Case Study in Improving the Health of a Remote Monitoring & Diagnostics Center

Sanjeev Heda[1]

[1]*General Electric Power Services Engineering, Atlanta, GA 30339*
sanjeev.heda@ge.com

ABSTRACT

This paper provides a case study where data analytics techniques are used to improve the health of a global remote monitoring & diagnostics (RM&D) center for power plants that is a key part of our industrial internet infrastructure. The "Industrial Internet" is being heralded as a transformative, disruptive technology that is part of the digitization of traditional industries all over the world. A key technical capability includes the ability to continuously monitor critical assets (like power generation equipment) with sensors and other measurements to get leading indicators of anomalous behavior and identify opportunities for optimizing the performance and life of this equipment. This allows our customers to maximize their asset's availability, reliability, and performance and is usually achieved by connecting these assets to dedicated RM&D centers that aggregate and analyze the data streaming in from all over the world.

An RM&D center is a complex system with hardware (data acquisition boxes, communication hardware, processing servers), software and analytics in place to ensure the generation of timely notifications and recommendations. There is a need to understand the health of this complex system and to quickly diagnose and mitigate issues before any disruption occurs that impacts the ability to monitor. What makes this case study unique is the combination of qualitative and quantitative input variables that need to be considered, which is different from traditional PHM applications which tend to be based on sensor derived numeric or binary data.

Details in the paper include extraction of key system health features (server health, data integrity, analytic robustness, etc.) from computer logs. These parameters, once collected, can be analyzed using various statistical techniques (Multivariate Outliers, Process Control, Mixture Distributions) to develop system health indices using methods like Principal Component Analysis, Factor Analysis as well as kernel PCA methods. We explore the development of anomaly detection methods using regression (Multilevel regression, Logistic regression), Clustering (K-Means, Hierarchic, Normal Mixtures and Latent Class Analysis) and Classification (Decision Trees, Bootstrap Forest, Discriminant) techniques followed by their comparison in terms of computational cost and classification accuracy. We then go further to pinpoint the likely cause (Bayesian Diagnosis) of disruptions, and demonstrate how this approach can be generalized for a variety of industrial assets. All of results/analysis are based on representative data from a global Remote Monitoring and Diagnostic center that is used to monitor a significant portion of the world's electric power generation fleet.

1. INTRODUCTION

Heavy duty gas turbines, steam turbines, and generators are widely used in the power generation plants worldwide. Remote Monitoring and Diagnostic (RM&D) services on these assets enable these units to maximize availability and reliability to produce power. A representation of the infrastructure used by GE Power Services RM&D that provides protection to assets can be seen in Figure 1.

Figure 1 shows that the data originates from the sensors on the assets. With the data sent to either a Mark-Series controller or a Distributed Control System (DCS), this can be store locally on the On-Site Monitor (OSM) device. Once available, the data can be transmitted via a firewall for all sites to a central location. Once present, the data is stored in a consolidated number of central historians. With the data present centrally, one service provided for monitoring assets, Orchestration Manager (OM), is available to understand what new data is available and determine which analytics to execute and to provide data to/from these analytics. Once executed, these analytics provide early warning notifications to the RM&D Operations team of abnormalities for a particular customer and to provide

Sanjeev Heda. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

recommendations to avoid any issues that could interrupt operations.

