Project Report

CMPT – 318 Spring 2024

SPECIAL TOPICS IN COMPUTING SCIENCE – CYBER SECURITY

Instructor - Uwe Glaesser

TA(s) - Fatemeh Movafagh and Seyed Amir Yaghoubishahir

By Group – 8

Members

Sahaj Karan (301386551)

Anshal Chopra (301384760)

Arshnoor Singh (301401444)

Sakshi Singh (301386720)

Table of Contents

Abstract

The following is a technical report focused on unsupervised intrusion detection through time series analysis and forecasting within supervisory control systems. The report combines knowledge from previous assignments done in our course and covers various different methods such as feature scaling, training and testing of Hidden Markov Models to detect anomalies in a electric power grid data. We utilize knowledge about Principal Component Analysis for feature selection and we use log-likelihood and Bayesian Information Criterion to evaluate the models performance. The analysis also includes comparing response variables and selecting ideal obervation windows, justifying model choice and finding anomalies in the dataset. The report aims to contribute to the field of cybersecurity by finding insights of critical infrastructure against potential threats.

Introducing the problem scope:-

Task 1 - Feature Scaling

Feature Scaling – also known as data normalization is a preprocessing step in machine learning in which the range of independent variables or features are adjusted in a dataset. The idea behind it is to handle varying magnitudes of units. It involves bringing all features to a similar scale which allows all the features to equally contribute to model performance and so that no single features dominate the learning algorithm ["].

Feature scaling is necessary since many machine learning models such as (e.g., k-NN, SVM) use distance-based calculations to make predictions. Feature scaling makes them perform better and faster.

(I) Normalization - Adjusts the scale of the data so that the range falls withing a specific interval usually being [0,1] or [-1,1]. Done through min-max scaling.

Effect of normalization on data and noise – It can make the model more sensitive to small variances in the data and potentially could even amplify noise specifically when the range of data is narrow. ["] Mathematically, for a value x, normalization is calculated as (x' is the normalized values.)

$$x' = \frac{(\max(x) - \min(x))}{(x - \min(x))}$$

(II) Standardization – rescales a feature value so that it has a distribution with 0 mean value and variance equal to 1. Particularly helpful in cases where the data is normally distributed, and we want to compare the score between certain features.

Effect of standardization on data and noise – it maintains outliers and can reduce the effect of small variances. It keeps the outliers but does not bound the data like normalization does. The formula for standardization is: where x_mean is the mean and σ is the standard deviation of the feature vector ["].

$$x' = \frac{(x - \bar{x})}{\sigma}$$

(III) For Hidden Markov Models (HMM) used in anomaly detection, standardization is often more appropriate because:

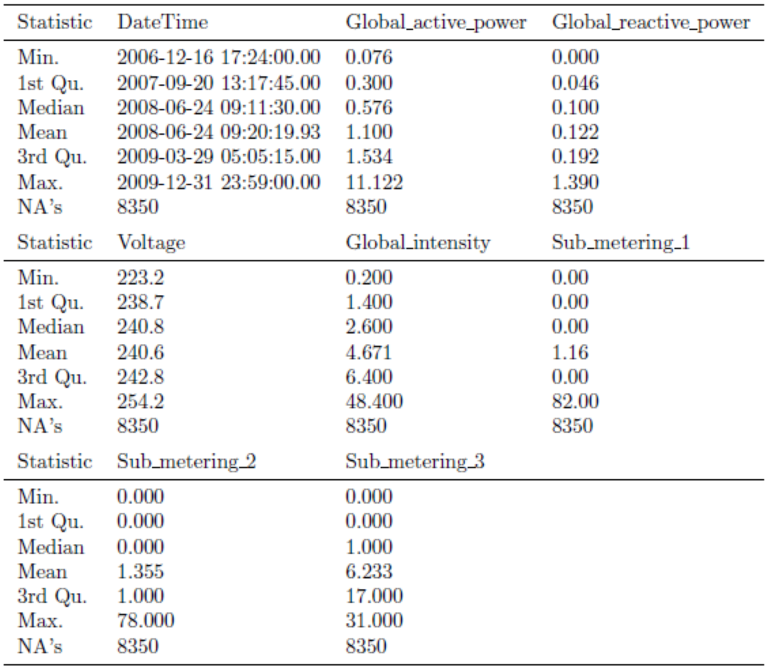
1.) Standardization makes the data more "normal" by aligning the mean and variance, which can be beneficial if Gaussian distributions are used in the HMM.

2.) Anomaly detection inherently deals with outliers. Standardization maintains the effect of outliers, which are crucial for detecting anomalies.

3.) Standardization might be more aligned with the goals of anomaly detection using HMMs due to its compatibility with Gaussian assumptions and its ability to preserve outliers. However, we should make the final decision based on specific characteristics of the dataset and the nature of the anomalies to be detected.

Task 2 - Feature Engineering

Before selecting subset of response variables for training of multivariate HMMs on normal electricity consumption data it is a good idea to inspect the data we have and do some initial analysis on it.



