Report for:

Grid Stability Analysis

Joe McData, Project Manager, Electric Supply LLC

Submitted by:
John Enrietto
10/30/2022

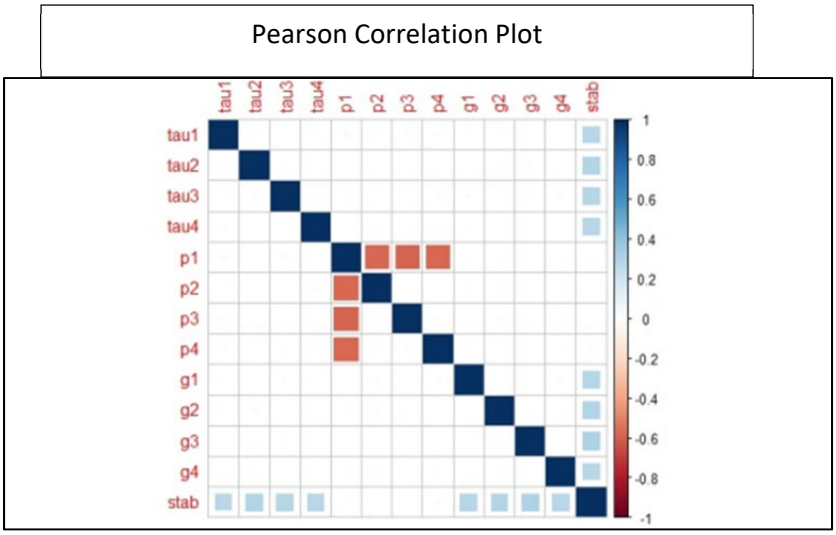
This report is to summarize analysis and results of data supplied by Electric Supply (ES) to determine if grid parameters can be used to predict grid performance and prevent faults. With multiple additional generating points being added to the grid that can be weather and time dependent, monitoring and predictive analysis of the grid is becoming imperative. With proper real time predictive analysis it may be possible to bring on additional generation and battery or capacitive loads to help stabilize the grid when generation systems or other anomalies create instability in the ES grid.

A single data file[(1)] was submitted by Mr. McData containing 10,00 instance of 15 data points each. Descriptions of the parameters are listed in the table below. To maintain trade secret information about exact parameter description, the information is described only as tau(x), p(x), and g(x) values.

ES Grid Stability Parameter definition	
1. tau[x]:	reaction time of participant (real from the range [0.5,10]s). Tau1 - the value for electricity producer.
2. p[x]:	nominal power consumed(negative)/produced(positive)(real). For consumers from the range [-0.5,-2]s^-2; p1 = abs(p2 + p3 + p4)
3. g[x]:.	coefficient (gamma) proportional to price elasticity (real from the range [0.05,1]s^-1). g1 - the value for electricity producer
4. stab	: the maximal real part of the characteristic equation root (if positive - the system is linearly unstable)(real)
5. stabf	: the stability label of the system (categorical: stable/unstable)

R programming language, with additional libraries such as caret machine learning algorithms were used to review the data. Summary and structure of the data are shown in appendix A.

A correlation matrix was run on the data to see if there was any direct correlation between any of the variables. From the plot below, you can see a slight correlation between the tau variables and the dependent stab variable. This is very low though, less than 30%, so no real predictive ability. The 4 p(x) variables do show some correlation amongst themselves, but they do not correlate to out dependent variable: stab. Raw data is shown in Appendix 2.

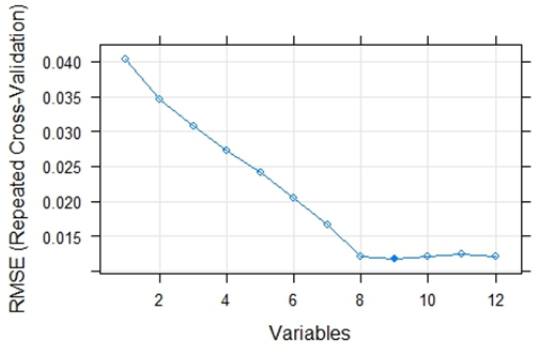


(1) Vadim Arzamasov (vadim.arzamasov '@' kit.edu), http://archive.ics.uci.edu/ml/datasets/Electrical+Grid+Stability+Simulated+Data+

Data was then analyzed using several different Machine Learning algorithms. Near Zero Variance and standard deviations were run on the data to eliminate any instances that did not show values with enough variance to possibly be predictive. None of the instances showed either low standard deviation or NZV score. The data appears to have good distribution at this point and we moved on to the RFE analysis.

