Anomaly Detection in Data Usage

[Link](https://nileuniversity-my.sharepoint.com/:f:/g/personal/m_abdallah2202_nu_edu_eg/EvQDx3G11K9HhOvSuvvxKdIBKUWSZwi2qEwCxtEHTthB7A?e=Ef7n95) For the Project

Ahmed Abdelmoneim   
Nile UniversityCairo, Egypt  
A.Mohamed2230@nu.edu.eg

NourEldeen Ayman  
Nile UniversityCairo, Egypt  
N.Ayman2201@nu.edu.egMaryam Mohamed  
Nile UniversityCairo, Egypt  
M.Abdallah2202@nu.edu.eg

Farida Salah  
Nile UniversityCairo, Egypt  
F.Sheref2223@nu.edu.egSeif Kassab  
Nile UniversityCairo, Egypt  
S.Usama2251@nu.edu.eg

*Abstract*— In this paper, we present a comprehensive analysis of anomaly detection in data usage patterns for three major cloud service providers: Google Web Services, Amazon Web Services, and Apple Web Services. The objective is to identify unusual data usage patterns that could indicate technical issues or security breaches. Our methodology includes the use of machine learning techniques to analyze historical data and detect anomalies. The results highlight significant patterns and provide actionable insights for improving data security and operational efficiency. (*Abstract*)

Keywords— Anomaly Detection, Data Usage Patterns, Google Web Services, Amazon Web Services, Apple Web Services, Machine Learning (key words)

# Introduction

The rise of cloud computing has led to an exponential increase in data usage. Monitoring data usage patterns is crucial for maintaining operational efficiency and ensuring security. Anomaly detection, which involves identifying patterns in data that do not conform to expected behavior, is a key technique in this context.  
  
This paper explores anomaly detection in data usage for Google Web Services, Amazon Web Services, and Apple Web Services. We employ various machine learning techniques to analyze historical data and identify anomalies, which could indicate potential security breaches or technical issues.

# Methodology

## Data Aggregation

## We aggregated data into 30-minute intervals to simplify analysis, reduce data volume, and capture meaningful trends, especially for time series data like data usage.

## Outlier Detection

## Outliers, which can skew analysis, were detected using statistical method like IQR. Not all outliers were removed, as some may indicate important anomalies.

## Data Transformation

## Summing usage columns and adding this total as a new column helps provide a clear view of total consumption, useful for knowing the total number of users for the—. Lastly, we converted date columns to a consistent datetime format using pandas' to\_datetime function for accurate time-based operations like sorting and resampling.

## Random Forest Regression

By Chance In machine learning, regression is an ensemble strategy that uses numerous decision trees along with a technique called Bootstrap and Aggregation, or bagging, to solve both regression and classification tasks. Random Forest Regression can be utilized not only for regression tasks but also for indirectly understanding clustering within the data and detecting anomalies by analyzing predictions error, The information can be valuable for identifying patterns, understanding data behavior and uncovering potential threats or security breaches in the dataset.

Figure 1 Dijon AWS

## Clustering

We used the powerful unsupervised machine learning model to find where the total usages are clustered, when they are clustered together and when exactly. showing points that are placed far from the clusters are considered an anomaly. Identifying possible threats and security breaches.

## KNN

We implemented a K-Nearest Regressor Distance. Based model for anomaly detection. Normal datapoints are near it’s neighbors while anomalies will be farther away. After predicting the total usages, we defined a threshold value of 5% to get the top 5% anomalies.

