CMPT-318 Term-Project Presentation

A Presentation by Group 2

Sanchit Jain: sja164@sfu.ca

Priyansh Sarvaiya: pgs3@sfu.ca

Daiwik Marrott: drm11@sfu.ca

Luvveer Singh Lamba: lsl11@sfu.ca

Table Of Contents

01 Problem Addressed

02 Characteristics & Rationale

03 Challanges

04 Lesson Learned

05 Conclusion

06 References

01 Problem Addressed

- Critical Infrastructures rely on automated control systems used for exploiting operational anomalies.
- Detecting anomalies is complex due to several external factors.
- In this study, we would be addressing about the challenges of designing and evaluating an unsupervised anomaly detection framework using Hidden Markov Model(HMM).

02

Characteristics and Rationale

Data Processing & Cleaning

- Missing values were computed using linear interpolation.
- Approx 7% of the data, were found outliers using Z-scores.
- All numeric features were standardized ensuring mean is 0 and standard deviation is 1.

- Missing data is handled in a way that maintains temporal integrity.

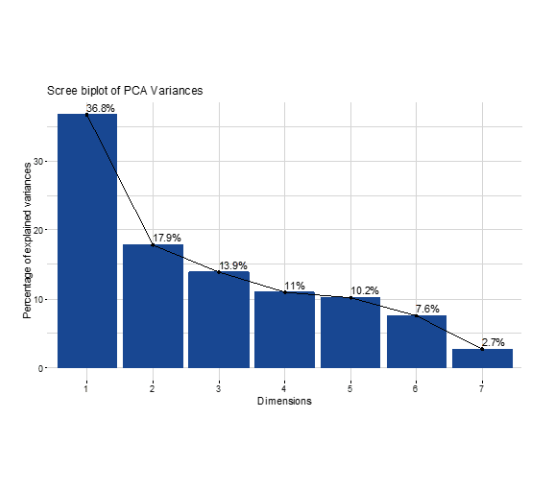
- Common scale of data can improve the accuracy and model performance.

- Standardization is performed due to gaussian distribution of data and it helps comparing the features effectively.

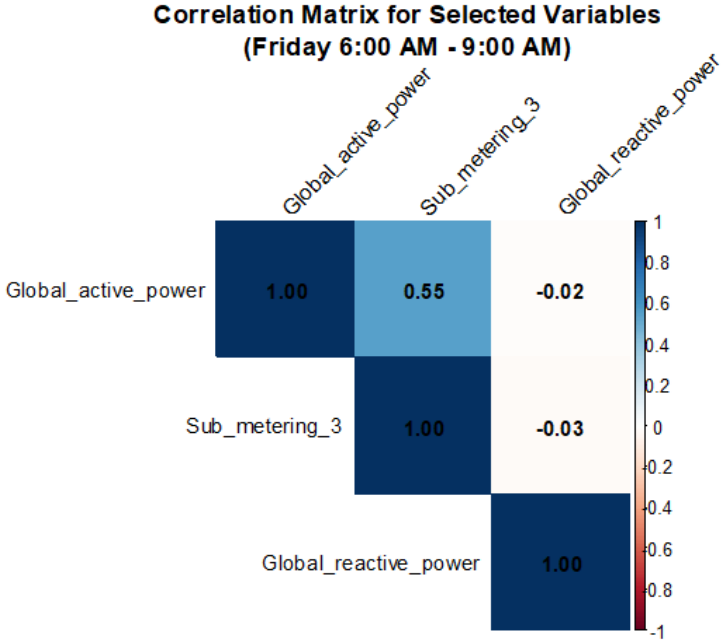
Feature Engineering

- Response Variables were selected using Principal Component Analysis(PCA).
- Based on PCA loadings, Global active power, Sub metering 3, and Global reactive power are the most significant features.