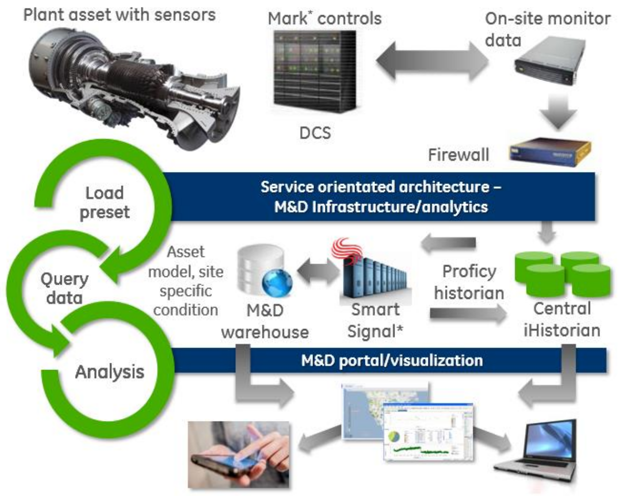


Figure 1. RM&D Architecture

The reliability and on-time execution of this process is critical to provide value-added recommendations. Data latency, calculation failures, system issues, and more could cause recommendations to not occur on time to where the issue would have already affected the customer. For the OM service, failures/slowness to respond could be caused by multiple reasons that can be related to the data repository (central historians), common databases used by analytics, the servers that execute the analytics themselves, or the analytics not behaving properly.

For monitoring the service enabled by analytics, a Sustaining team is assigned to monitoring the health of the system in recurring reports. These reports monitor the execution successes/failures, key parameters on the time to perform certain activities, and the health of the machines running these analytics. These reports help provide visibility into when something has occurred and to start the process of the root-cause of the issue and how to resolve as quickly as possible to minimize downstream impact to customers.

Even though this system operates well and there is a system in place to monitor the health of the system, it is typically retroactive and could last hours/days to identify the issue and provide a resolution. The system is highly complicated and issues could be related to multiple systems that takes multiple people to identify the root-cause in a timely manner. In addition, there could be certain latency in the system itself that may not be obvious until too late where it tips over the edge and becomes catastrophic.

This paper is focused on how to monitor the OM service and related databases and historians effectively to determine when the system is behaving different than before and to be more proactive than reactive. A holistic approach is applied using advance analytic methods on the same data used to monitor the OM service to determine the key features to monitor, statistical analysis for detecting when anomaly has occurred, and predictive diagnostics to quickly root-cause the issue.

2. DATA AND EXPLORATORY ANALYSIS

The following section describes the data obtained for this case study and some of the feature analysis performed on the data to obtain insights.

2.1. Description of Data

For this case study, data was obtained from the system to monitor the health of the system. The following quantitative measurements were captured:

- Execution Interval (seconds) – the amount of time-series data to be processed by an analytic
- Pull Time (milliseconds) – the amount of time to pull all data requested for a given analytic
- Execution Time (milliseconds) – the amount of time to process the data through the analytic
- Historian Write Time (milliseconds) – the amount of time to write resultant data from the analytic to the Historian
- SQL Write Time (milliseconds) – the amount of time to write key information from the analytic to a SQL database
- Insertion Time (milliseconds) – the total amount of time to write all output data into various downstream systems
- # of Successful Executions
- # of Failed Executions

To understand how well the system was performing, a new value was calculated as the % success of executions, which was calculated to be:

$$\%Success = \frac{\#\ of\ Successful\ Executions}{\#\ of\ Total\ Executions} \qquad (1)$$

The average of this data was calculated for a given day and was deconstructed to include the following parameters:

- Server of Execution
- Historian Database
- Analytic Name

This qualitative information is useful to help pinpoint the root cause of an issue when one occurs. For this case study, 6 months' worth of data was obtained for 1 analytic where a known system issue had occurred and the performance of 2 of the historian databases were performing slower than the rest.

2.2. Exploratory Data Analysis - Quantitative

For exploring the quantitative data, only the measurements captured in time are analyzed. Figure 2 shows the distribution of the amount of time to insert data after the analytic had executed.

