Cybersecurity Final Project Report

Group 5
George Yang 301343144
Jasper Quan 301255149
Kun Hyung(Arthur) Park 301262692

Cmpt 318 Spring 2022

Abstract

A statistical analysis on electricity consumption was performed to identify data points as either normal behavior or anomalous. Through principal component analysis, global active power and global intensity was chosen to be the most meaningful out of the 7 response variables in the dataset. Monday between 20:00 and 23:59 was chosen as the timeslot to train and test using the Hidden Markov Model as it had the most distinguishable pattern. State 12 was chosen as the best performing model considering the computation time and its BIC and log-likelihood values. Using the generated model, dataset 1,2 and 3 were tested to identify anomalous behaviors.

Table of Contents

Table of Figures

1. Introduction

In modern society, Electricity has become a necessity that powers our lifestyle. We have grown accustomed to this lifestyle over the past few decades and we are relying more and more on this energy source. Throughout North America, there is a complex network of power plants and transmission lines. Protection of these critical infrastructures is crucial to prevent power outages. Nowadays, critical infrastructures rely heavily on automated control to operate continuously. Automation maximizes efficiency and protects the critical assets in case of any internal or external disruptions. Near real-time multivariate continuous data analysis allows the system to detect the threat early and mitigate the impact of the threat. The main challenge is to recognize and differentiate various types of anomalies.

This project focuses on analyzing electrical data using Principal Component Analysis(PCA) and using it to train and test a Hidden Markov Model for anomaly detection.

2. Technical Background Information

Anomaly Detection Techniques

Behavioral anomaly detection defines normal patterns in network traffic or individual computer operations. Then, it scans continuously for patterns that depart from the norm sufficiently to cause information system operators to suspect malicious activity. This pattern is called the anomaly. (Glässer, 2022)

Types of Anomalies

Anomalies can typically be classified into three categories: point anomaly, contextual anomaly, and collective anomaly. Point anomaly is identified if an individual data instance is considered anomalous with respect to the rest of the data and is the simplest type of anomaly (Glässer, 2022). Contextual Anomaly, also referred to as the conditional anomaly is detected if a data instance is anomalous in a specific context, but not otherwise (Glässer, 2022). Finally, collective anomaly is detected when a collection of related data instances is anomalous with regards to the entire data set (Glässer, 2022).

Hidden Markov Model (HMM)

Hidden Markov Model is a statistical model to capture information that is not observable from observable sequential symbols. There are three fundamental problems in HMM design and analysis: Evaluation of the probability of a sequence of observations generated by a given HMM, Determination of the best sequest of model states, and adjustment of model parameters so as to best account for the observed signals. (Glässer, 2022).

3. Principal Component Analysis

Before anything, we had to make the dataset usable. We chose to replace every NA value in the dataset with the mean value of the column instead of deleting the row containing the NA value because it would give more accurate results. Next, we scaled each variable so that they were comparable with each other.

In order to understand the data better, we performed principal component analysis(PCA) on 7 response variables. Once the PCA was complete, we analyzed the standard deviation returned from PCA. Next, we calculated that PC1 accounted for 40.8% of the variation, which was more than double of PC2 value, which was 14.3%. As a result, when calculating loading scores for different response variables, we decided to only look at loading scores for PC1. Figure 3.1 shows the percentage of variation that each PC accounts for. Figure 3.2 shows the cluster of the data.