Statistic	DateTime	Global_active_power	Global_reactive_power
Min.	2006-12-16 17:24:00.00	0.076	0.000
1st Qu.	2007-09-20 13:17:45.00	0.300	0.046
Median	2008-06-24 09:11:30.00	0.576	0.100
Mean	2008-06-24 09:20:19.93	1.100	0.122
3rd Qu.	2009-03-29 05:05:15.00	1.534	0.192
Max.	2009-12-31 23:59:00.00	11.122	1.390
NA's	8350	8350	8350
Statistic	Voltage	Global_intensity	Sub_metering_1
Min.	223.2	0.200	0.00
1st Qu.	238.7	1.400	0.00
Median	240.8	2.600	0.00
Mean	240.6	4.671	1.16
3rd Qu.	242.8	6.400	0.00
Max.	254.2	48.400	82.00
NA's	8350	8350	8350
Statistic	Sub_metering_2	Sub_metering_3	
Min.	0.000	0.000	
1st Qu.	0.000	0.000	
Median	0.000	1.000	
Mean	1.355	6.233	
3rd Qu.	1.000	17.000	
Max.	78.000	31.000	
NA's	8350	8350	

Based on the summary statistics of the test data we observe that the data set we have has the following key points.

- **Time Span**: About 3 years of data.
- **Missing Data**: 8,350 missing values for each variable.

- **Voltage**: Average around 240.6V, indicating stable supply levels.

- **Global Intensity**: Wide range from 0.2 to 48.4, with a mean of 4.671, showing fluctuating power consumption.

- **Sub-meterings**: Highlight detailed consumption patterns across specific circuits, with Sub_metering_3 showing higher average consumption.

- **Data Distribution**: The mean higher than the median in some variables suggests a right-skewed distribution, likely due to periods of high consumption or outliers.

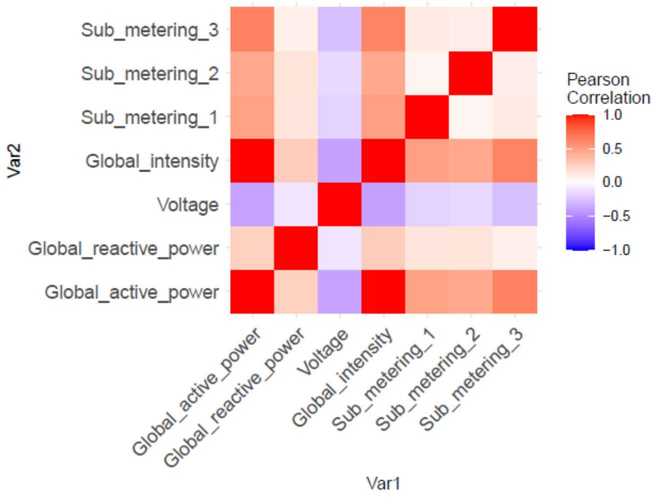
- **Outliers**: Significant differences between the maximum values and the third quartiles in some variables suggest outliers or very high consumption instances.

The next thing we analyze is a correlation matrix plot that we created to draw further insights on the data.

The key observation drawn from the correlation matrix plot were.

- **Global Active Power** is highly correlated with **Global Intensity**, which indicates that as the power consumption increases, the intensity of the usage increases as well.

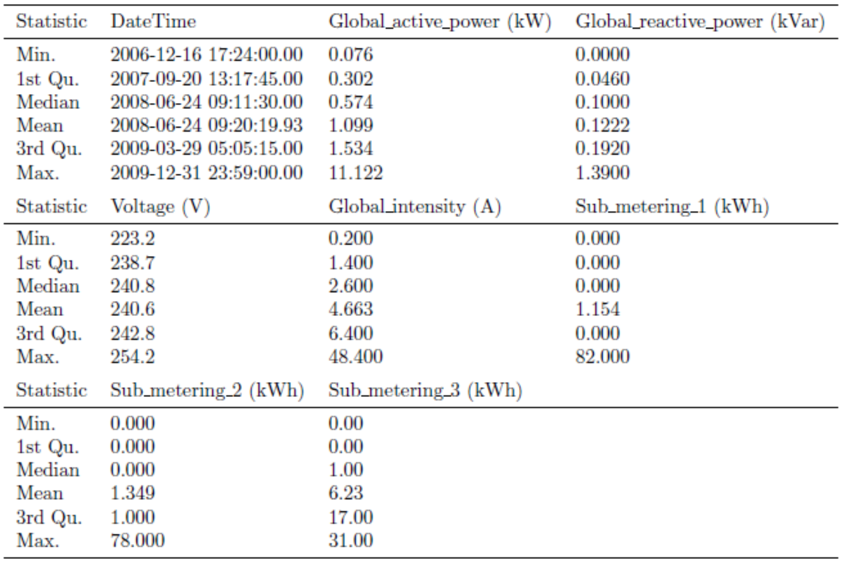
- Sub Metering 1, Sub Metering 2, and Sub Metering 3 show varying degrees of correlation with Global Active Power and Global Intensity, suggesting that these sub-meterings capture different aspects of the overall power usage.



- Voltage shows very little correlation with other variables, implying that voltage fluctuations have minimal direct association with the power consumption measures in this dataset.

Before proceeding further in our analysis from the previous summary statistics we had to address NAs in the dataset. To address the issue of 8,350 missing data points across all variables, interpolation methods were used, the dataset is now complete with estimated values filling the gaps. This makes it more robust for analysis, allowing for a comprehensive understanding of electricity usage patterns over the three-year period. Once we did that, we did look at the summary statistics for interpolated training dataset as well.

Another observation made on the data set is the magnitude of values are drastically different across variables. So, we created a box plot of means for all variables for more analysis.

