CMPT 318 - Cybersecurity

Simon Fraser University - Fall 2018

Project Report

Group 10

Team Members:

Fahd Chaudhry	- 301215679
Bowen Wang	- 301267523
Enhong Han	- 301324883
Orion Hsu	- 301283481
Rudy Kong	- 301215349

Abstract

Cybersecurity is a critical issue as more and more people and devices involved into the cyber network. The importance of cybersecurity becomes aware by people as the relationship between their digital information and life in real world. The technique of cybersecurity is evolving fast recent years, there are many kinds of methods to help in process of keeping the system secure in every step, in this report, behaviour-based intrusion detection methods will be used to find the anomalies from a dataset of household power consumption in north America. Through the analysis, different kinds of anomalies will be introduced such as point anomalies and contextual anomalies. Hidden Markov Models (HMMs) will be used for detecting contextual anomalies. HMMs is a method of analysing the causation of an observation occurring and probability transaction between each states of causation. The analysis of a large dataset is difficult, so the dataset will be split due to it being a time series which will result in more reliable and clear findings. RStudio will be used as the analysis program to help process the dataset. Furthermore, the HMM and moving average packages (depmixS4 and TTR) in RStudio are easy to use and help create models and identify anomalies.

Table of Contents

Table of Figures

Background

Cybersecurity

Definition: “Cybersecurity is the practice of protecting systems, networks, and programs from digital attacks.”

Cybersecurity Facts:

1. The increasing amount of large-scale, well-publicized breaches suggests that not only are the number of security breaches going up — they’re increasing in severity, as well.
2. 68% of funds lost as a result of a cyber-attack were declared unrecoverable
3. 99% of computers are vulnerable to exploit kits
4. The most expensive computer virus of all-time cost $38.5 billion
5. Cybercriminals will steal an estimated 33 billion records in 2023. That’s according to a 2018 study from Juniper Research. The compares with 12 billion records Junipers expect to be swiped in 2018.

From the facts above, one can see the importance of cybersecurity has increased tremendously because cyber-attacks have increased in significant numbers. So, what can anyone do to improve cybersecurity? The basic idea revolves around improving the accuracy of anomaly detection by monitoring a range of internet activity and detect anomalies. Services which do not match with normal data instances may be dealt with using the appropriate measures. Another useful way to improve anomaly detection technology is using better mathematical models and better detection methods. However, this is not easy. As defenders, all attacks must be dealt with and all conditions possible conditions must be assumed. However, an attacker only needs to succeed once and focus on one aspect.

Behavioural-based Intrusion Detection

The stakes for potential cybersecurity breaches are higher than ever before, and so has the motivation to commit attacks on the billions of devices that are connected to the internet. This has resulted in an arms race between the defenders and the attackers, with the attackers at an advantage. To counteract this, cybersecurity experts have devised behaviour-based intrusion detection in contrast with traditional signature-based detection in order to dynamically prevent and stop threats from occurring without prior knowledge of its signature. This allows for an adaptable system that can detect new zero-day exploits, and specialized threats. Deviations from normal behaviour are known as anomalies, and there are three main types of anomalies that are used in this paper to detect potential threats and anomalous behaviour. The three types of anomalies identified are point anomalies, which is known as a simple anomaly, and contextual anomalies and collective anomalies which are known as complex anomalies.

There are many uses for behavioural based intrusion detection in a wide variety of industries that require strong cybersecurity systems, and industries that require management and processing of a vast amount of data to find anomalous points, such as banks, and military surveillance applications. The application of behavioural-based anomaly detection in this paper shows how the data provided in a training data set can then be applied to find anomalous data points in certain features in a separate test data set.

Point Anomaly

A point anomaly is a type of simple anomaly that can be simply described as a single point of data in a wider dataset that is outside of a range of values in comparison with the rest of data that is deemed normal. If a single point in a dataset lies well beyond what other points are located at, then that point is considered to be a point anomaly. Figure [1] gives an example of a point anomaly in a plot of data. A real-world depiction of this could be the plots of a person's credit card spending. A point anomaly in this scenario would be the single point that is far away from where the rest of the points are.

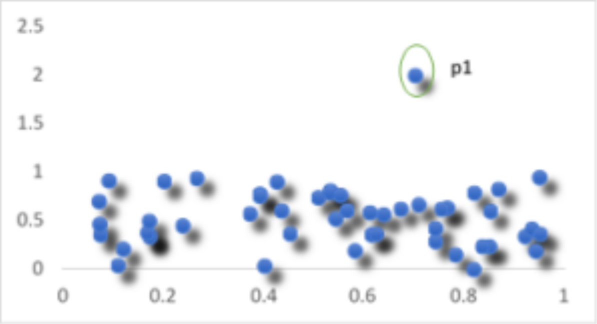

Figure [1].