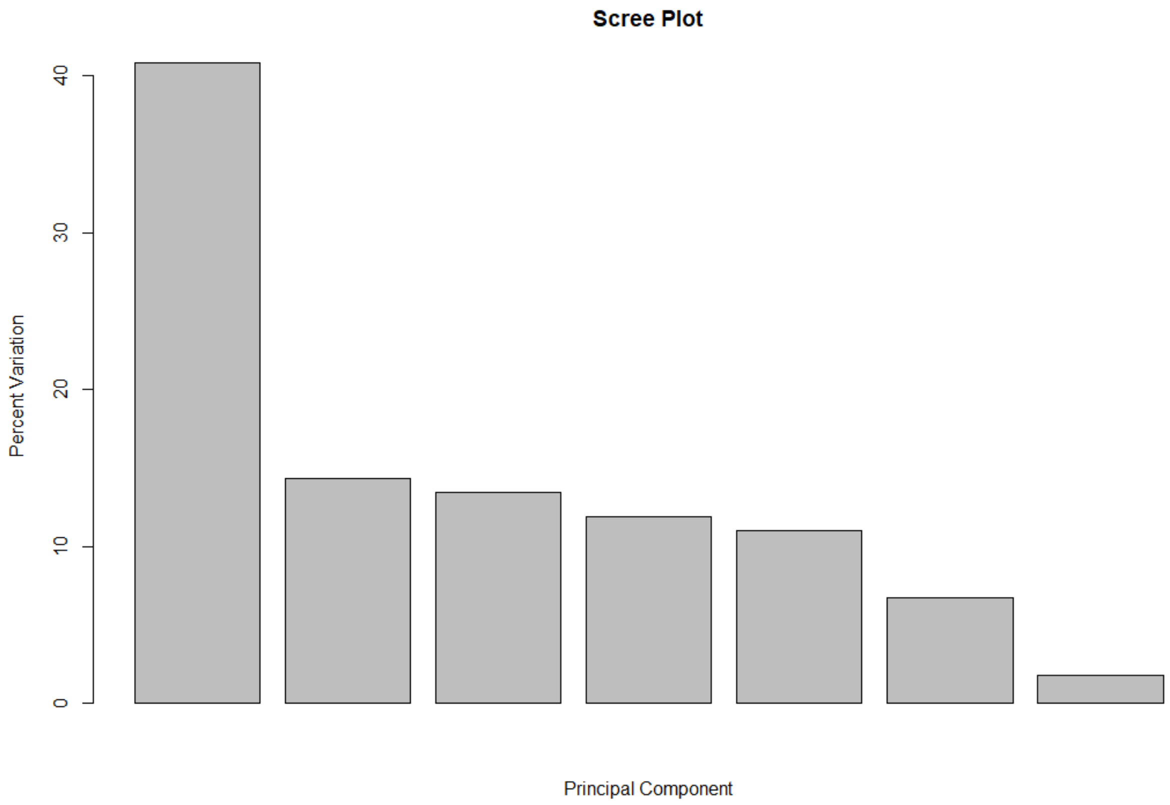

Figure 3.1 Scree plot

Next, we used the loading scores from PCA to determine which response variable had the largest effect on where samples are plotted in the PCA plot. Based on the result seen in Figure 3.3, response variables Global_Intensity and Global_active_power were chosen for HMM training and Testing as they had the largest absolute loading scores.