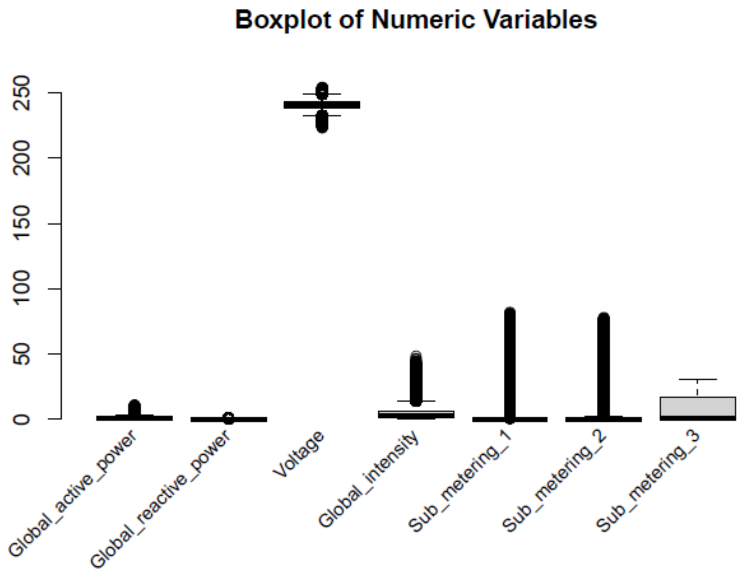


Statistic	DateTime	Global_active_power (kW)	Global_reactive_power (kVar)
Min.	2006-12-16 17:24:00.00	0.076	0.0000
1st Qu.	2007-09-20 13:17:45.00	0.302	0.0460
Median	2008-06-24 09:11:30.00	0.574	0.1000
Mean	2008-06-24 09:20:19.93	1.099	0.1222
3rd Qu.	2009-03-29 05:05:15.00	1.534	0.1920
Max.	2009-12-31 23:59:00.00	11.122	1.3900

Statistic	Voltage (V)	Global_intensity (A)	Sub_metering_1 (kWh)
Min.	223.2	0.200	0.000
1st Qu.	238.7	1.400	0.000
Median	240.8	2.600	0.000
Mean	240.6	4.663	1.154
3rd Qu.	242.8	6.400	0.000
Max.	254.2	48.400	82.000

Statistic	Sub_metering_2 (kWh)	Sub_metering_3 (kWh)
Min.	0.000	0.00
1st Qu.	0.000	0.00
Median	0.000	1.00
Mean	1.349	6.23
3rd Qu.	1.000	17.00
Max.	78.000	31.00

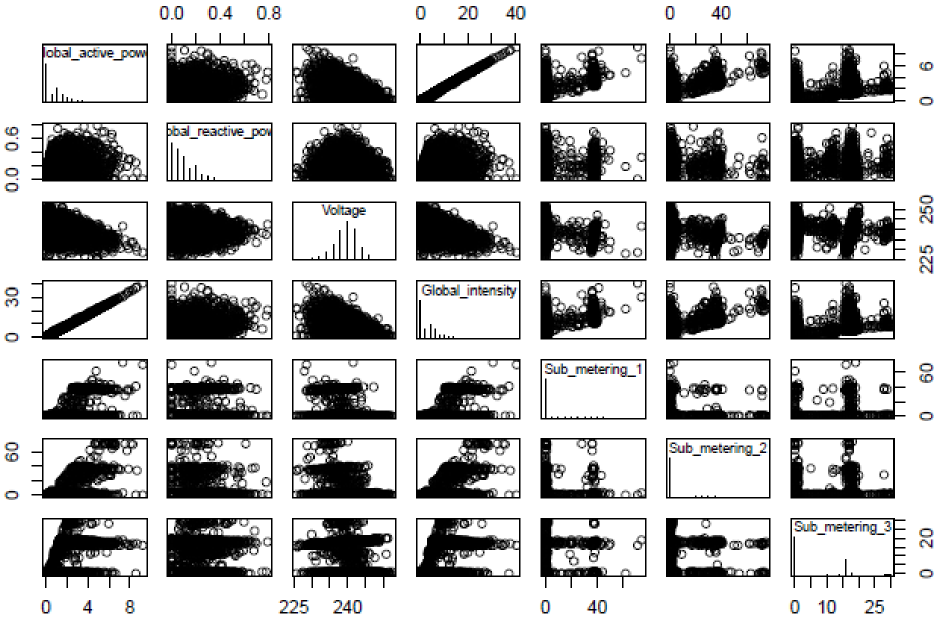
On initial observation we clearly observed that there is a scaling issue with the dataset. This is evident because the variables are measured on different scales, which is highlighted by the varying ranges of the boxplots:



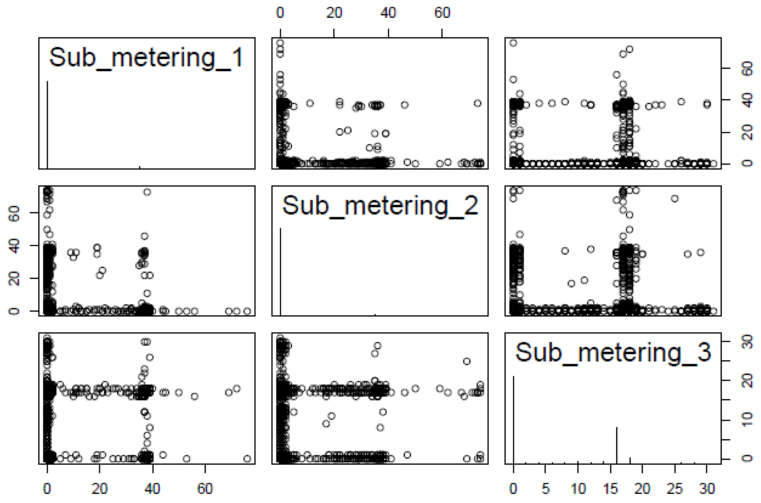
"Global_active_power" and "Global_intensity" have higher median values and a wider range, suggesting they are measured on a larger scale or simply have higher magnitude values. We also observe the presence of distinct outliers present in the variables.'Global_active_power' has several outliers above the upper whisker, and 'Global_intensity' has one outlier above the upper whisker.