Contextual Anomaly

A contextual anomaly in comparison to a point anomaly requires more investigation to determine if it is an actual anomaly, or whether it is contextually within normal behaviour given the context of the data point, which results in contextual anomalies being more complex in nature. This is due to the contextual anomalies requiring context in terms of the type of data, and the situations that the data point could be a part of, as well as the other data points in the data set. The restriction of determining contextual anomalies is that the data points must be related with each other, for example, a time series set of data.

Contextual anomalies can be caused by a variety of factors, but one is most clearly shown using an example regarding credit card fraud. If a credit card has an average low amount of spending, but in a short period of time increased its spending, the credit card's owner would be notified of this anomalous behaviour, however given context this may or may not be a contextual anomaly - if the high spending were to be during Christmas season, then the high spending would be within context of the time period. Figure [2] shows a data that can be interpreted as a contextual anomaly if, using the previous example of credit card fraud, this spike in spending was made mid-year, in June, where there would not be a trend of higher spending.

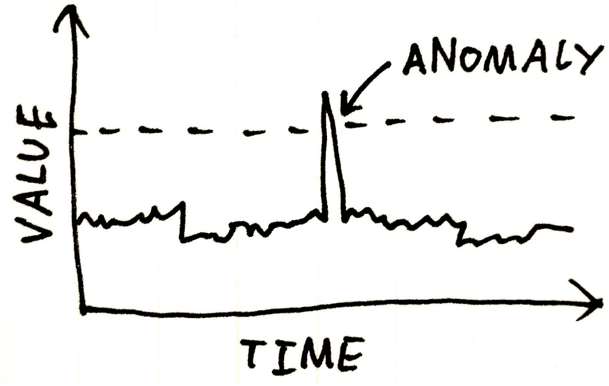

Figure [2].

To determine context for each particular instance of data, two sets of attributes must be determined. The first being contextual attributes, which is used to determine where along the data set a given instance is, an example being time in time-series data. The second being behavioural attributes, which determine the non-contextual parts of a data instance, which can be described using an example where the data set is a spatial data set describing the average rainfall worldwide, and the specific amount of rainfall in a given location is a behavioural attribute.

Collective Anomaly

The final type of anomaly to be defined is the collective anomaly. Collective anomalies are complex anomalies that is defined to be “a collection of related data instances [that] is anomalous with respect to the entire data set”, where each individual data point may not be anomalous, but when observed together are. This could potentially be a sequence of otherwise normal events, but when connected together become a collective anomaly.

An example of a collective anomaly is shown in Figure [3]. Figure [3] shows the rhythm of a patient connected to an electrocardiogram, or ECG/EKG, which shows the electrical signals that correspond to heart activity. The sequence data points shown in red highlight a sequence that contains measurements that would be seen as normal if viewed individually, but this specific sequence of heart activity is abnormal in comparison with the rest of the pulses, therefore it can be considered to be a collective anomaly.

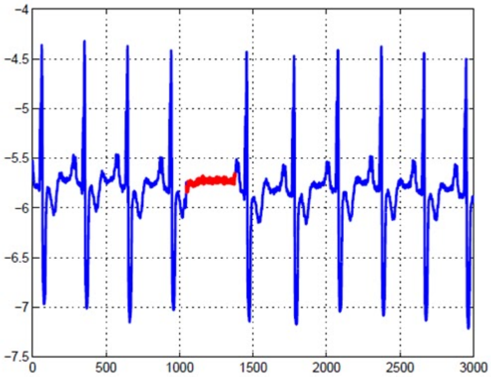

Figure [3].

Situational Awareness

Nowadays, it is mandatory for most companies, big or small, to have meaningful cyber situation awareness to safeguard sensitive data, sustain fundamental operations, and protect national infrastructure. Our growing reliance on the internet and the internet of things has greatly increased the need for situational awareness. This means that understanding your environment and accurately predicting and responding to potential problems that might occur. Systems that rely on the internet and networks are susceptible to vulnerabilities that present significant risks to both individual organizations as well as national security. By anticipating what might happen to these systems, companies can develop effective countermeasures to protect their critical infrastructure. Behaviour based anomaly detection is one the techniques that professionals can use to be situationally aware and safeguard their infrastructure. [5]

A car equipped with a GPS based navigation system to help the driver reach a particular destination is making the driver more situationally aware as they can see the different routes and traffic information and may avoid areas which may lead to traffic jams.