# Data Usage Trends

## A graph with a line Description automatically generated AWS

A graph with blue lines

Description automatically generated

Figure 2 Toulouse AWS

## A graph with numbers and lines Description automatically generatedGWS

**A blue line graph with white text

Description automatically generated**

Figure 3 KNN TOULOUSE Apple web services

Figure 4 KNN Dijon apple web service

A graph with a line going up

Description automatically generated

Figure 5 Dijon GWS

Figure 6 Toulouse GWS

## A graph with a line going up Description automatically generatedApple Web ServicesA graph with blue lines and numbers Description automatically generated

Figure 7 Dijon Apple WS

Figure 8 Toulouse Apple WS

The Yellow Vest protests in France led to a significant drop in application usage on March 17th. On 20/3, a day after Apple CEO Tim Cook attributed a downturn in the company’s outlook to below-expectation Chinese iPhone sales increasing usages. However, a culinary event called "A Menu for the Planet" in France on March 21st, promoting sustainable cuisine, likely increased usage of applications related to food, dining, and environmental awareness. This event highlights the diverse factors that can influence application usage patterns.

Figure 9 KNN DIJON GWS

# Future Trafic Using KNN

An effective method for identifying anomalies in traffic data is to forecast the future 12 hours based on the previous 12 hours. It makes use of temporal patterns and real-time monitoring to spot behavioral anomalies and enable prompt resolution of possible problems.

## **A screen shot of a sound wave Description automatically generated**Apple Web Services

Figures 7 and 8 show strong positive correlations between actual and predicted values, indicating the effectiveness of KNN regression for predicting usage. However, the Dijon model performs slightly better in terms of R2 score and errors. Both models show increasing variance at higher usage values.

## A blue sound wave Description automatically generatedA blue sound wave Description automatically generatedGoogle Web Services

Figure 10 KNN Toulouse GWS

The regression models in Figures 9 and 10 show strong positive correlations between actual and predicted values, indicating their effectiveness in predicting usage. The Toulouse model has a slightly higher R2 score, indicating better data explanation. However, both models show increasing variance at higher usage values, suggesting possible security breaches. The differences in MSE, MAE, and R2 scores between Toulouse and Dijon highlight regional variations in usage patterns.

## A screen shot of a sound wave Description automatically generatedA blue line graph with white text Description automatically generatedAmazon Web Services

Figure 11 Clustering analysis Toulouse amazon ws

Figure 12 KNN Toulouse amazon WS

Figure 13 KNN DIJON AMAZON WS

Figures 11 and 12 are effective for predicting usage, with strong positive correlations between actual and predicted values. The Dijon model has a higher R2 score, indicating better performance in explaining data variance. However, both models show increasing variance at high usage values, suggesting the need for further refinement or additional features. The differences in MSE, MAE, and R2 scores between Dijon and Toulouse highlight regional variations in usage patterns, possibly due to user abnormal data usages indicating possible anomalies.

## Summary

Strong positive correlations are shown by all models, indicating their ability to make accurate predictions. Out of all the models examined, the Dijon AWS model performs the best, as indicated by its highest R2 score. All models show increased variance at higher usage levels, indicating the need for additional features or model refinement to increase accuracy. Every model performs well in terms of identifying anomalies, especially when usage is higher, and variance is more noticeable. Regional variations in usage patterns are indicated by differences in MSE, MAE, and R2 scores between Dijon and Toulouse. These variations introduce outlier behaviour indicating possible abnormal behaviour or security breaches.

Figure 14 Clustering analysis Dijon amazon ws

1. *Clustering analysis*
2. A screenshot of a graph

   Description automatically generated*Amazon Web Services*

Figure 13 has 4 clusters with Cluster 2 representing the most data usages, while Clusters 2 and 0 have the least number of total usages. Anomalies can be detected in the second graph by identifying cluster points of low usages plotted among high usage values, representing possible data breaches at those times of the day. The Hotpoint of the day where security breaching could happen was at 11:00.