PCA Metric	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.6047	1.1187	0.9867	0.8761	0.8438	0.72858	0.4356
Proportion of Variance	0.3679	0.1788	0.1391	0.1096	0.1017	0.07583	0.0271
Cumulative Proportion	0.3679	0.5466	0.6857	0.7954	0.8971	0.97290	1.0000



Correlation of Responses



Statistical Metric	Global_active_power	Global_reactive_power	Sub_metering_3
Minimum	-1.197	-1.15	-0.709
Maximum	3.838	3.462	3.075
Mean	0.01576	0.008	0.004
PCA	-0.497	-0.687	-0.49
Range	High	Low	Medium

Global_active_power

Power consumption in the grid

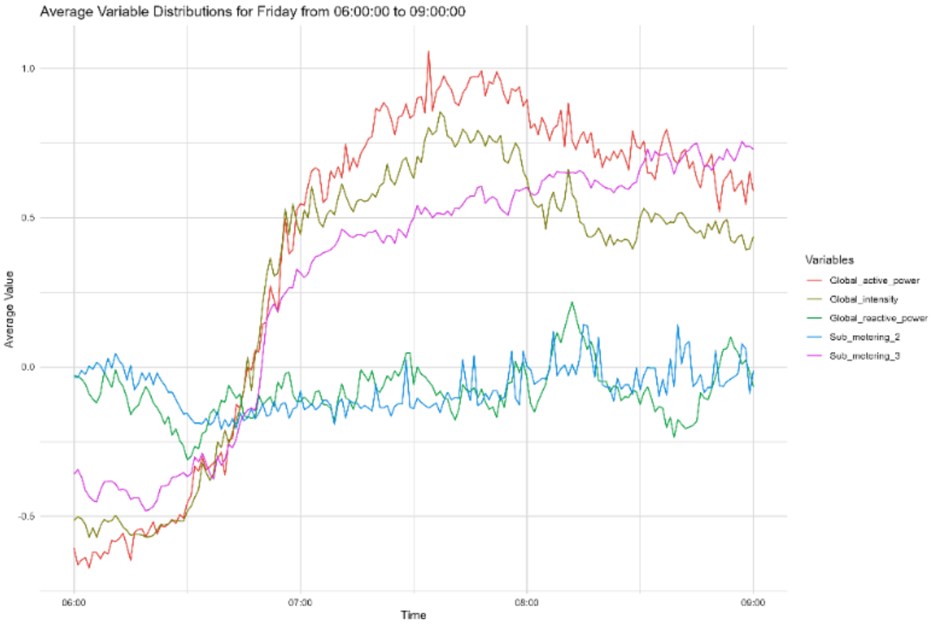
Global_reactive_power

Necessary for maintaining the voltage levels

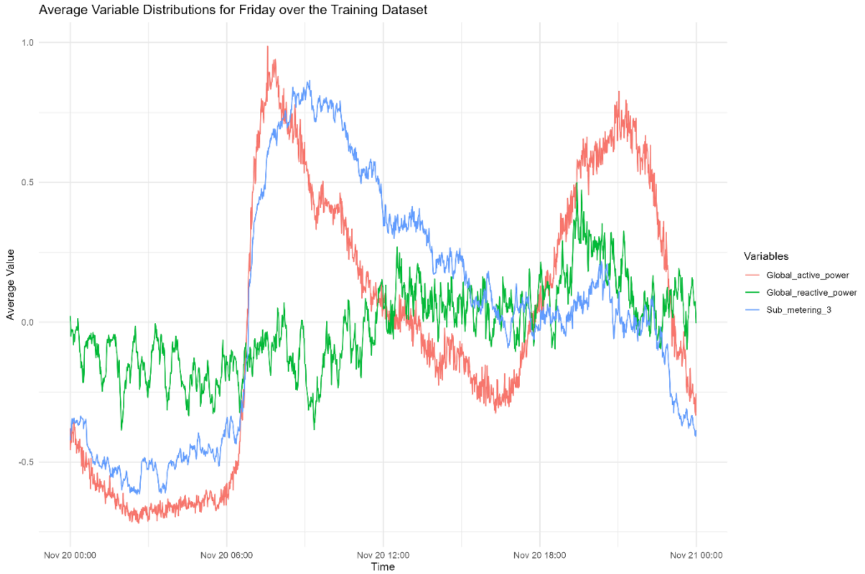
Sub_metering_3

Power use from a particular area

Time Window



Graph describing average variable distribution across chosen time window



Graph describing average chosen variable distribution across selected weekday

Splitting Of Dataset

Fridays
6:00 AM to
