ANOMALY DETECTION ON ELECTRIC POWER GRID DATA SET

ANTONIO LEUNG, JOHNY KUANG, FANG CHI (AMANDA) CHANG, TRUNG (LEON) TRIEU
Student No. 301006413, 301174274, 301278767, 301339741
acl3@sfu.ca, johnyk@sfu.ca, fcchang@sfu.ca, thtrieu@sfu.ca

1. ABSTRACT

This project consists of analysis and data exploration of a normal data set compiled from the monitoring of household power consumption. The data exploration is then followed by a series of anomaly detection techniques performed, on a test data set, in order to find any suspicious behavior. Some of the anomaly detection techniques used in this project include finding point anomalies which consists of methods such as finding points in the data set that are out of the range and the moving average. Contextual anomalies, in the test data set, are then discovered by building Hidden Markov Models and calculating its log-likelihood.

2. TABLE OF CONTENTS

3.INTRODUCTION

Interconnectivity is a global phenomenon that continues to be relied upon as society continues to advance. Even critical infrastructures nowadays depend on interconnected systems for the sake of efficient communication, along with other obvious advantages, when it comes to supervisory management. With advantages, comes disadvantages. The ability to be connected also arises the vulnerability for systems to be attacked. The total attack surface has never been larger, and the possibility of a major breach, of a critical infrastructure is not particularly inconceivable. One successful phishing attempt can be all it takes.

The U.S electric power grid is no exception to this phenomenon. Thousands of power companies are interconnected. Moreover, they all run and have the same control system architecture: supervisory control and data acquisition (SCADA) systems. The high-level idea of having similar systems was to maximize efficiency and simplicity for users; the negative by-product of this idea was the increase of simplicity for attackers to move laterally from system to system after infiltration. With attackers, and the types of attacks, getting more and more sophisticated, cybersecurity is shifting its detection practices from signature-only to behavior-based detection; this allows for higher coverage of attacks and less limitations. One of these methods is to find patterns in data that do not conform to expected behavior; this is also known as anomaly detection. With the availability of a dataset, compiled from the monitoring of household power consumption, it is possible for us to find and detect anomalies in the systems of the electric power grid.

The report is organized as follows: the first few sections will focus on the background of the problem, as well as definitions and methodology required before approaching the problem. The latter half of the report will describe our approach to the problem, our proposed methods and the implementation of them, as well as our results.

4.BACKGROUND

4.1 WHAT IS AN ANOMALY

Anomalies and outliers are the two most common terms that are used in anomaly detection, but can also be called discordant observations, exceptions, surprises, peculiarities or contaminants. There are three main categories of anomalies, point anomalies, contextual anomalies and collective anomalies. Point anomalies can be found when an individual data instance does not fit in with respect to the rest of the data. This is the simplest type of anomaly and is the focus of majority of research on anomaly detection (Chandola, Banerjee, & Kumar, 2009). For example, a credit card fraud detection, if the amount spend on a single transaction on a particular day is very high compared to the usual range of expenditure, that transaction is considered as a point anomaly. Contextual anomalies can be found when the data instance is anomalous in a specific context but normal in other situations (Chandola, Banerjee, & Kumar, 2009). Collective anomaly can be found when a particular set or a particular sequence of data instances is anomalous compared to the entire data set. The individual data instances in a collective anomaly themselves are not anomalies, only when it is in a certain occurrence together as a set is considered anomalous.

4.2 WHAT IS ANOMALY DETECTION

Anomaly detection refers to the process of finding patterns that do not conform well to the expected normal behavior. Anomaly detection is widely used in things such as fraud detection for credit cards, intrusion detection for cybersecurity, and criminal intelligence. There are three general techniques of anomaly detection supervised anomaly detection, semi-supervised anomaly detection and unsupervised anomaly detection.

Techniques trained in supervised mode assume the availability of a training data set that has labeled instances for normal as well as anomaly classes (Chandola, Banerjee, & Kumar, 2009). A predictive model for normal and anomaly classes is built, and any data that have not been seen is compared to the model to see which class it belongs to. Techniques that operate in a semi-supervised mode assumes that the training data does not contain any anomalies; it only contains normal data and possibly noise. The general approach used in this technique is to build a model for normal behaviour, and use the model to find anomalies in the test data. Techniques that are used in an unsupervised mode assumes that majority of the data instances are normal, and less anomalies. But unfortunately, if this assumption is

false, then this technique will suffer from high false alarm rate (also known as false positives, something that is incorrectly identified).

4.3 WHAT IS A MARKOV MODEL, HIDDEN MARKOV MODEL (HMM), UNIVARIATE & MULTIVARIATE HMM

A Markov model is a stochastic model used to monitor random changing systems where we assume it has the Markov property, which is when the future states depends upon the current states and not the states that precede it. A hidden Markov model implies that the Markov Model underlying the data is hidden or unknown to the observer.[1] This Markov model is "hidden" because the observer does not know how many states there are and can only guess the number of states by analyzing the sequence of observational data. A univariate hidden Markov model, is one emission corresponds to one variable, whereas in a multivariate hidden Markov model, one emission corresponds to several variables.[2]

4.4 WHAT ARE BIC AND LOG-LIKELIHOOD

Bayesian Information Criterion(BIC) is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup.[3] In this project, the likelihood ratio test was used to help compare the fit of two statistical models (our training and validation HMM, and also with the test data), this ratio expresses how many times more likely the data is under one model than the other. A high ratio value corresponds to a better fit whereas a smaller ratio meant a worse fit; ratio values range from 0 to 1.[4]

[1] http://www.statisticshowto.com/hidden-markov-model/

[2] https://www.quora.com/What-is-the-difference-between-a-univariate-and-multivariate-hidden-Markov-model

[3] https://methodology.psu.edu/AIC-vs-BIC

[4] https://en.wikipedia.org/wiki/Likelihood-ratio_test#Interpretation

5. PROBLEM DEFINITION

In this project, we are given two data sets: a training data set that is assumed to be “normal” and free of anomalies, and a test data set that contains anomalies.

The training set contains real data that is compiled from monitoring household power consumption for some parts of the U.S electrical power grid, the data is split in a one-minute sampling rate from December 16, 2006 to December 1, 2009. The goal is to be able to differentiate between anomalous and normal electrical power usage, when given another set of the U.S electrical power grid data.

6. DATA SET DESCRIPTION

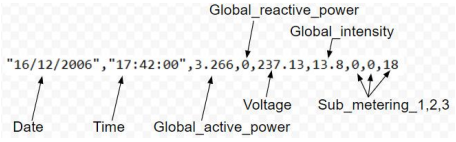


Figure 6.1 Sample instance from the data set

• Date, a character string in format dd/mm/yyyy.

• Time, a character string in format hh:mm:ss.

• Global_active_power, household global minute-averaged active power (in kilowatt).

• Global_reactive_power, household global minute-averaged reactive power (in kilowatt).

• Voltage, minute-averaged voltage (in volt).

• Global_intensity, household global minute-averaged current intensity (in ampere).

• Sub_metering_1, energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).

• Sub_metering_2, energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.

• Sub_metering_3, energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

7. METHODOLOGY

In this project since the dataset given is a normal data set with noise, a semi-supervised anomaly detection is favorable since the normal data set is labeled. This data is valuable for the creation of Markov model that represents normal behaviour, which could be compared to the unseen data in order determine if the data is anomalous or normal. The process of training the hidden Markov model (HMM) consisted of the analyzation of a specific time frame, the BIC and log-likelihood of the training and validation data. Each specified time frame data is split into two parts, 80% of data for training our model and 20% for validation data. The Markov model are trained using the depmix function where the parameters, response, data, and ntimes are specified. Response is a list of response models, for this project the response model is the Global active power. Data is the data that you want to use for the Markov model and ntimes is a vector used to separate the lengths of independent time series.[5]

Time frames are specified due to potential fluctuation of electricity consumption of households at different times of the day and different days of the week. With the time frames narrowed down, we can further classify the normal behaviour and have a trend that is consistent and unique. In our Markov model, an ideal log-likelihood is one that is higher since a greater likelihood of an observation sequence implies that the given HMM generates the particular sequence more likely than the other sequence. But there is no definite value to what a "good" log-likelihood should be, it is a value that is to be compared to one another. In general, a ratio that is as close to 1 is most desired, because that means the training log-likelihood and validation log-likelihood are relatively similar which implies a similar trend in the values. The BIC value needs be a lower value, since a lower BIC means that a model is considered to be more likely to be the true model.[6] These two values are used to rank which models are a well trained model, with a low BIC and log-likelihood, but we have to be careful to not overfit the model. If there is a model with a really low BIC and log-likelihood compared to the other models, there could be a chance our model is overfitted, which is not what we want. After determining the best model, the test data will be compared to the model to see if there are any anomalies.