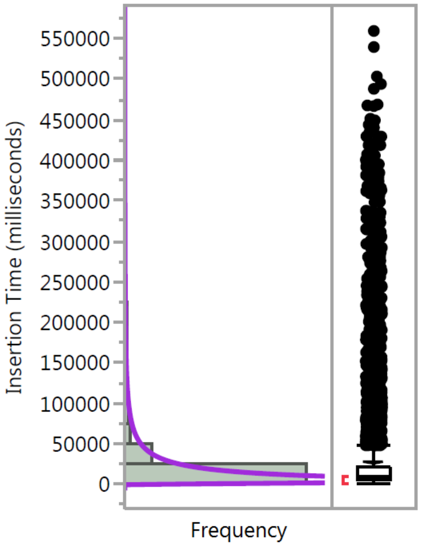


Figure 2. Distribution of Insertion Time

Figure 2 shows that most of the analytic executions had a small amount of time to insert the data with 90% of the data being less than 48,000 milliseconds. Figure 2 also shows that there are many points above this that occurred. To understand when this occurs, Figure 3 shows the amount of time to insert the data over the time.

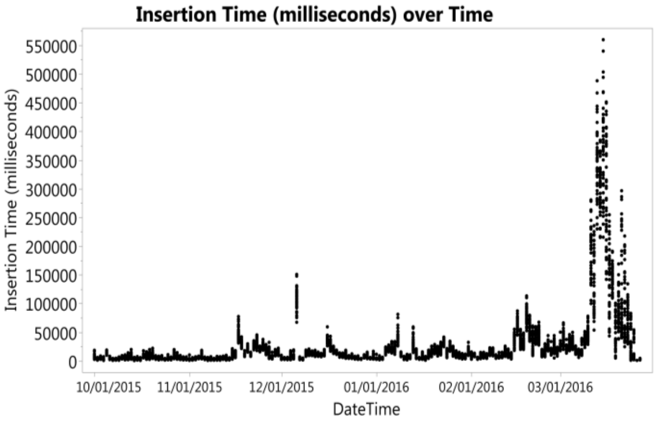


Figure 3. Insertion Time over 6 Month Period

Figure 3 shows that the behavior appears to be normal with some excursions before a significant shift in mid-March 2016. This is the abnormality seen in the system as well. This appears to be true on all of the other quantitative values shown measuring time for this analytic execution.

To understand further of the data distribution, multiple types of distribution were applied to expose the underlying behavior. When doing so, normal mixture models were found to be key to see that the data naturally separated different distributions. For the insertion time, a 3 normal mixture model helps understand the various categorization and behavior of the data. This is shown in Figure 2 with the blue line. This is helpful to know later on when performing a qualitative analysis on this data.

To understand the relationships between the measurements captured, Figure 4 shows the Principle Component Analysis (PCA) is performed for the 5 measurement times captured.

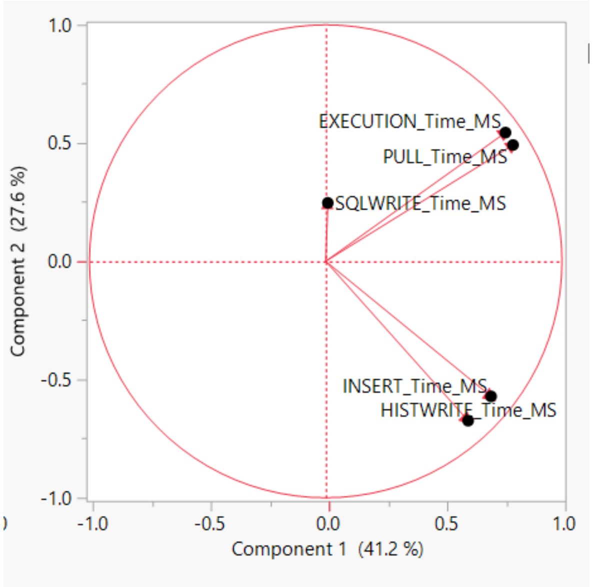


Figure 4. Principle Component Analysis of Measured Values

Figure 4 shows that the Execution / Pull Time and Insertion Time / Historian Write time are highly related to each other. The amount of time to write to SQL is more of an independent variable. Reviewing the Bartlett Test and Scree plot, it appears that the data can be simplified from 5 variables to 3 principle components. Given the low number of variables, this is not required. However, this is useful information to consider in the future of redundant variables when trying to understand when an issue occurs.