9:00 AM

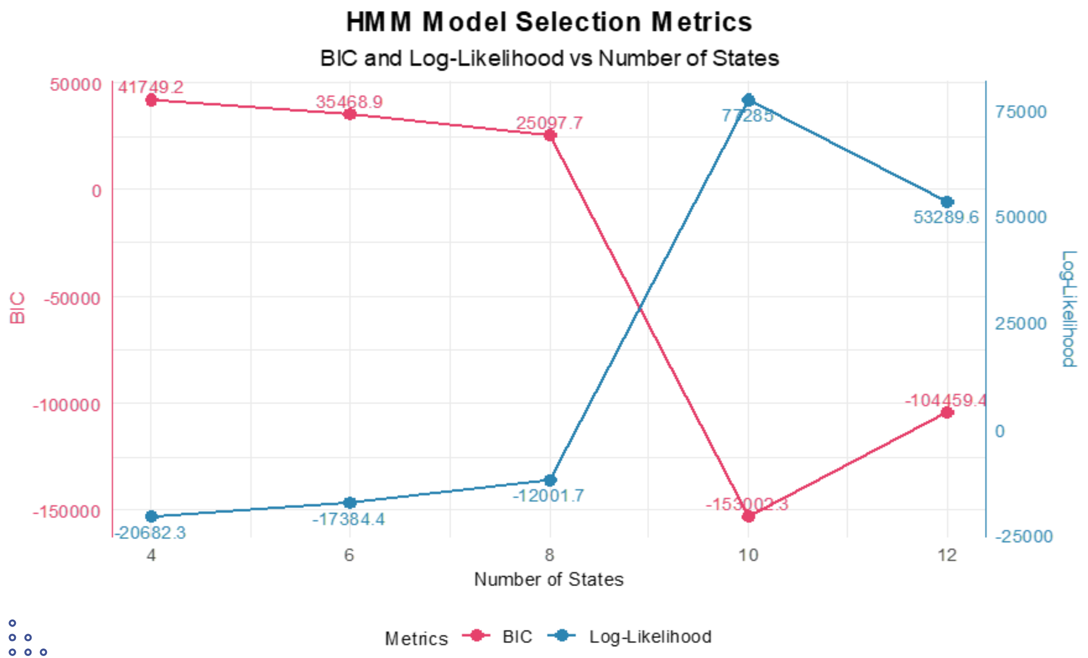
Testing Data
~ 44 Weeks
Jan. 25, 2009
onwards

Training Data
~ 110 Weeks
Dec. 16, 2006 – Jan. 24, 2009

Splitting Of Dataset…

- The larger training data allows the HMM model to learn diverse patterns across all temporal variations.
- Sufficient data fed to the HMM model reduces the risk of overfitting to short-term patterns.
- The test data time has to follow the training data time to test the model's performance under conditions of temporal drift.

Model Design



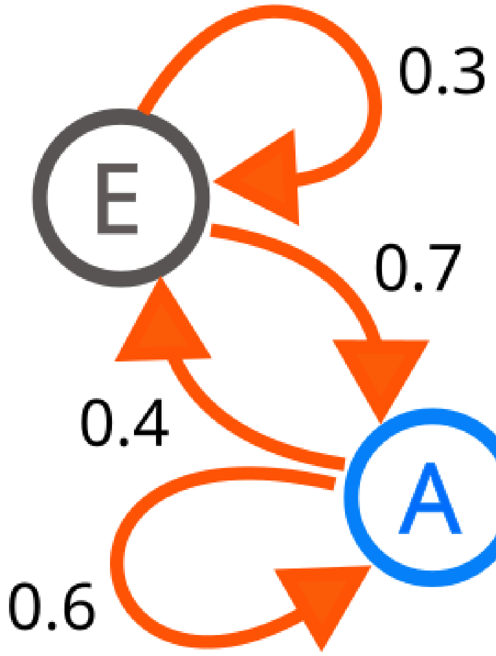
- The state counts (4, 6, 8, 10, 12) were explored using the log-likelihood and Bayesian Information Criterion (BIC).
- The 6th-state HMM was chosen for it's balance of performance and complexity, as it produced a log-likelihood of −17384.45 and BIC of 35468.91.
- The Gaussian Distribution was chosen.
- To ensure consistency we transferred the parameters form the trained model to the test model.

