The University Of Sheffield.

Department Of Computer Science

MSc Data Analytics

COM6012 Scalable Machine Learning

Adhiraj Banerjee
ACS23AB

May 2024

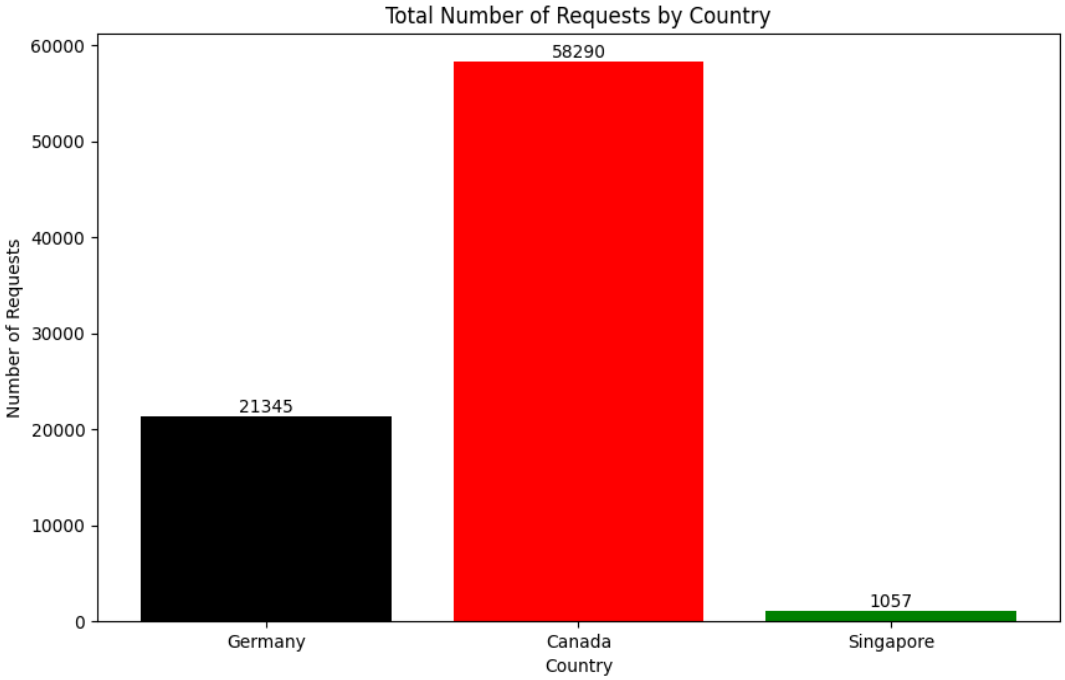
Question 1: Log Mining and Analysis

TASK-A:

Using code, the total number of requests were determined for all hosts for countries, Germany, Canada and Singapore. The numbers are reported below:

Country	Number of Requests for all hosts
Germany (.de)	21345
Canada (.ca)	58290
Singapore (.sg)	1057

The total number of requests have been visualised below:



The bar chart visualisation above clearly demonstrates Canada as the dominant country in terms of the total number of requests.

TASK-B:

Using the python code, the number of unique hosts for all the three countries were determined and are reported below:

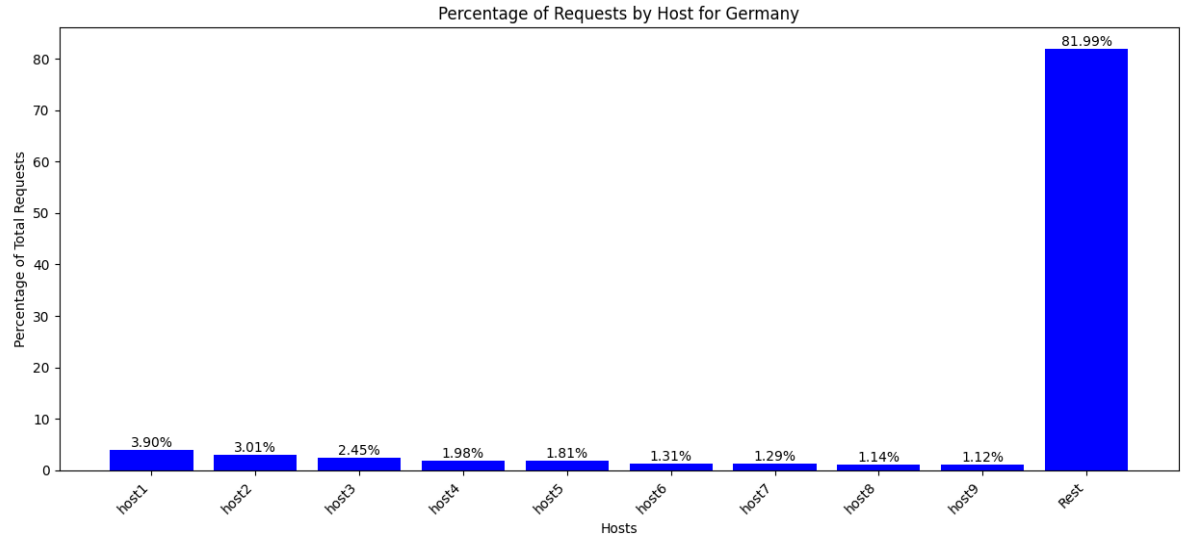
Country	Number of Unique Hosts
Germany (.de)	1138
Canada (.ca)	2970
Singapore (.sg)	78

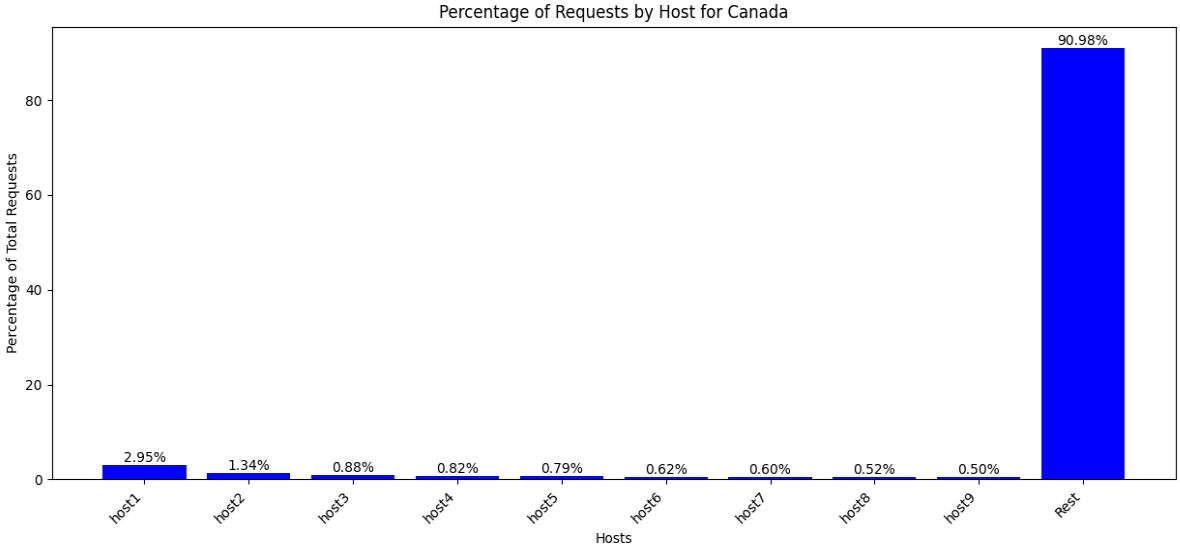
The top 9 hosts getting frequent requests for Germany, Canada and Singapore were recorded. They are presented below in tabular format:

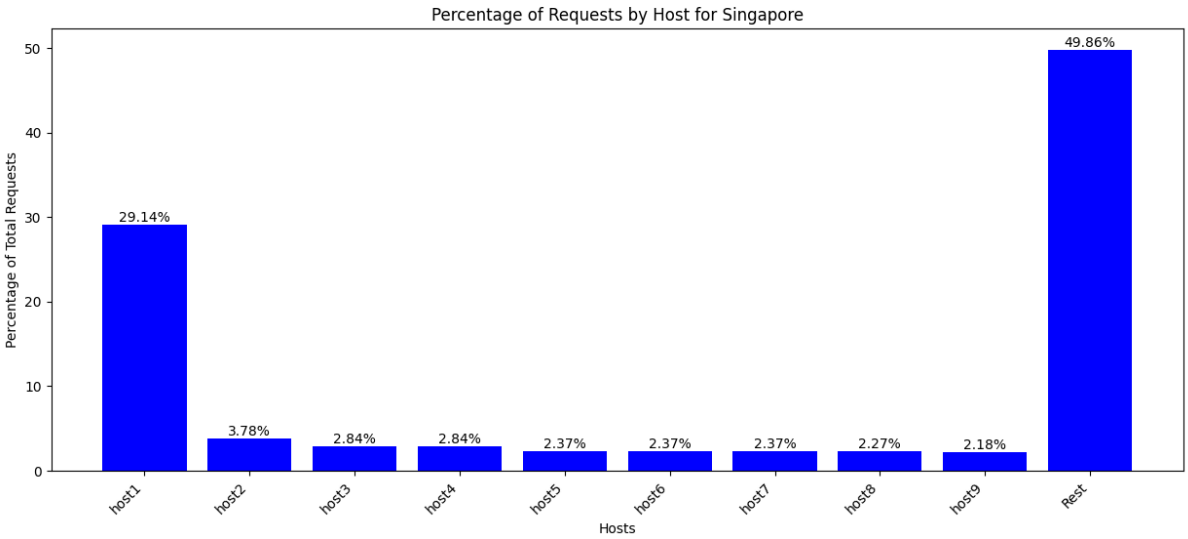
Top 9 frequent Hosts getting requests for each country				
Rank	**Host No.**	**Host Domains for Germany(.de)**	**Host Domains for Canada(.ca)**	**Host Domains for Singapore(.sg)**
1	Host 1	host62.ascend.interop.eunet.de, Requests: 832	ottgate2.bnr.ca, Requests: 1718	merlion.singnet.com.sg, Requests: 308
2	Host 2	aibn32.astro.uni-bonn.de, Requests: 642	freenet.edmonton.ab.ca, Requests: 782	sunsite.nus.sg, Requests: 40
3	Host 3	ns.scn.de, Requests: 523	bianca.osc.on.ca, Requests: 511	ts900-1314.singnet.com.sg, Requests: 30
4	Host 4	www.rrz.uni-koeln.de, Requests: 423	alize.ere.umontreal.ca, Requests: 479	ssc25.iscs.nus.sg, Requests: 30
5	Host 5	ztivax.zfe.siemens.de, Requests: 387	pcrb.ccrs.emr.ca, Requests: 461	scctn02.sp.ac.sg, Requests: 25
6	Host 6	sun7.lrz-muenchen.de, Requests: 280	srv1.freenet.calgary.ab.ca, Requests: 362	ts900-1305.singnet.com.sg, Requests: 25
7	Host 7	relay.ccs.muc.debis.de, Requests: 275	ccn.cs.dal.ca, Requests: 351	ts900-406.singnet.com.sg, Requests: 25
8	Host 8	dws.urz.uni-magdeburg.de, Requests: 244	oncomdis.on.ca, Requests: 304	ts900-402.singnet.com.sg, Requests: 24
9	Host 9	relay.urz.uni-heidelberg.de, Requests: 239	cobain.arcs.bcit.bc.ca, Requests: 289	einstein.technet.sg, Requests: 23

TASK-C:

The percentage (with respect to the total in a particular country) of requests by each of the top 9 most frequent hosts and the rest (i.e. 10 proportions in total) have been visualised below:

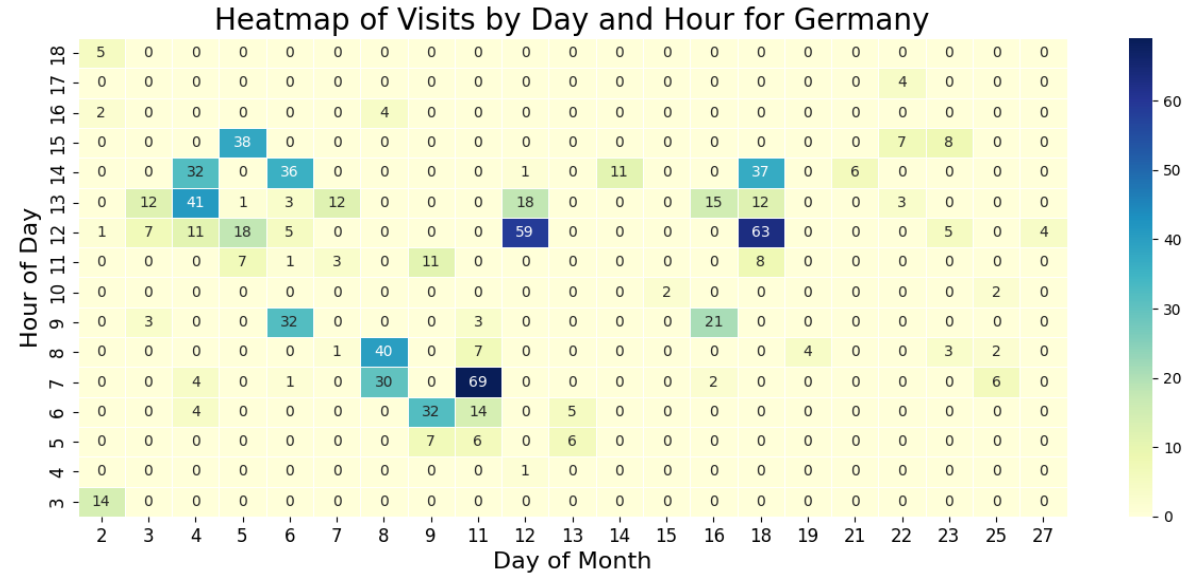


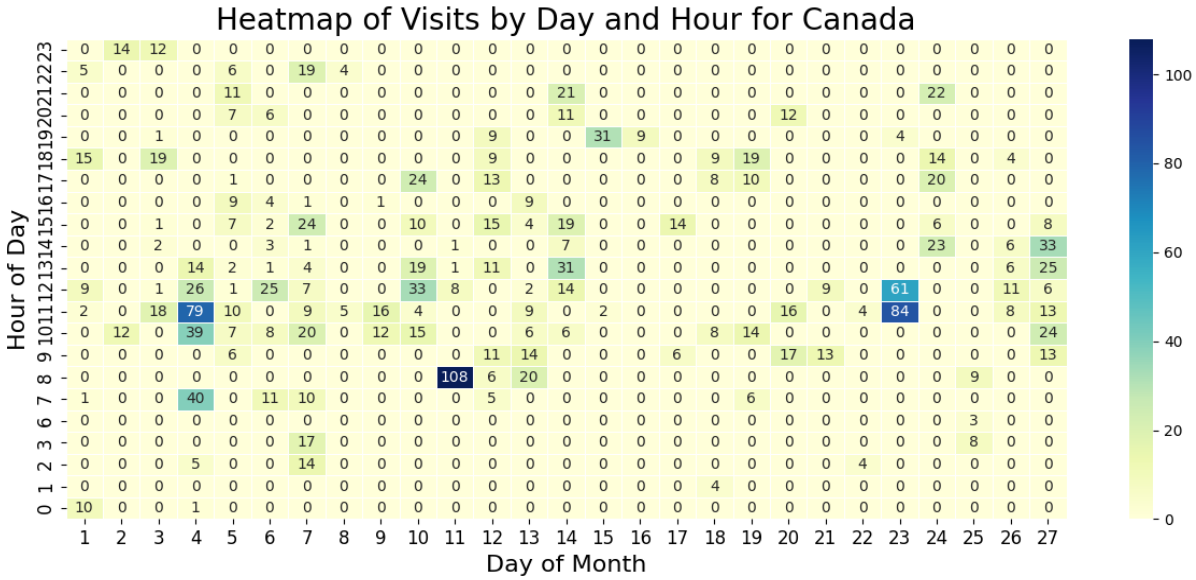


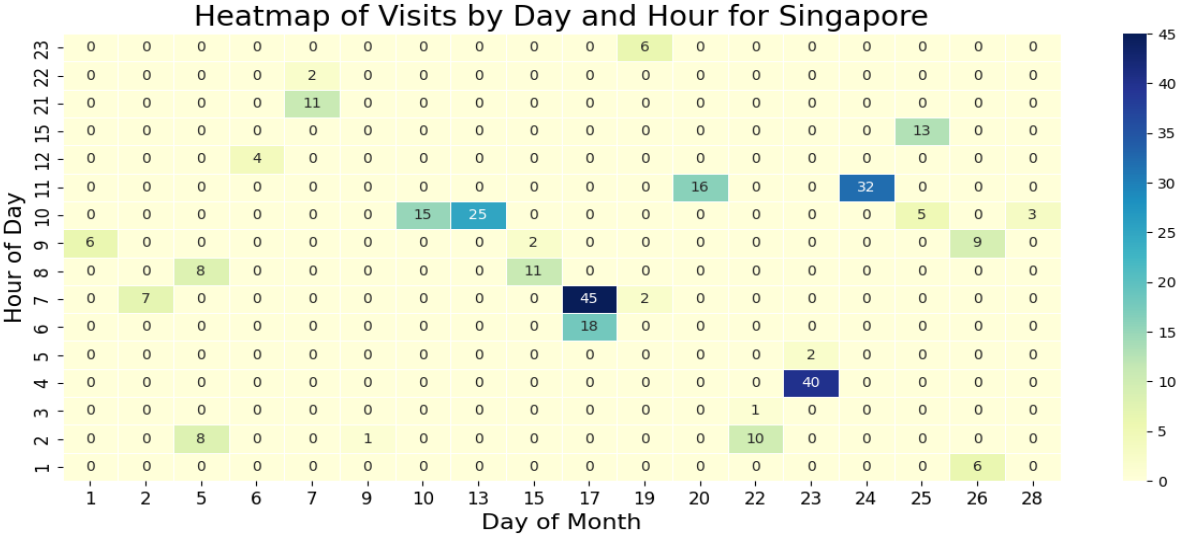


TASK-D:

The heatmaps for the most frequent host from each of the three countries, are given below:







TASK-E:

Observation 1: Sharp Drop in Requests Handled by Top 9 Hosts for Singapore

- **What is the observation?**

  Besides 'host1,' which manages 29.14% of the requests, all other hosts handle considerably fewer requests, mostly around 4%, with most hosts managing between 2-3%.

- **What are the possible causes of the observation?**

  This distribution suggests that 'host1' likely serves as a central hub for essential operations or high-demand services, whereas the other hosts are tasked with handling more niche or less frequently accessed services.

- **How useful is this observation to NASA?**

  The study of uneven web requests' distribution can assist NASA improve its online traffic management. By identifying traffic bottlenecks, NASA may improve its existing technologies, such as cloud-based services and traffic load balancers, to handle large volumes of online traffic during major events like live satellite launches. This ongoing improvement ensures that NASA's websites are efficient and stable, allowing them to transmit mission-critical information and efficiently communicate with the public during peak times. This method is critical for ensuring continuous access to NASA's educational and outreach information while not interfering with the functionality of its primary operating platforms.

Observation 2: Peak Traffic Periods as observed on the Heatmap of Visits by Day and Hour for Canada

- **What is the observation?**

  The heatmap for Canada illustrates high traffic surges at certain times, notably around mid-day and early evening on the 9th and 13th of the month, with traffic peaking at as many as 108 visits within a single hour.

- **What are the possible causes of the observation?**

  These peaks may correspond to specific operational tasks, such as submission deadlines, data or report releases, or scheduled events requiring a high level of user engagement.

- **How useful is this observation to NASA?**

  This insight is extremely useful given NASA's deliberate approach to regulating online traffic, particularly during peak hours. It enables NASA to plan for moments when their websites may suffer a sudden surge in visitors, such as during spacecraft launches or major scientific announcements. By recognising these traffic patterns,

NASA may better prepare its infrastructure, such as scalable servers, content networks, and traffic tools, to handle increased website traffic. This guarantees that their systems perform effectively and remain operational, giving the public continuous access to important events and information when it is most required. This proactive strategy is crucial for keeping users engaged and ensuring NASA's services remain reliable during peak periods.