[5] https://cran.r-project.org/web/packages/depmixS4/depmixS4.pdf
[6] https://methodology.psu.edu/AIC-vs-BIC

8. EXPERIMENTS, ANALYSIS, AND RESULTS

8.1 GENERAL DATA EXPLORATION

8.1.1– Checking for overall changes relative to the expected normal behaviour.

In order to help decide on what type of HMM to train (univariate or multivariate), an important aspect was to look at each feature individually as well as their relationships. The latter would be further explained in 8.1.2 – Understanding feature correlation. Three different time frames were chosen from the training data for comparison: 2007-2008 Weekends (Sat and Sun) 7am-3pm, 2007-2008 Mon, Wed, Fri 4pm-8pm, and 07-08' Monday 12pm-8pm.

Feature 1: Global active power

Global_active_power	Global_active_power	Global_active_power
Min. :0.0780	Min. :0.0780	Min. :0.0780
1st Qu.:0.5322	1st Qu.:0.3980	1st Qu.:0.4250
Median :1.3120	Median :0.5898	Median :0.6606
Mean :1.4254	Mean :0.9532	Mean :1.0296
3rd Qu.:1.9207	3rd Qu.:1.3680	3rd Qu.:1.4868
Max. :8.1660	Max. :9.5900	Max. :7.5660

Figure 8.1 : Global active power in different time frames for 2007-2008, Weekend 7am-3pm, Monday, Wednesday, Friday 4pm-8pm, and Monday 12pm-8pm respectively

The first feature being explored was global active power. As shown from Figure 8.1, the minimum for all three time frames were identical. However, the median, mean, and maximum varies. This indicate that when choosing the timeframe for the HMM, if global active power was used, the time frame chosen could significantly affect the outcome of the model. The large fluctuation in the values of global active power and the range (maximum versus minimum) is quite large therefore, this feature would be a good candidate to include for training the HMM. The fluctuations are visualized in a graphical representation shown in Figure 8.2.

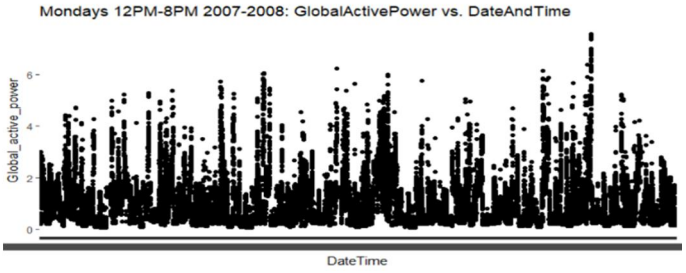


Figure 8.2 Global active power 2007-2008 Monday 12pm-8pm

To further visualize the fluctuation of global active power, the training dataset was separated into first 80% and last 20%, and compare the values to the results from the testing dataset using the same time windows.

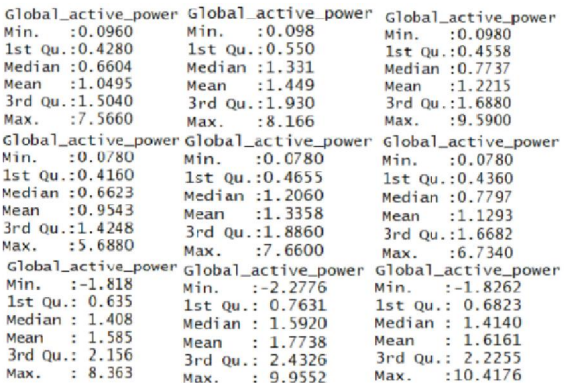


Figure 8.3 Global active power in 3 different time windows in 2007-2008, 1st column: Monday 12pm-8pm, 2nd column: Weekends 7am-3pm, 3rd column: Mon, Wed, Fri 4pm-8pm

From Figure 8.3, looking at the first two rows, which were the two partitions of the training data set in different windows, showed that the minimum was quite consistent. However, the maximum differed in all cases. More importantly, all the cases indicated that the range of the global active power was quite large. To understand the fluctuation to a better extent, the standard deviation for each case was calculated. In general, small standard deviation signaled that data are clustered around the mean.[7]

Figure 8.4 Standard Deviation (1) for each time window related to Figure 8.3

For example, using the standard deviation for the first time window, with 0.9001173 as 1 standard deviation from the mean, this translate to 85.766% (0.9001173 / 1.0495) change. This showed a huge change in the value with just 1 standard deviation. This percentage of change will be used to compare with the percentage change in other features.

Feature 2: Voltage

In order to use the results in some meaningful way that would help decide what data set to use to train the HMM, a new approach in partitioning the data was to partition the training dataset into 2 sets (first 80% and last 20% of the training data). In a way, this would represent our training data set compared to the validation data set.

Figure 8.5 Voltage 2007-2008 1st set of 3: Monday 12pm-8pm, 80% training data, 20% training data, and testing data respectively
2nd set of 3: Weekends 7am-3pm, 80% training data, 20% training data, and testing data respectively, 3rd set of 3: Mon, Wed, Fri 4pm-8pm, 80% training data, 20% training data, and testing data respectively

[7] https://www.pacificpower.net/ya/po/otou/ooh.html

The first takeaway from the results was that the three different time windows resulted in similar values for minimum, median, mean, and maximum. Looking at the time windows separately, the minimum slightly increased, but the % change (I.e. from 226.3 to 230.0) was not significant. The maximum was consistent. One of the key results that was consistent with all three time windows was that the maximum for testing was a lot higher in all three cases (421.2, 412.4, and 421.2) compared to both the testing and validation dataset. Another approach was looking at the standard deviation of the chosen time windows.

```
> sd(play1train$Voltage, na.rm = TRUE )
[1] 2.991427
> sd(play1validate$Voltage, na.rm = TRUE)
[1] 2.583324
> sd(play1$Voltage, na.rm = TRUE )
[1] 4.284597
> sd(play1train$Voltage, na.rm = TRUE )
[1] 2.546652
> sd(play1validate$Voltage, na.rm = TRUE)
[1] 2.681236
> sd(play1$Voltage, na.rm = TRUE )
[1] 3.862377
> sd(play1train$Voltage, na.rm = TRUE )
[1] 2.989401
> sd(play1validate$Voltage, na.rm = TRUE)
[1] 2.513686
> sd(play1$Voltage, na.rm = TRUE )
[1] 4.290695
```

Figure 8.6 Voltage Std Dev 3 different time windows, each column represent a time window, 80% training dataset, 20% training dataset, and testing dataset respectively

The results showed that the first and third time windows provided similar results for standard deviation. Nonetheless, the outcomes for training dataset were all between 2.50-3.00 standard deviation. Similar to global active power, it was important to compare the standard deviation with the mean, to see what the percentage change was. For example, in the first case where 1 standard deviation was 2.991427, the mean for the same time window was 240.5. This meant that for one standard deviation, there was a 1.244% change (2.991427 / 240.5). Comparing this with the percentage change of global active power (85.766%) for the same time window showed that the fluctuation or change for voltage was significantly less although the value itself was a lot bigger (I.e. mean for global active power in the same time window was 1.0495, whereas the mean for voltage was 240.5).

Feature 3: Global reactive power

```
Global_reactive_power  Global_reactive_power  Global_reactive_power
Min.   :0.0000         Min.   :0.0000         Min.   :0.0000
1st Qu.:0.0000         1st Qu.:0.0460         1st Qu.:0.0480
Median :0.1060         Median :0.1040         Median :0.1100
Mean   :0.1277         Mean   :0.1288         Mean   :0.1366
3rd Qu.:0.2060         3rd Qu.:0.2020         3rd Qu.:0.2140
Max.   :1.0060         Max.   :1.2180         Max.   :1.0360
NA's   :145            NA's   :1477           NA's   :244
Global_reactive_power  Global_reactive_power  Global_reactive_power
Min.   :0.0000         Min.   :0.0000         Min.   :0.0000
1st Qu.:0.0540         1st Qu.:0.0520         1st Qu.:0.0600
Median :0.0880         Median :0.1100         Median :0.1120
Mean   :0.1217         Mean   :0.1364         Mean   :0.1543
3rd Qu.:0.1860         3rd Qu.:0.2080         3rd Qu.:0.2260
Max.   :0.7980         Max.   :0.8480         Max.   :0.9180
NA's   :6              NA's   :45             NA's   :237
Global_reactive_power  Global_reactive_power  Global_reactive_power
Min.   :0.0000         Min.   :0.0000         Min.   :0.0000
1st Qu.:0.0580         1st Qu.:0.0620         1st Qu.:0.0600
Median :0.1040         Median :0.1180         Median :0.1080
Mean   :0.1358         Mean   :0.1486         Mean   :0.1433
3rd Qu.:0.2060         3rd Qu.:0.2200         3rd Qu.:0.2140
Max.   :0.8840         Max.   :0.8700         Max.   :1.0180
NA's   :480            NA's   :2800           NA's   :960
```