- The distribution has the ability to well-approximate data that shows natural variability, energy consumption.

- The number of states in an HMM essentially controls the granularity of the model's representation of underlying patterns in the data.

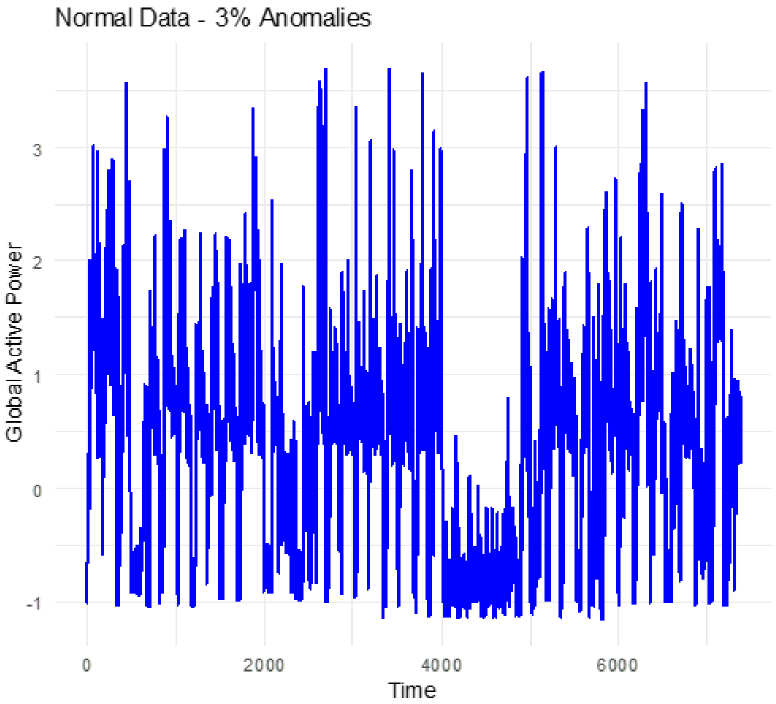
- The 6th-state model was chosen because of the best balance between complexity and fit.

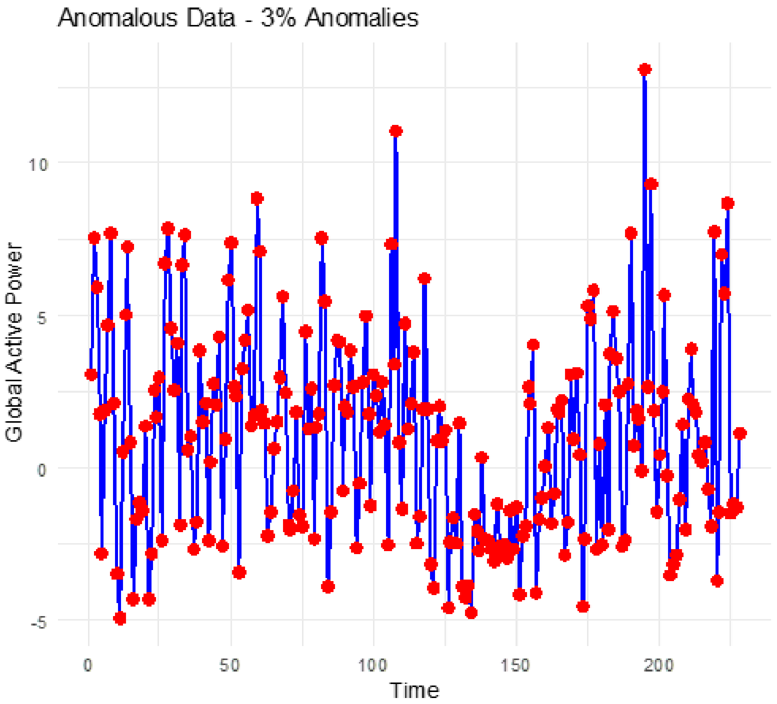
- To ensure consistency between training and testing,the parameters (state transition probabilities and emission distributions) were transferred from the trained to the test model using the `setpars()` function.