Since scaling can lead to issues where larger scales may disproportionately influence the results.

We further wanted to understand the shape of the data set, so we also created a matrix scatter plot of the data.



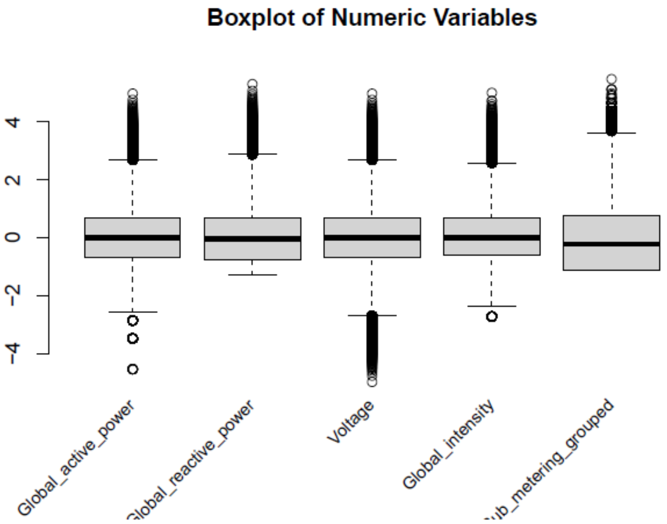
We see that most variables except voltage have a skewed shape. Whereas voltage is normally distributed.

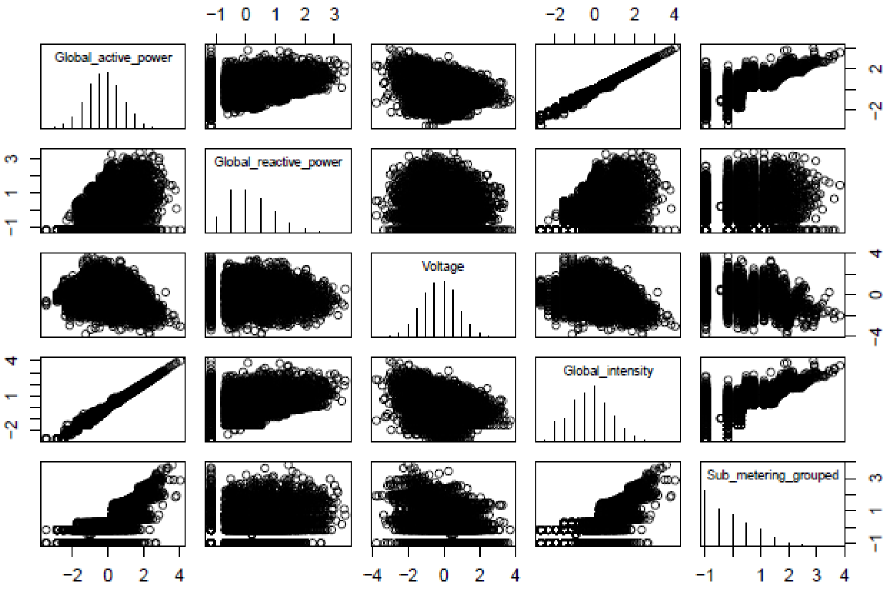


We also took a deep dive into the different Sub_metering groups. In our thorough examination of the different sub-metering groups, we observed distinct usage patterns that have implications for our anomaly detection model. Sub_metering_1 primarily measures the power consumption of the kitchen, and we noticed peaks during meal preparation times, which is expected. Sub_metering_2 tracks power usage in the laundry room, showing increased activity during early morning and late evenings, aligning with typical household behavior. Sub_metering_3, which captures the electric heater and air conditioner consumption, presented the highest average consumption, with spikes correlating with seasonal temperature changes. These insights into specific power usage behaviors by circuits are crucial for distinguishing between normal consumption variations and actual anomalies that could indicate a cybersecurity threat.

We decided to address the issue of scaling by conducting normalization on the dataset as it is better to use it instead of standardization in the case of skewed data. We established that the data is skewed confirmed using the summary statistics and the matrix scatterplots.

We then applied transformations to the training set and the testing set where we used the bestNormalize package which applies the best normalization to each column. So, we finished this process we created the same plots again.



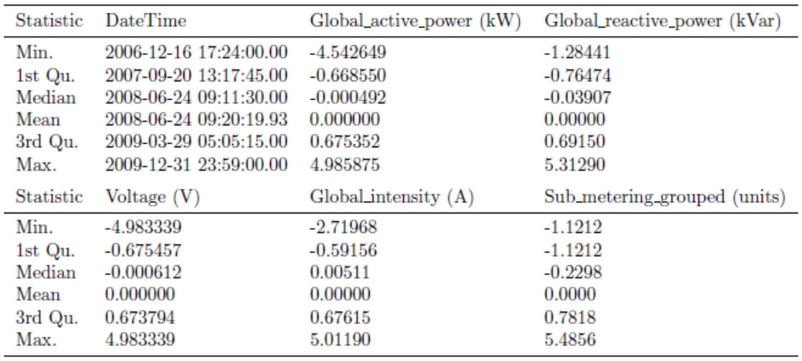


Observing the updated box plot of means we see that the scaling issue has been resolved.

