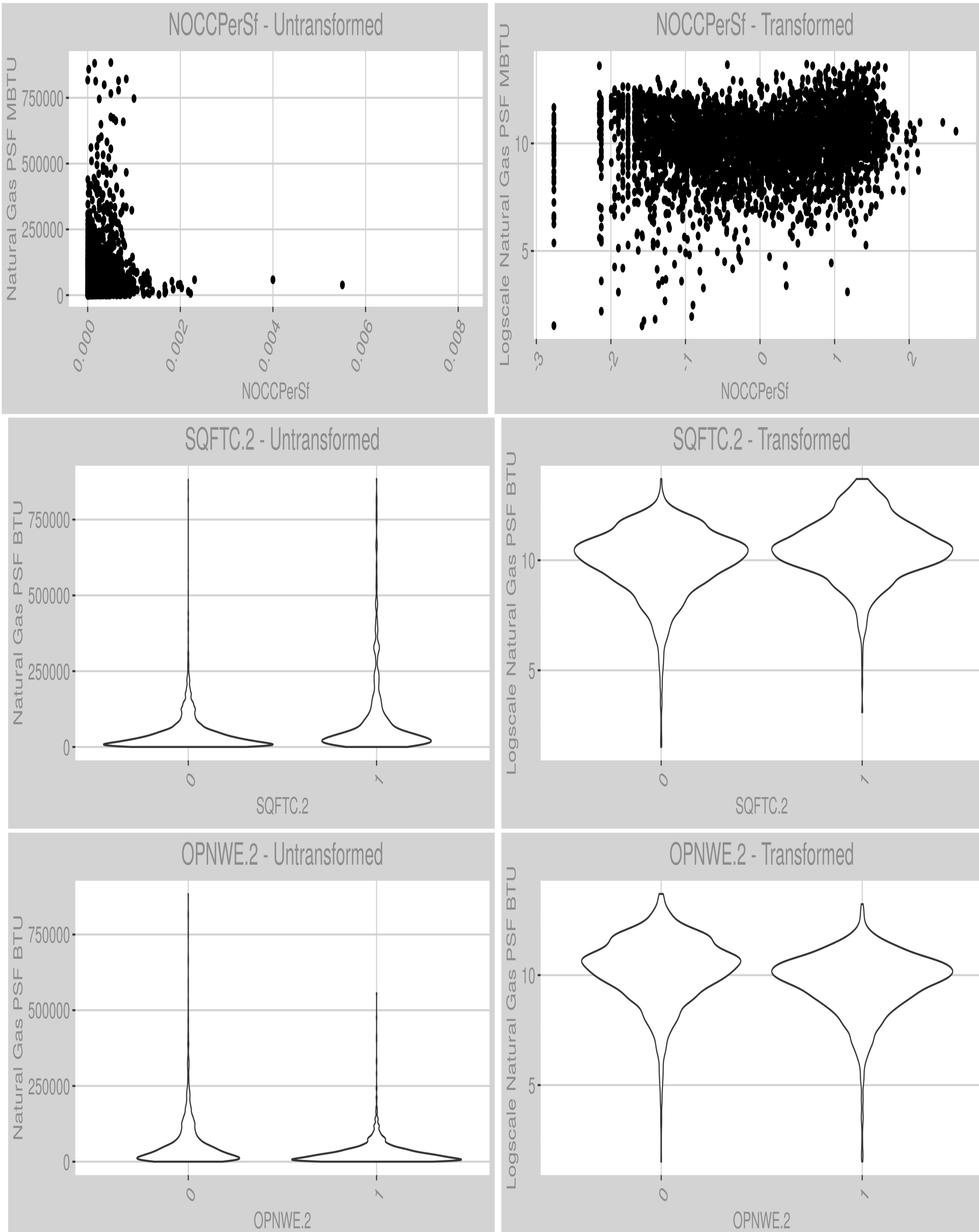


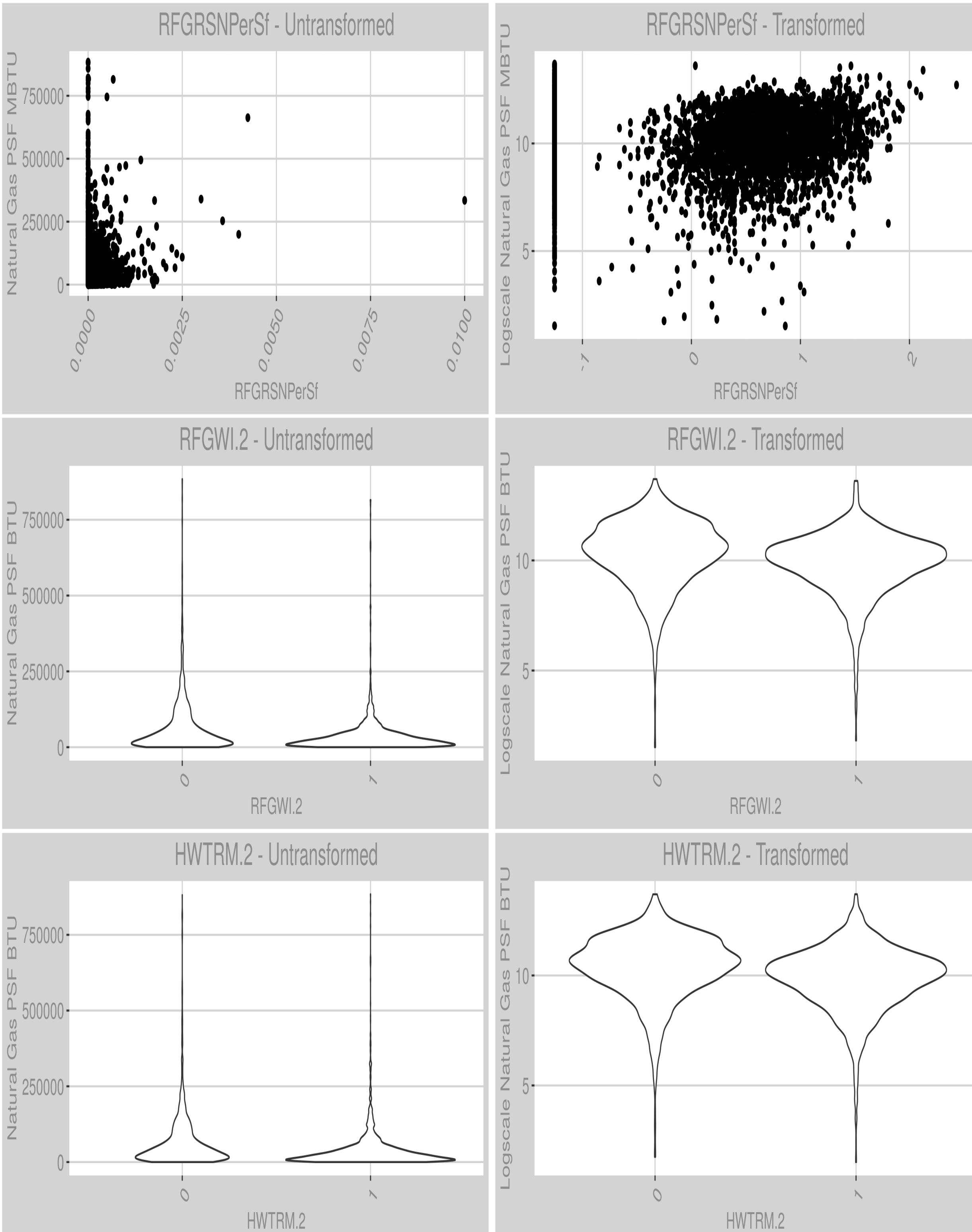

FDSEATPerSf - Untransformed
FDSEATPerSf - Transformed
RGSTRNPerSf - Untransformed
RGSTRNPerSf - Transformed
PBAPLUS.33 - Untransformed
PBAPLUS.33 - Transformed
Natural Gas PSF MBTU
Logscale Natural Gas PSF MBTU
Natural Gas PSF BTU
Logscale Natural Gas PSF BTU
FDSEATPerSf
RGSTRNPerSf
PBAPLUS.33

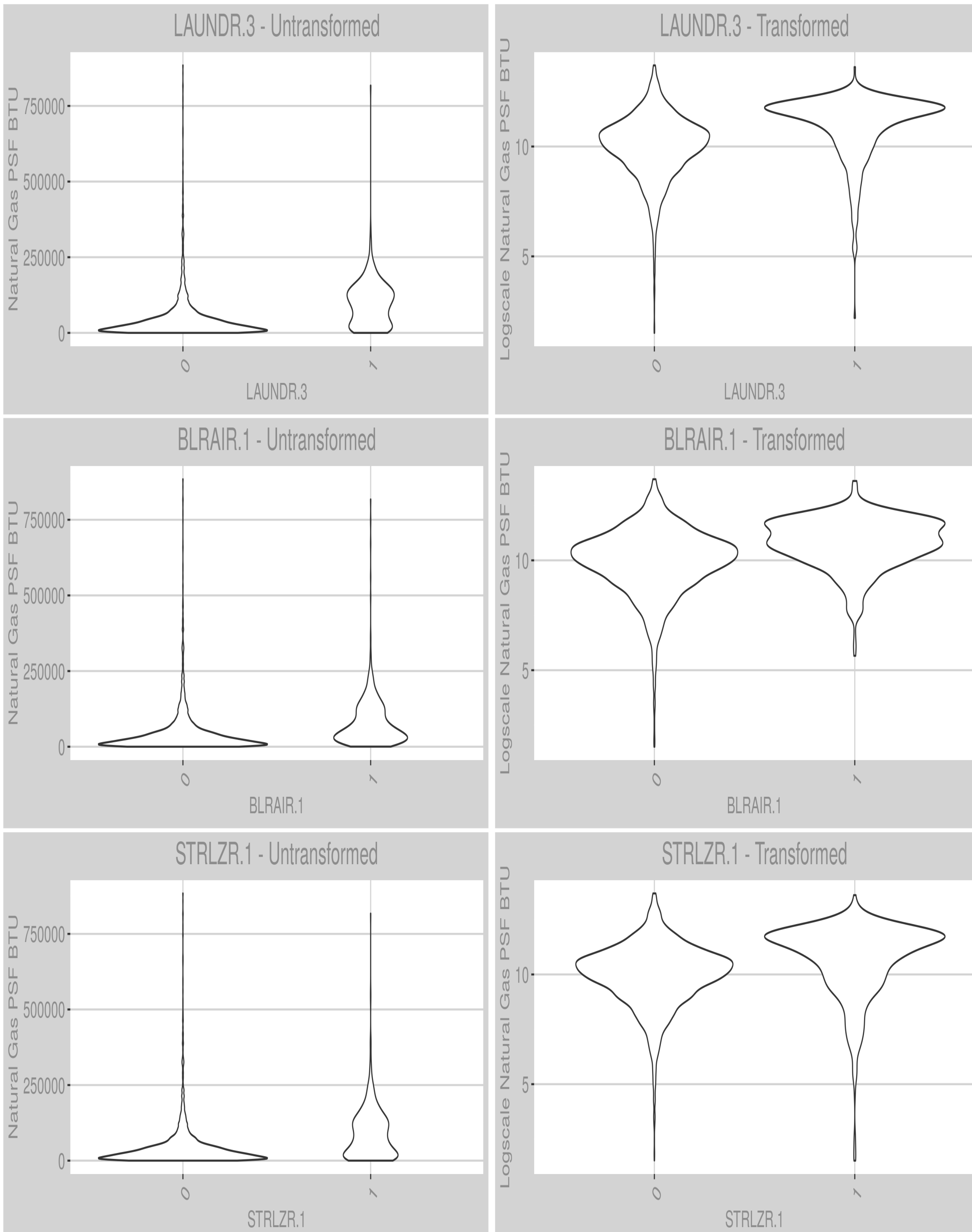

RFGWINPerSf - Untransformed
Natural Gas PSF MBTU
RFGWINPerSf
RFGWINPerSf - Transformed
Logscale Natural Gas PSF MBTU
RFGWINPerSf
DRYCL.1 - Untransformed
Natural Gas PSF BTU
DRYCL.1
DRYCL.1 - Transformed
Logscale Natural Gas PSF BTU
DRYCL.1
PBAPLUS.32 - Untransformed
Natural Gas PSF BTU
PBAPLUS.32
PBAPLUS.32 - Transformed
Logscale Natural Gas PSF BTU
PBAPLUS.32


NWKERPerSf - Untransformed
Natural Gas PSF MBTU
750000
500000
250000
0
0.00
0.05
0.10
0.15
NWKERPerSf
NWKERPerSf - Transformed
Logscale Natural Gas PSF MBTU
10
5
0
4
8
12
NWKERPerSf
TVVIDEONPerSf - Untransformed
Natural Gas PSF MBTU
750000
500000
250000
0
0.0000
0.0025
0.0050
0.0075
0.0100
TVVIDEONPerSf
TVVIDEONPerSf - Transformed
Logscale Natural Gas PSF MBTU
10
5
-1
0
1
2
TVVIDEONPerSf
NGCOOK.2 - Untransformed
Natural Gas PSF BTU
750000
500000
250000
0
0
1
NGCOOK.2
NGCOOK.2 - Transformed
Logscale Natural Gas PSF BTU
10
5
0
1
NGCOOK.2