Anomaly Detection

- Deviation was identified by comparing log-likelihood of test and training data.
- Normalized log-likelihood threshold defined on maximum deviation observed in test subsets.
- Test Data divided into 10 equal-size subsets to capture temporal patterns.
- A threshold of -17452.07 log-likelihood and −1.005208 for normalized log-likelihood was established.
- Synthetic anomalies were introduced to the test data which included point & temporal anomalies.


Normal Data - 3% Anomalies
Global Active Power
Time
0
2000
4000
6000
-1
0
1
2
3


Anomalous Data - 3% Anomalies
Global Active Power
Time
0
50
100
150
200
-5
0
5
10

- Low log-likelihood values indicate that model has not learned to represent the observation, detecting potential anomalies.

- Threshold was derived by calculating maximum deviation between log-likelihood of training and test data.

- Subset Analysis of Data reflects the need for adaptive anomaly detection

- By injecting synthetic anomalies, we tested for the frameworks ability to detect known deviations.

Challenges

01 Identifying most relevant features due to complexity of energy consumption patterns.

02 State Configuration Selection based on log-likelihood and BIC values.

03 Dynamic threshold for Anomaly detection to balance false positives and false negatives.

04 Realistic Anomaly Stimulation needed to be diverse enough.

Lesson Learnt

01

Future Scaling Matters

Dimensionality reduction techniques like PCA not only simplify the model but also enhance interpretability.

02

Model Configuration is Curtail

Choosing the optimal number of HMM states are a balance between complexity and performance.

03

Thresholds should be Data-Driven

Empirical methods for determining thresholds, such as calculating maximum log-likelihood deviation, are more effective than static thresholds.

04

Anomaly Injection Validates Robustness

Injecting synthetic anomalies provides an effective way to test and refine the model.

05

Iterative Refinement Yields Better Results

Each phase of the project—preprocessing, model training, threshold determination, and anomaly detection—benefited from iterative refinement.

— Conclusion

01 Effective data preprocessing and dimensionality reduction using PCA

02 Optimal 6-state HMM configuration balancing performance and complexity

03 Robust anomaly detection framework with data-driven thresholds

04 Successful identification of both natural deviations and injected synthetic anomalies

Thank You

References

[1] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” Proc. IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989, doi: 10.1109/5.18626.

[2] A. Nassar, “Answer to ‘Hidden Markov models and anomaly detection,’” Cross Validated. Accessed: Nov. 24, 2024. [Online]. Available: https://stats.stackexchange.com/a/135946

[3] I. Visser and M. Speekenbrink, “depmixS4 : An R Package for Hidden Markov Models,” J. Stat. Soft., vol. 36, no. 7, 2010, doi: 10.18637/jss.v036.i07.

[4] N. Goernitz, M. Braun, and M. Kloft, “Hidden Markov Anomaly Detection,” in Proceedings of the 32nd International Conference on Machine Learning, PMLR, Jun. 2015, pp. 1833–1842. Accessed: Nov. 24, 2024. [Online]. Available: https://proceedings.mlr.press/v37/goernitz15.html

[5] “Principal Component Analysis (PCA) in R Tutorial.” Accessed: Nov. 24, 2024. [Online]. Available: https://www.datacamp.com/tutorial/pca-analysis-r