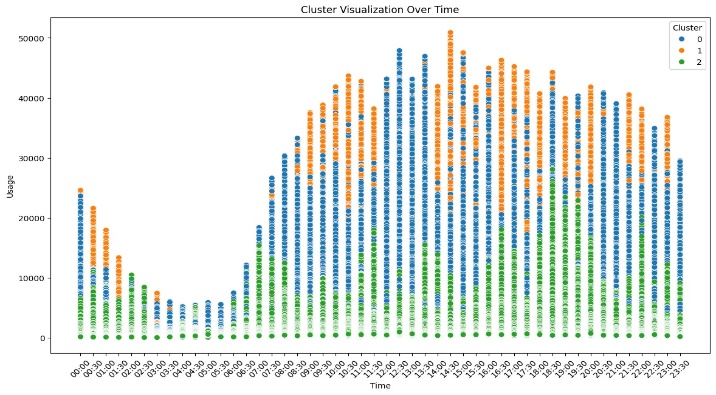
A screenshot of a graph

Description automatically generated

Figure 14 has 3 clusters, with Cluster 2 representing the highest data usage points. Cluster 2 dominated the scatter graph during the day, while Cluster 1 had low activity usage. Possible security breach times were identified at 11:00, 12:30, 13:30, and 22:00. The average data usage during the second part of the day was about 750 usages, so any usages exceeding that number pose a potential anomaly.

1. A screenshot of a graph

   Description automatically generated*Apple Web Services*

*A graph of data usage

Description automatically generated with medium confidence*Three clusters are shown in Figure 15, with Cluster 1 having the highest data usage. Contrary to the generally low usage trend for this cluster, the graph indicates a high number of security threats at 10:30 as many points belongings to Cluster 0 are spotted. Furthermore, the large number of application usages found after 21:00 raises concerns about anomalies and abnormal usage. At 21:30, a few points from Cluster 2 were observed, deviating from the cluster's typical trend and being recognised as anomalies.

Figure 15 Clustering analysis Toulouse Google web services

*A graph of different colored lines

Description automatically generated*

A graph showing the number of numbers

Description automatically generated with medium confidence

Figure 16 Clustering analysis Toulouse apple web services

Figure 16 is divided into 3 clusters, with Cluster 1 dominating the data usages. The graph suggests a high number of security threats as numerous points in Cluster 1 were observed at higher usages, contradicting the overall cluster trend. However, while plotting the scatter plot, high activity was detected in Cluster 0, indicating the highest security threat and data usage abnormality at 13:30. Additionally, Cluster 1 points were identified in the early morning at 3:00, 5:00, and 6:00, indicating anomaly points at these times of the day.

1. A graph showing the number of the number of the number of the number of the number of the number of the number of the number of the number of the number of the number of the number of

   Description automatically generated*Google Web Services*

Figure 17 has three clusters, with Cluster number 2 representing the highest data usage, while Clusters 1 and 0 have the lowest total usage. Anomalies can be detected in the second graph by identifying points of low usage within areas of high usage, indicating potential data breaches at those times of the day. The critical time of day for security breaches was 14:30.

Figure 17 Clustering Analysis Dijon apple web services

Figure 18 also has three clusters, with the second cluster also representing the highest data usage, clusters 1, 0 having the lowest total usage. The time at 13:30 is abnormal, which can indicate potential data breaches.

A line graph with numbers and lines

Description automatically generated

1. *Random Forest Regression*
2. A screenshot of a graph

   Description automatically generated*Toulouse AWS*

Figure 21 RFR Toulouse Apple WS

Figure 19 RFR Toulouse AWS

In Figure 19, Large Positive Errors (Over-predictions, High Usage):17/3/2019: 2100(Yellow)19/3/2019: 1200(Orange)

Large Negative Errors (Under-predictions, Low Usage): 16/3/2019: -1300 (Dark Purple) 1/3/2019: -2700(Dark)

Smaller Negative Errors (Under-predictions, Normal Usage): 18/3/2019: -44(Lightest Purple) 20/3/2019: 610 (Light Purple) 22/3/2019: -480(Lightest Purple)

Figure 22 RFR Dijon Apple WS

1. A graph of a number of numbers

   Description automatically generated with medium confidence*Dijon AWS*

In Figure 20,Large Positive Errors (Over-predictions, High Usage): 20/3/2019: 370(Yellow) 16/3/2019: 210(Orange) Large Negative Errors (Under-predictions, Low Usage): 21/3/2019: -390(Dark) 22/3/2019: -320(Dark) 17/3/2019: -300(Dark Purple) Smaller Negative Errors (Under predictions, Normal Usage): 19/3/2019: 71 (Lightest Purple) 18/3/2019: 180 (Light Orange)A screenshot of a graph