Finally, as an exploration of the data, a multivariate hierarchical cluster can be applied to the measured data. This provides additional insight of how the variables are linked and associated in particular when an abnormal issue has occurred. Figure 5 represents the dendogram produced for these clusters

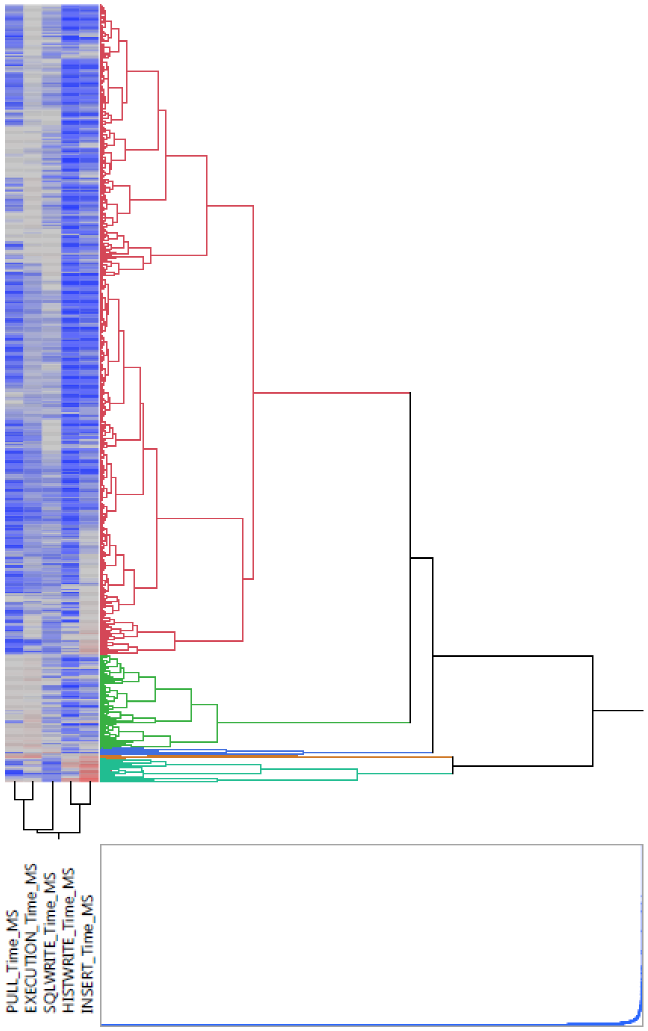


Figure 5. Multivariate Hierarchical Clustering

Figure 5 shows that most of the variation of the data can be best described in 5 clusters. For the majority of the data, there’s variation of execution of analytics that results in low measured times. There appears to be a shift when more time is required to pull the data but the insertion/write times to various databases are not affected. The bottom cluster is of most importance where a significant amount of time is required to process the data for insertion. This clustering is key to help relate variables as segments of operations and to understand which cluster shows abnormality.

This quantitative analysis has helped shown the various behaviors and relationships with the measured values calculated so far. However, none of these take into consideration of how does the machine or source database contribute to being a potential issue. With this, the focus shifts to perform data into a qualitative analysis.

2.3. Exploratory Data Analysis – Qualitative

To perform data into a qualitative analysis and use relevant methods, the measured values need to be assigned a qualitative value. To do and referring to Figure 2, the data can be broken by the underlying normal distribution applied to that measured value. With this for each of the 5 measured values, the values were binned by these distributions. Figure 6 is an example of how the Insertion Time was deconstructed into 3 segments.