Critical Infrastructure

Critical Infrastructure is a term that refers to processes, systems, facilities, technologies, networks, assets and services of a nation or country. In Canada, these attributes are required to ensure the health, safety, security and economic well-being of Canadians as well as the effective functioning of government. This infrastructure may be stand-alone or may be interconnected and interdependent within and across provinces, territories and national borders. A compromise within the critical infrastructure of Canada or any country for that matter could result in catastrophic loss of life, adverse economic effects and significant harm to public confidence. A risk-based approach for strengthening the resiliency of Canada’s vital assets and systems including food, supply, electric power grids, transportation, communications and public safety systems is established by The National Strategy and Action Plan for Critical Infrastructure. [6]

Introduction

Problem Scope

This project consists of an analysis of a dataset referring to household power consumption in a specific portion of the power grid. The aim of this project is to explore behaviour-based intrusion detection methods used for cyber situational analysis of automated control processes, such as Hidden Markov Models and anomaly detection methods. As the power grid is a critical infrastructure, not only individual consumers but also industry consumers can get affected by the fluctuation of it. The whole dataset will be split into specific period of times, Sunday mornings and Sunday nights. The challenges faced are that the data is not guaranteed to be perfect, there is a lack of ground truth in certain cases, no labels are added to categorize the data as normal or anomalous, and the various type of anomalies. These defects increase the difficulty of distinguishing anomalies from noise and simple error. In the rest of this report, there will be a list of different approaches and methods used in the analysis such as the min-max and moving average methods for point anomalies and Hidden Markov Models for contextual anomalies. Lastly, the report will go over the result of the approaches and discuss some problems occurred and encountered through the course of completing this project.

Our Approaches and Methods

Phase 1: General Data Exploration

Before using the dataset to be used for analysis, there are many preparations to be done before processing the data. The imperfection of the dataset requires the removal of noise and null values from it. Noise data would result as a point anomaly incorrectly and lead to an inaccurate Hidden Markov Model. Null data would increase the inconvenience of further data processing.

Once the dataset has been cleaned, the dataset is split into parts based on time periods to make the trends more clear, concise and easier for further analysis. The dataset contains power consumption data on each minute in three years. Assumptions were made that there would be some behaviour-based trends on specific days of the week and trends on specific times of the day. To test these assumptions, the dataset was split into Sunday mornings and Sunday nights respectively.

There are 9 columns of data, which are:

- Date: Date as dd/mm/yy
- Time: Time in the day as hh:mm:ss
- Global_active_power: Household global minute-averaged active power (in kilowatts)
- Global_reactive_power: Household global minute-averaged reactive power (in kilowatts)
- Voltage: Minute-averaged voltage (in volts)
- Global_intensity: Household global minute-averaged current intensity (in ampere)
- Sub_metering_1: First smaller partition of the bigger grid

- Sub_metering_2: Second smaller partition of the bigger grid
- Sub_metering_3: Third smaller partition of the bigger grid

There is one training dataset and five testing datasets provided. The training dataset contains 1,556,444 rows of data and each test dataset has 518,816 rows of data.

Use *as.POSIXIt()* function to get the date format set as *%d/%m/%y*, and use *weekdays()* function to set the day of week on each day for later split. The standard Time format *%H:%M:%S* is used.

Phase 2. Anomaly detection approach

Approach 1: Finding Point Anomalies

Value of data is assumed to be normally distributed and use the 99.5% confidence interval of the normal model to define the threshold for anomalies detection, all the data above or below the interval bound would be regarded as anomaly. To get the upper and lower bound of the interval, the mean (μ) and standard deviation (σ) were calculated for each test dataset. Specific value matches were found with the confidence level needed from the normal distribution table, 2.81, for which the interval can be calculated by $\mu \pm 2.81\sigma$.

After the anomaly threshold have been obtained, the moving average method will be used to identify point anomalies. The moving average is used to smoothen the curve and reduce the effect created by the noise value. In this project, 7 observations for the window size were used, which means the average of 7 nearby observations was calculated for each data point. The size of the window affects the smoothness of the curve and the intensity of avoiding noise. If the size too big, the intensity of avoiding noise would be high which may lead to miss some anomaly as normal, if the size too small, the intensity of avoiding noise would be low and lead to normal value identified as anomaly by mistake.

Approach 2: Building HMMs and calculating log-likelihood

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modelled is assumed to be a Markov process with unobserved (i.e. hidden) states. (Wikipedia, 2018). In this project, the observations of electricity consumption features, neither the states nor transactions are known or visible, therefore HMMs were used to find proper amount of states and transactions through them.

As the observations are continuous values of electricity consumption features, continuous HMM rather than discrete HMM were used (which needs the observations to be isolated and no numerical relationship to each other).