Figure 8.7 Global reactive power in 3 different time windows in 2007-2008, 1st column: Monday 12pm-8pm, 2nd column: Weekends 7am-3pm, 3rd column: Mon, Wed, Fri 4pm-8pm

```
> sd(play1train$Global_reactive_power, na.rm = TRUE) > sd(play1train$Global_reactive_power, na.rm = TRUE) > sd(play1train$Global_reactive_power, na.rm = TRUE)
[1] 0.1168215                                       [1] 0.118931                                        [1] 0.1235306
> sd(play1validate$Global_reactive_power, na.rm = TRUE)> sd(play1validate$Global_reactive_power, na.rm = TRUE) > sd(play1validate$Global_reactive_power, na.rm = TRUE)
[1] 0.1085462                                       [1] 0.1197451                                       [1] 0.1413387
> sd(play1$Global_reactive_power, na.rm = TRUE) > sd(play1$Global_reactive_power, na.rm = TRUE) > sd(play1$Global_reactive_power, na.rm = TRUE)
[1] 0.1161446                                       [1] 0.1245018                                       [1] 0.1216181
```

Figure 8.8 Global reactive power Std Dev 3 different time windows, each column represent a time window, 80% training dataset, 20% training dataset, and testing dataset respectively

One thing that was consistent across all the cases was that the minimum value was 0.00. In regards to the maximum, the value was higher for the latter 20% of the training dataset compared to the 1st 80% of the training dataset. When compared to the testing dataset (3rd row), the third time window seemed to better represent the testing dataset. Since the mean of each time window for the training dataset range from 0.1277 to 0.1543, with the standard deviation range from 0.1086 to 0.1413, the percentage change for all cases were around 100% change for 1 standard deviation from the mean. However, one key thing to note was that the range was generally only between 0 to 1. Another thing to note was that the latter (last 20%) of the training dataset better represent the test dataset where the maximum and mean are closer in values. The value of 1 standard deviation seemed to be the closest with the first time window between the training dataset and testing dataset.

Feature 4: Global intensity

```
Global_intensity  Global_intensity  Global_intensity
Min.   : 0.200    Min.   : 0.200    Min.   : 0.200
1st Qu.: 1.400    1st Qu.: 1.600    1st Qu.: 1.400
Median : 2.200    Median : 5.200    Median : 3.000
Mean   : 4.044    Mean   : 5.377    Mean   : 4.912
3rd Qu.: 5.800    3rd Qu.: 7.000    3rd Qu.: 6.800
Max.   :34.000    Max.   :39.000    Max.   :41.200
NA's   :145       NA's   :1477      NA's   :244
Global_intensity  Global_intensity  Global_intensity
Min.   : 0.400    Min.   : 0.400    Min.   : 0.600
1st Qu.: 1.400    1st Qu.: 1.800    1st Qu.: 1.800
Median : 2.200    Median : 5.200    Median : 4.200
Mean   : 3.587    Mean   : 5.399    Mean   : 5.145
3rd Qu.: 5.400    3rd Qu.: 7.400    3rd Qu.: 7.000
Max.   :22.800    Max.   :34.400    Max.   :33.000
NA's   :6         NA's   :45        NA's   :237
Global_intensity  Global_intensity  Global_intensity
Min.   : 0.600    Min.   : 0.600    Min.   : 0.600
1st Qu.: 1.600    1st Qu.: 1.600    1st Qu.: 1.600
Median : 3.200    Median : 5.400    Median : 3.400
Mean   : 4.547    Mean   : 5.282    Mean   : 4.638
3rd Qu.: 6.200    3rd Qu.: 7.200    3rd Qu.: 6.400
Max.   :32.200    Max.   :39.400    Max.   :32.200
NA's   :480       NA's   :2800      NA's   :960
```

Figure 8.9 Global intensity in 3 different time windows in 2007-2008, 1st column: Monday 12pm-8pm, 2nd column: Weekends 7am-3pm, 3rd column: Mon, Wed, Fri 4pm-8pm

```
> sd(play1train$Global_intensity, na.rm = TRUE) > sd(play1train$Global_intensity, na.rm = TRUE) > sd(play1train$Global_intensity, na.rm = TRUE)
[1] 3.969414                                    [1] 4.688528                                    [1] 4.661781
> sd(play1validate$Global_intensity, na.rm = TRUE)> sd(play1validate$Global_intensity, na.rm = TRUE) > sd(play1validate$Global_intensity, na.rm = TRUE)
[1] 3.11436                                     [1] 4.433668                                    [1] 4.189148
> sd(play1$Global_intensity, na.rm = TRUE) > sd(play1$Global_intensity, na.rm = TRUE)> sd(play1$Global_intensity, na.rm = TRUE)
[1] 3.947017                                    [1] 4.432422                                    [1] 3.832028
```

Figure 8.10 Global intensity Std Dev 3 different time windows, each column represent a time window, 80% training dataset, 20% training dataset, and testing dataset respectively

Reading the results from Figure 8.9 showed that global intensity had a wide range (0.2 to 34.4 for training dataset). The maximum value had decreased for all three time windows between the first 80% of training dataset to last 20% of training dataset. This was also reflected in the standard deviation. For example, in the first time window, the first 80% of the training dataset had a 3.969 value for 1 standard deviation, and decreased to 3.114 for the last 20%.

Comparing the training data (first 80%) to the testing data, the result seemed to be most consistent with the first window column. Although the results were different, it was important to note that there was a slight increase for every attribute (minimum, maximum, median, mean, and standard deviation) from the results of the training dataset to the testing dataset. That was not the case for the other two time windows. This could potentially indicate that the first time window would contain data that could better train a HMM.

Feature 5, 6, and 7: Sub Metering 1, 2, and 3

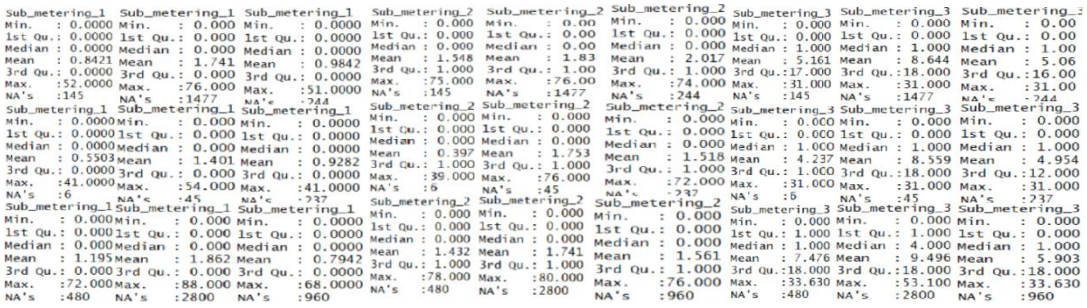


Figure 8.11 sub metering 1, 2, and 3 respectively, in 3 different time windows in 2007-2008, 1st column: Monday 12pm-8pm, 2nd column: Weekends 7am-3pm, 3rd column: Mon, Wed, Fri 4pm-8pm

The reason why sub metering 1, 2, and 3 was discussed in the same section was because they share similar results. First, all the results had minimum value of 0.00. Second, there are no consistent trend between the maximum of training dataset and testing dataset. The first and third time window seemed to have similar values for maximum but the second time window differ significantly. Whereas for sub metering 2, the values seemed to be consistent other than the maximum for the last 20% of the training dataset, where the maximum was only 39.00. With sub metering 3, the maximum seemed to be consistent for training dataset (31.00), but the testing dataset provided different maximum value for the second time window (maximum was 53.10). The mean values vary largely for each sub metering. For example, for sub metering 1, the mean for the first 80% of the training dataset were different (0.8421, 1.741, and 0.9842 respectively), the mean for the latter 20% of the training dataset were different (0.5503, 1.401, 0.9282 respectively), and the mean for the testing dataset were different (1.195, 1.862, and 0.7942 respectively). The results of means for sub metering 2 and 3 varies did not show any trend as well. Comparing different time windows within the same sub metering also showed no consistency. By looking at the results from sub metering 1, the mean of the first window for first 80% of training dataset, latter 20% of training dataset, and the testing dataset were 0.8421, 0.5503, and 1.195

respectively. For the third time window of the same sub metering (1), the results were 0.9842 ,0.9282, and 0.7942. Again, the results did not show any trend. Looking at the other two sub metering in a similar fashion also showed no trend.

