Simon Fraser University

Spring 2024 CMPT 318 - D100

Application of Hidden Markov Models in Electric Power Consumption Anomaly Detection

Anna Rusinova

Tyler Oh

Daniel Kim

Brandon Quon

April 7, 2024

Abstract

The motivation for this study was to identify anomalies within electric power consumption data. To achieve this, we employed a Hidden Markov Model (HMM), hypothesizing its potential to differentiate between normal and anomalous consumption patterns effectively. Given an initial dataset with point anomalies and noise, the task was to perform data cleaning and feature scaling to prepare for a principal component analysis (PCA). Afterwards, we obtained test and train data that would be inputted into various HMMs, which would be compared to determine the best model for the task. In particular, it was discovered that certain numbers of states (nstates) within the GMM were particularly effective in preventing both overfitting and underfitting, which enhanced the model's robustness. Drawing from these findings, we conducted anomaly detection utilizing the chosen HMM model and threshold determined from the maximum deviation of log-likelihood values. Finally, the analysis produced promising results, confirming the effectiveness of the HMM approach in addressing anomaly detection challenges within electric power consumption data.

Table of Contents

Table of Figures

1 Introduction

1.1 Background

Cybersecurity is paramount in automated control systems due to their critical role in managing industrial processes and infrastructure. As these systems become increasingly interconnected, they are exposed to sophisticated cyber threats, including unauthorized access and sabotage, making them vulnerable. Anomaly detection plays a pivotal role by continuously monitoring system behavior for deviations from normal operation, which is crucial for promptly identifying and mitigating potential threats. By integrating robust cybersecurity measures and employing advanced anomaly detection techniques, critical infrastructures and industrial processes are safeguarded against potential disruptions and hazards, ensuring operational continuity and security.

1.2 Project Scope

This project is focused on developing an anomaly detection system for electric power consumption data using Hidden Markov Models (HMMs). Attacks on electrical systems are exceptionally critical due to their extensive reach across critical infrastructure. The potential for widespread disruption to essential services make them a paramount concern for public safety. By training the model on historical data to understand normal consumption behaviors, the system aims to accurately identify deviations. Through meticulous feature scaling and engineering, the project enhances the model's ability to interpret complex datasets effectively. Then, the best performing HMM model will be selected to identify anomalous observations. Selecting the most accurate HMM model for anomaly detection is crucial for identifying and responding to anomalous observations promptly. The system's adaptability to new and evolving threats makes it an invaluable component of a comprehensive cybersecurity strategy for protecting critical energy infrastructure.

1.3 Problem Statement

In the context of cybersecurity for automated control systems, there is a critical need for effective anomaly detection methods to identify deviations in electric power consumption data. Specific threats that this project aims to detect include unauthorized usage, meter tampering, ransomware attacks, and energy theft. Unauthorized access and usage leads to financial losses and disruptions for utility providers. Meter tampering alters consumption readings, resulting in inaccurate billing and revenue losses. Similarly, energy theft involves the illegal acquisition of energy resources through fraudulent means, leading to revenue loss and operational challenges. Ransomware attacks occur through infiltrating networks of the energy infrastructure and could encrypt data or disrupt operations. This leads to excessive payment in exchange for restoring access or functionality.

2 Methodology

2.1 Dataset Description

The dataset used in this project provides detailed measurements of electric energy consumption within a household, recorded at one-minute intervals over a period of time. It comprises a total of nine variables, each representing a different aspect of energy usage or the time at which the measurement was taken. Below is an overview of these variables, including their names, data types, and a brief description:

- **Date**: Recorded in the format **dd/mm/yyyy**. Although categorical in nature, this variable is treated as sequential in the context of time-series analysis, as it marks the progression of days over the period of study.
- **Time**: Captured in the format **hh:mm:ss**. Similar to the Date variable, Time is categorically structured but is treated as sequential, representing the minute-by-minute flow of measurements throughout each day.

- **Global_active_power**: Measured in kilowatts (kW), this continuous variable represents the total active power consumed by the household at a given moment. Active power is the portion of electricity that is used to perform work, such as running appliances.

- **Global_reactive_power**: Also measured in kilowatts (kW), this continuous variable indicates the total reactive power consumed. Reactive power is the portion of electricity that is used to maintain the voltage levels necessary for active power to perform work.

- **Voltage**: This continuous variable is measured in volts (V) and represents the minute-averaged voltage at which the electric energy is supplied.

- **Global_intensity**: Measured in amperes (A), this continuous variable reflects the total current intensity or the rate at which electric power is consumed.

- **Sub_metering_1**: Expressed in watt-hours of active energy (Wh), this continuous variable specifically accounts for the energy consumption in the kitchen, including devices such as the dishwasher, oven, and microwave. Notably, hot plates are powered by gas and are therefore not included.

- **Sub_metering_2**: This variable measures the watt-hours of active energy consumed in the laundry room, capturing the usage of appliances like the washing machine, tumble dryer, refrigerator, and a light.

- **Sub_metering_3**: Also measured in watt-hours, Sub_metering_3 tracks the consumption related to an electric water heater and air-conditioner, highlighting another significant area of energy use within the household.

2.2 Data Preprocessing

Description: df [25,979 × 9]

	Date <chr>	Time <chr>	Global_active_power <chr>	Global_reactive_power <chr>	Voltage <chr>	Global_intensity <chr>	Sub_metering_1 <chr>	Sub_metering_2 <chr>	Sub_metering_3 <dbl>
6840	21/12/2006	11:23:00	?	?	?	?	?	?	NA
6841	21/12/2006	11:24:00	?	?	?	?	?	?	NA
19725	30/12/2006	10:08:00	?	?	?	?	?	?	NA
19726	30/12/2006	10:09:00	?	?	?	?	?	?	NA
41833	14/1/2007	18:36:00	?	?	?	?	?	?	NA
61910	28/1/2007	17:13:00	?	?	?	?	?	?	NA
98255	22/2/2007	22:58:00	?	?	?	?	?	?	NA
98256	22/2/2007	22:59:00	?	?	?	?	?	?	NA
142589	25/3/2007	17:52:00	?	?	?	?	?	?	NA
190498	28/4/2007	00:21:00	?	?	?	?	?	?	NA

1–10 of 25,979 rows

Previous 1 2 3 4 5 6 … 100 Next

Table 1. Whole dataset before preprocessing

After inspecting the dataset, we found 25,979 rows containing missing values in the column ‘Sub_metering_3’. These missing rows were represented as “?” symbols in all other columns, indicating gaps in the recorded data. To maintain the integrity and continuity of the dataset, we took the following steps to address these missing values:

- We identified all occurrences of "?" in the dataset and replaced them with NA (Not Available) values to standardize the representation of missing data across all columns.
- To address the missing values without discarding valuable data points, we employed linear interpolation. This method allowed us to estimate and fill in missing values based on the available data points before and after the gaps. This approach was particularly suited for time-series data like ours, as it ensured no disruption in dates and times, preserving the chronological order and continuity of the dataset.

