COM6012 Assignment

ACP22ABJ

03 May 2024

1 Log Mining and Analysis

1.1 Introduction

This lab report presents the analysis of log data obtained from NASA's access logs for July 1995. The analysis includes tasks such as finding the total number of requests from different countries, identifying unique hosts and top frequent hosts, visualizing the data, and discussing observations.

1.2 Data Description

The dataset used in this analysis is the NASA access log for July 1995, downloaded from the provided FTP link. Each log entry contains information such as the host, date, and time of the request.

1.3 Methodology

Apache Spark was utilized for processing and analysing the log data. Regular expressions were employed to extract relevant information from the log entries. Matplotlib library in Python was used for data visualization.

1.4 Total Number of Requests

- Total requests from Germany (.de): 21345
- Total requests from Canada (.ca): 58290
- Total requests from Singapore (.sg): 1057

1.5 Number of Unique Hosts and Top 9 Most Frequent Hosts

1.5.1 Germany (.de)

- Number of unique hosts: 1138
- Top 9 most frequent hosts:
  - host62.ascend.interop.eunet.de
  - aibn32.astro.uni-bonn.de
  - ns.scn.de
  - www.rrz.uni-koeln.de
  - ztivax.zfe.siemens.de
  - sun7.lrz-muenchen.de
  - relay.ccs.muc.debis.de
  - dws.urz.uni-magdeburg.de
  - relay.urz.uni-heidelberg.de