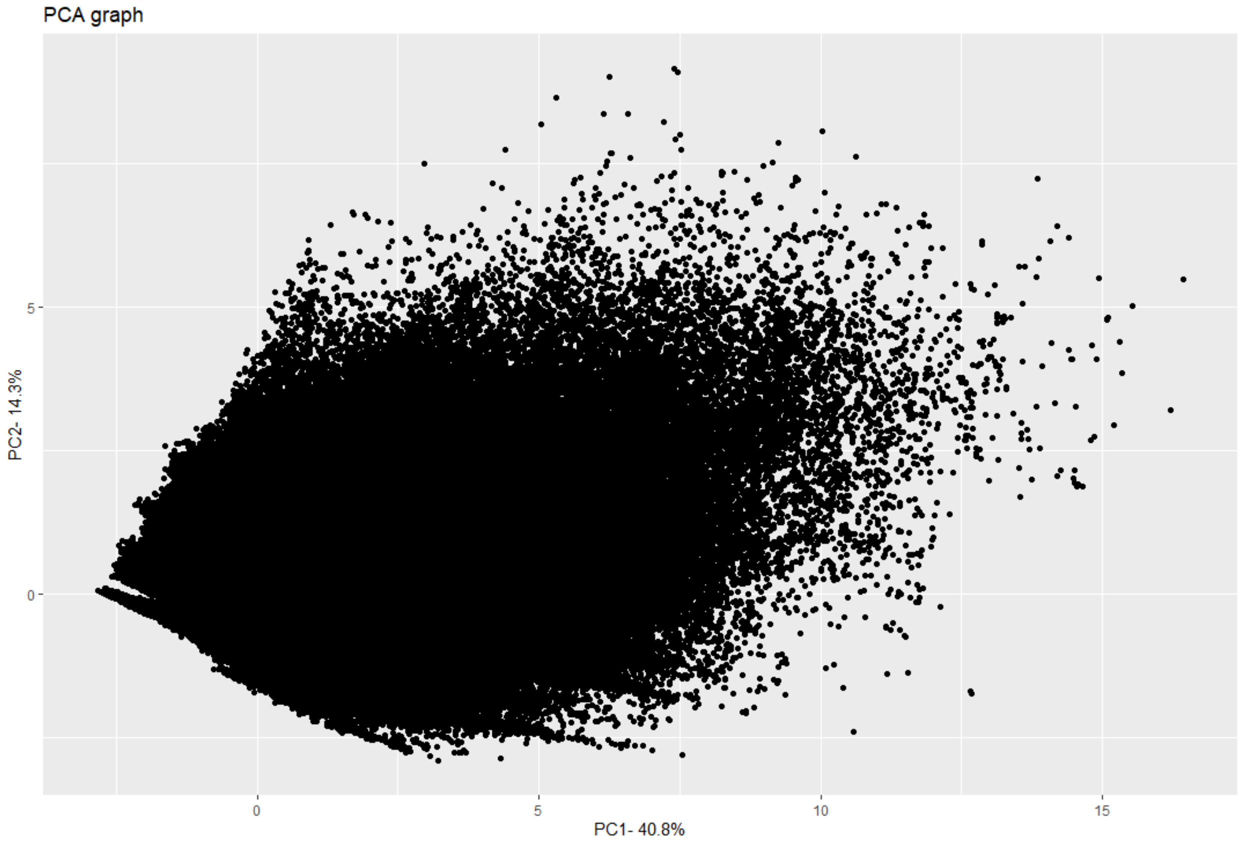

Figure 3.2 PCA graph

Global_ Intensity	Global_ Active_ Power	Sub_ metering_ 3	Voltage	Sub_ metering_1	Sub_ metering_2	Global_ Reactive_ Power
0.5597804	0.4687937	0.3876345	-0.3296146	0.2991213	0.2839871	0.1945506

Figure 3.3: Loading scores for different response variables

In order to choose the time window for HMM training and testing, we first separated the dataset by weekday. Next, we graphed the Global_intensity and

Global_active_power as seen in Figure 3.4 to see which dataset we wanted to work with. We wanted to choose a day where there was a clear pattern for both Global_intensity and Global_active_power within a 4 hour time window. We chose to work on day 2, time slot 20:00 - 23:59.