2.2.1 Feature Scaling

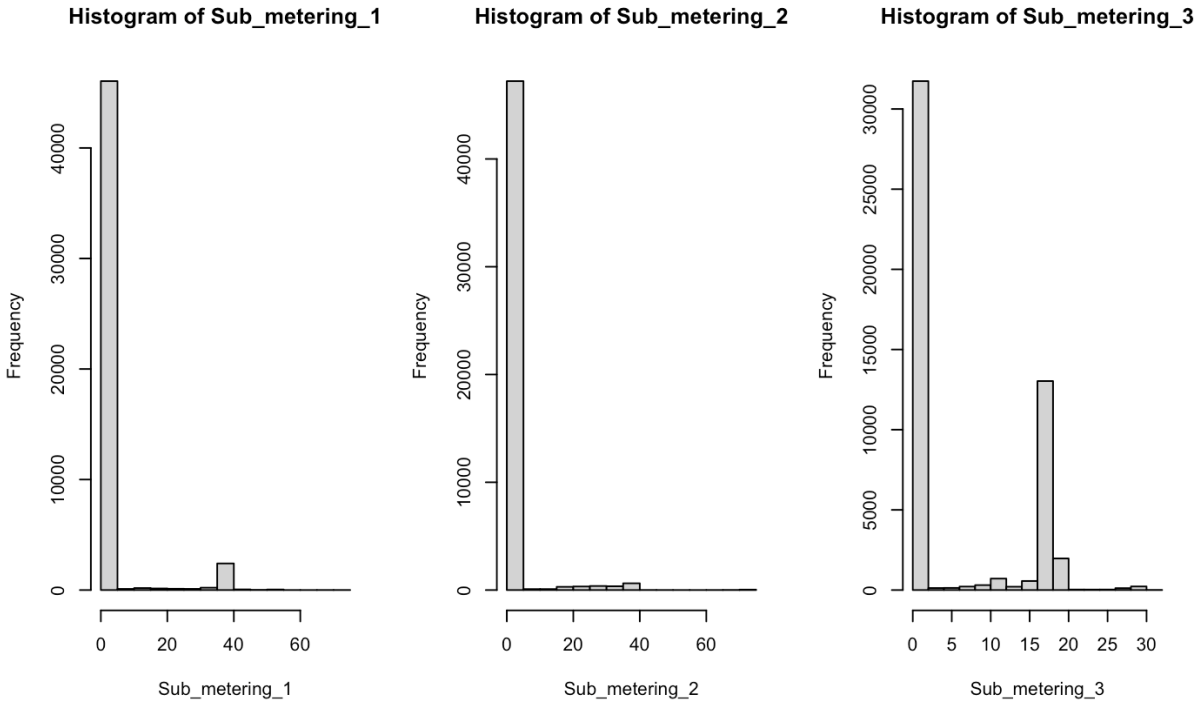

Figure 1. Histograms of Discrete Variables

Upon careful examination of the dataset, we observed that the variables related to Sub_metering (Sub_metering_1, Sub_metering_2, and Sub_metering_3), which track the energy consumption in different household areas, predominantly exhibit distributions with a large number of 0 values. IThis occurrence can be attributed to periods of no energy consumption in the corresponding areas, such as when appliances are turned off. While these observations might initially seem to provide insights into household energy usage patterns, their overwhelmingly zero-valued distributions suggest a limited variability that may not significantly contribute to distinguishing normal from anomalous energy consumption patterns in the context of our cybersecurity objectives. Given the characteristics of the Sub_metering variables, and the potential for these to introduce noise rather than meaningful information into our anomaly detection model, a decision was made to reassess their inclusion in our feature set. It was imperative to focus on features that offer substantial

variance and potential insights into energy consumption patterns that could be indicative of cybersecurity threats.