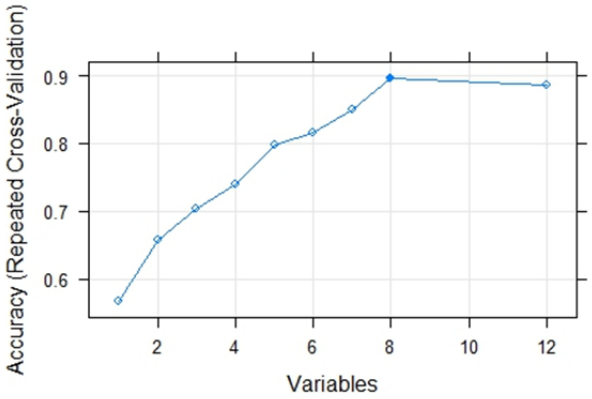
RFE analysis did show some promising analysis. Using variable 13 as dependent regression variable, the RFE analysis showed very low RMSE scores when using 8 through 12 variables, but at cost of very high computing time and oversampling. Your Root Mean Error is very low, but we are not trying to match an exact value, we are looking for a yes / no to the question: is the grid stable. Below is the output graph of # of variables vs RMSE. At 8 variables, the curve levels off. The analysis gave us 9 optimal variables: listed in Appendix C with all the output data.

RMSE output plot



The data frame was then analyzed using a classification model in RFE with all 12 independent variables and "stabf" as the dependent classification. Results show a projected accuracy of 89.7% using 8 optimal variables. Computer time is measured in just a couple minutes, not several hours. We will move forward with those criteria: 8 independent variables, and "stabf" as the independent. Data frame will be reduced to 8 independent variables and 1 dependent variable with all 10,000 instances. Optimal variable selection for this setup includes: [1] "tau1" "tau3" "tau2" "tau4" "g3" "g1" "g4" "g2"

Accuracy plot for classification analysis

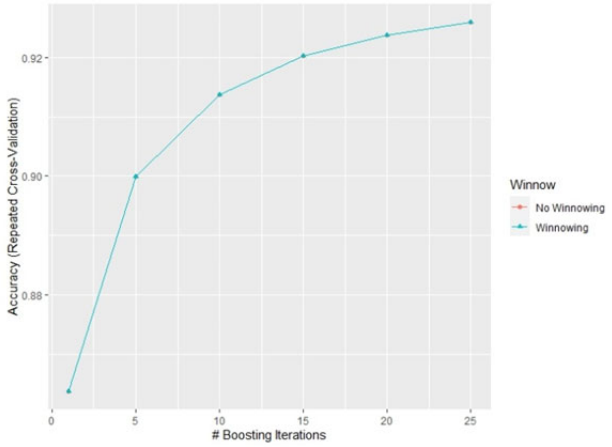


Caret ml algorithms are used next to develop a predictive algorithm. Test and train data frames are developed using a 20/80 split of the data frame. We will use the **c5.0, Random Forest, and SVMLinear2** algorithms and find the best predictive model.

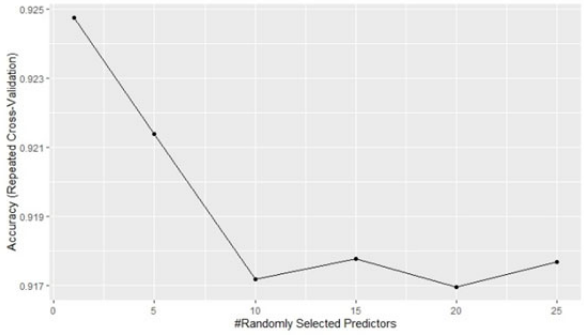
The train set of the data was run through each of the 3 algorithms to train the models. Varying parameters were used to determine the best model set up. Top accuracy, run time, and tuning parameters for the algorithms are listed in the table below. Train Results and confusion matrix output data are listed in appendix D, E, F for C5.0, Random Forest, and SVMLinear2 respectively.