RFGICNPerSf - Untransformed
Natural Gas PSF MBTU
750000
500000
250000
0
0.000
0.001
0.002
0.003
RFGICNPerSf
RFGICNPerSf - Transformed
Logscale Natural Gas PSF MBTU
10
5
0
1
RFGICNPerSf
NGWATR.2 - Untransformed
Natural Gas PSF BTU
750000
500000
250000
0
0
1
NGWATR.2
NGWATR.2 - Transformed
Logscale Natural Gas PSF BTU
10
5
0
1
NGWATR.2
NFLOORPerSf - Untransformed
Natural Gas PSF MBTU
750000
500000
250000
0
0.0000
0.0025
0.0050
0.0075
0.0100
NFLOORPerSf
NFLOORPerSf - Transformed
Logscale Natural Gas PSF MBTU
10
5
-2
0
2
NFLOORPerSf

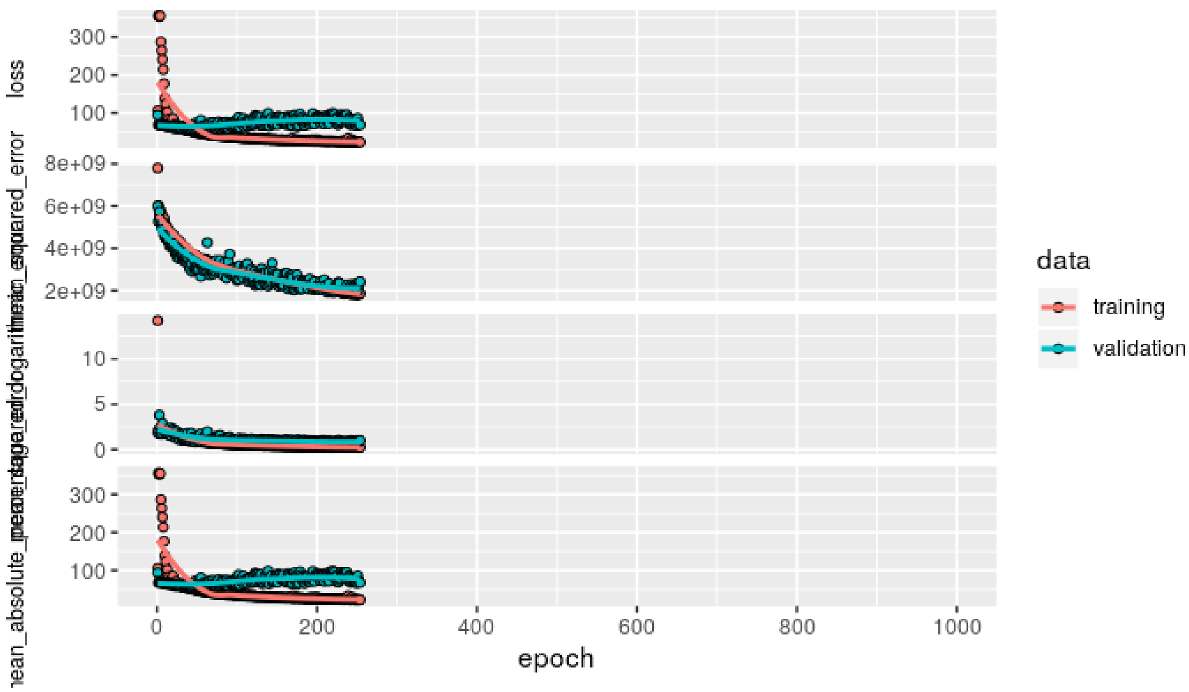

NOCCPerSf - Untransformed
Natural Gas PSF MBTU
750000
500000
250000
0
0.000
0.002
0.004
0.006
0.008
NOCCPerSf
NOCCPerSf - Transformed
Logscale Natural Gas PSF MBTU
10
5
-3
-2
-1
0
1
2
NOCCPerSf
SQFTC.2 - Untransformed
Natural Gas PSF BTU
750000
500000
250000
0
0
1
SQFTC.2
SQFTC.2 - Transformed
Logscale Natural Gas PSF BTU
10
5
0
1
SQFTC.2
OPNWE.2 - Untransformed
Natural Gas PSF BTU
750000
500000
250000
0
0
1
OPNWE.2
OPNWE.2 - Transformed
Logscale Natural Gas PSF BTU
10
5
0
1
OPNWE.2

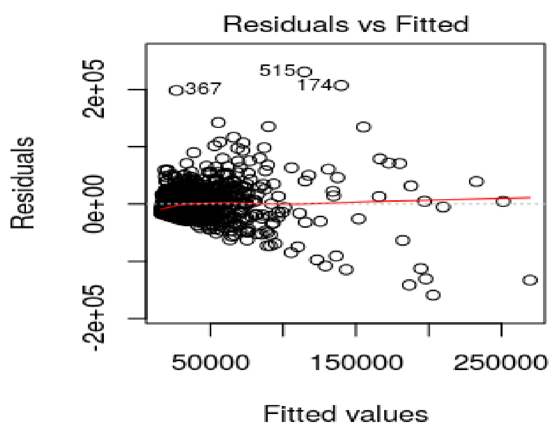

RFGRSNPerSf - Untransformed
Natural Gas PSF MBTU
750000
500000
250000
0
0.0000
0.0025
0.0050
0.0075
0.0100
RFGRSNPerSf
RFGRSNPerSf - Transformed
Logscale Natural Gas PSF MBTU
10
5
-1
0
1
2
RFGRSNPerSf
RFGWI.2 - Untransformed
Natural Gas PSF BTU
750000
500000
250000
0
0
1
RFGWI.2
RFGWI.2 - Transformed
Logscale Natural Gas PSF BTU
10
5
0
1
RFGWI.2
HWTRM.2 - Untransformed
Natural Gas PSF BTU
750000
500000
250000
0
0
1
HWTRM.2
HWTRM.2 - Transformed
Logscale Natural Gas PSF BTU
10
5
0
1
HWTRM.2

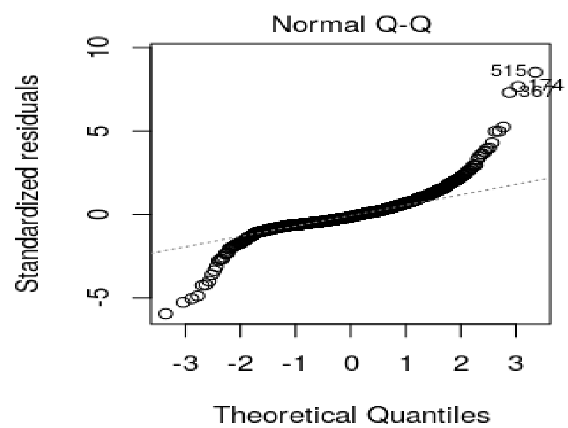

LAUNDR.3 - Untransformed
Natural Gas PSF BTU
750000
500000
250000
0
0
1
LAUNDR.3
LAUNDR.3 - Transformed
Logscale Natural Gas PSF BTU
10
5
0
1
LAUNDR.3
BLRAIR.1 - Untransformed
Natural Gas PSF BTU
750000
500000
250000
0
0
1
BLRAIR.1
BLRAIR.1 - Transformed
Logscale Natural Gas PSF BTU
10
5
0
1
BLRAIR.1
STRLZR.1 - Untransformed
Natural Gas PSF BTU
750000
500000
250000
0
0
1
STRLZR.1
STRLZR.1 - Transformed
Logscale Natural Gas PSF BTU
10
5
0
1
STRLZR.1