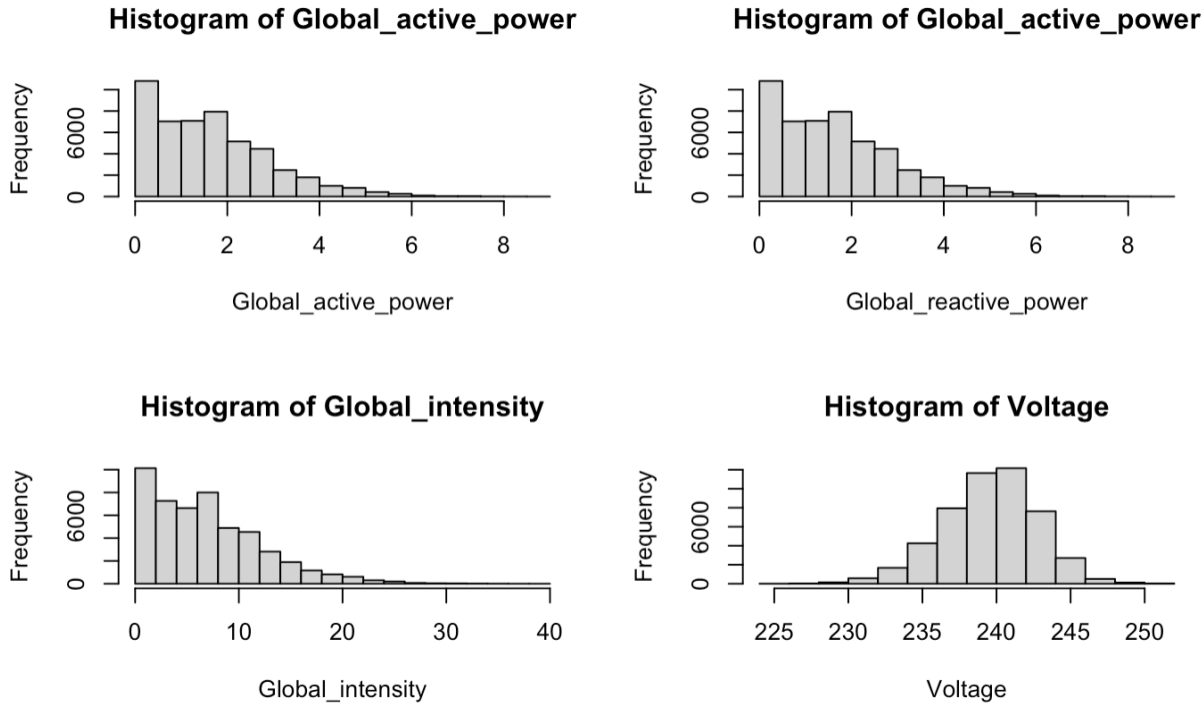

Figure 2. Histograms of Continuous Variables

The initial assessment of our dataset's variables revealed non-normal distributions for most, with the exception of the Voltage variable, which displayed a distribution closer to normality. The histograms of Global_active_power, Global_reactive_power, and Global_intensity showed skewed distributions with a high frequency of lower values, consistent with periods of low consumption. This suggests further scaling or normalization methods are needed.

i. When providing data to models, it is imperative that we submit parsable data to make analysis easier for machines. It ensures that the dataset's variables have comparable scales, which facilitates pattern analysis and aids in the detection of anomalies. Scaling is particularly crucial when models, such as the one we plan to use in the **depmix** function, make assumptions about the data distribution, specifically Gaussian distribution in our case.

ii. Normalization and standardization are two prevalent feature scaling techniques. Normalization is the process of rescaling the data such that the smallest value is mapped to

the lower end of the range and the largest value is mapped to the higher ends. It can be achieved using this min-max scaling equation:

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

Any noise after normalization may become more apparent to machine learning models, which can help with detection of the underlying signals.

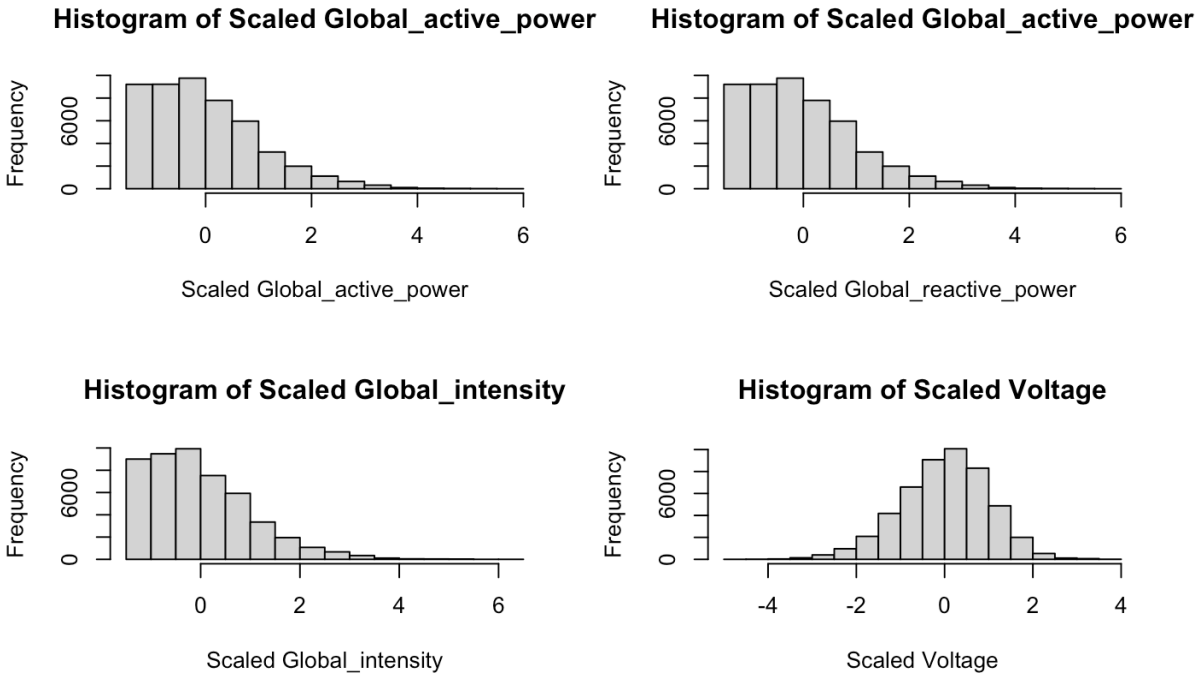

Figure 3. Histograms of Scaled Continuous Variables

Standardization, on the other hand, centers the dataset around its mean μ, and applies scaling based on its standard deviation σ. Any data point x_i that undergoes standardization outputs its own z-score, which can be calculated with this equation:

$$z = \frac{x_i - \mu}{\sigma}$$

By centering the data around its mean, we are able to prevent machine learning models (such as PCA) from using extreme weighting for features with high variances.