Global Active Power is the amount of electricity consumed by households, which best reflects the actual consumption of electricity without the noise from electricity transaction. Due to the nature, time windows were built by splitting the dataset into Sunday mornings and Sunday nights. The given dataset is assumed to be the electricity consumption of houses. So, due to normal working schedule and lifestyle of people, majority of people work on weekdays so that less electricity would be consumed by household appliances during day time. However, during weekends, people have more time at home and may have different preferences of using electricity which make the weekends an interesting time frame for analysis.

In this project, the program trains a model with the number of states ranging from 2 to 15 to find the most proper HMM and compares the log-likelihood and BIC values. The model with a high log-likelihood and a low BIC would be regarded as the most suitable to do further tests with the given test datasets.

Our Solution

Finding Point Anomalies

Max-min

The table below shows the minimum value and the maximum value of Global active power for Sunday mornings and Sunday nights from the training dataset.

	Sunday Morning	Sunday Evening
Min value	0.078	0.080
Max value	7.552	8.592

Table [1].

Moving Average

With a fixed size window of 7 observations, selecting only Global active power as the feature, the threshold for each test dataset was calculated. Next, the difference of the value of the observation and the calculated average was calculated. When the difference is either above or below the threshold, that observation can be considered a point anomaly of the feature. Difference of the selected feature, Global Active Power, was plotted with the moving average of that feature against time in minutes. In the project, Test 1 and Test 3 datasets produced very similar results. Analysis on each of the test datasets was performed and Test 5 was found to be the most anomalous among the five datasets. The red points on each of the plots below represent the point anomalies in the respective dataset and the green points represent normal data.

Figure [4] shows the difference in Global Active Power for test 5 and its moving average with a window size of 7 for Sunday Mornings.

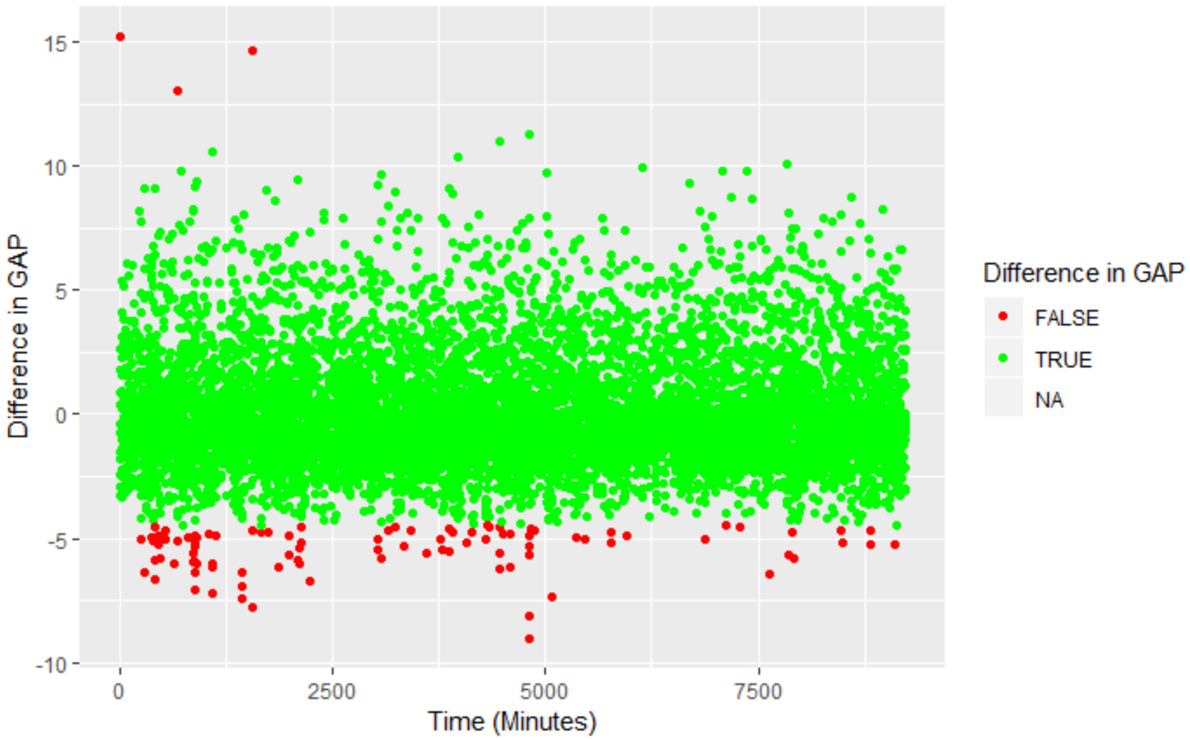

Figure [4].

Figure [5] shows the difference in Global Active Power for test 5 and its moving average with a window size of 7 for Sunday Nights.