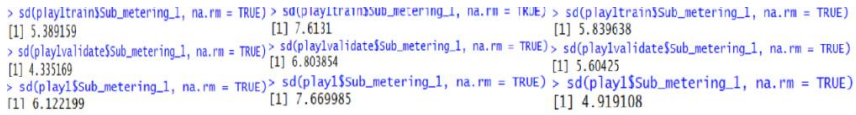


Figure 8.12 Sub metering 1 Std Dev 3 different time windows, each column represent a time window, 80% training dataset, 20% training dataset, and testing dataset respectively

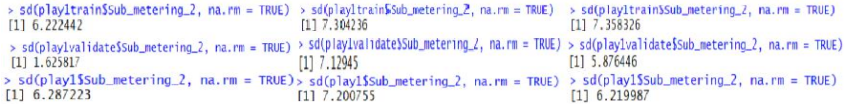


Figure 8.13 Sub metering 2 Std Dev 3 different time windows, each column represent a time window, 80% training dataset, 20% training dataset, and testing dataset respectively

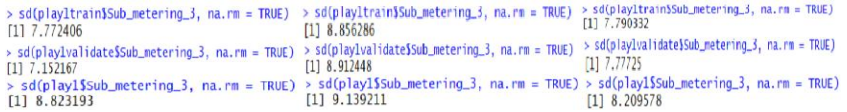


Figure 8.14 Sub metering 3 Std Dev 3 different time windows, each column represent a time window, 80% training dataset, 20% training dataset, and testing dataset respectively

The results from Figure 8.12, 8.13 and 8.14 confirmed the same thing about each of the sub metering, which was that there was no trend on any of the data. For all three sub metering, standard deviation could be increasing from training dataset to testing dataset for one time window, and decreasing for another. Comparing different time windows for the same sub metering also did not show any trend, which was the same observation from the means of all three sub metering. Therefore, one could question the accuracy on using any of the sub metering as a feature for training the HMM might not yield a good model, especially when none of the sub metering seemed to follow any type of trend.

8.1.2 – Understanding Feature Correlation

In understanding the dataset feature correlation, it was assumed that the correlation results between different features from the first dataset given would be fairly consistent to the results from the training dataset. Therefore, for understanding feature correlation the calculations and results were based on the first given dataset. Correlation is a statistical measurement that indicate the extent to which two or more variables (feature) fluctuate together. A positive correlation indicate the extent to which variables increase or decrease in parallel; a negative correlation indicate the

extent to which one variable increases as the other decreases.[8] One important distinction to point out was that correlation does not indicate causation; a high positive correlation does not imply that an increase in the value of one feature caused the increased in the other feature. Nonetheless, it was important to analyze the correlation between the features because by looking at the correlation coefficient, it could provide information that could help decide which feature to use or proceed with in training, validating, and testing for our Hidden Markov Model.

Wednesday 4-8pm	Friday 9-12pm
Global_active_power & global_reactive_power =0.182013	Global_active_power & global_reactive_power = -0.2746126
Global_active_power & global_intensity = 0.9962178	Global_active_power & global_intensity = 0.9967875
Global_active_power & voltage= -0.2721188	Global_active_power & voltage= -0.055234488
Global_active_power & sub_metering_1 =0.107927	Global_active_power & sub_metering_1 =0.4424803
Global_active_power & sub_metering_2= 0.7268048	Global_active_power & sub_metering_2= -0.2026798
Global_active_power & sub_metering_3 =0.449147	Global_active_power & sub_metering_3 = 0.9468868
Global_active_power & sub_metering_sum(1,2,3) =0.708217	Global_active_power & sub_metering_sum(1,2,3) =0.8191631

Figure 8.15 Correlation coefficient between global active power and other features (2 different time windows)

Since the results from basic data exploration seemed to indicate that global active power would be a good candidate for training the HMM, global active power was used as the main feature that was used to test for correlation with the rest of the features. Based on the result, it showed that the correlation between global active power and global intensity was highly positive (linear relationship). A graphical representation of the correlation between global active power and global intensity was created and shown below.

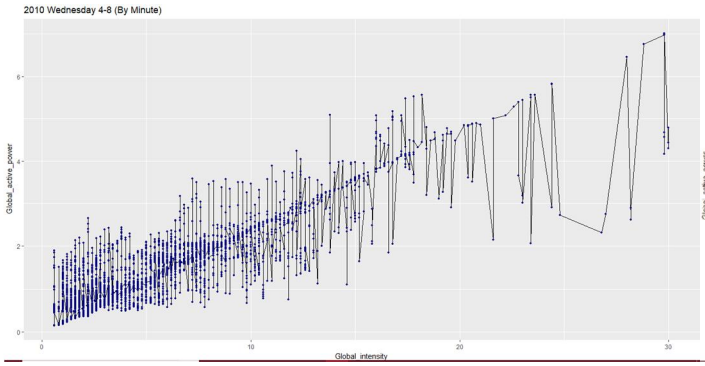


Figure 8.16 2010 Wednesday 4 pm – 8 pm (by Minute)

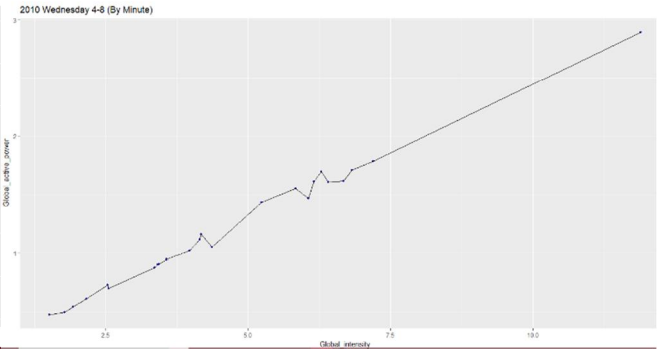


Figure 8.17 2010 Wednesday 4 pm – 8 pm (by Mean of each Wednesday)

From Figure 8.16, although the number varies, in general, there was a positive upward relationship between global active power and global intensity. To reduce possible noise in the graph, we used the mean of the same time frame to plot the relationship between global active power and global intensity. As shown in Figure 8.17, by converting the graph

[8] http://whatis.techtarget.com/definition/correlation

to use the means of each Wednesday, it helped reduce the possible noise and by doing so, it created a clear graphical representation of the positive relationship between global active power and global intensity. This meant, when one increase, the other will increase in a similar percentage. This was consistent with the correlation coefficient from Figure 8.15. Because of the high positive correlation between global active power and global intensity, using one of the two features would be sufficient for training the Hidden Markov Model. On the other hand, the result seemed to differ when a different time window was used to find the correlation between global active power and other features. For reactive power, there was a low positive linear relationship (~0.18) for the first time window, and a low negative linear relationship (~ -0.27) for the second time window. For voltage, both time window resulted in a negative correlation, but the results from both were weak (~-0.272 and -0.055 respectively). This meant that it was important to further analyzed both reactive power and voltage to decide whether to incorporate these features in training the HMM. For the sub metering, the results between the two time windows were inconsistent as well. However, when the values were added together, the correlation between global active power and summation sub metering seemed to be positive. A few more time windows were tested, and the positivity level varied, but all the results turned out to be a positive linear relationship. A graphical representation of the relationship between global active power and summation of all 3 sub meters were shown in Figure 8.18.

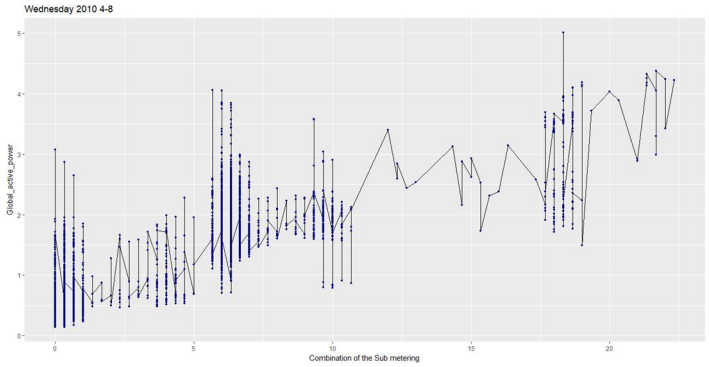


Figure 8.18 Global active power vs summation of all Sub metering

8.2 FINDING POINT ANOMALIES

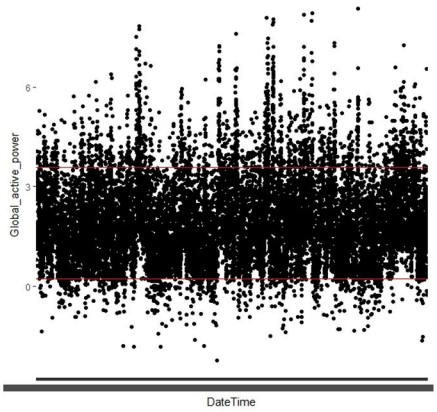
In simple and common scenarios, a well utilized approach in the cybersecurity industry was detecting for obvious outliers in the data set. This technique of analysis was most effective when dealing with datasets that could be empirically represented since the approach was to observe for instances that did not fit in relative to the rest of the data

(i.e. point anomalies). Techniques such as out of range and moving average detection were used to visualize and record point anomalies within the dataset. The feature used to compare and plot data was decided to be global active power, which based on our data exploration had the most significant trends and would yield the best results.