	Top Accuracy	Parameters	Run Tim (min)
C5.0	0.9259	trials = 25, model = tree, winnow = TRUE	5.40
RF	0.9248	mtry = 1	26.22
SVMLinear2	0.8156	cost = 0.25	0.78

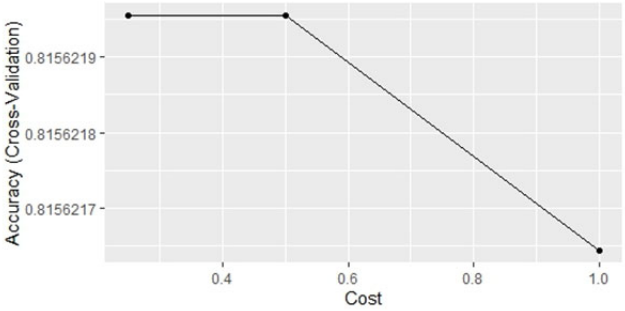
C5.0 Accuracy vs # Iterations
(Winnowing matches no Winnowing)



Random Forest Accuracy vs
# of predictors



SVMLinear2 cost vs Accuracy



After running all the algorithms, the C5.0 algorithm was selected as the best algorithm to use with this data set. The test data set was input into the predict function with the train C5.0 model and generated an accuracy of 0.91. This is slightly below the Random forest accuracy of .92, but has a much quicker run time.

Full output test file containing the 8 independent variables, one dependent variable (stabf) and the predicted dependent variable from the C5.0 algorithm are included in the github submission.

Appendix A
ES Data summary and structure

Structure of Grid data supplied

```
> str(grid_df)
'data.frame':     10000 obs. of  14 variables:
 $ tau1 : num  2.959 9.304 8.972 0.716 3.134 ...
 $ tau2 : num  3.08 4.9 8.85 7.67 7.61 ...
 $ tau3 : num  8.38 3.05 3.05 4.49 4.94 ...
 $ tau4 : num  9.78 1.37 1.21 2.34 9.86 ...
 $ p1   : num  3.76 5.07 3.41 3.96 3.53 ...
 $ p2   : num  -0.783 -1.94 -1.207 -1.027 -1.126 ...
 $ p3   : num  -1.26 -1.87 -1.28 -1.94 -1.85 ...
 $ p4   : num  -1.723 -1.255 -0.92 -0.997 -0.554 ...
 $ g1   : num  0.65 0.413 0.163 0.446 0.797 ...
 $ g2   : num  0.86 0.862 0.767 0.977 0.455 ...
 $ g3   : num  0.887 0.562 0.839 0.929 0.657 ...
 $ g4   : num  0.958 0.782 0.11 0.363 0.821 ...
 $ stab : num  0.05535 -0.00596 0.00347 0.02887 0.04986 ...
 $ stabf: Factor w/ 2 levels "stable","unstable": 2 1 2 2 2 1 2 2 1 2 ... (changed from “Char” to “Factor”)
```

Summary of Grid data supplied

```
> summary(grid_df)
      tau1             tau2              tau3              tau4              p1        
 Min.   :0.5008   Min.   : 0.5001   Min.   :0.5008   Min.   :0.5005   Min.   :1.583  
 1st Qu.:2.8749   1st Qu.: 2.8751   1st Qu.:2.8755   1st Qu.:2.8750   1st Qu.:3.218  
 Median :5.2500   Median : 5.2500   Median :5.2500   Median :5.2497   Median :3.751  
 Mean   :5.2500   Mean   : 5.2500   Mean   :5.2500   Mean   :5.2500   Mean   :3.750  
 3rd Qu.:7.6247   3rd Qu.: 7.6249   3rd Qu.:7.6249   3rd Qu.:7.6248   3rd Qu.:4.282  
 Max.   :9.9995   Max.   : 9.9998   Max.   :9.9994   Max.   :9.9994   Max.   :5.864  
      p2                p3                p4                g1                 g2          
 Min.   :-1.9999   Min.   :-1.9999   Min.   :-1.9999   Min.   :0.05001   Min.   :0.05005  
 1st Qu.:-1.6249   1st Qu.:-1.6250   1st Qu.:-1.6250   1st Qu.:0.28752   1st Qu.:0.28755  
 Median :-1.2500   Median :-1.2500   Median :-1.2500   Median :0.52501   Median :0.52500  
 Mean   :-1.2500   Mean   :-1.2500   Mean   :-1.2500   Mean   :0.52500   Mean   :0.52500  
 3rd Qu.:-0.8750   3rd Qu.:-0.8750   3rd Qu.:-0.8751   3rd Qu.:0.76243   3rd Qu.:0.76249  
 Max.   :-0.5001   Max.   :-0.5001   Max.   :-0.5000   Max.   :0.99994   Max.   :0.99994  
      g3                g4                stab                stabf          
 Min.   :0.05005   Min.   :0.05003   Min.   :-0.08076   Length:10000  
 1st Qu.:0.28751   1st Qu.:0.28749   1st Qu.:-0.01556   Class :character  
 Median :0.52501   Median :0.52500   Median : 0.01714   Mode  :character  
 Mean   :0.52500   Mean   :0.52500   Mean   : 0.01573                     
 3rd Qu.:0.76244   3rd Qu.:0.76243   3rd Qu.: 0.04488                     
 Max.   :0.99998   Max.   :0.99993   Max.   : 0.10940                     
```