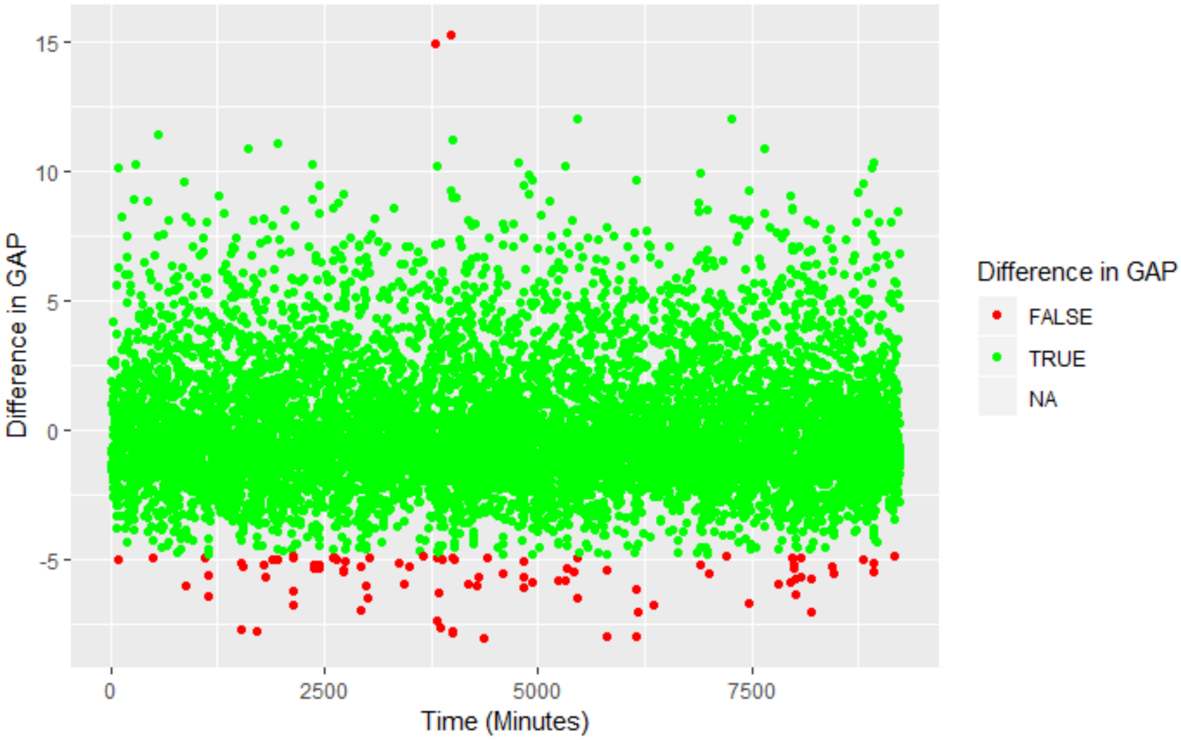

Figure [5].

For Contextual Anomalies

Building HMMs

HMMs were trained using the Global Active Power as the feature.

Sunday Mornings

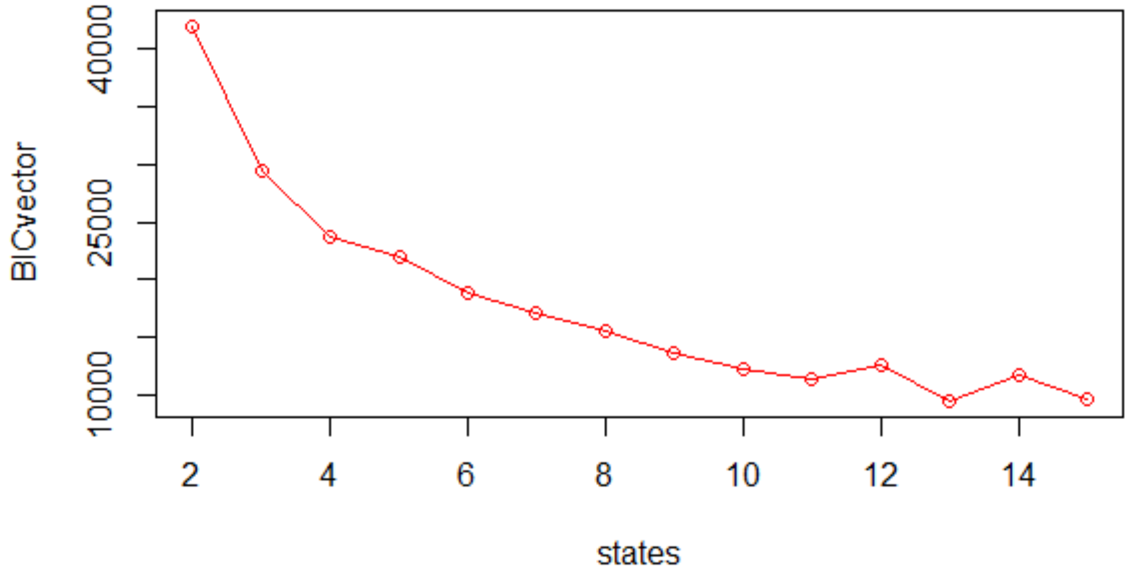

Figure [6].