Description automatically generated

Figure 20 RFR Dijon AWS

1. A screenshot of a computer

   Description automatically generated*Toulouse Apple Web Services*

Figure 18 Clustering analysis Dijon Google web services

In Figure 21,Large Positive Errors (Over-predictions, High Usage): 21/3/2019: 10000 (light Orange) 16/3/2019: 17000 (Yellow) Large Negative Errors (Under-predictions, Low Usage): 17/3/2019: -17000 (Dark) 19/3/2019: -5800 (Dark Purple) 20/3/2019: -4100 (Dark Purple) Smaller Negative Errors (Under-predictions, Normal Usage): 18/3/2019: -640 (Purple) 22/3/2019: 72 (Lightest Purple)

1. A screenshot of a graph

   Description automatically generatedDijon Apple Web Services

In Figure 22,Large Positive Errors (Over-predictions, High Usage): 21/3/2019: 2800 (Orange) 16/3/2019: 7900 (Yellow) Large Negative Errors (Under-predictions, Low Usage): 17/3/2019: -7700 (Dark) 20/3/2019: -1700 (Dark Purple) Smaller Negative Errors (Under-predictions, Normal Usage): 18/3/2019: -250 (Light Purple) 21/3/2019: -560 (Light Purple)22/3/2019: -950 (Purple)

1. A screenshot of a computer

   Description automatically generated*Toulouse GWS*

Figure 23 RFR Toulouse GWS

In Figure 23, Large Positive Errors (Over-predictions, High Usage): 16/3/2019: 9400 (Orange) 19/3/2019: 14000 (Yellow) Large Negative Errors (Under-predictions, Low Usage): 17/3/2019: -9800 (Dark) 20/3/2019: -10000 (Dark) Smaller Negative Errors (Under-predictions, Normal Usage): 18/3/2019: -2800 (Dark Purple) 21/3/2019: -2600 (Dark Purple) 22/3/2019: -1100 (Light Purple)

1. A screenshot of a graph

   Description automatically generatedDijon GWS

Figure 24 RFR Dijon GWS

In Figure 24,Large Positive Errors (Over-predictions, High Usage): 16/3/2019: 4200 (Yellow) 19/3/2019: 3800 (Yellow) Large Negative Errors (Under-predictions, Low Usage): 17/3/2019: -4200 (Dark) 20/3/2019: -300 (Dark Purple) Smaller Negative Errors (Under-predictions, Normal Usage): 18/3/2019: -1200 (Purple) 22/3/2019: -1500 (Purple) 21/3/2019: -80(Lightest Purple)

1. *Discussion*
2. *Key Findings*

* The study focuses on anomaly detection in data usage across three major cloud service providers: Google Web Services (GWS), Amazon Web Services (AWS), and Apple Web Services (Apple WS). Through the application of machine learning techniques, particularly Random Forest Regression and K-Nearest Neighbors (KNN).
* The study successfully identifies anomalies that indicate potential security breaches and technical issues. The significant patterns observed in the data provide actionable insights for enhancing data security and operational efficiency.

1. *Combining Data and Finding Outliers*

* To implement feature reduction, the data was combined into 30-minute chunks. For time-series data analysis, this interval selection is crucial because it strikes a compromise between the need to reduce data volume and the granularity required to spot trends and anomalies in the data.
* The Interquartile Range (IQR) method was used to identify outliers. While this method is useful for detecting statistical anomalies, it's vital to remember that not all outliers were eliminated because some may have represented significant abnormalities rather than errors.