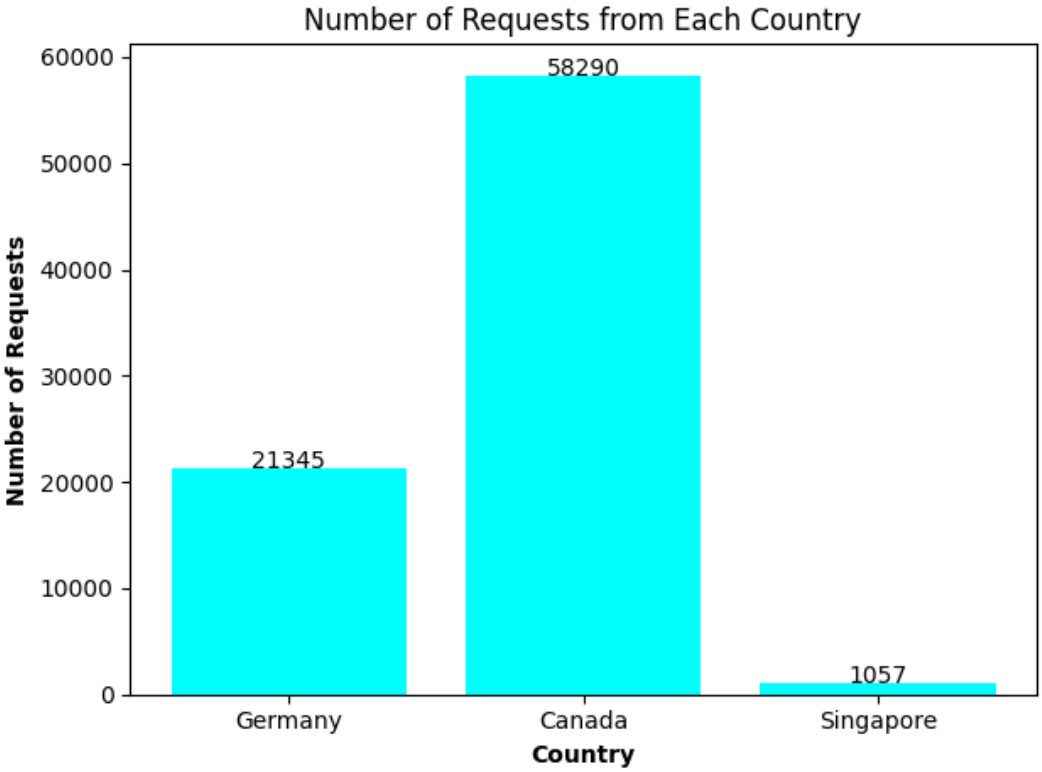


Figure 1: Requests By Countries

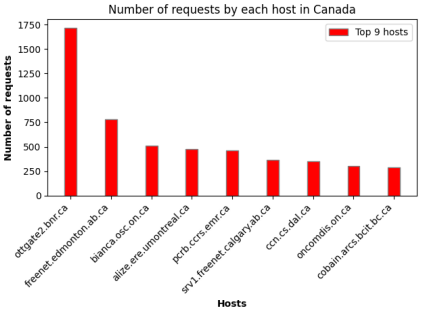
1.5.2 Canada (.ca)

- Number of unique hosts: 2970
- Top 9 most frequent hosts:
  - ottgate2.bnr.ca
  - freenet.edmonton.ab.ca
  - bianca.osc.on.ca
  - alize.ere.umontreal.ca
  - pcrb.ccrs.emr.ca
  - srv1.freenet.calgary.ab.ca
  - ccn.cs.dal.ca
  - oncomdis.on.ca
  - cobain.arcs.bcit.bc.ca

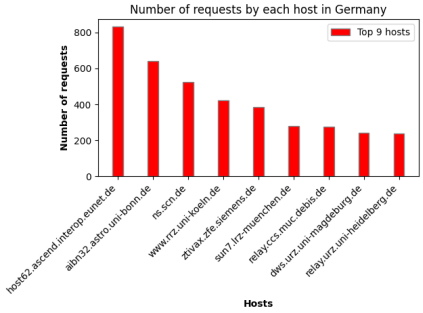
1.5.3 Singapore (.sg)

- Number of unique hosts: 78
- Top 9 most frequent hosts:
  - merlion.singnet.com.sg
  - sunsite.nus.sg
  - ts900-1314.singnet.com.sg
  - ssc25.iscs.nus.sg
  - scctn02.sp.ac.sg
  - ts900-1305.singnet.com.sg
  - ts900-406.singnet.com.sg

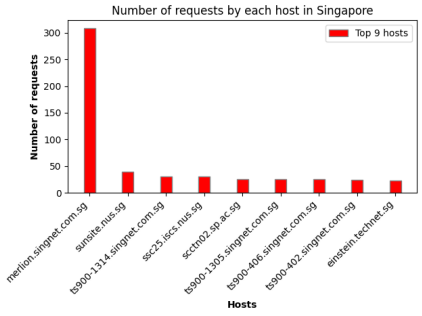
– ts900-402.singnet.com.sg
– einstein.technet.sg



(a) Canada



(b) Germany



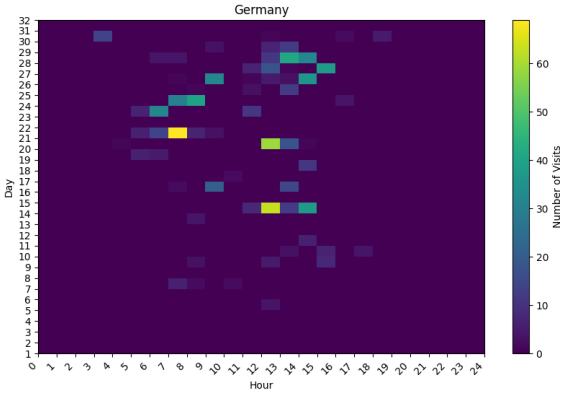
(c) Singapore

Figure 2: Visualizations of requests from different countries.