8.2.1 Out of Range

Using the training data set, the minimum and maximum global active power of each Monday were recorded. Afterwards, the minimum and maximum values were averaged across the entire dataset and were calculated as follows:

Average Minimum of Mondays (12-8PM, 2007-2008) : 0.2411429
Average Maximum of Mondays (12-8PM, 2007-2008): 3.580115
These two parameters were used in comparison against the testing dataset, where data instances in the testing set that are above the maximum or below the minimum criteria were considered anomalous points.



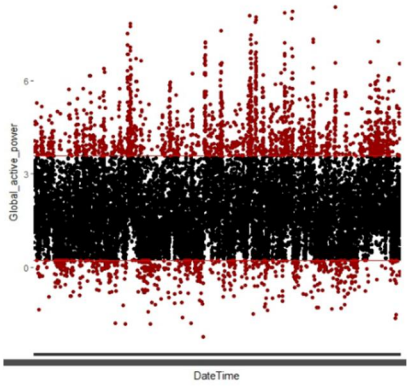


Figure 8.19 The above graphs are data plots of each minute for every Monday between 12-8PM of 2007 & 2008 with it's respective global active power.
On the right is the visualization of the considered anomalous points that are out of range.

In figure 8.19 the points black represent each point of data from the testing data set. The average minimum and maximum threshold were set as a red line across the y-axis and points above the maximum line and the points below the minimum line were considered anomalous.

Evidently at the start it revealed there are anomalous points for values that are negative, which was appropriate for the context of the dataset (impossible to have a negative global active power). Also, it showed points of data above the maximum to be anomalous which may accurately detect the anomalies extreme maximum points of the graph. However, it may also involve a lot of misidentified data points as the points of data are closer to the maximum line. A clear issue with this technique was that it assumed the data was also simple.

8.2.2 Moving Average

Using the Simple Moving Average function in the TTR package (R Studio), the moving average was calculated based on the time window of observations as a parameter. It was crucial to use the best fit moving average that represents the data the most, so experimentation of time windows was needed. With a data set of roughly 24,000 points, the first trial was using a time window of 480 (N = 480) observations to encompass each Monday separately.

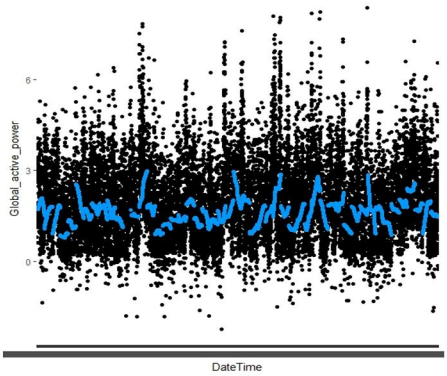


Figure 8.20 The moving average is indicted by the blue line on the plot graph. As depicted, with a high N value (480).

As shown in figure 8.20 the moving average was sparse and did not fully represent the graph in an appropriate scale. By changing N to various ranges could improve and smoothen the curve. A clear pattern of lower n value producing a smoother curve was evident:

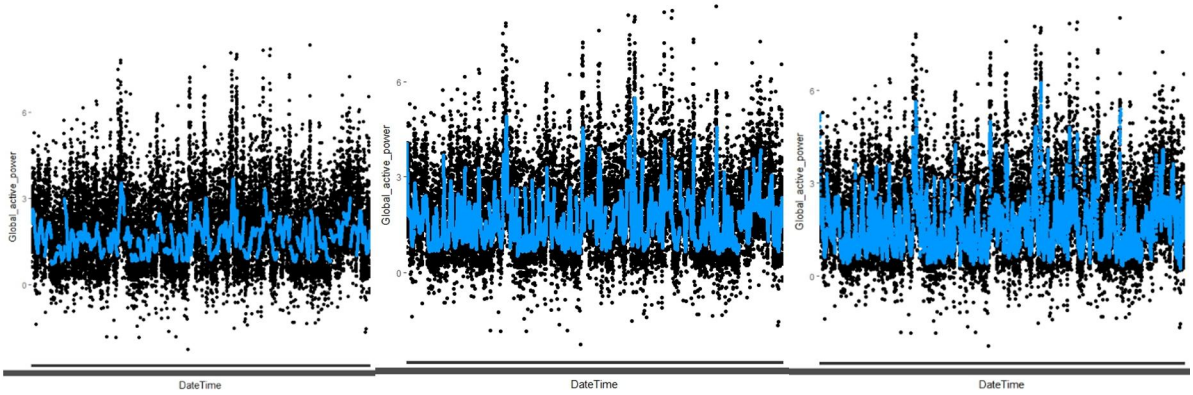


Figure 8.21 Moving averages with N = 240, 60, 20 respectively.

.

After manipulating the time window with various N values, it was determined that n = 20 was depicted to be the best fit as the moving average for the data set. With higher N's there is evident underfitting visible and with a lower N it begins to overfit the data set. Therefore used in the moving average calculations, the N parameter was set to 20 observations

per time window. Afterwards, in the moving average analysis, a threshold must be declared to determine whether the data points are within the range of it's corresponding moving average. Based on the training data's average minimum and maximum in comparison to the testing data's mean (1.0296), the threshold used in the detection was set to +/- 1.75. Points of data outside of it's threshold is considered anomalous.

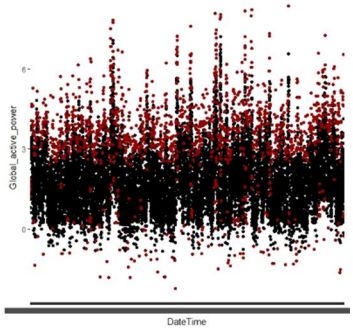


Figure 8.22 Using N = 20 as the window of observations, the anomalous points that are outside of the moving average threshold is highlighted in red.

Contrast to the out of range anomaly detection, the moving average performs better in finding contextual patterns to detect anomalies. Unfortunately, this comes as a trade-off as more obvious anomalous points may be ignored or undetected (such as all the negative points) due to the curving of the moving average.

Based on the visual graphs and results of the simple point anomaly detection techniques, it is difficult to not only recognize patterns of interest, but it is also challenging and time consuming to interpret the data as a whole. Furthermore, these simple detection techniques are only best used in data science scenarios where the data can be easily represented by single data points and involve no context. This is not the case on our data set and thus a more in-depth approach is required (HMMS).

8.3 BUILDING HMMs AND CALCULATING LOG-LIKELIHOOD

In deciding whether to proceed building the Hidden Markov Model by using multiple features (multivariate) or a single feature (univariate), the decision was based on the results from general data exploration, correlation coefficients, understanding the meaning behind specific features, and certain criteria that were placed by our group. The decision was to train a univariate HMM based on global active power. By understanding the meaning on the object of measurement was an essential step toward measurement. This step was known as the first step of analysis. For

example, looking at global reactive power, a better understanding of the meaning was required to make a decision on whether to include it as a feature for training the HMM. In general, reactive power regulate voltage and ensure that the voltage available was sufficient enough for active power to be used.[9] In a sense, it represent or help maintain the minimum voltage level required at all times. Reactive power represented the portion of electricity that help establish and sustain the electric and magnetic fields required by alternating current equipment. The amount of reactive power present in an AC circuit would depend upon the phase shift or phase angle between the voltage and the current. Reactive power would be positive when if supplied, and negative if consumed.[10] In general, global reactive power does not seem to have a highly positive or direct impact on the household power consumption.