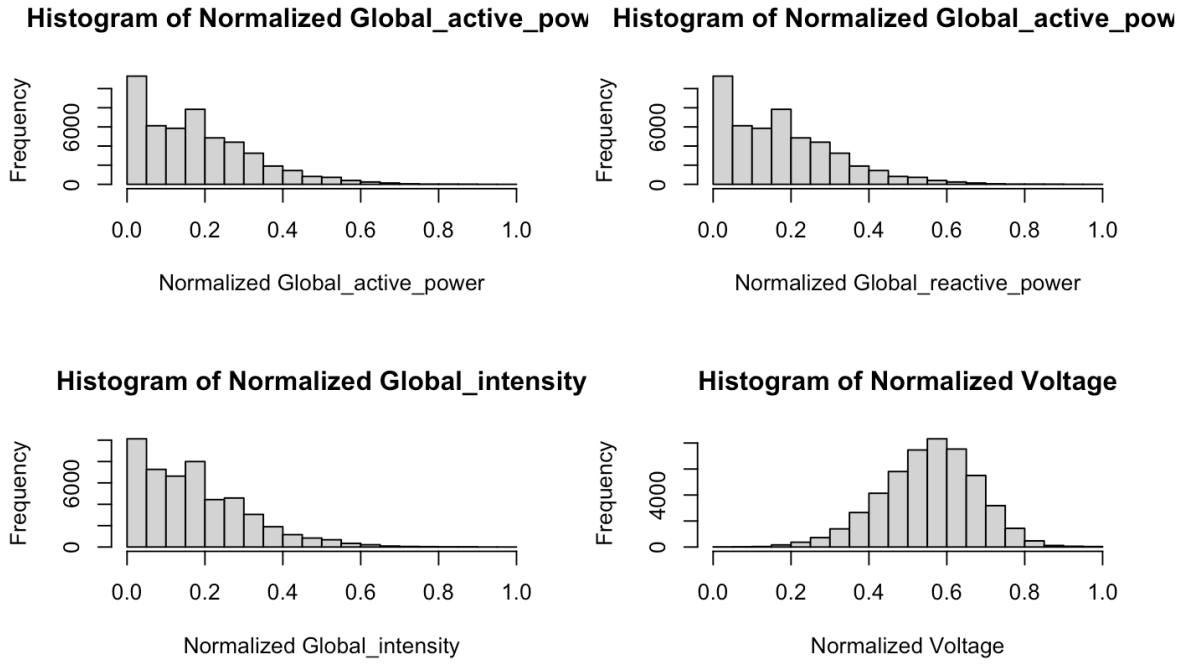

Figure 4. Histograms of Normalized Continuous Variables

iii. Although both methods have their merits, standardization was chosen for its compatibility with our modeling assumptions and the nature of our dataset. By centering the data around the mean, standardization offers a critical advantage: it aligns the scaled data with the assumptions of Gaussian distributions, which is a prerequisite for our chosen statistical model. This alignment is fundamental because it allows the model to accurately infer the data structure, which is paramount for detecting anomalies indicative of cybersecurity threats. By applying standardization, we have effectively prepared our dataset for further analysis with the depmix model, setting a solid foundation for reliable and accurate anomaly detection in the subsequent stages of our methodology.

2.2.2 Feature Engineering

Our dataset contains several variables that record different aspects of electricity usage. To determine which of these variables would be most indicative of anomalous behavior, we

performed a Principal Component Analysis (PCA). PCA is a statistical technique that reduces the dimensionality of the data while retaining most of the variability in the dataset. It allows us to identify the combinations of variables—or principal components—that account for the most variance in the dataset. The PCA was conducted on the standardized dataset to prevent features with larger variances from dominating the principal components. The analysis resulted in a set of principal components, each a linear combination of the original variables, ranked by the amount of variance they capture from the data. After examining the PCA results, we selected the principal components that collectively accounted for a substantial portion of the variance, ensuring that we retained the features most representative of significant consumption patterns. This selection process aimed to maximize the model's sensitivity to anomalies while minimizing the inclusion of redundant or irrelevant information.

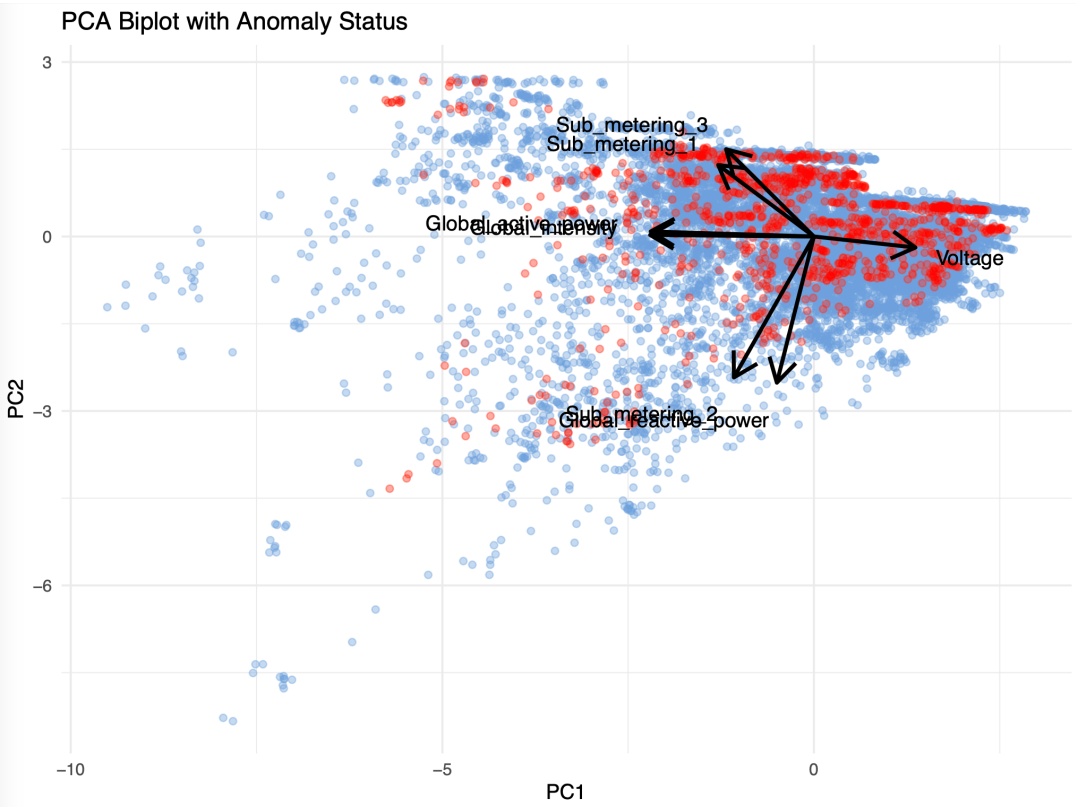

Figure 4. Biplot with Anomalies

The Principal Component Analysis (PCA) of our standardized variables yielded seven principal components (PCs). The contribution of each variable to the first four PCs is as follows:

- **PC1**: Primarily influenced by Global_active_power and Global_intensity, indicating a strong negative contribution, suggesting these features are important for defining the principal component.
- **PC2**: Dominated by Global_reactive_power with a significant negative contribution and Sub_metering_3 with a positive contribution, highlighting their relevance in capturing energy consumption patterns distinct from PC1.
- **PC3** and **PC4**: Show a more complex mix of contributions from the Sub_metering variables, which capture different aspects of household energy usage.