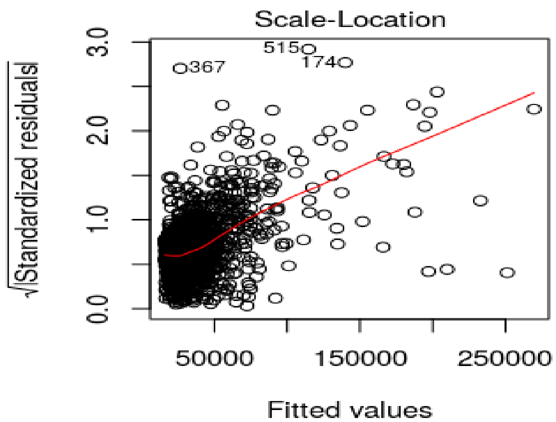
Appendix - Neural Networks

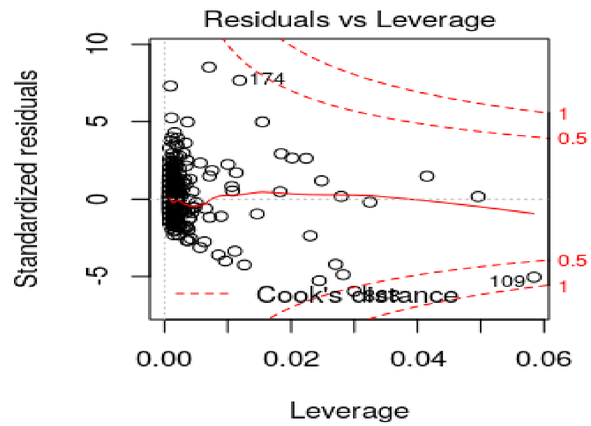
Electricity

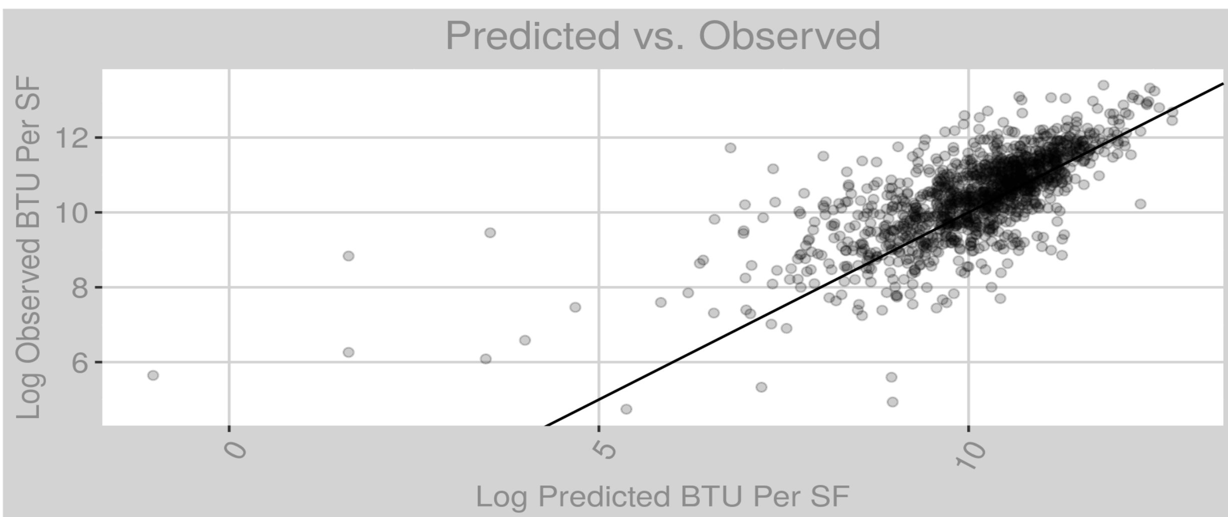


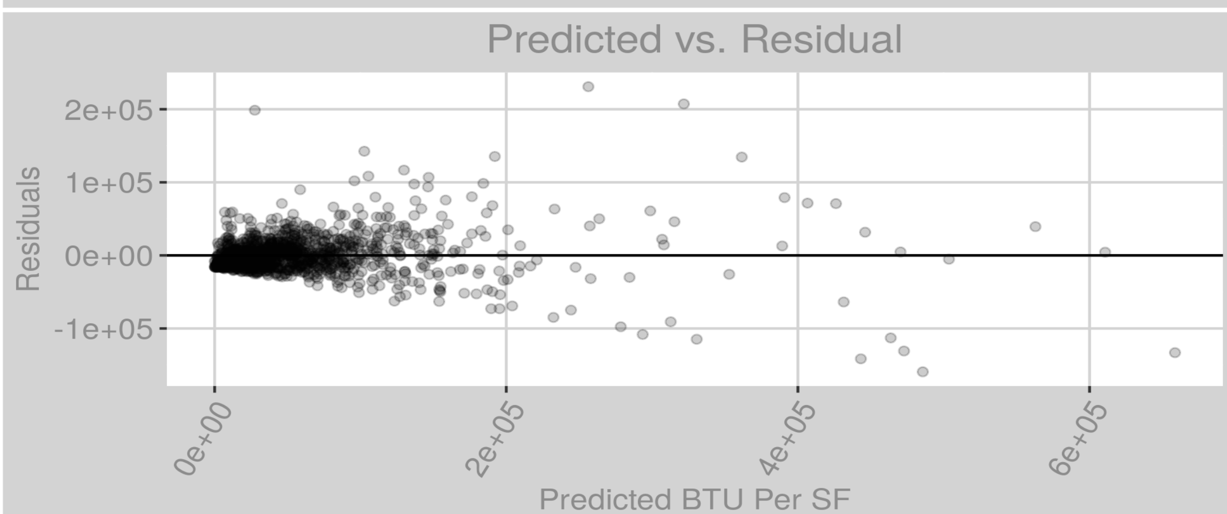


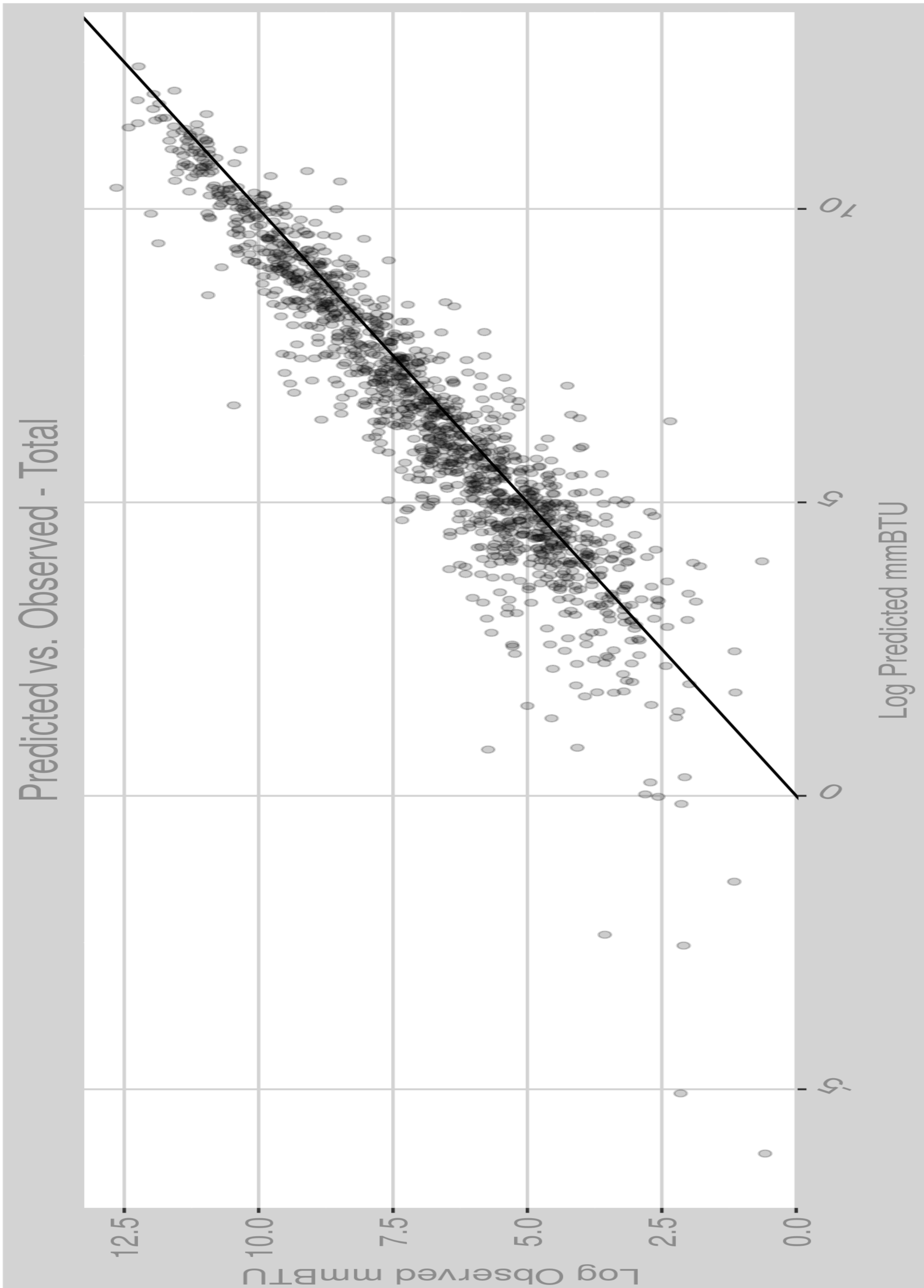



Scale-Location
√|Standardized residuals|
3.0
2.0
1.0
0.0
367
515
174
50000
150000
250000
Fitted values


Residuals vs Leverage
Standardized residuals
10
5
0
-5
174
109
1
0.5
0.5
1
Cook's distance
0.00
0.02
0.04
0.06
Leverage


Predicted vs. Observed
Log Observed BTU Per SF
12
10
8
6
0
5
10
Log Predicted BTU Per SF


Predicted vs. Residual
Residuals
2e+05
1e+05
0e+00
-1e+05
0e+00
2e+05
4e+05
6e+05
Predicted BTU Per SF


Predicted vs. Observed - Total
Log Predicted mmBTU
Log Observed mmBTU
-5
0
5
10
0.0
2.5
5.0
7.5
10.0
12.5