Whereas observing the updated matrix scatter plot we see that the data is now less skewed and follows a more normally distributed.

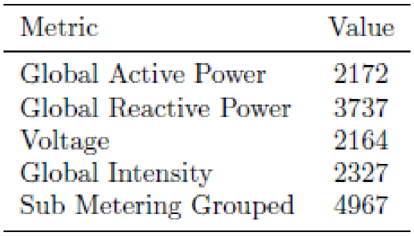
Once we had the data set scaled, we created summary statistics for it again so we could proceed further with our analysis.

After this point we created a function to detect anomalies across our dataset. The way we achieved this was by counting the number of data points that had a Z-score greater then 3 for the entire dataset. In the entire dataset we found that a total of 11446 values satisfied the condition.



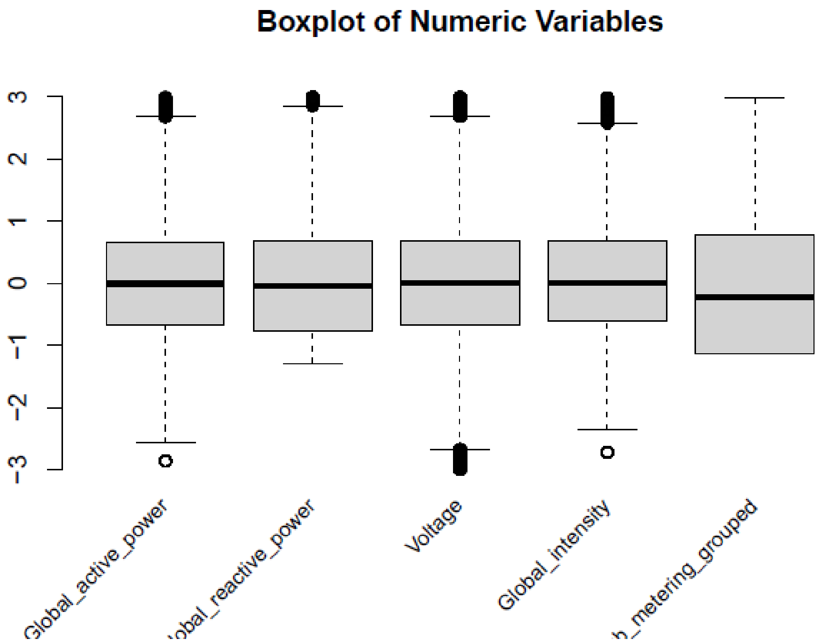
Statistic	DateTime	Global_active_power (kW)	Global_reactive_power (kVar)
Min.	2006-12-16 17:24:00.00	-4.542649	-1.28441
1st Qu.	2007-09-20 13:17:45.00	-0.668550	-0.76474
Median	2008-06-24 09:11:30.00	-0.000492	-0.03907
Mean	2008-06-24 09:20:19.93	0.000000	0.00000
3rd Qu.	2009-03-29 05:05:15.00	0.675352	0.69150
Max.	2009-12-31 23:59:00.00	4.985875	5.31290
Statistic	Voltage (V)	Global_intensity (A)	Sub_metering_grouped (units)
Min.	-4.983339	-2.71968	-1.1212
1st Qu.	-0.675457	-0.59156	-1.1212
Median	-0.000612	0.00511	-0.2298
Mean	0.000000	0.00000	0.0000
3rd Qu.	0.673794	0.67615	0.7818
Max.	4.983339	5.01190	5.4856

Number of data points with Z-score > 3 for each feature are as follows.



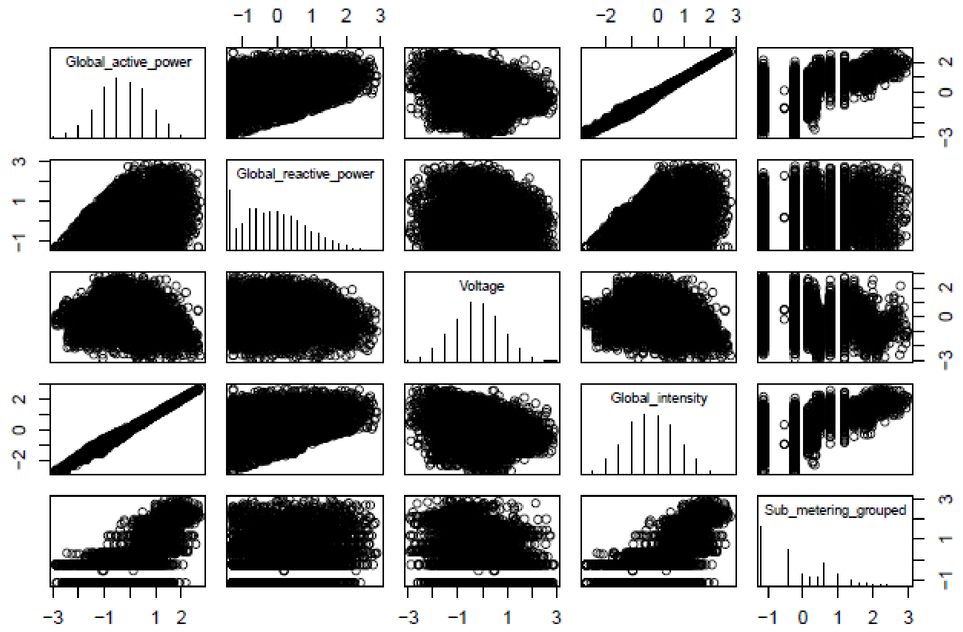
Metric	Value
Global Active Power	2172
Global Reactive Power	3737
Voltage	2164
Global Intensity	2327
Sub Metering Grouped	4967

We also calculated the percentage of anomalies in the entire dataset which came out to be 71.5 % Following this we shifted our focus to attempt to reduce the number of outliers in the data set based on the bounds that we calculated.