1. *Machine Learning Models*

* Random Forest Regression: This model was instrumental in predicting data usage and understanding clustering within the dataset. The ability to analyze prediction errors helped in identifying patterns and potential threats. The results demonstrated strong positive correlations between actual and predicted values, although increasing variance at higher usage levels indicated the need for further refinement.
* KNN Regression: By defining a threshold value of 5%, we effectively identified the top 5% anomalies. The KNN model highlighted that normal data points are near their neighbors while anomalies are farther away. This method proved efficient in pinpointing significant variance from expected usage patterns.
* Clustering Analysis: The clustering models successfully identified groups of similar data usage patterns and highlighted outliers. Clusters with high usage variability were particularly useful in spotting potential security breaches. Regional variations in data usage were evident, indicating that local events or behaviors significantly influence cloud service utilization.

1. *Limitations*

* Data Quality: The accuracy of our findings is contingent on the quality and completeness of the data from the NetMob 2023 Challenge dataset. Any gaps or inaccuracies in the data set could affect our results.
* Model Assumptions: The predictive models assume that historical traffic patterns will continue in the future. Significant changes in user behavior or external factors (e.g., major events, new regulations) could impact the accuracy of our predictions.
* Geographical Limitation: The study is limited to Paris and Marseille, and the findings may not be directly applicable to other cities with different demographics or internet usage patterns.

1. *Conclusion*

In this work, we give a comprehensive examination of anomaly detection in data usage patterns from three major cloud service providers: Google Web Services, Amazon Web Services, and Apple Web Services. We detected anomalous data usage patterns using a variety of machine learning approaches, including Random Forest Regression, KNN, and clustering, which could suggest potential security breaches or technical concerns.

*Key Conclusions:*

* Effectiveness of Machine Learning Models: The machine learning models used, particularly Random Forest Regression and KNN, proved to be effective tools for spotting anomalies in data usage patterns. These algorithms accurately projected consumption and detected anomalies that could indicate security breaches or technical concerns.
* Importance of Data Aggregation and Transformation: Accurate analysis and anomaly detection need data to be aggregated into meaningful intervals and suitably transformed.

*Regional variations:*

* Local events and behaviors have a substantial impact on cloud service consumption trends. Understanding these variations allows providers to customize their security and operational methods to individual regions.
* Identification of High-Risk Periods: The study identified specific times of day when anomalies are more likely to occur, providing actionable information for improving monitoring and security procedures.
* The use of powerful machine learning techniques to detect anomalies in data usage patterns provides cloud service providers with significant insights and practical applications. Future research should concentrate on refining these models and adding new characteristics to improve their accuracy and resilience. In an increasingly data-driven environment, we can improve the security and optimization of cloud services by constantly improving our understanding of data usage trends and anomalies.

1. *Future Research and Development*

Encourage ongoing research and development in the field of anomaly detection for cloud services. Invest in exploring new machine learning algorithms, deep learning techniques, and hybrid models that can offer improved accuracy and efficiency in detecting anomalies.

By implementing these recommendations, cloud service providers can enhance their ability to detect and respond to anomalies in data usage patterns, ultimately improving data security, operational efficiency, and user trust.

##### References

1. Azzedine Boukerche, Lining Zheng, and Omar Alfandi. 2020. Outlier Detection: Methods, Models, and Classification. ACM Comput. Surv. 53, 3, Article 55 (May 2021), 37 pages. <https://doi.org/10.1145/3381028>
2. Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2020). A review of local outlier factor algorithms for outlier detection in big data streams. Big Data and Cognitive Computing, 5(1), 1.
3. Escalante, H. J. (2005, June). A comparison of outlier detection algorithms for machine learning. In Proceedings of the international conference on communications in computing (pp. 228-237).
4. Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine learning for anomaly detection: A systematic review. Ieee Access, 9, 78658-78700.
5. Boukerche, A., Zheng, L., & Alfandi, O. (2020). Outlier Detection: Methods, Models, and Classification. ACM Computing Surveys, 53(3).
6. Martínez-Durive, O. E., Mishra, S., Ziemlicki, C., Rubrichi, S., Smoreda, Z., & Fiore, M. (2023). The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography. arXiv preprint arXiv:2305.06933. Retrieved from <https://arxiv.org/abs/2305.06933>
7. Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2020). A review of local outlier factor algorithms for outlier detection in big data streams. Big Data and Cognitive Computing, 5(1), 1.