The first criteria that was placed in choosing the features was that the values of the features must have a wide range. For example, for global reactive power, the minimum and maximum value were around 0 to 1, therefore, this feature was excluded. Second, the percentage change in value for 1 standard deviation from the mean cannot be small. For example, when deciding whether to include voltage as one of the features, looking at a specific time window, because the percentage change for 1 standard deviation from the mean was 1.244% change only (whereas the % change for global active power was 85.766%), the decision was made to exclude voltage as a feature in training the HMM. Third, if there was a high positive or negative linear relationship between global active power and the feature being considered, then that feature would be excluded. For example, for the global intensity, as shown in Figure 8.15 from understanding feature correlation, it showed that global active power and global intensity are highly positively correlated. Therefore, it was excluded for training our HMM. Lastly, a decision was made to exclude all the sub metering for training the HMM because each sub metering represents specific household power consumption, which most likely would not be significant enough individually to impact the entire trend or movement of power consumption. In addition, looking at the correlation between global active power and the summation of all sub metering (Figure 8.15 from understanding feature correlation), it showed that they are positively correlated. Also, household power consumption

[9] https://www.quora.com/What-is-reactive-power-What-is-the-use-of-reactive-power
[10] https://www.electronics-tutorials.ws/accircuits/reactive-power.html

for different equipment in the house would differ based on season, time, temperature, needs, etc. The power consumption of each equipment would be different for each household too. Assuming these sub metering represent different grouping of equipment that used power, it was assumed that the values would not have a consistent trend at all. The values could differ by day, by time, by year, etc. Therefore, based on various results from our data exploration and research into the different features, the decision was to use global active power as the single feature to train the univariate HMM. Once the decision was made to train a univariate HMM, the next step was to make some educated guess in regards to what time window and number of states to use in training the HMM.

Of the provided training data set, many of the row instances pertaining to the year 2006 contained missing (NA) values; also, only the last half of December was included in data set. For these reasons, we decided not to include 2006 as part of our training data. When it comes to choosing data for training the model, it was important not only to account for the completeness of the data set, but also the external factors that could have influenced the data set. Year 2008 was the year of the worst financial crisis since "The Great Depression" of the 1930s. Even though the economy was already shrinking near the start of 2008, experts, along with the general public, did not fully realize this until the third quarter of 2008.[11] It could be assumed that consumer spending habits during dire times are much different when compared with when the economy was more stabilized; this include consumption of electricity. The recession continued till mid 2009 with the lowest points of the downturn being in the first quarter of 2009. Even with the end of the recession, unemployment rates continued to rise to the highest it has been in 26 years.[12] It is with the fact that the recession was not in full throttle and fully realized till the third quarter of 2008 that we choose to include it in our training data. The exclusion of 2009, from the training data, was due to its volatility. Also, the 2009 electric power consumption of the US was upward trending which was inconsistent and contrasting with the slight downward trend of 2010 (the year of the testing data); this could be observed from Figure 8.23 provided below.

Electric power consumption (kWh per capita)

[11] https://www.thebalance.com/the-great-recession-of-2008-explanation-with-dates-4056832
[12] https://www.csmonitor.com/Business/2013/0908/Timeline-on-the-Great-Recession

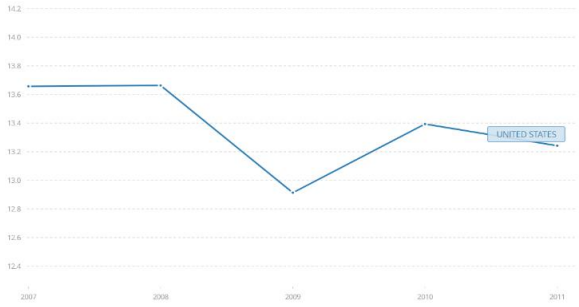


Figure 8.23 Graph of electric power consumption (kWh per capita) from years 2007 to 2011

Once the years (2007-2008) for training the model were decided, the next step was to come up with different days and times that could be used to train the HMM and would provide a better log-likelihood comparison between the training data (approximately 80% of the training dataset) and the validation data (approximately 20% of the training data set) as well as having a low BIC. More than thirty different time windows were tested and validated. As a trial with using the depmix function, the entire data set was used to train a HMM. Because of the size of the dataset, the result took a long time to compute, and the results were not ideal. Even when the number of states was increased to a larger number, the outcome was not as good as other time windows that was tested later. In the beginning, without fully understanding the different functions or packages that could help in sectioning or partitioning the training dataset, the approach was to hardcode the dates one by one. This approach was very time consuming and when the first two time windows used were not providing a persuasive model, it was important to change our approach, and began exploring different packages and functions that could help partition and filter the dataset according to the time windows. Packages that were used include chron, depmixS4, dplyr, and lubridate. This allowed us to speed the process in filtering the dataset into only data that match our time windows, and for finding the ideal number of states for each time window.

8.3.1 Starting Models

The first time frame analyzed was between 4-8pm time on a weekday. The reason for choosing this specific period was based information obtained from the official website of Pacific Power (one of the larger electric power companies in the western United States) which provided a breakdown of the cost of electricity consumption during a

specific time of the day.[13] The "on peak" time represent the highest electricity consumption period, and thus the cost of electricity being higher. "Off peak" time corresponds to a low electricity consumption time, where most people were not using as much electricity relative to the "on peak" time. This is represented in the figures below.

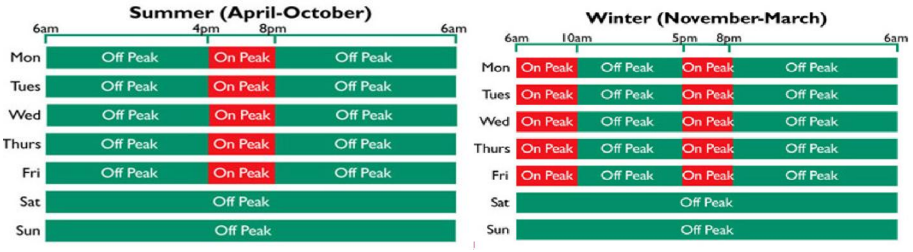


Figure 8.24 (from source 3) "On Peak" time frames by the week and season.

Wednesday was picked to represent one of the weekdays in exploring the trend of the data for “on peak” 4-8pm time window. As a trial, the dataset was partitioned to only include year 2008, which would be compared to another model with the same time frame but include year 2007 as well.

Wed 16:00-19:59 08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
3	-3539.423	-84.27197619	7207.22	-1916.57	-174.2336364	3940.63	0.483672257
4	-2795.2	-66.55238095	5801	-1547.535	-140.685	3271.66	0.473059537
5	-2330.897	-55.49754762	4973.557	-1384.092	-125.8265455	3029.232	0.441063906
8	-1001.864	-23.85390476	2728.121	-980.0264	-89.09330909	2566.604	0.267740698
9	-428.547	-10.2035	1755.707	-885.6255	-80.51140909	2523.682	0.126733591

Figure 8.25 Log-likelihood & BIC for Wednesday 4pm-8pm 2008

Based on the results, the log-likelihood in both the training and validation models remained very consistent, and as the number of states increased, the log-likelihood increased as well. Also, the BIC remained relatively consistent. However, after normalizing both training and validation log-likelihoods and calculating the log-likelihood ratios, it was discovered that the normalized log-likelihoods of the training and validation data are quite far from each other, which was undesirable. When running the same time window with the inclusion of year 2007 (2 years in total). The results were as follows:

Wed 16:00-19:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
12	-685.9017	-8.263875904	3025.016	-505.4596	-22.97543636	2442.39	0.359683088
13	-117.517	-1.41586747	2155.533	-636.5176	-28.93261818	2935.941	0.048936721
14	positive						

Figure 8.26 Log-likelihood & BIC for Wednesday 4pm-8pm 2007-2008

Like the previous model, the eventual log-likelihood values obtained were good, but the associated ratios were terrible which led to the exploration and production of further models.

[13] https://www.pacificpower.net/ya/po/otou/ooh.html

8.3.2 Second Model

The first model had very low normalized log-likelihoods and consistent numbers, the second Markov model analyzed the morning "on peak" time to verify if there was a similar trend with all "on peak" times. More days of the week were included to increase the data, Tuesday, Wednesday, and Thursday of 7am-11am were the selection of choice.

TWT 7:00-11:00 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
8	-14726.59	-59.14293173	30322.03	-4705.18	-72.38738462	10173.11	0.817033687
12	-16776.46	-67.37534137	35389.6	-6516.303	-100.2508154	14645	0.672067764
13	-8620.873	-34.62197992	19375.38	-3299.318	-50.75873846	8471.71	0.682089054
17	-4582.217	-18.40247791	12705.82	-3185.63	-49.00969231	9480.178	0.375486502
20	-1324.022	-5.31735743	3526.044	-1693.09	-26.04753846	7624.736	0.204140496
22	-2998.385	-12.04170683	11792.76	-2793.814	-42.98175385	10675.83	0.280158573
23	positive						

Figure 8.27 Log-likelihood & BIC for Tue, Wed, Thu 7am-11am 2007-2008

Based on the results above, the training log-likelihood, validation log-likelihood and BIC values were significantly larger compared to the very first models, but there was still a general decrease in all three of those values. However, as the log-likelihood values became greater, the ratios became worse.

8.3.3 Third Model

So far with the "on peak" times, the normalized log-likelihoods have been considerably low, which was not ideal, so in our third Markov model, we chose the time window of every other weekday: Monday, Wednesday, and Friday from 10am-2pm.