The proportion of variance explained by each PC is critical for understanding their importance:

- **PC1**: Accounts for a substantial 45.21% of the variance, indicating its strong representation of the dataset's overall structure.
- **PC2**: Contributes an additional 14.82% to the variance, cumulatively reaching 60.03% with PC1.
- **PC3**: Adds 13.73%, taking the cumulative proportion to 73.76%.
- **PC4**: Adds another 11.78%, bringing the cumulative total to 85.55%.

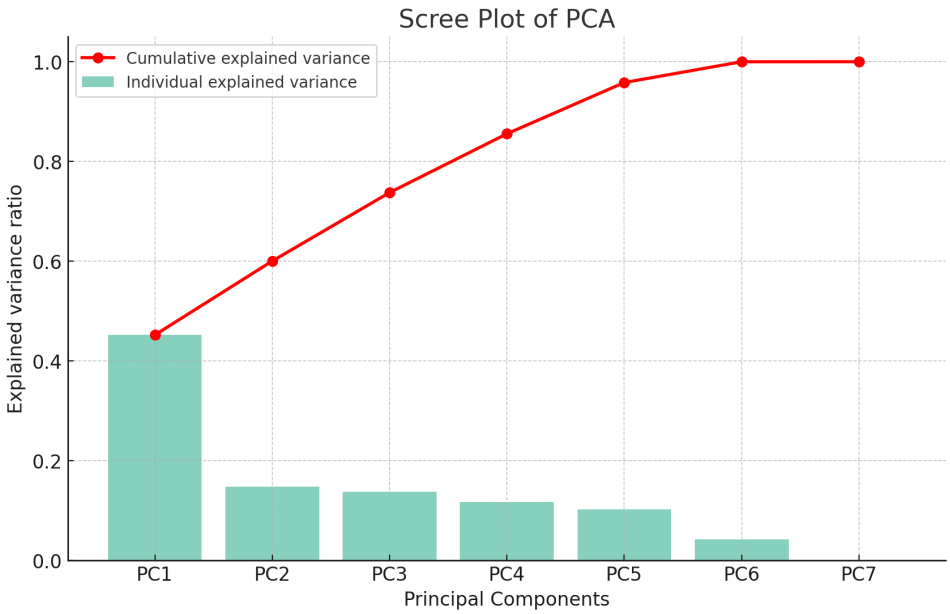

Figure 5. Scree Plot

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Global_active_power	-0.5416541	0.062149264	-0.006757640	0.005023745	-0.1206580	0.44065567	-0.702809774
Global_reactive_power	-0.1730789	-0.704049026	0.225868419	-0.607574804	0.2243954	-0.06080285	-0.012029007
Voltage	0.3436370	-0.083588050	0.074919072	-0.309856361	-0.8503254	0.22388457	0.011189803
Global_intensity	-0.5437165	0.045426848	-0.001855589	0.006038667	-0.1037204	0.43100140	0.711160155
Sub_metering_1	-0.3066339	0.006838445	0.731295974	0.291202716	-0.2854714	-0.45258633	-0.002781252
Sub_metering_2	-0.2558798	-0.495509720	-0.584122930	0.345291178	-0.3244706	-0.35111261	-0.002966716
Sub_metering_3	-0.3216235	0.495807654	-0.259476554	-0.575113180	-0.1204244	-0.48798016	0.004817699

Table 2. Loading Scores of PCAs

2.3 Model Training and Testing

2.3.1 HMM Training

For the development and validation of our Hidden Markov Models (HMMs), we adopted a partitioning strategy that separates the data into a **training set**, encompassing the years **2006-2008**, and a **testing set** for the year **2009**. The selection of the time window for model training is a strategic decision that impacts the model's ability to learn distinct consumption patterns. We focused on the time frame between **7 PM and 11 PM on Wednesdays**. This

period was chosen for its potential to reflect typical residential energy consumption behaviors during weekday evenings, as well as to detect anomalies potentially arising from industrial processes or infrastructure maintenance activities that may occur during these hours.

In our analysis, we used the scaled variables: Global_active_power, Global_intensity, and Voltage, which were identified as significant through the PCA and feature engineering processes. Our exploration involved varying the number of states (nstates) for each iteration, ranging from 4 to 20. For the 'ntimes' parameter within the HMM function, we used 'rep(240, each=107)', signifying that our dataset represents minute-by-minute observations from 7 PM to 11 PM, which totals 240 observations per hour. With 107 Wednesdays available in our training data, this parameter setting aligns with the structure of our dataset. We chose the Gaussian family for the emission distribution, acknowledging the continuous nature of our variables and the previous standardization step, which aligned the data distribution closer to the Gaussian assumption.

2.3.2 Model Selection

In the process of Hidden Markov Model (HMM) training and testing, we tried to identify the optimal number of states. As instructed, we used a range of states from 4 to 20 and fitted HMMs using three key variables from the feature engineering part: Global_active_power, Global_intensity, and Voltage. For each model, we evaluated its performance using two primary criteria:

- **Bayesian Information Criterion (BIC)**: This criterion is crucial for our model selection as it incorporates a penalty term for the number of parameters in the model, thus counteracting the risk of overfitting. By preferring models with lower BIC values, we aim to strike a balance between model complexity and goodness of fit, ensuring that the model is neither too simplistic to capture the data's structure nor too complex to generalize well to unseen data. Yonekura, et al (2020)., also remarks that BIC

values “are strongly consistent” in the context of nested HMMs, further validating our use of BIC as a measure of model fitness.

- **Log-likelihood**: The log-likelihood is a measure of how well the model describes the observed data. A higher log-likelihood indicates that the model's assumed state structure and parameter estimates are more likely to have generated the observed sequence of data. In contrast to BIC, log-likelihood alone does not penalize for model complexity, which is why it is considered in tandem with the BIC.

To select the optimal model, we plotted both BIC and log-likelihood scores against the number of states. The combination of BIC and log-likelihood allowed us to identify a model that adequately captures the underlying patterns without overfitting, thus ensuring its predictive robustness in practical cybersecurity applications. The chosen model is expected to discern regular consumption patterns from potential anomalies with a high degree of accuracy.