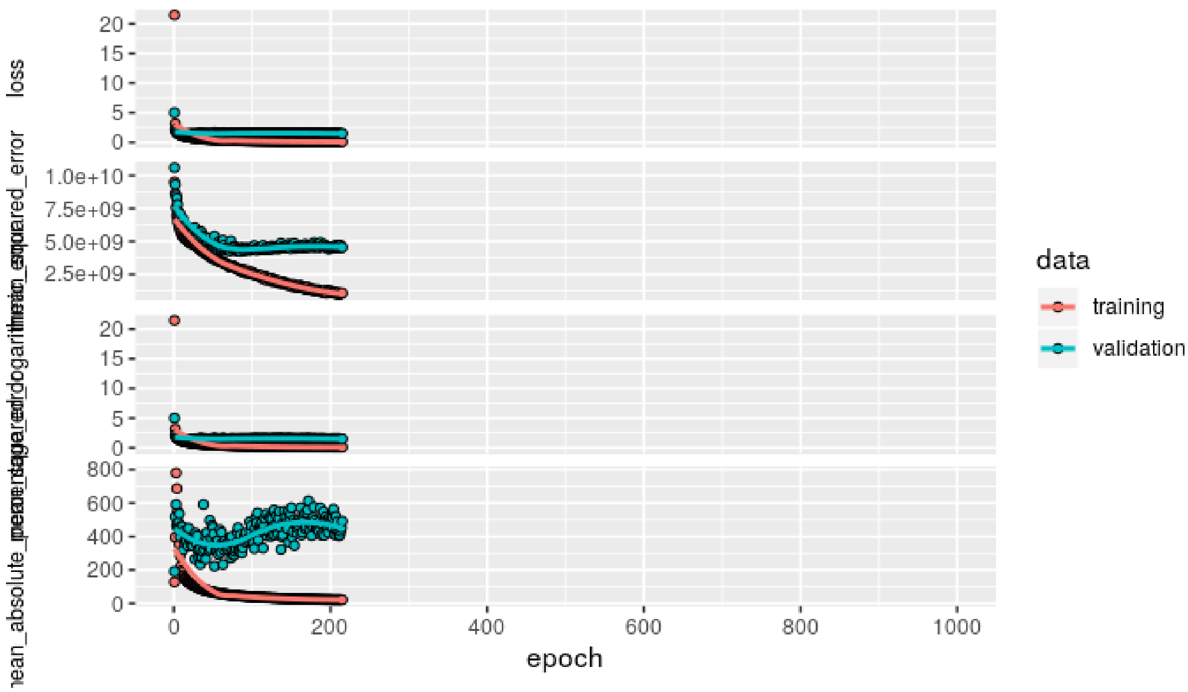
Selected Variables

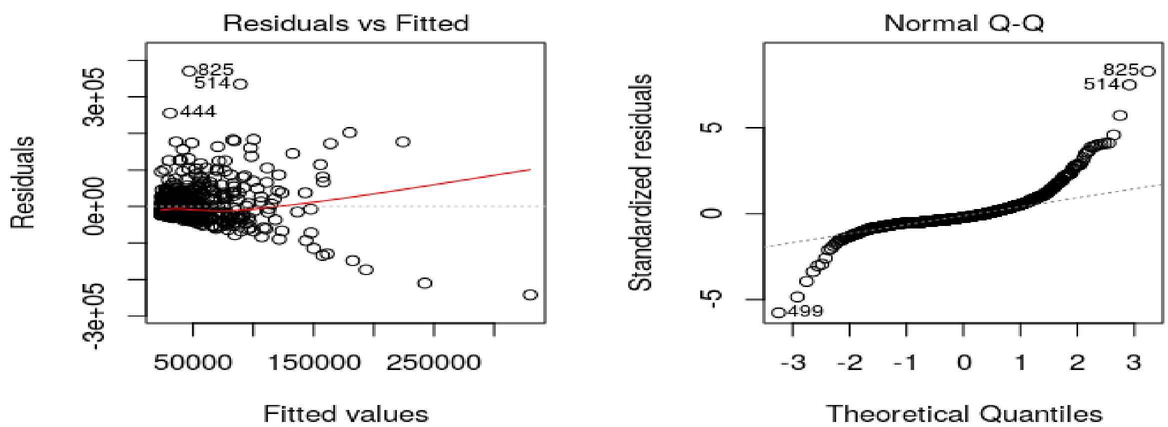
RFGWIN	Number of walk-in units	TVVIDEON	Number of TV or video displays	WKHRSC	Weekly hours category	PKLT	Lighted parking area
NWKER	Number of employees	HWTRM	Large amounts of hot water	GLSSPC	Percent exterior glass	FAX	FAX machines
RFGWI	Walk-in refrigeration units	RFGWI	Walk-in refrigeration units	CLVOAS	Cooling ventilation: Dedicated outside air system	OPNMF	Open during week
FDSEAT	Food service seating capacity	OPNMF	Open during week	LAUNDR	Laundry onsite	NGCOOK	Natural gas used for cooking
RGSTRN	Number of cash registers	EVAPCL	Evaporative or swamp coolers	CFLRP	Percent lit by compact fluorescent	RFGCOMPN	Number of compact refrigerators
PBAPLUS	More specific building activity	NWKERC	Number of employees category	RFGVEN	Refrigerated vending machines	FKHT2	Fuel oil used for secondary heating
RFGICN	Number of ice makers	ELCOOK	Electricity used for cooking	PBAPLUS	More specific building activity	COPIERN	Number of photocopiers
RFGICE	Commercial ice makers	STRLZR	Sterilizers or autoclaves	FACACT	Type of complex	FKTYPE	Specify fuel oil, diesel, or kerosene
RGSTR	Cash registers	MAINT	Regular HVAC maintenance	PCTERM	Computers used	LNHRPC	Lit off hours category
RFGCLN	Number of closed case refrigeration units	WKHRSC	Weekly hours category	VACANT	Completely vacant	NWKERC	Number of employees category

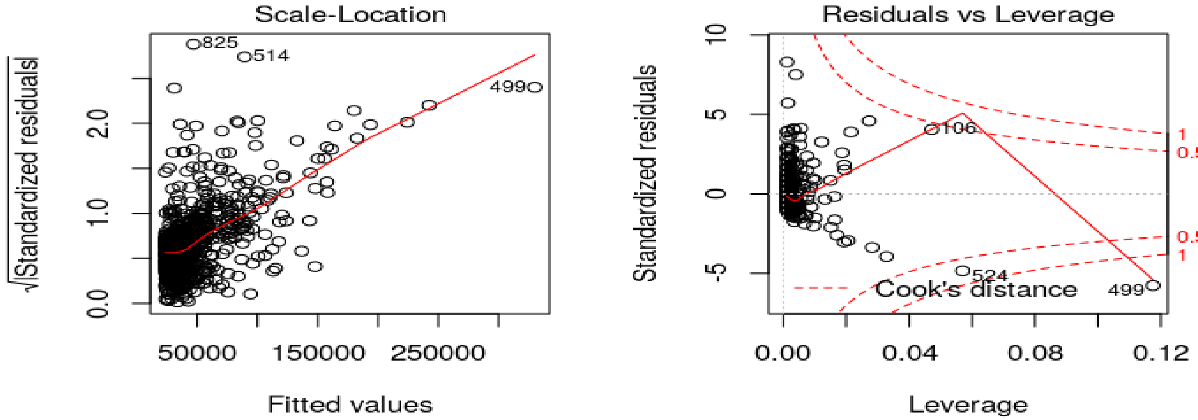
RFGCL	Closed case refrigeration units	TVVIDEO	TV or video displays	OTLT	Other type of bulbs	HCBED	Licensed bed capacity
COOK	Energy used for cooking	RFGRES	Full-size residential-type refrigerator	FASTFD	Fast food or small restaurant	CFLR	Compact fluorescent bulbs
NGCOOK	Natural gas used for cooking	GENUSE	Use of generated electricity	DCNTRSFC	Data center or server farm sqft category	ELEVTR	Elevators
PCTERMN	Number of computers	HEATP	Percent heated	RFTILT	Roof tilt	CHLPKG	Chiller system: Packaged unit
LNHRPC	Lit off hours category	MRI	MRI machines	TVVIDEON	Number of TV or video displays	CHLAIRCL	Chiller type: Air-cooled
PBA	Principal building activity	BOOSTWT	Booster water heaters	PBAPLUS	More specific building activity	LTEXPC	Percent of exterior lighted
RFGOPN	Number of open case refrigeration units	LTNR24	Lights off during 24 hours	LINACC	Linear accelerators	RFGCOMP	Half-size or compact refrigerators
PRNTRN	Number of printers	SERVERN	Number of servers	LABEQP	Laboratory equipment	KITCHN	Small kitchen area
RFGOP	Open case refrigeration units	XRAYN	Number of X-ray machines	WHRECOV	Waste heat recovery	TRIM	High-end trimming or light-level tuning
PBA	Principal building activity	ANYEGY	Any energy used	RFGICE	Commercial ice makers	GLSSPC	Percent exterior glass