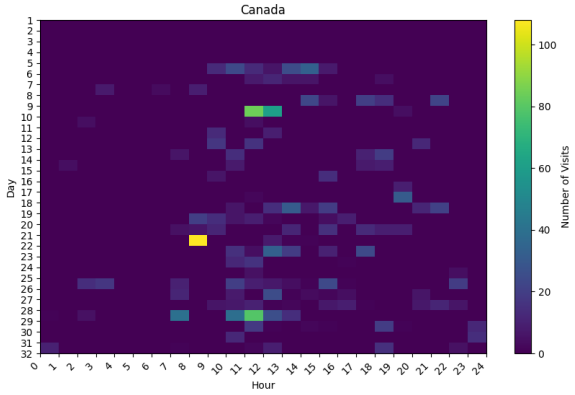
1.6 Observations

1. Observation A:
   - The number of requests from Canada (.ca) is significantly higher than the number of requests from Germany (.de) and Singapore (.sg).
   - Possible causes: Canada has a larger population and more internet users than Germany and Singapore. Canada may have more internet infrastructure and a higher level of internet penetration than Germany and Singapore. Canada may have more websites and online services that users access, leading to more requests.
   - Usefulness: This observation is useful to NASA as it provides insights into the distribution of website visits across different countries. It can help NASA understand the geographic distribution of its online audience and tailor its online services and content to better serve users in different regions.
2. Observation B:
   - The top 9 most frequent hosts in Canada (.ca) have a higher number of requests compared to the top 9 most frequent hosts in Germany (.de) and Singapore (.sg).
   - Possible causes: The top hosts in Canada may host popular websites or online services that attract a large number of visitors. Canada may have a more developed online ecosystem with a higher number of popular websites and online services. The top hosts in Canada may have better internet connectivity and server infrastructure, leading to faster response times and more requests.
   - Usefulness: This observation is useful to NASA as it provides insights into the popularity and traffic patterns of different websites in Canada, Germany, and Singapore. It can help NASA identify popular websites and online services that attract a large number of visitors and potentially collaborate with them to reach a wider audience. It can also help NASA understand the online landscape in different countries and tailor its online outreach and engagement strategies accordingly.
3. Observation C:
   - The heatmap plots show clear peaks in visitation throughout the day, with a peak in visits in the middle of the day (around hour 12) and a smaller peak in the later evening hours (around hour 18).

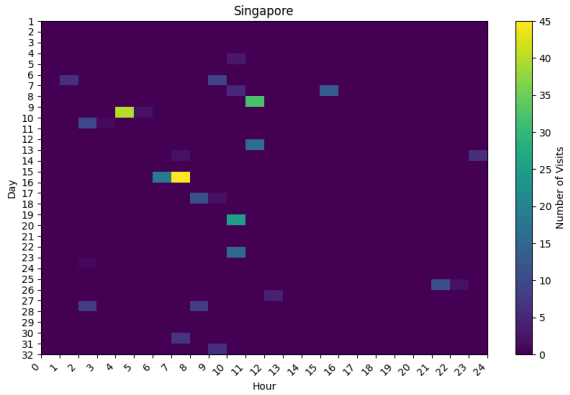
- Possible causes: The peaks in visitation could be due to users accessing the website during working hours or breaks, leading to increased traffic during the middle of the day. The smaller peak in the later evening hours could be due to users browsing the internet after work or during leisure time. The peaks in visitation could also be influenced by factors such as time zones, internet usage patterns, and user behavior.
- Usefulness: This observation is useful to NASA as it provides insights into the daily traffic patterns of its website visitors. It can help NASA optimize its website performance and content delivery to accommodate peak traffic times and ensure a smooth user experience. It can also help NASA schedule online events, updates, and maintenance activities during off-peak hours to minimize disruptions and reach a larger audience.



(a) Germany



(b) Canada



(c) Singapore

Figure 3: Heatmap visualizations for different countries.

1.7 Conclusion

In conclusion, the analysis of NASA's access logs provides valuable insights into the geographic distribution, traffic patterns, and popularity of websites across different countries. These insights can help NASA optimize its online services and content delivery to better serve its global audience.

2 Liability Claim Prediction

2.1 Pre-processing

2.1.1 Conversion to PySpark DataFrame

The dataset was converted to a PySpark DataFrame, and a new column `hasClaim` was created to indicate the presence or absence of a claim. The value is set to 1 if `ClaimNb > 0`, otherwise 0.