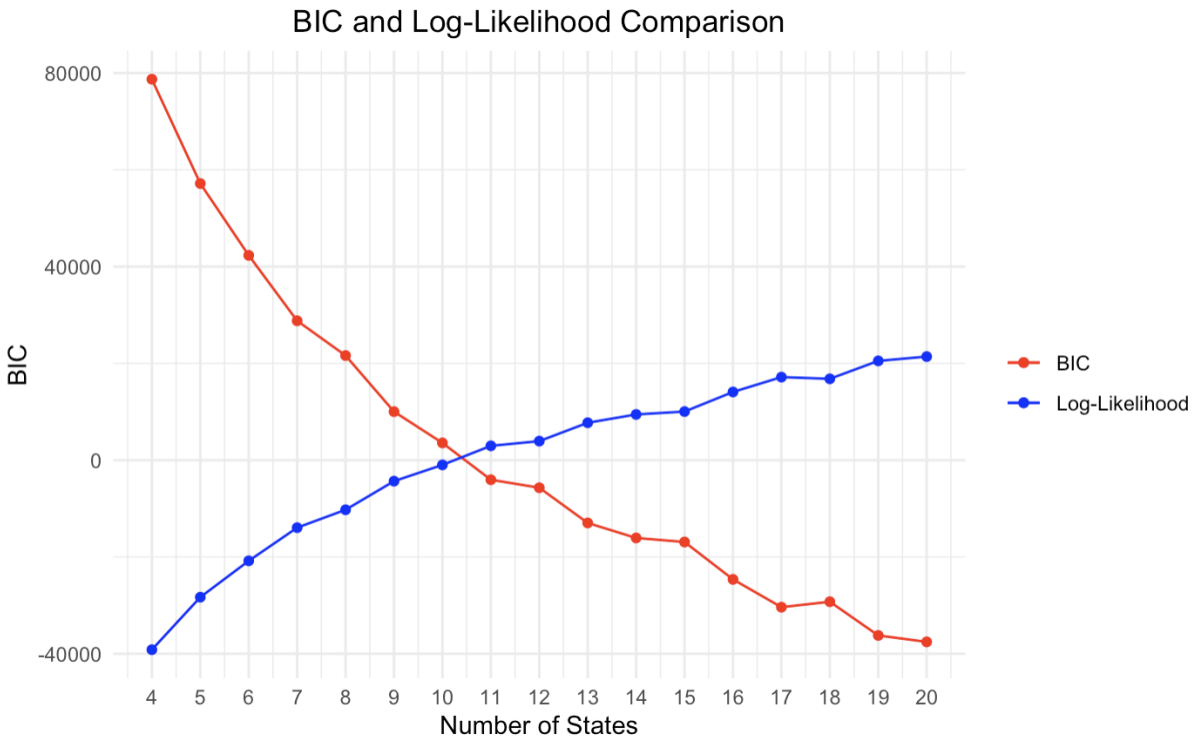

Figure 6. BIC and Log-Likelihood Comparison by Number of States

3 Anomaly Detection

In the domain of cybersecurity, anomaly detection is a fundamental task for identifying potential threats and vulnerabilities. Our approach utilized Hidden Markov Models (HMMs) to model normal behavior patterns in energy consumption data, thereby enabling the detection of deviations that could signify cyber threats.

The core of our anomaly detection method lies in calculating the log-likelihood of new data sequences. Log-likelihood measures how well new observations conform to the statistical model we have developed. Anomalies are detected when the log-likelihood of a new sequence is significantly lower than what the model expects, indicating a pattern that is markedly different from what has been learned as "normal." To classify observations as normal or anomalous, we established a threshold. This threshold was derived from the distribution of log-likelihoods computed for a validation set.

The test dataset was divided into 10 subsets, each representing a week's worth of data. We calculated the log-likelihood for each subset using the optimal 18-state HMM identified in our model fitting phase. By comparing these log-likelihood values to the threshold, we could identify which sequences were likely to be anomalies.

The maximum log-likelihood deviation observed in the training phase was 1.795. Consequently, this value was adopted as our threshold. Sequences that yielded log-likelihood values deviating more than this threshold from the training model's log-likelihood were flagged as anomalous. Upon applying this threshold to our test data, we observed several instances where the log-likelihood of a weekly subset deviated the established threshold, signaling

potential anomalies.