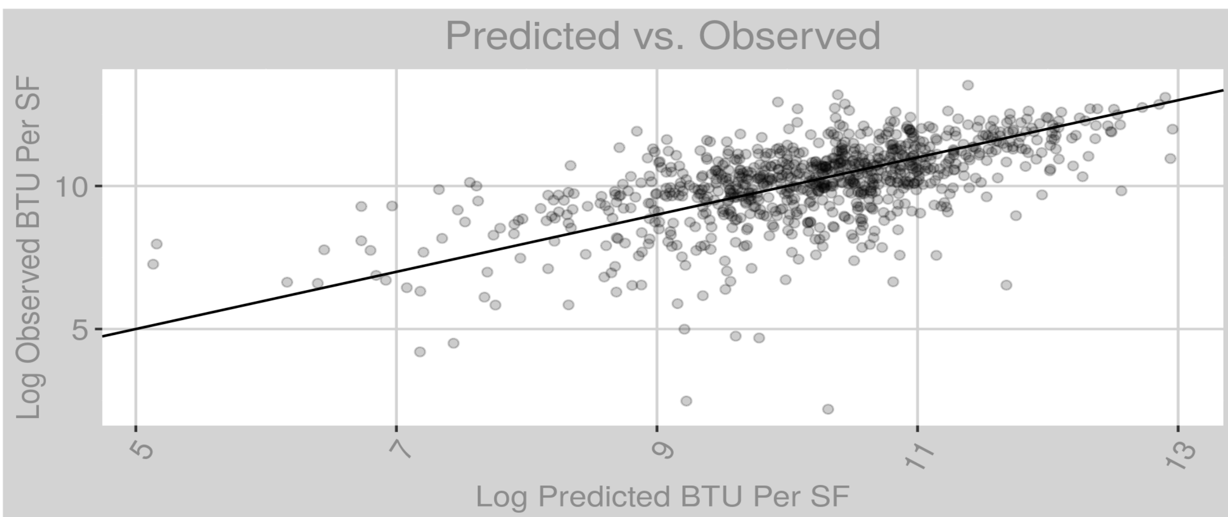
WKHRSC	Weekly hours category	LEDP	Percent lit by LED	MCHEQP	Machine equipment	RFGSTO	Large cold storage areas
OPNWE	Open on weekend	RFGVNN	Number of refrigerated vending machines	MONUSE	Months in use	CHLFNCL	Chiller system: Fan coil units in rooms
LOHRPC	Lit when open category	PBAPLUS	More specific building activity	OWNTYPE	Building owner	HTVVAV	Heating ventilation: Central air handling with VAV
PBAPLUS	More specific building activity	RWSEAT	Religious worship seating capacity	RFGSTO	Large cold storage areas	NELVTR	Number of elevators
MCHEQP	Machine equipment	PBAPLUS	More specific building activity	LAUNDR	Laundry onsite	NGHT1	Natural gas used for main heating