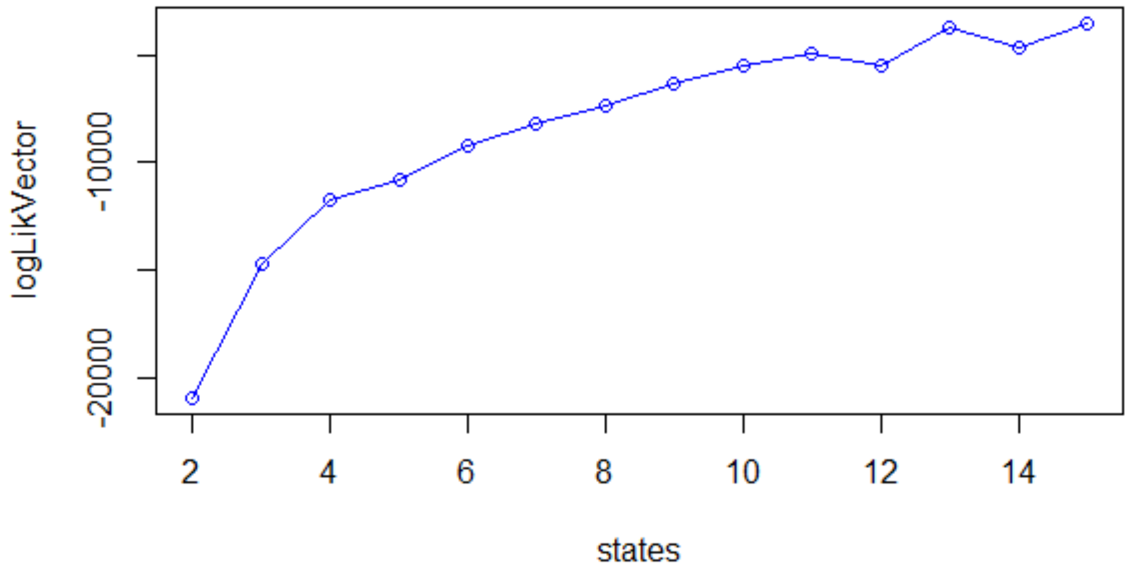

Figure [7].

The training data set was split into the time frame of Sunday mornings which are from 8 AM to 11 AM inclusively. The *depmix* function with 2 to 15 states was run on the dataset and the results were plotted on a graph with the respective BIC values and the negative log-likelihood values.

From the plots obtained, the ideal number of states for the Sunday mornings dataset is 11. The reason 11 states were chosen is because the BIC value goes up and the log-likelihood value goes down with 12 states. Any number of states higher than 11 seem to be overfitting the model.

The exact BIC value when the model has 11 states is: 11372.767.

The exact log-likelihood when the model has 11 states is: -4959.207.