The resulting boxplot seems much better in terms of outliers as compared to previous iterations.

Also, after reducing the outliers our updated matrix scatter plot is as follows.



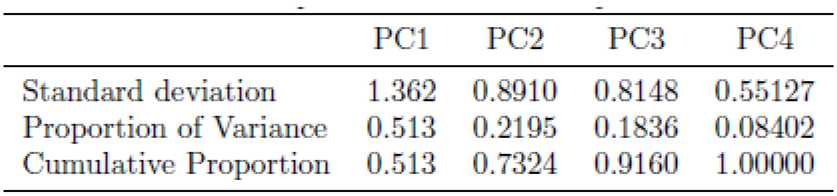
Also, our updated correlation matrix was as follows.

	Global Active Power	Global Reactive Power	Voltage	Global Intensity	Sub Metering Grouped
Global Active Power	1.0000	0.2999	−0.2874	0.9953	0.6730
Global Reactive Power	0.2999	1.0000	−0.1077	0.3310	0.2612
Voltage	−0.2874	−0.1077	1.0000	−0.3182	−0.2696
Global Intensity	0.9953	0.3310	−0.3182	1.0000	0.6807
Sub Metering Grouped	0.6730	0.2612	−0.2696	0.6807	1.0000

Observing the correlation matrix, we observed that Global_Intensity is highly correlated with Global_Active_Power hence we removed it. Upon analyzing the correlation matrix, we found a strong correlation coefficient of 0.95 between Global_Intensity and Global_Active_Power. This high degree of correlation suggests redundancy, as both variables contribute similar information to the model, with Global_Intensity being a derivative of Global_Active_Power and Voltage. In predictive modeling, such redundancy can lead to multicollinearity, where highly correlated predictors can distort the importance of feature weights, making the model unstable and sensitive to small fluctuations in the data. To mitigate this, we chose to exclude Global_Intensity from our feature set. This decision not only simplifies the model but also helps in enhancing the interpretability and generalizability of our anomaly detection approach. We also considered other multicollinearity remedies, such as principal component regression and ridge regression. However, given the nature of our dataset and the goal of interpretability in the context of anomaly detection, feature exclusion was deemed the most suitable approach.

With all preprocessing done we were ready for

Feature Engineering Using PCA – On conducting PCA we on the preprocessed data set we retrieved the following results.

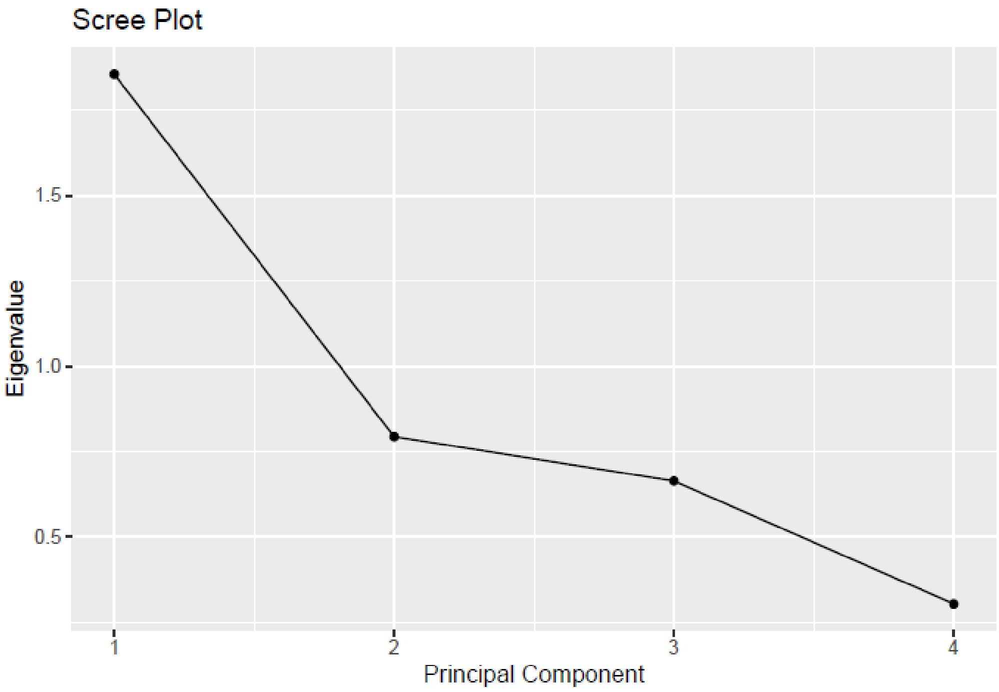


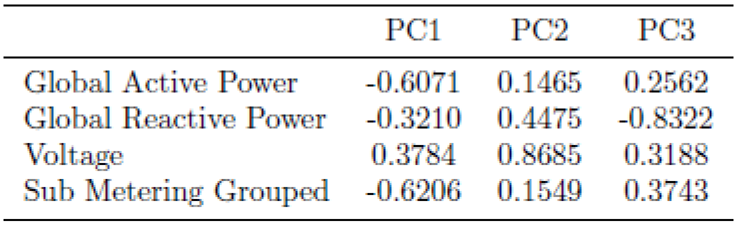
	PC1	PC2	PC3	PC4
Standard deviation	1.362	0.8910	0.8148	0.55127
Proportion of Variance	0.513	0.2195	0.1836	0.08402
Cumulative Proportion	0.513	0.7324	0.9160	1.00000

The first component is the most important as it captures the majority of the variance in the data. The first two components together account for a significant majority (73.24%) of the total variance, indicating that they might be sufficient to represent the original dataset for some purposes. All four components together describe the entire variance in the dataset.