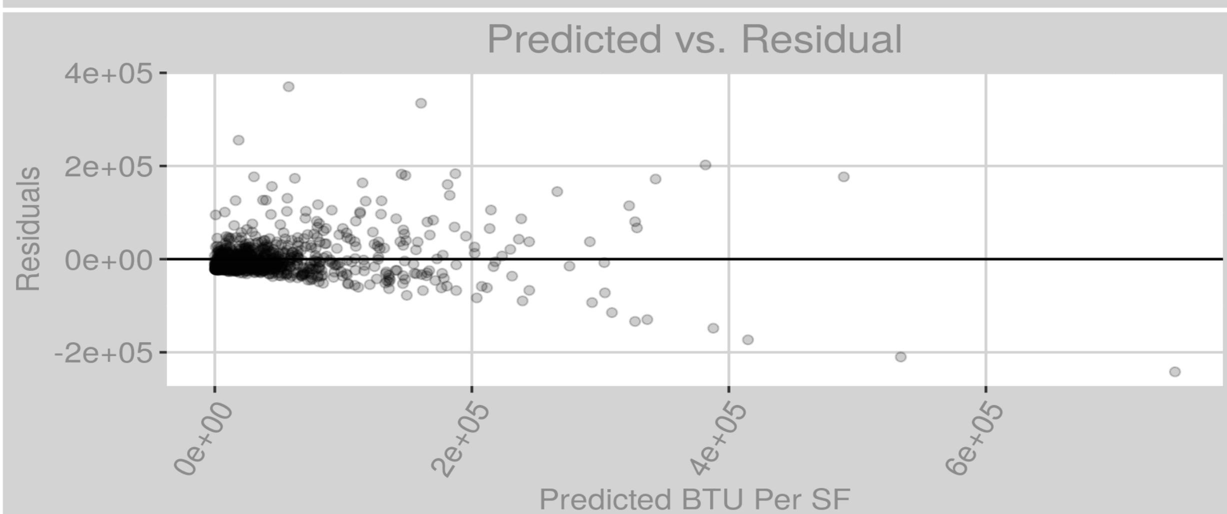
Natural Gas

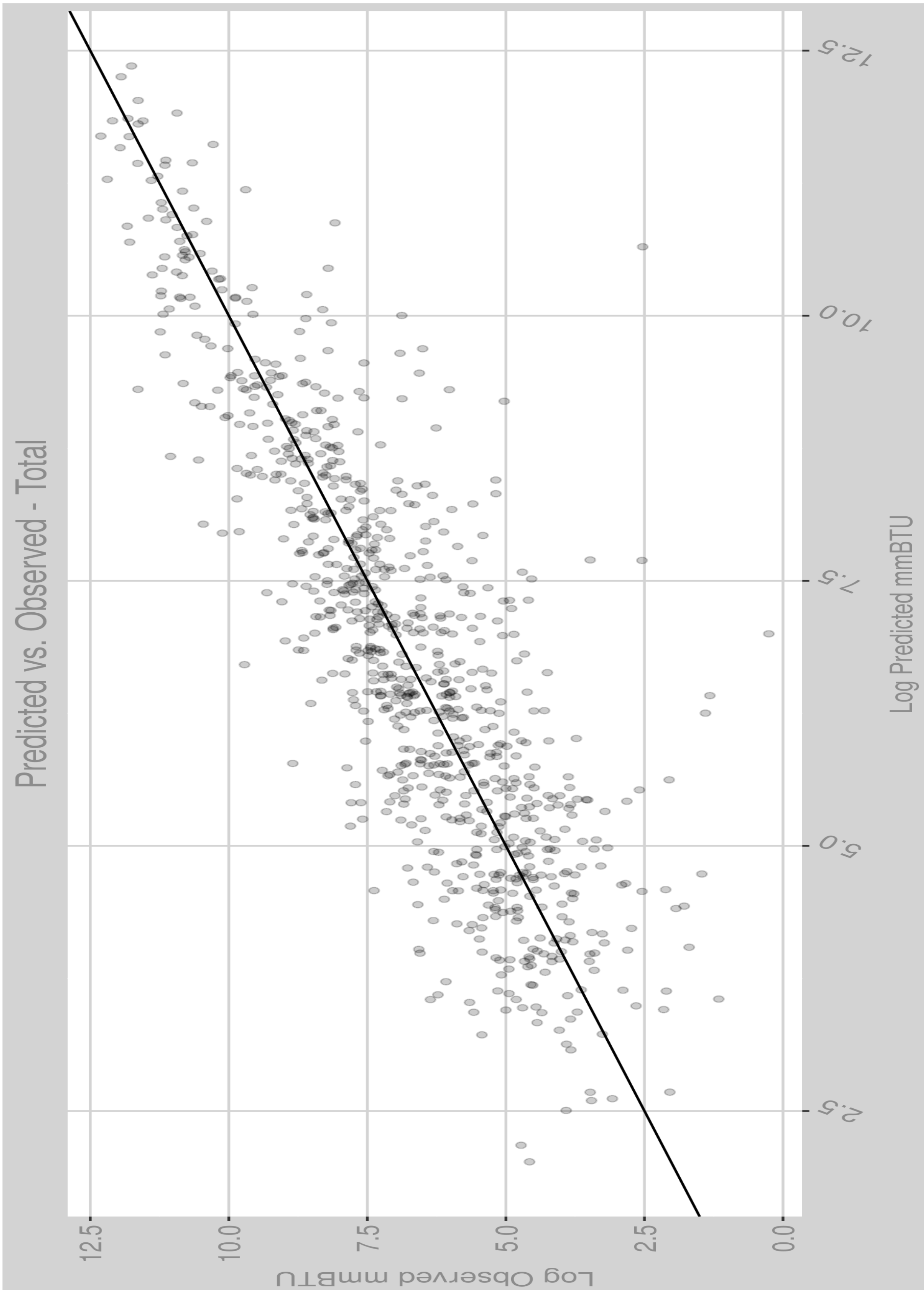

loss
epoch
data
training
validation
0
200
400
600
800
1000


Residuals vs Fitted
Residuals
Fitted values
825
514
444
Normal Q-Q
Standardized residuals
Theoretical Quantiles
499


Scale-Location
√|Standardized residuals|
825
514
499
0.0
1.0
2.0
50000
150000
250000
Fitted values
Residuals vs Leverage
Standardized residuals
10
5
0
-5
106
524
499
1
0.5
0.5
1
Cook's distance
0.00
0.04
0.08
0.12
Leverage


Predicted vs. Observed
Log Observed BTU Per SF
10
5
5
7
9
11
13
Log Predicted BTU Per SF


Predicted vs. Residual
Residuals
4e+05
2e+05
0e+00
-2e+05
0e+00
2e+05
4e+05
6e+05
Predicted BTU Per SF


Predicted vs. Observed - Total
Log Predicted mmBTU
2.5
5.0
7.5
10.0
12.5
Log Observed mmBTU
0.0
2.5
5.0
7.5
10.0
12.5

Selected Variables

FDSEAT	Food service seating capacity	FACACT	Type of complex	HCBED	Licensed bed capacity	CDD65	Cooling degree days (base 65)
RGSTRN	Number of cash registers	RGSTR	Cash registers	DHPKG	District heat system: Packaged unit	WTHTEQ	Water heating equipment
PBAPLUS	More specific building activity	PBA	Principal building activity	BLRRAD	Boiler system: Radiators	LNHRPC	Lit off hours category
RFGWIN	Number of walk-in units	XRAYN	Number of X-ray machines	STRLZR	Sterilizers or autoclaves	DHFNCL	District heat system: Fan coil units in rooms
DRYCL	Dry cleaning onsite	BLRFNCL	Boiler system: Fan coil units in rooms	PBAPLUS	More specific building activity	WTHTEQ	Water heating equipment
PBAPLUS	More specific building activity	BLRDUCT	Boiler system: Duct reheat	WKHRSC	Weekly hours category	NGCOOL	Natural gas used for cooling
NWKER	Number of employees	PRNTRN	Number of printers	PCTRMC	Number of computers category	LINACC	Linear accelerators
TVVIDEON	Number of TV or video displays	RFGCLN	Number of closed case refrigeration units	DHDUCT	District heat system: Duct reheat	BLRRAD	Boiler system: Radiators
NGCOOK	Natural gas used for cooking	LTNR24	Lights off during 24 hours	BLDPLT	Central plant in building	NGOTH	Natural gas for some other use
RFGICN	Number of ice makers	COPIER	Photocopiers	MCHEQP	Machine equipment	WKHRSC	Weekly hours category