2.1.2 Splitting the Dataset

The dataset was split into training (70%) and test (30%) sets using a stratified split on `hasClaim` to handle the imbalanced dataset.

2.2 Training Predictive Models

2.2.1 Feature Selection

Ten features were selected for training the models: Exposure, Area, VehPower, VehAge, DrivAge, BonusMalus, VehBrand, VehGas, Density, and Region.

2.2.2 Model Training: Poisson Regression

A small subset (10%) of the training set was sampled, and cross-validation was performed to determine the best value of `regParam` for modeling the number of claims (`ClaimNb`) conditionally on the input features via Poisson regression. The optimal `regParam` was found to be 0.01. The RMSE for the test set was 0.248.

2.2.3 Model Training: Logistic Regression

Cross-validation was also used to determine the best value of `regParam` for modeling the relationship between `hasClaim` and the input features via Logistic regression, with L1 and L2 regularization respectively. The optimal `regParam` for L1 regularization was found to be 0.001, and for L2 regularization, it was 0.01. The AUC for the test set was 0.5 for L1 regularization and 0.629 for L2 regularization. The accuracy for both L1 and L2 regularization was 0.949.

2.3 Model Coefficients and Performance Analysis

2.3.1 Poisson Regression

The coefficients for the Poisson regression model were as follows:

$$\begin{bmatrix} 0.398 & 0.079 & -0.221 & 0.079 & 0.289 \\ 0.064 & -0.018 & -0.039 & -0.013 & -0.063 \end{bmatrix}$$

The RMSE for the test set was 0.248, indicating a good fit of the model.

2.3.2 Logistic Regression

For Logistic regression with L1 regularization, the coefficients were all zero, suggesting that L1 regularization effectively reduced the model complexity. However, for L2 regularization, the coefficients were non-zero:

$$\begin{bmatrix} 0.293 & 0.045 & -0.170 & 0.050 & 0.212 \\ 0.042 & -0.010 & -0.040 & -0.010 & -0.040 \end{bmatrix}$$

The AUC for L1 regularization was 0.5, indicating poor model performance, while for L2 regularization, it improved to 0.629. The accuracy for both L1 and L2 regularization was 0.949.

2.4 Observations

1. The Poisson regression model showed a reasonable fit with an RMSE of 0.248, indicating that it can effectively model the number of claims based on the input features.
2. The Logistic regression model with L1 regularization effectively reduced the model complexity by setting all coefficients to zero, suggesting that it might be useful for feature selection. However, its performance, as indicated by the AUC, was poor (0.5).
3. The Logistic regression model with L2 regularization performed better in terms of AUC (0.629) compared to L1 regularization, indicating that L2 regularization helped improve the model's predictive performance.

3 Searching for exotic particles in high-energy physics using ensemble methods

3.1 Introduction

In this report, we explore the use of supervised classification algorithms to identify Higgs bosons from particle collisions, utilizing the HIGGS dataset. We employ Random Forests, Gradient Boosting, and Neural Networks and evaluate their performance using classification accuracy and area under the curve (AUC).

3.2 Methodology

We first download the HIGGS dataset using wget and preprocess it. Then, we split the data into training and testing sets. For parameter tuning, we use a 1% randomly chosen subset of the data with proper class balancing. We employ pipelines and cross-validation to find the best configuration of parameters for each model. Finally, we compare the performance of the algorithms on the full dataset.

3.3 Parameter Tuning on Subset Data

Classifier	**Parameter**	**Value**	**AUC**
Random Forest	numTrees	30	0.690
	maxDepth	7	
	minInstancesPerNode	1	
Gradient Boosted Tree	maxIter	150	0.736
	maxDepth	6	
	stepSize	0.1	
Multilayer Perceptron	layers	[28, 40, 2]	0.679
	blockSize	64	
	stepSize	0.03	
Random Forest	numTrees	100	0.672
	maxDepth	10	
	minInstancesPerNode	3	
Gradient Boosted Tree	maxIter	100	0.718
	maxDepth	5	
	stepSize	0.1	
Multilayer Perceptron	layers	[28, 20, 2]	0.687
	blockSize	128	
	stepSize	0.1	
Random Forest	numTrees	30	0.690
	maxDepth	7	
	minInstancesPerNode	1	
Gradient Boosted Tree	maxIter	150	0.736
	maxDepth	6	
	stepSize	0.1	
Multilayer Perceptron	layers	[28, 40, 2]	0.679
	blockSize	64	
	stepSize	0.03	