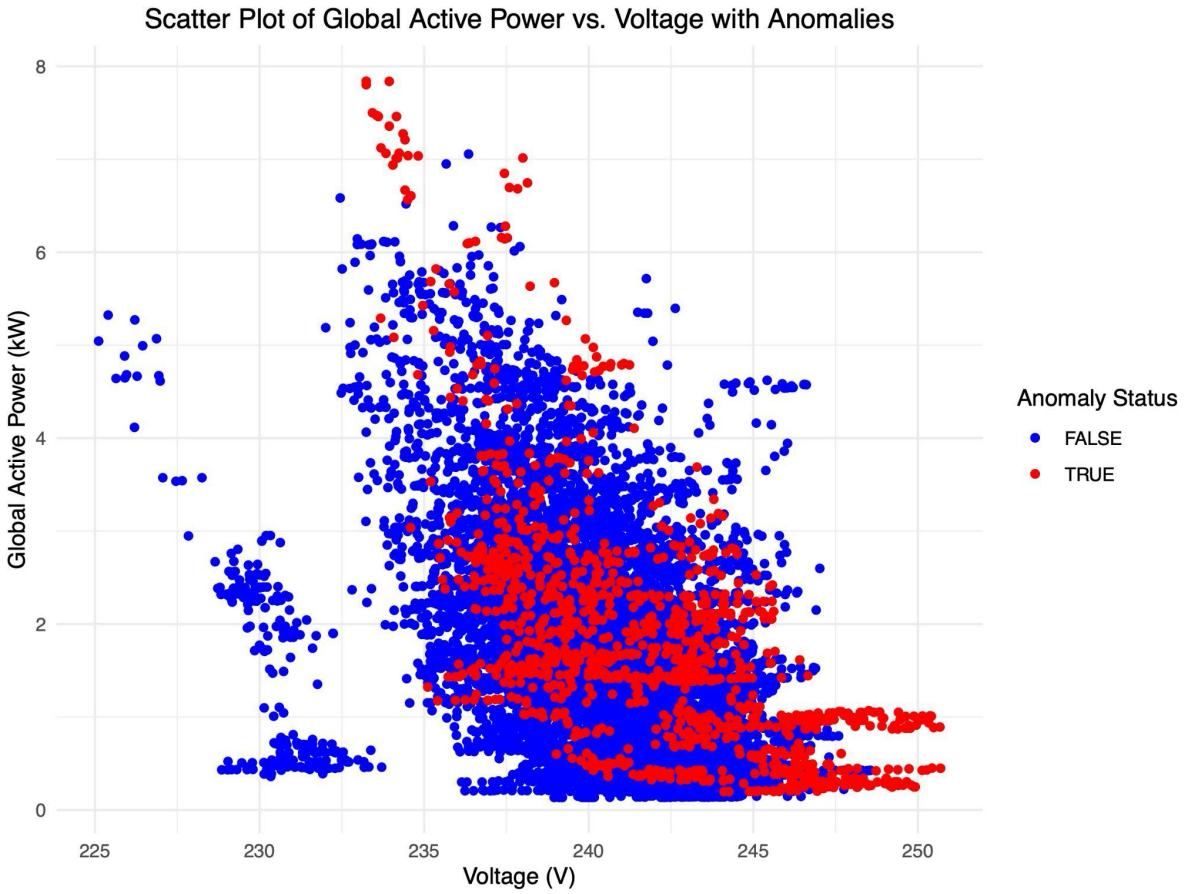

Figure 7. Scatter Plot of Global Activer Power vs. Voltage

The scatter plot depicts the relationship between Global_Active_Power (kW) and Voltage (V), with anomalies highlighted in red. We can observe that most voltage values fall within the normal range. On the other hand, anomalies are detected for higher values of the Voltage values range approximately from 235 to 250V.

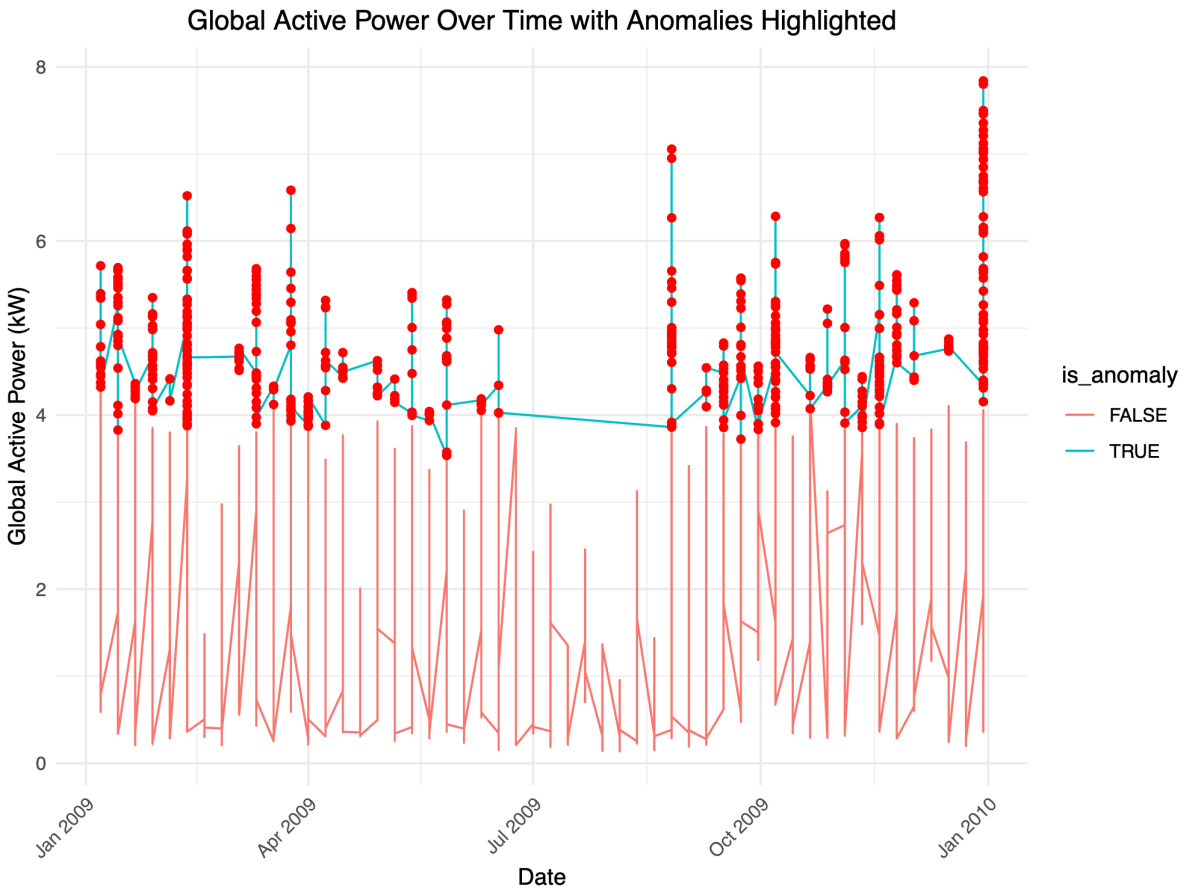

Figure 9. Time Series Plot of Global_active_power with Anomalies

This plot displays the trend of Global_active_power over time, with anomalies highlighted in red. Most anomalies are detected during periods when the power consumption exceeds approximately 4 kW. Also, between July 2009 and early September 2009, there is an absence of anomalies, indicating a period of lower power consumption during the time frame.