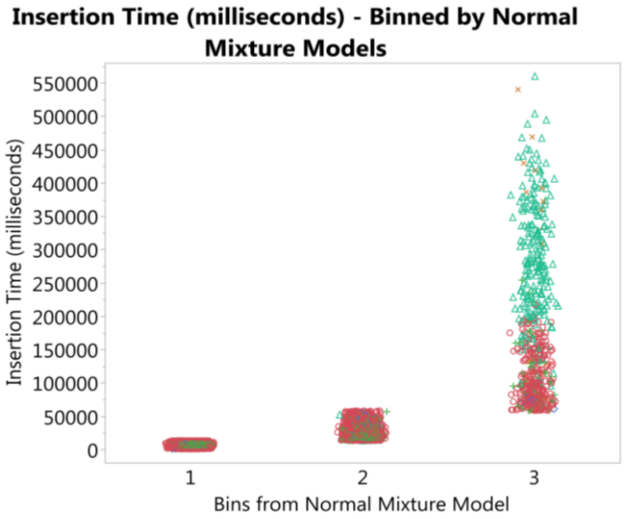


Figure 6. Insertion time binned into 3 levels

Figure 6 shows the insertion time into 3 bins. This method of creating these bins is preferred as it is based on the underlying data behavior and not arbitrary limits set by a person. With this, now we have a total of 7 qualitative variables that can be explored.

For this, the first analysis leveraged is the latent class analysis. Figure 7 shows one of the best results of clustering this data.

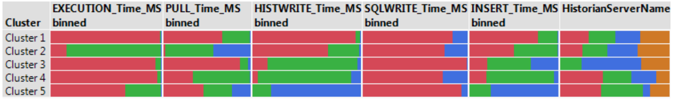


Figure 7. Latent Class Analysis results

Figure 7 shows some interesting insight in the behavior. For cluster #3, when the pull time takes longer than required, there is real no noticeable effect of the rest of the components. For cluster #4 when the insertion time is significantly longer than the rest, there is more of a contribution/impact related to 2 of the 4 databases that stored the data (as seen in Historian Server Name). There is real no difference in the operation of the machines running the analytic across all of the clusters. This helps provide some evidence that the historian server is related to the performance of the system.

Figure 8 shows a Multiple Correspondence Analysis can also be performed to show the relation of these variables to others similar to the PCA done previously.

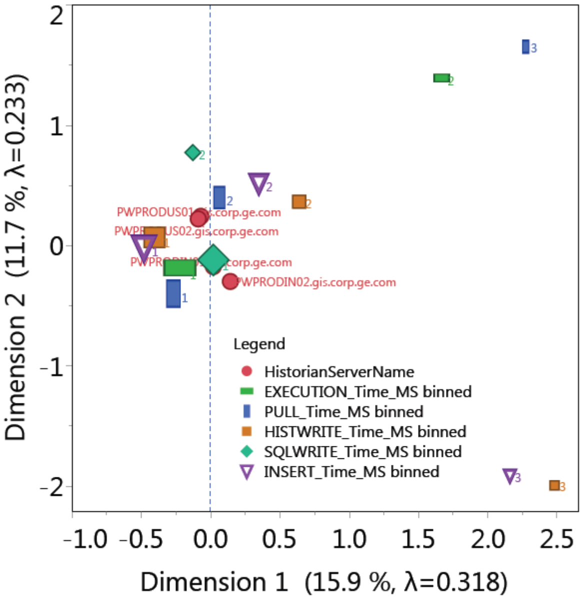


Figure 8. Multiple Correspondence Analysis

This analysis shows similar behaviors that we saw previously in the PCA, where the insertion/write times to databases and the amount of time to pull data/execute the analytic are related. Closer in the center, there is a distinct separation between the 4 historian databases separating into 2 populations. For the machines where the analytics execute, there appears to be no difference in relationship to these machines vs. others.

All of this analysis is helpful to visualize how is the system behaving today and how the parameters (both quantitatively and qualitatively) relate to each other. To be proactive, an anomaly has to be identified using this data to help with understanding when to do such an analysis on the data.