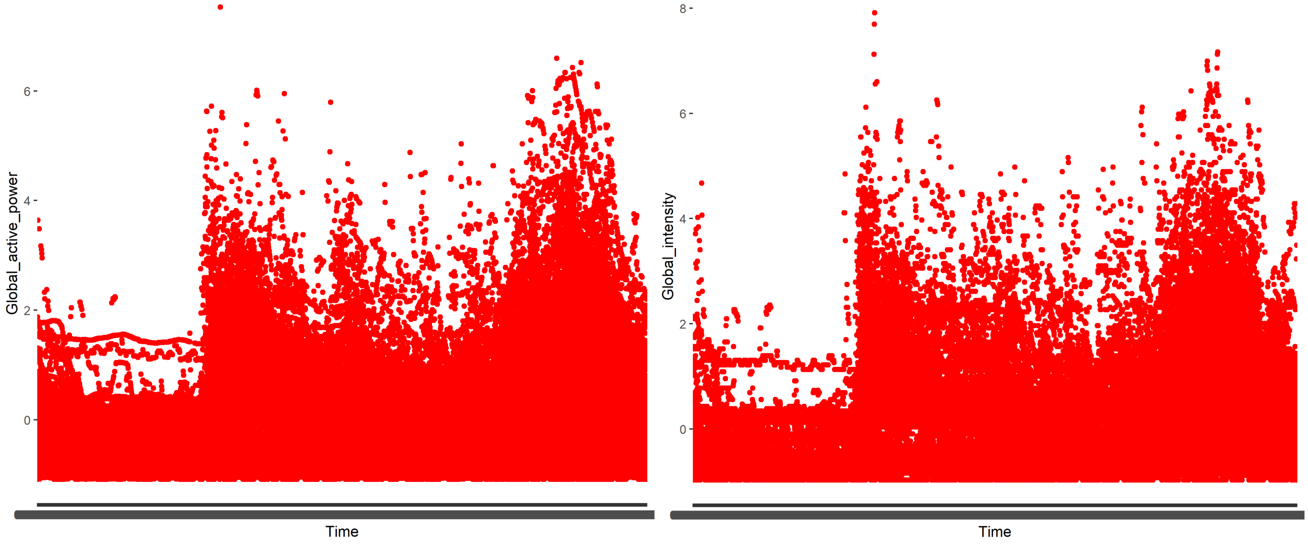

Figure 3.4 Global_active_power and Global_intensity plot for day 2

4. HMM Training and Testing

Before performing HMM training, we had to divide the scaled data into training data and test data. We decided to allocate 30% data for testing and the rest as training data as there is no validation data. Next, we used this data and the two response variables to train various multivariate HMM with different numbers of states. We started with models with 4 states and incremented by 4 until we reached 24 states.

States	BIC	LogLik	Normalized BIC	NormalizedLogLik
4	98130.554	-48901.6136	917.10798	-457.024426
8	61180.072	-30088.4866	571.77637	-281.200810
12	34444.242	-16213.7429	321.90880	-151.530307
16	21588.247	-9109.9738	201.75932	-85.139942
20	11125.963	-3034.1172	103.98096	-28.356235
24	7479.003	-196.9793	69.89722	-1.840928

Figure 4.1 HMM results for training data

States	LogLik	NormalizedLogLik
4	-22979.399	-478.73748
8	-15582.120	-324.62750
12	-10377.003	-216.18756
16	-7498.149	-156.21143
20	-4431.307	-92.31890
24	-3376.227	-70.33806

Figure 4.2 HMM results for test data

Based on the computed results and its performance, we decided that the best performing model was state 12 even though higher states have higher log-likelihood values and lower BIC values. For states 20 and 24, HMM sometimes failed to converge. For states larger than 12, the computation did not complete in a timely manner. In a critical Infrastructure where near real-time computation is required, this can lead to problems.

5. Anomaly Detection

After we decided which hidden markov model yielded the best results we used it to model the data sets with anomalous data points that were provided as we did with the test data from above.

Anomaly Number	LogLik	NormalizedLogLik
1	-24178.90	421.15
2	-24178.16	421.14
3	-36850.74	722.56

Figure 5.1 HMM applied onto potentially anomalous data

We then compared the log likelihood of the potentially anomalous data to the log likelihood of our training data for 12 states that we calculated above in figure 4.1.