Further we created a Scree Plot based on the eigen values calculated.

From the Scree plot we observe a sharp bend after 3 principal components, so we decided to retain 3 principal components. The eigen vectors of the 3 principal components retained are as follows.

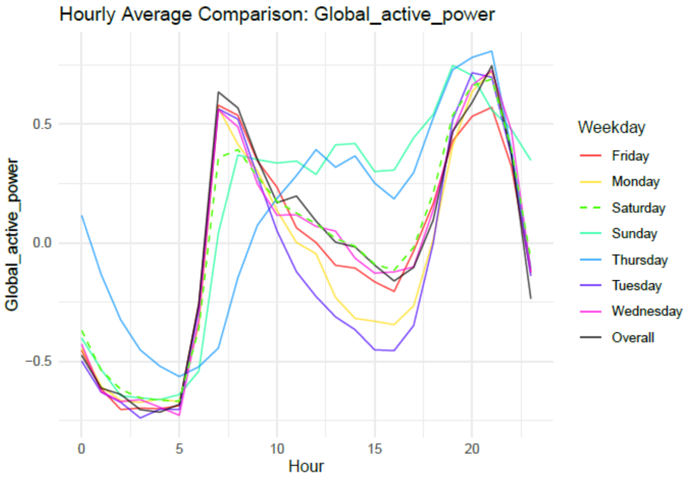




	PC1	PC2	PC3
Global Active Power	-0.6071	0.1465	0.2562
Global Reactive Power	-0.3210	0.4475	-0.8322
Voltage	0.3784	0.8685	0.3188
Sub Metering Grouped	-0.6206	0.1549	0.3743

In summary, PC1 seems to be strongly influenced by power consumption variables (Global_active_power and Sub_metering_grouped), but inversely. PC2 is most strongly associated with Voltage, indicating that it might represent the aspect of the data that is most related to voltage variation. PC3 appears to differentiate most strongly Global_reactive_power from the other variables.

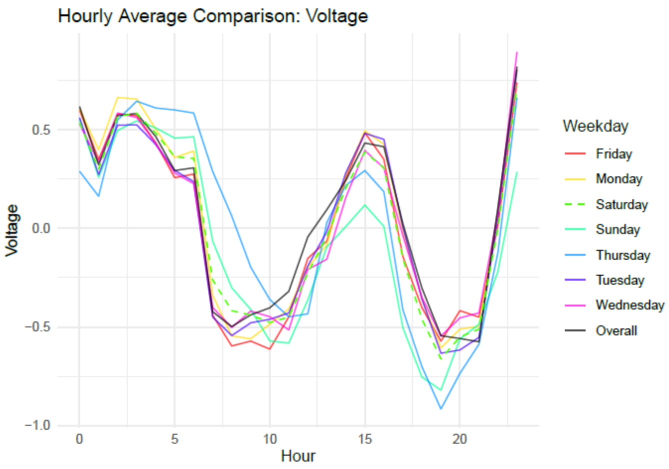
Filtering a Time Window – our next step was to create an hourly average comparision plot for all the different days of the week



From the created plots we filter the data where the time is between 12:00 – 16:00 on Saturday from the training dataset. And the time between 18:00 and 22:00 on Saturday for the test dataset.

From the generated plots we observe

- **Global Active Power:** Peaks around morning and evening hours, indicating higher consumption times.
- **Global Reactive Power:** Follows a similar pattern to active power but with less pronounced peaks.
- **Voltage:** Exhibits dips coinciding with the peaks in power consumption, suggesting a possible inverse relationship with demand.



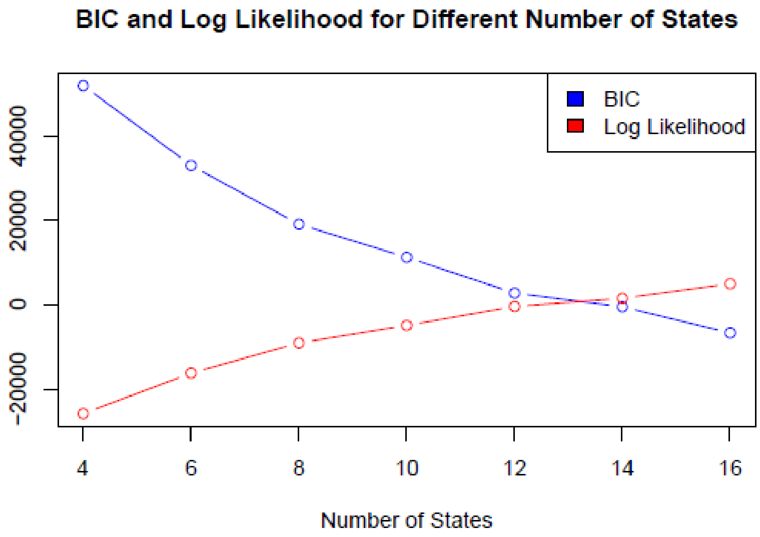
- **Sub-metering Grouped:** Shows a significant spike in the evening hours, potentially indicating a specific time-dependent usage of appliances or systems.