NGWATR	Natural gas used for water heating	HCBED	Licensed bed capacity	NGSRC	How purchase natural gas	ELHT1	Electricity used for main heating
NFLOOR	Number of floors	BLRPKG	Boiler system: Packaged unit	DATACNTR	Data center or server farm	BLRINDC	Boiler system: Induction units
NOCC	Number of businesses	HEATP	Percent heated	PCTERMN	Number of computers	RFGOPN	Number of open case refrigeration units
SQFTC	Square footage category	DHRAD	District heat system: Radiators	STDNRM	Student or public computer center	TRNGRM	Computer-based training room
OPNWE	Open on weekend	WKHRSC	Weekly hours category	BOOSTWT	Booster water heaters	PBAPLUS	More specific building activity
RFGRSN	Number of residential refrigerators	LINACC	Linear accelerators	RFGCOMPN	Number of compact refrigerators	ELCOOK	Electricity used for cooking
RFGWI	Walk-in refrigeration units	BLRLOOP	Boiler system: Water loop heat pump	WBOARDS	Interactive whiteboards	RFGOP	Open case refrigeration units
HWTRM	Large amounts of hot water	ELWATR	Electricity used for water heating	PUBCLIM	Building America climate region	CUBELOC	Location of open plan
LAUNDR	Laundry onsite	CUBELOC	Location of open plan	RFGRES	Full-size residential-type refrigerator	GLSSPC	Percent exterior glass
BLRAIR	Boiler system: Central air handler	PBAPLUS	More specific building activity	DHFNCL	District heat system: Fan coil units in rooms	FDPREP	Commercial or large kitchen

STRLZR	Sterilizers or autoclaves	BOOSTWT	Booster water heaters	SERVERN	Number of servers	STDNRM	Student or public computer center
STCOOK	District steam used for cooking	RFGICE	Commercial ice makers	LABEQP	Laboratory equipment	DHDUCT	District heat system: Duct reheat
PBA	Principal building activity	LAPTPN	Number of laptops	PBAPLUS	More specific building activity	LAPTPC	Number of laptops category
BOOSTWT	Booster water heaters	TVVIDEON	Number of TV or video displays	CHLDUCT	Chiller system: Duct reheat	CHLFNCL	Chiller system: Fan coil units in rooms
STWATR	District steam used for water heating	LABEQP	Laboratory equipment	LNHRPC	Lit off hours category	CUBEC	Percent open plan