MWF 10:00-13:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
7	-4623.019	-18.64120565	9370.039	-576.8778	-8.740572727	1753.314	0.468884518
8	-3954.22	-15.94443548	8776.971	-1723.065	-26.10704545	4210.082	0.610733049
9	-11.00976	-0.044394194	1099.438	106.6644	1.616127273	734.36	-0.027469491

Figure 8.28 Log-likelihood & BIC for Mon, Wed, Fri 10am-2pm 2007-2008

As seen from the results above, the log-likelihood value jumped radically from state 8 to 9. In state 9, the log-likelihood values for both training and validation point to a possibly overfitted model; the ratio for this state was incoherent anyways. Furthermore, the ratios shown in the final column were far less than stellar.

8.3.4 Fourth Model

Since, the previous model, which only included weekdays, did not turn out favorable. It was decided that the next model was to include the entire week instead.

ALL DAYS 10:00-13:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
9	-13374.24	-23.13882353	27908.81	-3097.357	-20.2441634	7224.799	0.874900289
10	-12898.98	-22.31657439	27206.94	-3553.303	-23.22420261	8357.425	0.960918864
13	-4345.402	-7.51799654	9078.804	-3410.649	-22.29182353	8860.446	0.337253546
14	-1322.774	-2.288536332	3091.549	-104.3524	-0.68204183	2552.675	0.298025345
15	1650.873	2.856181661	-294.3312	671.2543	4.387283007	1327.305	0.651013772

Figure 8.29 Log-likelihood & BIC for All days 10am-2pm 2007-2008

First, the normalized log-likelihood of training data was not as consistent as before, there seemed to be a discrepancy between the log-likelihood of 13 states and 14 states. When there were 13 states, the training log-likelihood was roughly 0.3 times larger than the validation log-likelihood but when there were 14 states, training log-likelihood was 3 times larger than the validation log-likelihood. Judging from the trends from the other models, this model did not have a consistent trend. This time frame was not as suitable to model 'normal' behaviour.

8.3.5 Final Model

Finally, after training with various time windows and different days of the week, Mondays, 12pm-8pm of 2007-2008 was chosen to be the best model. The idea behind choosing the 12pm-8pm time window was that these 8 hours encompass both “mid-peak” and “on peak” hours throughout the entire year as opposed to just using the “on peak” hours like the first models. The idea of this time window came from pricing information provided by Portland General Electric[14]. In order to determine which day to choose, it was necessary to individually train and find the best model for each day of the week.

[14] https://www.portlandgeneral.com/residential/power-choices/time-of-use/time-of-use-pricing

Mon 12:00-19:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	-19678	-237.0843373	39430.15	-4340.496	-197.2952727	8745.846	0.832173373
5	-2919.447	-35.17406024	6199.043	-339.0766	-15.41257273	993.1575	0.438180086
6	-319.0362	-3.843809639	1135.926	-114.9913	-5.226877273	665.4296	0.73539313
7	1589.869	19.15504819	-2522.995	456.81	20.76409091	-339.2006	0.922508396

Tues 12:00-19:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	-32773.64	-394.8631325	65621.42	-6821.28	-310.0581818	13707.4	0.785229504
6	-12533.65	-151.0078313	25565.15	-2210	-100.4545455	4855	0.66522739
8	-7483.172	-90.1586988	15803.16	-1278.711	-58.12322727	3289.34	0.644676865
12	-3838.78	-46.25036145	9446.529	-24.2274	-1.101245455	1595.68	0.023810526
14	-4862.457	-58.58381928	12087.07	-39.45315	-1.793325	2144.96	0.030611268
15	positive						

Wed 12:00-19:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	-26089	-314.3253012	52252.14	-8673.365	-394.2438636	17411.58	0.797286477
6	-8319.358	-100.2332289	17136.57	-3442.964	-156.4983636	7321.376	0.640474613
10	-1999.841	-24.09446988	5260.205	-2482.161	-112.8255	6066.837	0.213555179
11	-654.7305	-7.888319277	2813.614	-3074.441	-139.7473182	7464.487	0.056447017
12	positive						

Thu 12:00-19:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	-16099.9	-193.9746988	32273.94	-3823.409	-182.0670952	7711.35	0.938612594
4	-3739.995	-45.06018072	7723.621	-677.1251	-32.24405238	1566.27	0.71557752
5	-1534.507	-18.48803614	3429.163	-207.9034	-9.900161905	729.229	0.535490186
6	1187.508	14.3073253	-1877.163	241.5677	11.50322381	-49.875	0.804009385

Fri 12:00-19:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
6	-3308.664	-40.34956098	7067.353	-1078.645	-49.02931818	2592.737	0.822968021
8	-3073.006	-37.47568293	6902.436	-1296.645	-58.93840909	3325.212	0.635844834
10	-1796.535	-21.90896341	4732.128	-322.8599	-14.67545	1748.234	0.669837715
11	-1766.154	-21.53846341	4891.956	-396.7697	-18.03498636	2109.145	0.83733858

Sat 12:00-19:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
5	-21963.71	-267.850122	44287.15	-4252.27	-193.285	8819.55	0.721616248
10	-15730.21	-191.8318293	32719.51	-2890.063	-131.3665	6882.64	0.68480033
12	-14075.2	-171.6487805	29917.35	-2576.033	-117.0924091	6699.29	0.682162779
15	-11386.68	-138.8619512	25460.82	-2484.556	-112.9343636	7322.38	0.813285156
20	-9025.868	-110.071561	22696.58	-1448.214	-65.82790909	6963.69	0.598046475

Sun 12:00-19:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	-46291.68	-557.7310843	92657.42	-11968.69	-544.0313636	24002.23	0.975436691
6	-24049.38	-289.7515663	48596.05	-4353.747	-197.8975909	9142.941	0.682990582
23	-12355.6	-148.8626506	30784.4	-1600.867	-72.76668182	8519.745	0.488817588

Figure 8.30 Log-likelihood & BIC for each weekday from 12pm-8pm 2007-2008. The row that is highlighted green contains the values for final model that was chosen.

After getting the training results for all 7 days, it was discovered that Monday, with 6 states, provided the best results (highlighted green in Figure 8.30 above). This was attributed to the fact that it had superior values compared to the other models in terms of quality; the log-likelihood, BIC, and ratio values were also more balanced compared to the other models. Some of the models, of the other days, may have had better individual values, but the chosen model was the only one that had good values for both BIC and log-likelihood while still maintaining a good ratio.

8.3.6 Other Models

Along with the many models above, there were many other models that were part of the trial and error process which took place before arriving at the final model. The results of some of these other models are shown below:

Weekdays 20:00-23:59 07'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
3	-31102.76	-75.12743961	62352.58	-21621.41	-198.3615596	43385.2	0.378739912
6	-22989.23	-55.52954106	46072.46	-15030.21	-137.8918349	30538.5	0.40270362
12	-17100.43	-41.30538647	35955.11	-10942.57	-100.3905505	23583.9	0.411446957
15	-16062.33	-38.79789855	34793.81	-10858.02	-99.61486239	24299.7	0.389479016
18	-13657.33	-32.98871981	31085.78	-9750.845	-89.45729358	23153.4	0.368765011
20	-12910.85	-31.18562802	30433.18	-8325.66	-76.38220183	21116.8	0.408283962

Weekdays 20:00-23:59 07-08'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
3	-83350.99	-201.3308937	166863.1	-18822.31	-172.6817431	37787	0.85770117
4	-70916.67	-171.2963043	142098	-15676.32	-143.8194495	31592.6	0.839594585
6	-58209.82	-140.60343	116960.4	-12865	-118.0275229	26208.1	0.839435588
10	-48069.63	-116.1102174	97508.54	-9968.713	-91.45608257	21147.9	0.787666104
15	-39803.97	-96.14485507	82530.59	-8054.059	-73.89044954	18691.8	0.768532539

Figure 8.31 Log-likelihood & BIC for Weekdays 8pm-12am 2007 (top) and
Log-likelihood & BIC for Weekdays 8pm-12am 2007-2008 (bottom)

The hours from 8pm-12am were chosen because they represented the night hours (after dinner and/or bedtime hours). Firstly, the weekdays of only one year, 2007, was tested. By state 20, training the model had already taken too much time, and the values obtained were not exactly promising either. Afterwards, the same time window of the all weekdays was chosen once again, but this time with two years. Again, training the models started to take too long, and the values weren't promising either. Because of this, training of higher states was also aborted.

MWF 00:00-06:59 07'

States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	positive						
5	positive						

Figure 8.32 Log-likelihood & BIC for Mon, Wed, Fri 12am-7am 2007-2008

Since nighttime was tested, it was only reasonable to try out the hours of the middle of the night as well. The results for Figure 8.32 above show that this time window, of in the middle of the night, was overfitted from the start; this is potentially due to the lack of electricity usage during these hours.