Fitting the model

1. Data Preparation and Train-Test Split: The dataset is divided into training and test sets based on a date that is 124 weeks from the earliest date in the data. This division is designed to validate the models that will be developed.

2. Weekly Aggregation: The training data is aggregated on a weekly basis. The number of records per week (ntimes) is then extracted and used in the model fitting process.

3. Model Fitting: A loop is established to fit Hidden Markov Models (HMMs) with a varying number of hidden states, ranging from 4 to 16, in increments of 2. Each model is fitted using the depmix function, which incorporates a Gaussian distribution for both 'Global_active_power' and 'Voltage'. Upon convergence of each model, the Bayesian Information Criterion (BIC) and log-likelihood values are computed.



1. Plot Interpretation: The plot illustrates the Bayesian Information Criterion (BIC) and log-likelihood values for models with different numbers of states. As the number of states increases, the BIC initially decreases, then slightly increases. This pattern suggests a

point of complexity beyond which adding more states does not improve the model according to the BIC criterion. On the other hand, the log-likelihood continues to increase with the number of states, indicating a better fit with more complex models.

2. Normalized Log-Likelihood: For model evaluation, the normalized log-likelihood (which is the log-likelihood divided by the number of rows in the dataset) is calculated for both the training data (state 10) and test data. The results suggest that the model performs better on the training data than on the test data, indicating potential overfitting or a change in the underlying process.

Selection of 10 States: The model with 10 states was chosen because it provided a good balance between model complexity and fit. The BIC, which penalizes model complexity, was relatively low for this model, suggesting that it was not overly complex. At the same time, the log-likelihood was relatively high, indicating a good fit to the data.

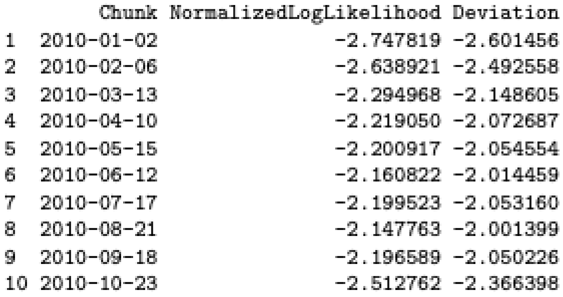
In our analysis, we evaluated the normalized log-likelihood values for both the train and test data. We found a value of -0.14 for the train data and a significantly lower value of -0.35 for the test data. This discrepancy suggests that our model, which consists of 10 states, fits the training data better than the test data, indicating a potential overfitting issue.

We preprocessed and divided our test data into 10 chunks, each representing different time periods. For each chunk, we calculated a normalized log likelihood and a deviation value. The maximum deviation observed was 2.60, indicating the largest discrepancy between how our model fits the training data versus any chunk of the test data.

We have also analyzed our test data in chunks, each representing different time periods

The deviation values represent the difference between the normalized log likelihood of our fitted model and the normalized log likelihood of these test data chunks. The maximum deviation observed was -2.601456, indicating the largest discrepancy between how our model fits the training data versus any chunk of the test data.

In our anomaly detection experiment, we introduced anomalies into one chunk of the test data. The deviation of the normalized log likelihood from the fitted model for this anomaly-containing chunk was 2.95. This value exceeds the maximum deviation observed in the test data (-2.601456), suggesting that our model is capable of recognizing and flagging data that deviate significantly from the expected pattern.



```
        Chunk NormalizedLogLikelihood Deviation
1  2010-01-02               -2.747819 -2.601456
2  2010-02-06               -2.638921 -2.492558
3  2010-03-13               -2.294968 -2.148605
4  2010-04-10               -2.219050 -2.072687
5  2010-05-15               -2.200917 -2.054554
6  2010-06-12               -2.160822 -2.014459
7  2010-07-17               -2.199523 -2.053160
8  2010-08-21               -2.147763 -2.001399
9  2010-09-18               -2.196589 -2.050226
10 2010-10-23               -2.512762 -2.366398
```

Conclusion

The HMM fitted on the training data demonstrates a consistent and reliable performance, as evidenced by the convergence diagnostics and the out-of-sample evaluation metrics. The selection of 10 states appears justified based on the BIC analysis, and the uniformity in the initial state probabilities and transition matrix suggests that the model is capturing the inherent randomness in the data’s structure. The consistent normalized log-likelihood values across the test data chunks further support the model’s validity and robustness.

Challenges and Lessons Learned

- One notable challenge when transferring trained model parameters to the testing phase was, we faced discrepancies in the lengths of parameters that required careful debugging and troubleshooting. This experience highlighted the importance of rigorous testing and validation procedures to ensure seamless integration between model training and deployment.
- Our code ended up requiring substantial computational resources to run efficiently, leading us to optimizing our code by making strategic use of available computing power.
- Additionally, we implemented standardization to scale our data but later realized that normalization was better suited for the skewed nature of our dataset.
- We also ended up creating a better model very late in the project and had to readjust out report.

Despite these challenges, we learned the importance of robustness and adaptability in designing effective intrusion detection systems for critical infrastructure.

References

Pickl, Shivani. “What Is Feature Scaling and Why Does Machine Learning Need It” *Medium*, 17 Sept, 2019, https://medium.com/@shivanipickl/what-is-feature-scaling-and-why-does-machine-learning-need-it-104eedebb1c9

Yaghoubishahir, Seyed Amir. “R Tutorials” Tutorial presented in CMPT 318 - Special Topics Cmpt. Science, Simon Fraser University, BC, Spring 2024

OpenAI. "ChatGPT." GPT-3.5, 2022. URL: https://openai.com/chatgpt.