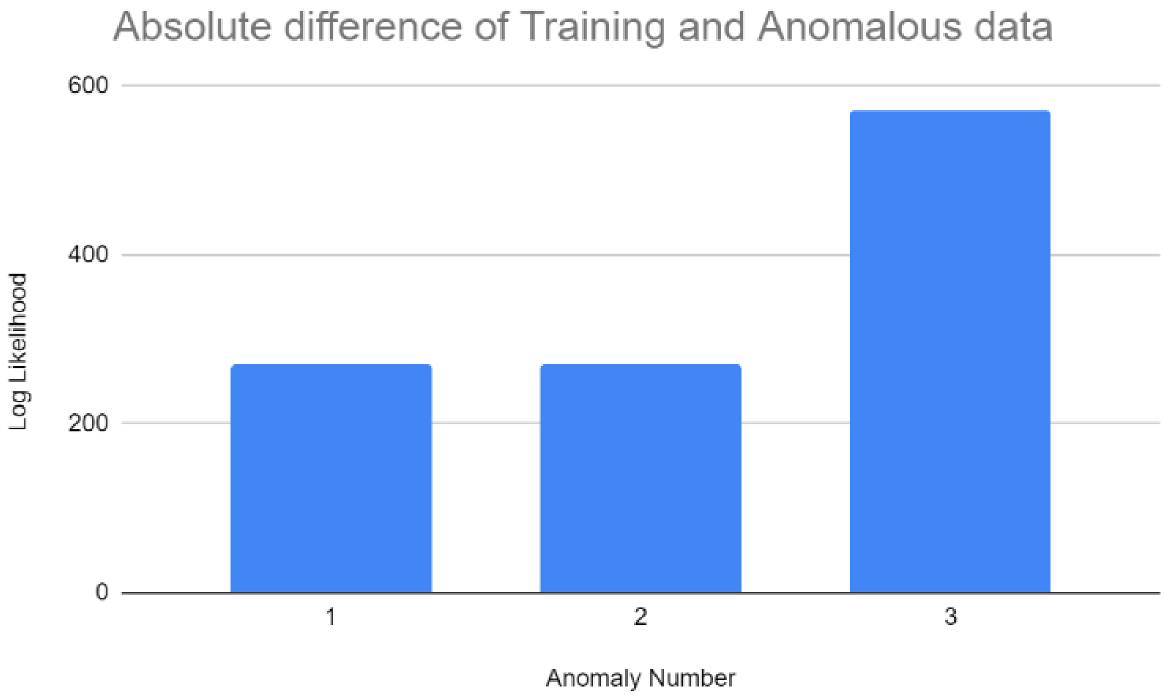

Figure 5.2 Absolute difference between Training and Anomalous data

From figure 5.2 we can see that data set number 3 has a substantially higher difference in log likelihood compared to both data set number 1 and 2. This suggests that data set number 3 is an Anomalous dataset which may contain anomalous data within it. A note to make here is that both dataset 1 and 2 both had the exact same log likelihood which is surprising because of the amount of unique data that is in both data sets.

6. Problems Encountered

Long training and testing time for models

For HMM training, as the number of states increased, the computation started to take a significant amount of time to train and test the model. For states 20 and 24, it took more than an hour for the script to complete and after a long time they would sometimes fail to converge.

7. Conclusion

During this project the team was able to experience what it would be like to work with large datasets and the challenges that come with real life datasets. Working with imperfect data that contained various types of anomalies made anomaly detection harder. On top of that, we realized how much the quality of training and test data impacts the statistical models.

8. Lesson Learned

Real Life data is noisy and messy

Data originated from the operation of a real-world system has many external factors and can be challenging to work with. Some characteristics, such as the imperfections in the data, outliers make it hard to detect anomalies and prevent false alarms. To ensure that the results were accurate and did not provide any misinformation, mean values were imputed into rows containing NA values. This gave us more accurate results compared to omitting the rows all together.

Applications of HMM

Hidden Markov Models are powerful predictive tools that can be used in a wide range of situations where the evolution of observable events depend on internal factors which are not directly visible. They have an efficient learning algorithm which allows them to perform a wide variety of operations such as structural analysis, pattern discovery, anomaly detection, etc. In our case, HMM was appropriate as the data was distributed stochastically .

References

Glässer, U. (2022). CMPT318: Cybersecurity, section 1 slides [PDF slides].

Glässer, U. (2022). CMPT318: Cybersecurity, section 2 slides [PDF slides].

Glässer, U. (2022). CMPT318: Cybersecurity, section 3 slides [PDF slides].