Appendix B
ES Pearson Correlation Data

	tau1	tau2	tau3	tau4	p1	p2	p3	p4	g1	g2	g3	g4	stab
tau1	1.000	0.016	-0.006	-0.017	0.027	-0.015	-0.016	-0.016	0.011	0.015	-0.001	0.005	0.276
tau2	0.016	1.000	0.014	-0.002	-0.005	0.007	0.008	-0.006	-0.002	0.015	0.017	-0.012	0.291
tau3	-0.006	0.014	1.000	0.004	0.017	-0.003	-0.009	-0.018	-0.012	0.008	0.015	-0.011	0.281
tau4	-0.017	-0.002	0.004	1.000	-0.003	0.011	0.006	-0.011	-0.004	0.008	0.003	0.000	0.279
p1	0.027	-0.005	0.017	-0.003	1.000	-0.573	-0.585	-0.579	0.001	0.015	0.001	-0.015	0.010
p2	-0.015	0.007	-0.003	0.011	-0.573	1.000	0.002	-0.007	0.016	-0.018	0.008	0.020	0.006
p3	-0.016	0.008	-0.009	0.006	-0.585	0.002	1.000	0.013	-0.003	-0.012	-0.006	-0.010	-0.003
p4	-0.016	-0.006	-0.018	-0.011	-0.579	-0.007	0.013	1.000	-0.014	0.003	-0.004	0.018	-0.021
g1	0.011	-0.002	-0.012	-0.004	0.001	0.016	-0.003	-0.014	1.000	0.008	-0.006	0.012	0.283
g2	0.015	0.015	0.008	0.008	0.015	-0.018	-0.012	0.003	0.008	1.000	-0.013	-0.015	0.294
g3	-0.001	0.017	0.015	0.003	0.001	0.008	-0.006	-0.004	-0.006	-0.013	1.000	0.007	0.308
g4	0.005	-0.012	-0.011	0.000	-0.015	0.020	-0.010	0.018	0.012	-0.015	0.007	1.000	0.279
stab	0.276	0.291	0.281	0.279	0.010	0.006	-0.003	-0.021	0.283	0.294	0.308	0.279	1.000

Legend:
Low correlation, not usable for prediction
Some correlation, but p(x) to p(x) gives no useful info

Appendix C
ES RFE Data

> rfe_grid

Recursive feature selection
Outer resampling method: Cross-Validated (10 fold, repeated 5 times)
Resampling performance over subset size:

Variables	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD Selected
1	0.0404	0.0294	0.033025	0.000625	0.010093	0.0005451
2	0.03452	0.1569	0.028266	0.000685	0.019549	0.0006484
3	0.03078	0.3091	0.024987	0.000622	0.022616	0.0005585
4	0.02723	0.4565	0.021925	0.000516	0.021144	0.0004772
5	0.02405	0.5801	0.019448	0.00041	0.0177	0.0003766
6	0.02051	0.6946	0.016468	0.000397	0.014016	0.0003703
7	0.0166	0.8099	0.01333	0.00031	0.009762	0.0002668
8	0.01199	0.9237	0.009388	0.000247	0.004482	0.0001746
9	0.01166	0.9206	0.009039	0.000241	0.00451	0.0001699
10	0.01202	0.9189	0.009371	0.000245	0.004625	0.000179
11	0.01241	0.9169	0.009733	0.000248	0.004655	0.0001854
12	0.01212	0.9152	0.009444	0.000255	0.004799	0.0001855