3. STATISTICAL ANOMALY DETECTION

To understand the health of the system and how well is it operating today versus previously, there are multiple multivariate techniques that can combine parameters to determine abnormality. These could be techniques including Hoteling-T^2, Mahalanobis distances, Jack-Knife, etc. For this analysis, based on literature reviews and industry practice, the multivariate Mahalanobis distance statistics was selected to detect abnormality.

For the Mahalanobis analysis, typically a limit can be calculated from the results that can be used as a fixed threshold. However, as seen previously when analyzing the measured data, it is beneficial to provide more insight of how the data could be behaving differently than before. Figure 9 shows the distribution of the Mahalanobis distances calculated for this data set using the measured values. To transform this data from quantitative to qualitative, the underlying distribution is analyzed. It is seen that a 3 Normal Mixture model would best fit the data (as indicated in Figure 9 with the blue line).

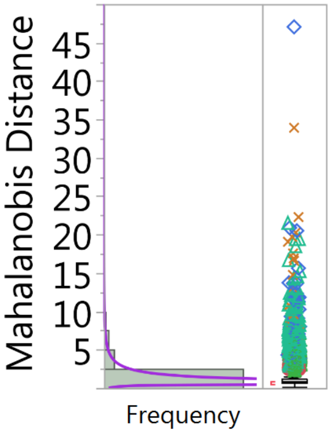


Figure 9. Mahalanobis distance histogram

With this understanding, data can be classified into the 3 regions. With this, Figure 10 shows the Mahalanobis distance calculated over time with the color coding of the bin where the data falls into.

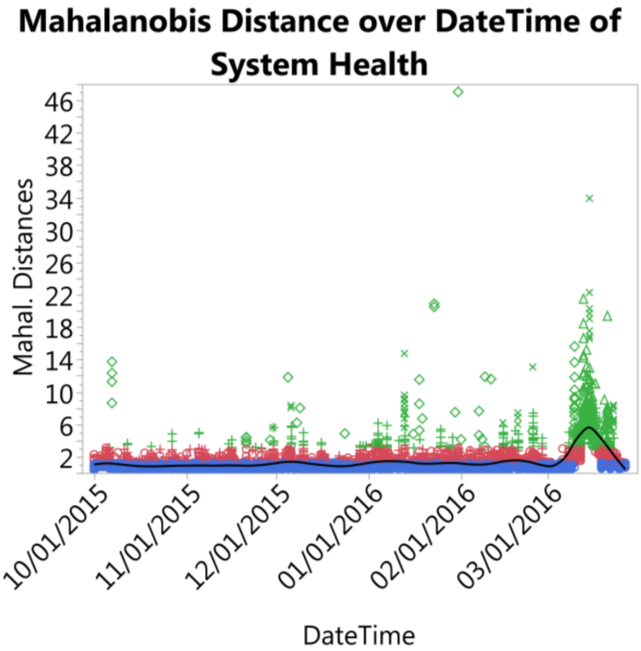


Figure 10. Mahalanobis Distance with Normal Mixture Model Color Coding

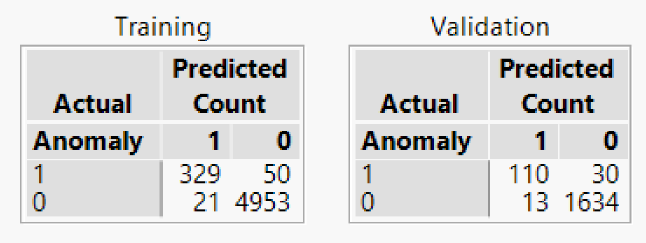
Figure 10 provides much insight into when an anomaly could have occurred. Over this time period, the distances show that the system was behaving fairly normal with some excursions to higher levels that was different than normal. As the calculation approaches to early/mid-March, there is a sudden increase in the frequency and the distance itself to indicate the system is behaving significantly differently. This is useful and can be used to help provide the initial alert to downstream users that something has changed in the system and the potential severity. An alert can be created by combining both the persistence and severity of the change. However, with this notification, there is still a need to investigate to understand what is the reason for the change. There needs to be predictive diagnostics on the system to help quickly pinpoint where the issue is.