Table 1: Best Parameters and AUC for Each Classifier on Subset Data

4 Model Evaluation on Full Dataset

Model	AUC on Full Dataset
Random Forest	0.688
Gradient Boosting	0.723
Multilayer Perceptron	0.623

Table 2: Area Under the Curve (AUC) on Full Dataset

4.1 Discussion

The Gradient Boosting algorithm achieved the highest AUC on the full dataset, indicating its superior performance in identifying Higgs bosons from particle collisions. However, further analysis is required to determine the significance of these findings and their implications for high-energy physics research.

5 Movie Recommendation and Cluster Analysis

5.1 Splitting Data

For the time-split recommendation, the data was sorted by timestamp, and splits were performed according to the sorted timestamp. Three splits were considered with training data sizes of 40%, 60%, and 80%. For that it was important to modify the format of timestamp into date-time, so that the splitting is purely on the basis of time.

5.2 ALS Settings

Two versions of ALS were studied for each split:

- Setting 1: The same ALS setting as used in Lab 7, which has been considered as base.
- Setting 2: Based on the results from Setting 1, another ALS setting was chosen to potentially improve the results, where Rank=30 has been considered.

5.3 Justification for ALS Parameter Adjustments

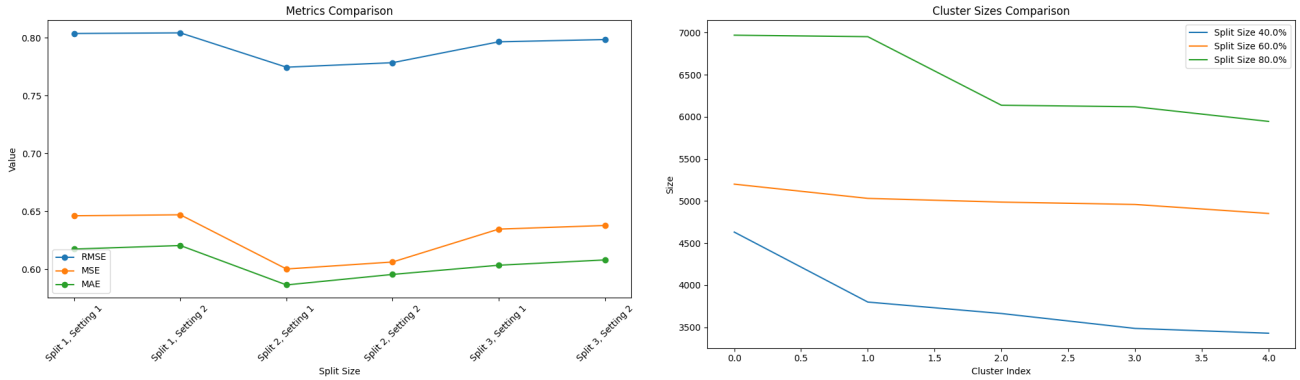
- Increasing the rank parameter to 30 can potentially improve the model's ability to capture complex patterns in the data, resulting in more accurate recommendations. This was observed in the provided output, where increasing the rank led to slightly lower RMSE values across different split sizes, indicating better model performance.
- Experimentation with the regularization parameter (*regParam*) was also conducted. However, decreasing the regularization did not significantly improve performance and, in fact, led to an increase in RMSE. This suggests that the original regularization level was appropriate or that other factors, such as the rank parameter, were more influential in determining model performance.
- Overall, adjusting the rank allows for a more expressive model that better represents nuanced user preferences and item characteristics, ultimately improving recommendation accuracy.