The top 5 variables (out of 9) were:
tau2, tau3, tau1, tau4, g3

Appendix D
Output Data C5.0 Train and Confusion Matrix

C5.0 Train output

8000 samples
8 predictor
2 classes: 'stable', 'unstable'
No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 7200, 7200, 7200, 7199, 7200, 7200, ...
Resampling results across tuning parameters:

winnow	trials	Accuracy	Kappa
FALSE	1	0.8637379	0.7037245
FALSE	5	0.8998751	0.7818591
FALSE	10	0.9137261	0.8114081
FALSE	15	0.9202379	0.8258776
FALSE	20	0.9238004	0.8334067
FALSE	25	0.9259123	0.8381394
TRUE	1	0.8637379	0.7037245
TRUE	5	0.8998751	0.7818591
TRUE	10	0.9137261	0.8114081
TRUE	15	0.9202379	0.8258776
TRUE	20	0.9238004	0.8334067
TRUE	25	0.9259123	0.8381394

Tuning parameter 'model' was held constant at a value of tree
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were trials = 25, model = tree and winnow = TRUE.

> confusionMatrix(tst_grid2$stabf, c50predict)

Confusion Matrix and Statistics

Prediction	Reference: stable	Reference: unstable
stable	621	103
unstable	71	1205

Accuracy : 0.913
95% CI : (0.8998, 0.925)
No Information Rate : 0.654
P-Value [Acc > NIR] : < 2e-16
Kappa : 0.8098
Mcnemar's Test P-Value : 0.
Sensitivity : 0.8974
Specificity : 0.9213
Pos Pred Value : 0.8577
Neg Pred Value : 0.9444
Prevalence : 0.3460
Detection Rate : 0.3105
Detection Prevalence : 0.3620
Balanced Accuracy : 0.9093
'Positive' Class : stable

Appendix E
Output Data RF Train and Confusion Matrix

Random Forest train output

8000 samples
8 predictor
2 classes: 'stable', 'unstable'
No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 7199, 7201, 7200, 7200, 7200, 7200, ...
Resampling results across tuning parameters:

mtry	Accuracy	Kappa
1	0.9247618	0.8330906
5	0.9213993	0.8281122
10	0.9171865	0.8189568
15	0.9177862	0.8202917
20	0.9169613	0.8184869
25	0.9176740	0.8200040

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 1.

> confusionMatrix(tst_grid2$stabf, rfpredict)
Confusion Matrix and Statistics
Reference

Prediction	stable	unstable
stable	604	120
unstable	37	1239

Accuracy : 0.9215
95% CI : (0.9088, 0.9329)
No Information Rate : 0.6795
P-Value [Acc > NIR] : < 2.2e-
Kappa : 0.8257
Mcnemar's Test P-Value : 5.977e-11
Sensitivity : 0.9423
Specificity : 0.9117
Pos Pred Value : 0.8343
Neg Pred Value : 0.9710
Prevalence : 0.3205
Detection Rate : 0.3020
Detection Prevalence : 0.3620
Balanced Accuracy : 0.
'Positive' Class : stable

Appendix F
Output Data SVMLearn2 Train and Confusion Matrix

svm_grid2 output

Support Vector Machines with Linear Kernel
8000 samples
  8 predictor
  2 classes: 'stable', 'unstable'
Pre-processing: centered (8), scaled (8)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 7201, 7200, 7201, 7200, 7201, 7199, ...
Resampling results across tuning parameters:

cost	Accuracy	Kappa
0.25	0.8156220	0.5926445
0.50	0.8156220	0.5927126
1.00	0.8156216	0.5927718

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cost = 0.25.

Confusion Matrix and Statistics

Prediction	Reference: stable	Reference: unstable
stable	500	224
unstable	152	1124

Accuracy : 0.812
95% CI : (0.7942, 0.8289)
No Information Rate : 0.674
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.584
Mcnemar's Test P-Value : 0.0002507
Sensitivity : 0.7669
Specificity : 0.8338
Pos Pred Value : 0.6906
Neg Pred Value : 0.8809
Prevalence : 0.3260
Detection Rate : 0.2500
Detection Prevalence : 0.3620
Balanced Accuracy : 0.8003
'Positive' Class : stable