Sunday Nights

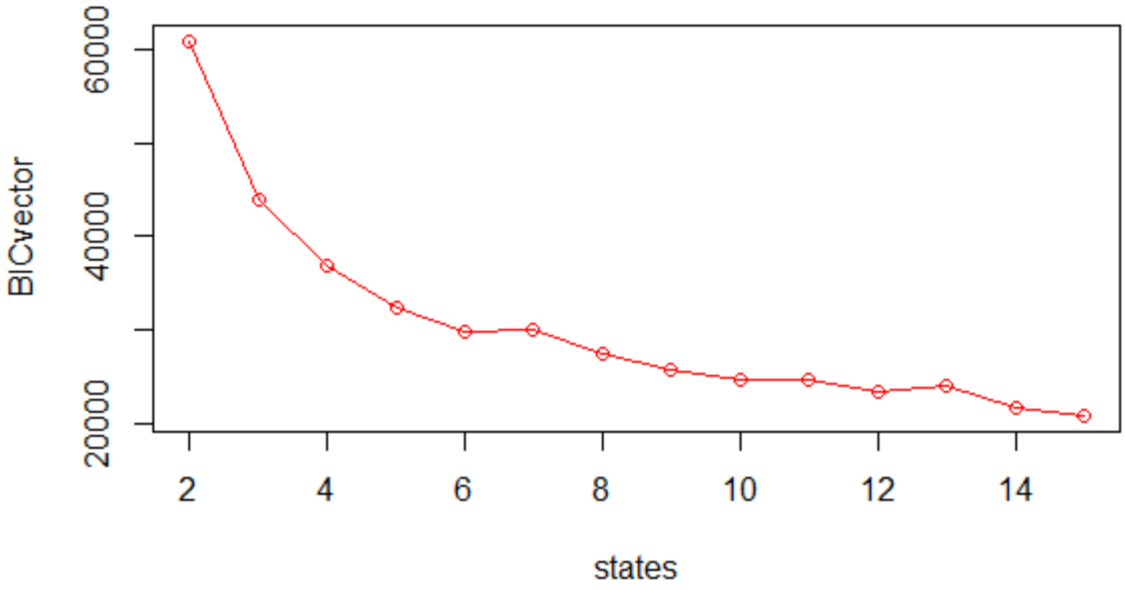

Figure [8].

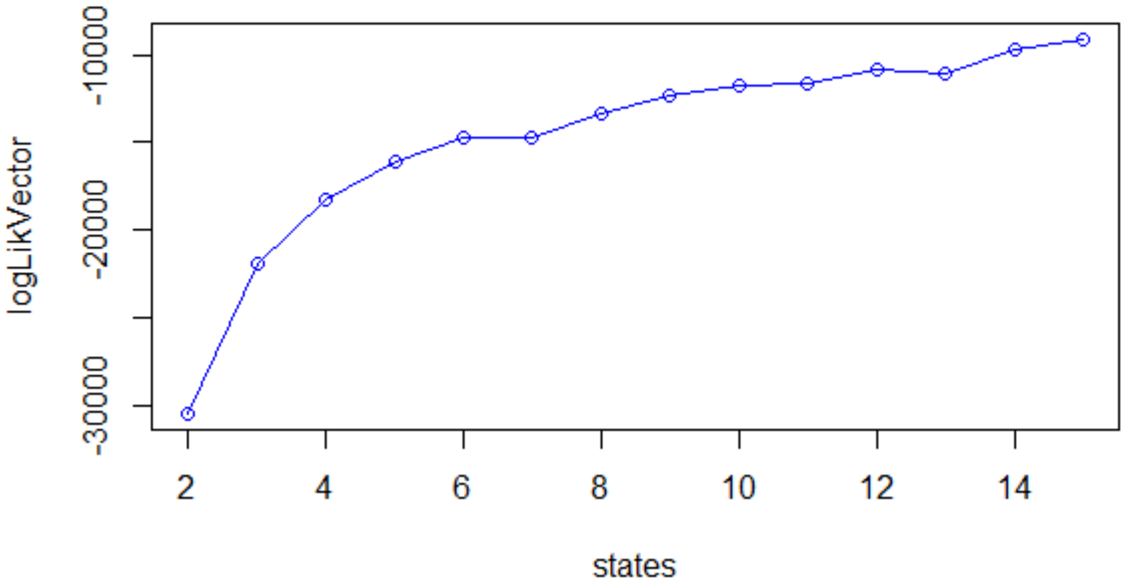

Figure [9].

The training data set was split into the time frame of Sunday nights which are from 9 PM to 12 AM inclusively. The *depmix* function with 2 to 15 states was run on the dataset and the results were plotted on a graph with the respective BIC values and the negative log-likelihood values.

From the plots obtained, the ideal number of states for the Sunday nights dataset is found to be 6. The reason 6 states were chosen is because the BIC value goes up and the log-likelihood value goes down with 6 states. Any number of states higher than 6 seem to be overfitting the model.

The exact BIC value when the model has 6 states is: 29801.86

The exact log-likelihood when the model has 6 states is: -14660.247

Calculating log-likelihood

The table below shows the log-likelihoods for Sunday mornings and Sunday nights from the training dataset and testing datasets.

	Sunday Mornings			Sunday Nights		
	logLik	No. of weeks	logLik/week	logLik	No. of weeks	logLik/week
Train	-4959.207	155	-31.99488	-14660.25	155	-94.58224
Test1	-9583.574	51	-187.9132	-11394.53	51	-223.4221
Test2	-9643.900	51	-189.0961	-11668.45	51	-228.7931
Test3	-9583.572	51	-187.9132	-11394.53	51	-223.4221
Test4	-19097.10	51	-374.4530	-20655.62	51	-405.0121
Test5	-19146.52	51	-375.4220	-20916.11	51	-410.1198

Table [2].

The log-likelihoods of Test 1, Test 2, and Test 3 are closer to the log-likelihood of the training dataset compared to Test 4 and Test 5. This tells us that there are much more anomalies in Test 4 and Test 5.