4. PREDICTIVE DIAGNOSTICS

Once an anomaly has been detected within the operation of the system, the next step is to identify what has changed and what actions are necessary, if any, to remediate any issues. With Figure 10, anomalous behavior can be seen when applying multivariate statistical analysis and to deconstruct the data into 3 separate levels based on the distribution of the data. With this, when the data is considered to be in the "highest" level, this will be considered anomalous.

With this classification, models on the data can be applied to understand what helps differentiate between normal/abnormal behavior. For this data set, there is 7% of anomalous behavior and 93% of normal behavior in the system. Various models can be used to help predict this behavior and apply key insights of what is driving this.

Following standard modeling approaches, the first model applied is ordinal logistic regression on the qualitative data. All of the data applied to the model is qualitative so that all data can be treated equally and give insight of what the potential issue is. The model is created with the measured values listed above that has been converted into a qualified measurement, the machine name, and the historian database. A validation data set of 30% was applied in the creation of the models. Figure 11 shows the confusion matrix of the model for both testing and validation



Training

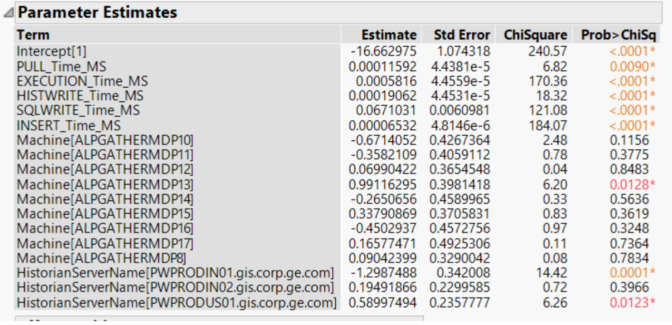
	Predicted Count	
Actual Anomaly	1	0
1	329	50
0	21	4953

Validation

	Predicted Count	
Actual Anomaly	1	0
1	110	30
0	13	1634

Figure 11. Ordinal logistic regression confusion matrix

The resultant model created had a 0.0241 misclassification rate in the validation data set. The parameters used in creating the model (Figure 12) provide additional interesting insights.



Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[1]	-16.662975	1.074318	240.57	<.0001*
PULL_Time_MS	0.00011592	4.4381e-5	6.82	0.0090*
EXECUTION_Time_MS	0.0005816	4.4559e-5	170.36	<.0001*
HISTWRITE_Time_MS	0.00019062	4.4531e-5	18.32	<.0001*
SQLWRITE_Time_MS	0.0671031	0.0060981	121.08	<.0001*
INSERT_Time_MS	0.00006532	4.8146e-6	184.07	<.0001*
Machine[ALPGATHERMDP10]	-0.6714052	0.4267364	2.48	0.1156
Machine[ALPGATHERMDP11]	-0.3582109	0.4059112	0.78	0.3775
Machine[ALPGATHERMDP12]	0.06990422	0.3654548	0.04	0.8483
Machine[ALPGATHERMDP13]	0.99116295	0.3981418	6.20	0.0128*
Machine[ALPGATHERMDP14]	-0.2650656	0.4589965	0.33	0.5636
Machine[ALPGATHERMDP15]	0.33790869	0.3705831	0.83	0.3619
Machine[ALPGATHERMDP16]	-0.4502937	0.4572756	0.97	0.3248
Machine[ALPGATHERMDP17]	0.16577471	0.4925306	0.11	0.7364
Machine[ALPGATHERMDP8]	0.09042399	0.3290042	0.08	0.7834
HistorianServerName[PWPRODIN01.gis.corp.ge.com]	-1.2987488	0.342008	14.42	0.0001*
HistorianServerName[PWPRODIN02.gis.corp.ge.com]	0.19491866	0.2299585	0.72	0.3966
HistorianServerName[PWPRODUS01.gis.corp.ge.com]	0.58997494	0.2357777	6.26	0.0123*