8.3.7 Testing the Models

Given the final trained model using the data partition of Mondays 12-8PM of 2007 and 2008, the final step was to input the test data to the model and record the outputs. The test data was partitioned in to the same time window as the training data, with the difference being in the year of the data set (2010).

Firstly, the whole test data set of Mondays yielded a result of a log-likelihood of –38044.04. Compared to the training model's log-likelihood of –319.0362, there is a disparity of ~37,600. With this given result it is concluded that the test data, according to the trained hidden Markov model, is considered an anomalous data set. This is because the

model interprets the trained modelling data and recognizes a significant pattern within the data set to create a scholastic model based on probabilities and state sequences. Therefore the test data does not conform or fit along this pattern, thus yielding a significant contrast of log-likelihoods.

To further examine analyze the test data set using the model, the test data was partitioned into separate months with it's own corresponding Mondays.

Monday 12:00-19:59 2007-2008 Model					
Month	Train LogLike	Norm. Train LogLike	Test LogLike	Norm. Test LogLike	Norm. LogLike Ratio
Dec 09'	-319.0362	-3.843809639	-2976.96	-744.24	0.005164745
Jan 09'	-319.0362	-3.843809639	-2844.231	-711.05775	0.005405763
Feb 10'	-319.0362	-3.843809639	-2822.686	-705.6715	0.005447024
Mar 10'	-319.0362	-3.843809639	-3530.595	-706.119	0.005443572
Apr 10'	-319.0362	-3.843809639	-2847.655	-711.91375	0.005399263
May 10'	-319.0362	-3.843809639	-3607.563	-721.5126	0.005327432
Jun 10'	-319.0362	-3.843809639	-3215.242	-803.8105	0.004781985
Jul 10'	-319.0362	-3.843809639	-3141.011	-785.25275	0.004894997
Aug 10'	-319.0362	-3.843809639	-3854.993	-770.9986	0.004985495
Sep 10'	-319.0362	-3.843809639	-2900.847	-725.21175	0.005300258
Oct 10'	-319.0362	-3.843809639	-3391.731	-847.93275	0.004533154
Nov 10'	-319.0362	-3.843809639	-2943.52	-735.88	0.005223419
ALL	-319.0362	-3.843809639	-38077.04	-746.6086275	0.00514836

Figure 8.33 The log-likelihood of training, test, and the ratio of each individual month.

The key values to be looked at, once again, were the ratios in the last column. As seen in Figure 8.33, the ratios obtained were far lower and worse than any of the ratios obtained earlier during the training process of the models (comparing of log-likelihood values of training data versus validation data). The subset that fit the least with the best model was August of 2010. Since it was known that the provided test data set was anomalous beforehand, the above results were expected.

9. CONCLUSION

Throughout the course study on cybersecurity, the importance of anomaly detection was illustrated and that not only is it important within it's own field, but to those that have a global application to data science everywhere else in the world. To demonstrate the basic concepts and understanding of the test study, simple approaches were used to find simple point anomalies in the data set. From the observable results, it was clear and determined that with a large and complex data set, more complex approaches were needed to be involved to fully dissect the data set and analyze the data set.

The goal was to be able to differentiate between anomalous and normal electrical power usage, when given another set of the U.S electrical power grid data. One of the first challenges while working on the project was the data

set, due to large amounts of data and 7 different attributes, it was difficult to judge what was important and what was not. The other challenge that appeared was training Markov models, this took up the most time as there was a lot of experimenting and comparison involved to find the best model. With each model, the parameters were constantly adjusted to have a better log-likelihood and BIC; we stop adjusting the parameters (states, time frames) once we hit a positive log-likelihood since that indicates an overfitted model.

To understand the dataset, each feature in the dataset was examined in detail. The mean, standard deviation, maximum, and minimum values were compared against the training dataset, validation dataset, and testing dataset. In order to see whether the training dataset differ from the testing dataset, and whether the dataset followed more closely toward the earlier or latter part of the dataset. Looking at the results of mean and standard deviation provided additional insight that allowed us to decide which features should be considered in training the HMM. Once a decision was made to include global active power in training for the HMM, the next step was to use this feature and find the correlation coefficient with other features. The main feature that had a strong positive correlation was global intensity. This allowed us to exclude global intensity as a feature for training the HMM.

After we decided to use global active power to train the univariate HMM, the next step was to choose the time window that would most accurately represent the trend of the dataset. The early stages for deciding the hours, weekday/s, and year/s for our models were dependent on a number of factors including intuition as well as statistics obtained from external sources. Some of the more prominent statistics found included the peak hours of electricity usage which was the main influence of our early models. Ultimately, it came down to trial and error mixed with our newly acquired knowledge of electricity usage hours; this eventually led us to our best model. After finding the best model, the observations in the anomalous test data were put through our best model, and what we found was that these observations did not fit at all with our model (according to the ratios) which complied with our expectations.

10. GROUP CONTRIBUTIONS

In regards to group contribution, it was important to mention that all group members showed up to all group meetings. Every group member did contribute to training different HMM and manipulating the datasets. Some members focused on manually "hardcoding" the datasets while others spent more time figuring out how to use different packages and functions in R. Antonio primarily focused on understanding and using different functions and packages from R (I.e. how to manipulate dataset efficiently), creating the structure of the report, writing table of contents, checking for overall changes relative to the expected normal behaviour, understanding feature correlation, 8.3 (HMM). Johny primarily focused on manipulating datasets, using ggplots, creating excel tables, writing for 8.3 (HMM), abstract, introduction, dataset description, why we decided on 2007-2008 and conclusion. Leon primarily focused on writing moving average, out of range, 8.3 (HMM), and conclusion. Amanda primarily focused on data manipulation, finding correlation coefficients, using ggplots, writing background, problem definition, methodology, 8.3 (HMM) and conclusion. It was all agreed that each member put in a fair amount of work and effort into doing research and generating the report.

11. REFERENCES

Anomaly Detection: A Survey. Chandola, V., Banerjee, A., and Kumar, V. ACM Computing Survey 41, 3, Article 15 (July 2009), 58 pages.

PennState College of Health and Human Development. (2018). AIC vs. BIC. [online]
Available at: https://methodology.psu.edu/AIC-vs-BIC [Accessed 12 Feb. 2018].

World Bank Group. (2018). Electric power consumption (kWh per capita) | Data. [online]
Available at: https://data.worldbank.org/indicator/EG.USE.ELEC.KH.PC?end=2011&locations=US&start=2007&view=chart [Accessed 15 Mar. 2018].

Statistics How To. (2017). Hidden Markov Model: Simple Definition & Overview. [online]
Available at: http://www.statisticshowto.com/hidden-markov-model/ [Access 18 Mar. 2018].

Wikipedia. (2017). Likelihood-ratio test. [online]
Available at: https://en.wikipedia.org/wiki/Likelihood-ratio_test#Interpretation [Accessed 23 Feb. 2018]

Ontario Ministry of Energy. (2015). Smart Meters and Time-of-Use Prices. [online]
Available at: http://www.energy.gov.on.ca/en/smart-meters-and-tou-prices/ [Accessed 20 Feb. 2018]

Electronic Tutorials. (2014). Reactive Power. [online]
Available at: https://www.electronics-tutorials.ws/accircuits/reactive-power.html [Accessed 15 Mar. 2018]

The Balance. (2017). The Great Recession of 2008: What Happened, and When? [online]
Available at: https://www.thebalance.com/the-great-recession-of-2008-explanation-with-dates-4056832 [Accessed 18 Mar. 2018]

Pacific Power. (2018). Time of Use Hours & Pricing. [online]
Available at: https://www.pacificpower.net/ya/po/otou/ooh.html [Accessed 17 Feb. 2018]

Portland General Electric. (2018). Time of Use Pricing. [online]
Available at: https: //www.portlandgeneral.com/residential/power-choices/time-of-use/time-of-use-pricing# [Accessed 17 Feb. 2018]

The Christian Science Monitor. (2013). Timeline on the Great Recession. [online]
Available at: https://www.csmonitor.com/Business/2013/0908/Timeline-on-the-Great-Recession [Accessed 15 Feb. 2018]

Whatis.com. (2016). What is correlation? – Definition from WhatIs.com. [online]
Available at: http://whatis.techtarget.com/definition/correlation [Accessed 15 Feb. 2018]

Quora. (2015). What is reactive power? What is the use of reactive power? [online]
Available at: https://www.quora.com/What-is-reactive-power-What-is-the-use-of-reactive-power [Accessed 16 Feb. 2018]

Quora. (2012). What is the difference between a univariate and multivariate hidden Markov model? [online] Available at: https://www.quora.com/What-is-the-difference-between-a-univariate-and-multivariate-hidden-Markov-model [Accessed 16 Feb. 2018]