Major Problems Encountered

Detecting Anomalies

It can be difficult to discern what an anomaly is and what is still considered a normal value. To decide what the threshold is in approach 1 given the training dataset, and then looking at the equivalent plots in each testing dataset is difficult because even the training dataset is not completely perfect; there are still some noise and outliers that can affect the moving average values as a whole. Having too big of a threshold, then many anomalies will be missed due to them being included within the range of acceptable values, too small of a threshold, and there will be many false positives in the detection of anomalies.

Comparing log-likelihood

The log-likelihoods of the test datasets are different from the log-likelihood of the training dataset. This does not really tell us where the anomaly exactly is.

Conclusion

Our Findings

In our analysis, a variety of methods were used to analyse the dataset of household power consumption. Detecting anomalies in a large group of datasets and how to use Hidden Markov Models to observe a pattern about power consumption in time frames such as weekends, mornings and nights was learned in this project. Anomaly detection, which can influence the result directly was the most difficult part in the project because of two reasons. First, when different methods were used to detect anomalies, the results obtained are different from other methods using different parameters. This means that no one method can find all anomalies in a dataset. Therefore, if an accurate analysis of a dataset is to be required, several types of detection techniques must be used. Secondly, because the datasets being used are large, it is not easy to check whether all anomalies have been found. The method used to deal with this problem is using a different method to check the type of anomaly in each dataset. If similar results were obtained from each method of one dataset, it means majority of the anomalies in the dataset are found. Hidden Markov Models is arguably the most valuable. All anomaly detection methods were used for creating a better HMM to analyse people's behaviour. In this part of the project, the key was to choose suitable response, *ntimes* and states. The difficult part here is how to decide *ntimes* because the number of records in each Sunday is different. A suitable number for *ntimes* needs to be used. In the project, number of observations were calculated for each weekend and the result was used as the *ntimes* for each week.

Future Work

In this project, the work done was around anomaly detection and building HMMs. First, in anomaly detection, although points can be identified as outliers, they cannot be automatically removed from the dataset. Second, moving average was used to detect anomalies. A problem of moving average is when all the observations in a window are anomalies. The method does not work very well, and this problem was not solved in this project. Third, collective anomaly was not checked, only point anomaly and contextual anomaly were identified. Real world dataset is messier and have more types of anomalies. Fourth, the threshold used in anomaly detection was decided by normal distribution. The problem is that the dataset does not necessarily fits to a normal distribution. Finally, only Hidden Markov Model was used to fit our dataset. No other models were used to check if there is a better model that fits the training dataset.

Summing up all of those points above, there is room for more work. Given more time and more resources, further tests could be done. More experiments and more detections could have been explored. The moving average could be experimented furthermore making sure that a window does not consist of only anomalies. Investigating collective anomalies could be done in the future. Deciding on a better threshold could be more work as well. Having multiple features could also be explored in the future.

Lessons Learned

1. R is very useful to analyse data. There are lots of packages like *ggplot2*, *TTR*, *depmixS4* to help us analyse dataset by plotting, checking for outliers, create a fitting model using HMMs.
2. Real world data is more complex than imagined. More than 100 NA values and more than 100 outliers and noise were found in each test datasets.
3. Hidden Markov Model is very useful in finding a pattern. HMM can help us get log-likelihood of a feature in a dataset. This means two patterns in two datasets can be checked to see if they are similar or not.

References

1. "What Is Cybersecurity? - Cisco." https://www.cisco.com/c/en/us/products/security/what-is-cybersecurity.html.

2. "60 Must-Know Cybersecurity Statistics for 2018 - Varonis." 18 May. 2018, https://www.varonis.com/blog/cybersecurity-statistics/.

3. “10 Alarming Cybersecurity Fact” 01 May 2018 https://www.pcworld.idg.com.au/article/636083/10-alarming-cybersecurity-facts/

4. "10 cyber security facts and statistics for 2018 - Norton." https://us.norton.com/internetsecurity-emerging-threats-10-facts-about-todays-cybersecurity-landscape-that-you-should-know.html.

5. "Cybersecurity Situation Awareness | The MITRE Corporation." https://www.mitre.org/capabilities/cybersecurity/situation-awareness.

6. "Critical Infrastructure." 22 May. 2018, https://www.publicsafety.gc.ca/cnt/ntnl-scrt/crtcl-nfrstrctr/index-en.aspx.

7. "Anomaly Detection : A Survey - CUCIS." http://cucis.ece.northwestern.edu/projects/DMS/publications/AnomalyDetection.pdf.

8. Glässer, Uwe. "318 Script SECTION 3 slides 14-85." *Simon Fraser University*, https://vault.sfu.ca/index.php/s/oBuzplgh1Uhip1s#pdfviewer

9. "Anomaly Detection Simplified - 'Factspan's." https://factspan.com/anomaly-detection-simplified/. Accessed 24 Nov. 2018.