5.4 Metrics

For each split and each ALS setting, three metrics were computed: RMSE, MSE, and MAE.

Split Size	Setting	RMSE	MSE	MAE
40%	1	0.8038	0.6462	0.6174
	2	0.8093	0.6549	0.6212
60%	1	0.7747	0.6001	0.5863
	2	0.7739	0.5990	0.5858
80%	1	0.7966	0.6346	0.6034
	2	0.8034	0.6455	0.6068

Table 3: Metrics for Time-split Recommendation



(a) Metrics Comparison

(b) Cluster Size Comparison

Figure 4: Comparison of Cluster Sizes and Metrics

5.5 User Analysis

5.5.1 User Clustering

K-means clustering with $k = 25$ was applied to cluster all users based on the user factors learned with ALS Setting 2.

5.5.2 Top Five User Clusters

The sizes of the top five user clusters were reported for each split.

	40%	60%	80%
Cluster 1	4631	5200	6969
Cluster 19	3800	5031	6119
Cluster 12	3664	4987	5945
Cluster 15	3487	4959	6137
Cluster 23	3430	4852	6119

Table 4: Sizes of Top Five User Clusters

5.5.3 Movie Analysis within Largest User Cluster

For each split, the movies rated by users in the largest user cluster were analyzed:

- All movies with an average rating greater than or equal to 4 were identified.
- Genres for these top-rated movies were determined using the movies dataset.
- The top ten most popular genres were reported.

Split Size	Top 5 Movies
40%	Innocent Voices (Voces inocentes) (2004) Nothing to Lose (1997) Lars and the Real Girl (2007) Breezy (1973) McLintock! (1963)
60%	Best Boy (1979) Deathmaker, The (Totmacher, Der) (1995) Electile Dysfunction (2008) Black Rain (Kuroi ame) (1989) Ronja Robbersdaughter (Ronja Rövardotter) (1984)
80%	How to Survive a Plague (2012) Black Dynamite (2009) Game of Death, The (Le Jeu de la Mort) (2010) Brother of Sleep (Schlafes Bruder) (1995) Private Function, A (1984)

Table 5: Top 5 Movies for Each Split

Split Size	Genre	Count
40%	Drama	436
	Comedy	210
	Romance	154
	Thriller	136
	Action	117
	Crime	97
	Adventure	86
	Documentary	62
	Sci-Fi	62
	Fantasy	49

(a) 40% Split

Split Size	Genre	Count
60%	Drama	2096
	Comedy	943
	Romance	590
	Documentary	524
	Thriller	443
	Crime	423
	Action	343
	Adventure	299
	War	248
	Mystery	192

(b) 60% Split

Split Size	Genre	Count
80%	Drama	717
	Comedy	368
	Romance	219
	Thriller	147
	Documentary	147
	Action	145
	Adventure	139
	Crime	126
	War	104
	Mystery	79

(c) 80% Split

Figure 5: Top 10 Genres within Largest User Cluster for Different Splits

5.6 Discussion

1. **Observation 1:** The RMSE, MSE, and MAE values tend to decrease as the size of the training data increases.
   - **Possible Causes:** With more training data, the model has more information to learn from, leading to better predictions. Additionally, larger training datasets can capture a more comprehensive representation of user preferences and item characteristics.
   - **Usefulness to Netflix:** This observation underscores the importance of continuous data collection and model refinement. Netflix could benefit from regularly updating their recommendation models with new data to improve accuracy.
2. **Observation 2:** There may be noticeable shifts in the popularity of genres over different time periods. For example, certain genres like sci-fi or fantasy might gain traction during specific years due to the release of blockbuster movies or influential TV shows.
   - **Possible Causes:** Changes in societal interests, advancements in technology influencing storytelling capabilities, and the emergence of new sub-genres could contribute to shifts in genre preferences.
   - **Usefulness to Netflix:** Understanding genre trends over time can help Netflix anticipate audience preferences and strategically acquire or produce content to cater to evolving tastes, thereby maximizing viewership and subscriber retention.