Figure 12. Ordinal logistic regression parameters

Figure 12 shows that all of the measured values are important to consider. When looking at the machine and historian database, it is interesting to see that one historian, PWPRODIN01 stands out as significant. Some potentially abnormal behavior can be seen on another historian database and one machine name. This model gives us insights of what drives the behavior and can help direct remediation activities.

In this example, we have demonstrated results using a standard method, logistic regression, which provided adequate results and actionable insights. However, in practice we have seen problems where additional classifiers like neural networks (with sensitivity analysis), tree-based methods (Classification and Regression Trees, Random Forests) and regularized regression techniques (Ridge Regression, Lasso and Elastic Net applied to a Binomial model with quantitative predictors) were applied, and provided good results. In this case, the additional computational costs and model complexities were not warranted as a simple logistic model was adequate,

5. CONCLUSION

This paper introduces the idea that methods for monitoring computing assets need to include data-driven statistical classifiers that include both quantitative and qualitative predictors. Most data-driven methods tend to neglect qualitative methods, and miss out on insights that can be obtained using techniques like Latent Class Analysis clustering and Multiple Correspondence Analysis. These are complementary to traditional classifiers and we recommend they be included in standard diagnostics toolkits.

In general, the use of statistical analysis provides visibility, insight, and actions into the health of our monitoring & diagnostics system, which in turn, increases opportunities to maximize customer reliability and availability in the Power Generation business.

The data collected in the system today can provide key insights into how the system is performing. The data needs to be reviewed over long-time periods and be aggregated in a way to include both measured parameters as well as metadata. The data exploratory analysis shows usable techniques into understanding the behavior of the system and how the variables relate to each other. New techniques to convert measured values to qualified variables to apply segmentation analysis such as latent class and multiple correspondence analysis enables insight into how these variables relate to the metadata.

Once the data was explored, multivariate techniques was identified as a useful way to monitor the normality of the system and the ability to create alarms in the system to when the system has changed significantly. This alarm can include both the severity and persistence to alert the appropriate people. Once identified, models can be applied to the data to both confirm that the abnormality can be modeled and what are the driving parameters. The driving parameters can be used as an initial description of the cause of the issue to help quickly identify and resolve the issue.

ACKNOWLEDGEMENT

The authors of the paper would like to acknowledge data, domain insights and general support provided by Power Generation GE Fleet Services and tools support (Beta versions of JMP/Pro V13) provided by the JMP division of SAS Institute. All data has been anonymized to protect proprietary information.

NOMENCLATURE

PCA	Principal Component Analysis
LCA	Latent Class Analysis
MCA	Multiple Correspondence Analysis
DCS	Distributed Control System
RM&D	Remote Monitoring & Diagnostics

REFERENCES

Gardner S., and Jayson G., Building Better Models with JMP Pro, (2015) SAS Institute

Jain, R., (1991), The Art of Computer Systems Performance Analysis. Wiley

Trivedi, K. S., and Sahner, R. (1995), Performance and Reliability Analysis of Computer Systems, Wiley

Trivedi, K.S., (2001), Probability and Statistics with Reliability, Queuing and Computer Science Applications, (Wiley)

BIOGRAPHY

Sanjeev Heda is the Engineering Technical Leader for Data Science & Analytics, Global Fleet Services department of GE Power Services. He has been with GE since 2007. He leads a global team of data scientists and engineers tasked with developing new analytics to increase customer reliability, availability, and performance and to provide fleet intelligence to determine emerging trends amongst the fleet. Sanjeev has a BS and MS in Mechanical Engineering from Georgia Institute of Technology and currently resides in Atlanta.