4 Results and Discussion

4.1 Comparison of Outcomes

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7
Standard deviation     1.7989 1.0257 0.9677 0.8961 0.8393 0.51726 0.02311
Proportion of Variance 0.4623 0.1503 0.1338 0.1147 0.1006 0.03822 0.00008
Cumulative Proportion  0.4623 0.6126 0.7463 0.8611 0.9617 0.99992 1.00000
```

Table 3. Summary of Principal Component Analysis

The PCA analysis revealed that PC1 explains a significant 46.23% of the total variability in the dataset, with Global_active_power, Global_intensity, and Voltage showing high absolute loadings. Following PC1, PC2 added to the cumulative variance, making it the next most informative component. Therefore, we will choose the results from PC1 and PC2 for variable selection.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Global_active_power	0.5347195	-0.05165373	0.07160005	-0.06207259	0.13576796	0.43558352	0.703037284
Global_reactive_power	0.1585286	0.67714875	-0.42168328	0.56662653	0.12979513	-0.02217758	0.010826725
Voltage	-0.3436598	0.03687903	0.10451465	-0.07591656	0.92072666	0.12658422	-0.009490564
Global_intensity	0.5367786	-0.03751821	0.05964126	-0.05908934	0.12414957	0.42721159	-0.710977906
Sub_metering_1	0.3091109	-0.17990237	-0.66934775	-0.43721486	0.22149893	-0.42874666	0.004108678
Sub_metering_2	0.2715577	0.56485960	0.51858438	-0.39242765	0.04621216	-0.42675099	0.002975373
Sub_metering_3	0.3367747	-0.42965400	0.29260910	0.56627750	0.22443761	-0.49527946	-0.003996406

Table 4. Loading Scores of Principle Components

In PC1, variables such as Global_active_power, Global_intensity, and Voltage show relatively high absolute loadings, indicating their importance in defining this component. In PC2, Global_reactive_power, Sub_metering_2, and Sub_metering_3 reveal notable loading scores.

```
          Date                Time   Global_active_power Global_reactive_power              Voltage      Global_intensity        Sub_metering_1
        " 0.00"             " 0.00"              " 0.00"               "18.26"             " 0.00"             " 0.00"              "78.48"
Sub_metering_2      Sub_metering_3
        "66.79"             "40.58"
```

Figure 10. The Percentage of Zeros in Each Columns

However, upon examining the table detailing the number of 0s in each column, it is evident that Sub_metering_1, 2 and 3 columns contain a considerable number of 0s. Moreover, given that PC1 explains almost half of the variability, we will focus on Global_active_power, Global_intensity, and Voltage for further model training and testing.

4.2 Model Fitting

Using the criteria mentioned in 2.2.2, we eventually determined the optimal number of states based on the model with the lowest BIC value, which in our case was 20 states. The optimal model indicates a BIC value of -37544.971 and has a corresponding log-likelihood of 21407.311. (From only BIC and Log-likelihood)

nstates	Train	Test
4	-1.525	-1.500
5	-1.102	-1.108
6	-0.810	-0.862
7	-0.543	-0.617
8	-0.399	-0.378
9	-0.169	-0.249
10	-0.038	-0.091
11	0.115	-0.025
12	0.153	0.045
13	0.301	0.197
14	0.368	0.330

15	0.391	0.365
16	0.549	0.783
17	0.669	0.477
18	0.655	0.657
19	0.799	0.637
20	0.834	0.700

Table 5. Normalized log-likelihood scores by the Number of States

Based on the table above, models with a number of states ranging from 14 to 20 show normalized log-likelihood scores for both training and testing dataset. However, after consideration, we finally selected nstates parameter as **18** as **the best model**. This decision was based on its robust performance across both training and testing datasets, which minimize overfitting and underfitting issues in prediction. By choosing **nstates=18**, we aim to strike a balance between model complexity and predictive accuracy.

5 Conclusion and Reflections

In our comprehensive analysis of household power consumption data, we have employed a multifaceted approach that integrated various statistical methods and machine learning techniques. A key focus of our project was the exploration of Hidden Markov Models (HMMs) to discern patterns of electricity usage across different time periods, such as weekdays and evenings, which are often prime targets. Our project demonstrated the application of Hidden Markov Models (HMMs) for anomaly detection in electric power consumption data, which is of paramount importance for cybersecurity in automated control systems. Through data preprocessing, feature engineering, and meticulous model selection, we have developed a model that effectively flags deviations that may indicate cyber threats.

The selected 18-state HMM, grounded on a solid foundation of statistical theory and empirical analysis, has shown a promising ability to discern between normal behavior and potential anomalies. By establishing a threshold based on the training data's log-likelihood distribution, our approach has proven to be both methodical and adaptable, catering to the dynamic nature of normal behavior patterns.The HMM proved to be an invaluable tool, providing a structured framework for capturing behavioral patterns.

5.1 Future Work

While the current model shows promise, future work will aim to refine the threshold-setting process, perhaps incorporating adaptive thresholding techniques to account for evolving normal patterns. Additionally, the integration of anomaly detection results with other cybersecurity measures could offer a more robust defense mechanism against a broader spectrum of threats. Moreover, different anomaly detection methods often yield disparate results. This variability underscores the fact that no single method is exhaustive in identifying all anomalies within a dataset. To ensure a thorough analysis, we should think about utilizing a combination of techniques, thereby enhancing the robustness of findings.

5.2 Lessons Learned

The intricacies of working with real-world datasets have been a critical learning curve for our team. Unlike datasets typically used in academic settings, real-world data is often messy, incomplete, and large-scale, presenting several challenges:

- The preprocessing required to clean and structure real-world data for analysis can be extensive. In our case, dealing with missing values and ensuring accurate interpolation demanded meticulous attention to detail and considerable effort.
- Real-world data is replete with noise and outliers, which can obscure the underlying patterns we sought to model. Distinguishing between noise, outliers, and actual

anomalies was a nontrivial task that required careful consideration and robust statistical methods.

- One of the most significant challenges was the absence of a clear ground truth against which to validate our anomaly detection results. Without verified labels for anomalies, assessing the accuracy of our model was reliant on indirect methods, such as cross-validation and comparison with known behaviors or expected patterns.

Despite these challenges, the practical experience gained from working on this project has been invaluable. It has provided us with a realistic perspective on the complexities of data analysis in a cybersecurity context and has better prepared us for future data-driven projects.

References

Rabiner, L. R. (1989). *A tutorial on hidden Markov models and selected applications in speech recognition.* Proceedings of the IEEE, 77(2), 257–286. https://doi.org/10.1109/5.18626

Yonekura, S., Beskos, A., & Singh, S. S. (2020, March 30). *Asymptotic analysis of model selection criteria for general hidden markov models*. arXiv.org. https://arxiv.org/abs/1811.11834