Question 2: Liability Claim Prediction

The evaluation results and the final coefficients for Poisson Regression model(modelled on the number of claims (ClaimNb)), Logistic Regression models both for L1 and L2 regularisation are provided below:

Generalized Linear Poisson Regression RMSE: 0.23763542051774328

GLR Coefficients: [-0.0340878860814766,0.0022905994298538716,0.016734227743965047,-0.04753962831558597,-0.03319259634028304,0.017902164105792606,-0.005525856542322053,0.012228814951627785,-0.05386428030634081,-0.04508271275270079,0.027142245078206218,0.031239942714323888,-0.0028895924561898037,-0.016224322582753158,-0.007710421207962572,-0.015415828253565916,0.051313557023413295,0.020645501525344504,0.019001750819841 22,0.007984943577689044,0.02600444327698353,0.002083533355131499,0.00439532197842671,0.01049185192198241,0.014150941297814421,0.0052550443328267206,0.04437345614672775,0.05224980767101895,0.02720043221484024,-0.008228940667687657,-0.009573419893622787,0.04071431902412208,0.009831540784477366,-0.00915892934902316,-0.018999263642334868,-0.02435500003868532,0.0015622189021025988,-0.02720062554734062,-0.019551754379463784,0.007199014572413663,-0.0018982946359533792,-0.0295084055964305,0.011655492955644971,-0.037897955004335133,0.01748143830265687,0.007466455553255172,0.005090603683684039,0.005769094389488467,0.38444228825479654,-0.1900206260835634,0.12548907924809352,0.3198638387332218,0.014472742786697825]

Logistic Regression L1 Accuracy: 94.8917%

LR L1 Coefficients: (53,[2,3,8,9,10,11,26,27,28,31,33,34,35,37,41,43,48,49,50,51,52],[0.010723821738860306,-0.005232366384774111,-0.04377577515904755,-0.028867896913506705,0.011365556657372895,0.012074863462696744,0.032323195536105125,0.015467536743908413,0.011932144396529103,0.019094939813479404,-0.005276694634005105,-0.00235308987320976,-0.004658646184020652,-0.004739376252600268,-0.005605819483199246,-0.009412843192980467,0.3859905984011764,-0.1715908360402255,0.10809800647092085,0.312336304566976,0.011584300732213068])

Logistic Regression L2 Accuracy: 94.8927%

LR L2 Coefficients: [-0.02253242181201102,0.008940152529781873,0.021531863574525162,-0.04083905393981155,-0.02364368460470035,0.02536910742006618,0.0002289862556393209 7,0.020301134106821 16,-0.05640076003913854,-0.044621171618751364,0.031773474843075744,0.03518436160933384,-0.00019736151318077856,-0.014606232843077459,-0.006352358304367991,-0.01281500037000605,0.03488860418803326,0.0049597000890074855,0.007779540070955677,0.00036530608813584164,0.02004431101291715,-0.00461621832818712,0.0014786665911737794,0.006425981637126773,0.010907332448448461,-9.511137479527757e-05,0.05031111041734716,0.053119844200511544,0.026716227994619465,-0.016553944332368887,-0.014621104434432765,0.04259697626840239,0.010215861047339203,-0.026797550154164945,-0.023816103702884152,-0.030427777403254,0.0006777551188464748,-0.029587553304040747,-0.019830653270284732,0.009871949449397859,-0.002255540817313577,-0.032228994821380086,0.013599865710642717,-0.03639513132995874,0.016887196911067614,0.0014226208237855956,0.0026360408951473864,0.005474653769135977,0.39744808547322474,-0.19869334416277798,0.13110500475068118,0.3332640344461646,0.018931469453400875]

Three observations which can be drawn after analysing the evaluation results along with the final model coefficients are:

- **Observation 1: Magnitude and Influence of GLR Poisson Coefficients**
  The Generalized Linear Poisson Regression (GLR) model has a low RMSE of 0.2376, showing it does a good job of fitting the data. The coefficients, which indicate how much each feature affects the outcome, vary widely from about -0.19 to 0.38. This shows that while some features greatly influence the result either positively or negatively, others have minimal effect. The larger coefficients by magnitude, specifically 0.384 positive and -0.190 negative, point out key traits that greatly influence the predictions of the model.

- **Observation 2: Impact of Regularization on Coefficients**

  L1 regularisation, also known as Lasso regularisation, simplifies models by reducing some coefficients to zero, thereby eliminating less important features. This results in a more interpretable model that focuses on the most significant features, which is especially useful in models where feature reduction is desired. On the other hand, L2

regularisation, or Ridge regularisation, reduces the magnitude of all coefficients towards zero but doesn't eliminate them completely. This approach keeps all features in the model but with reduced influence, which helps in capturing complex patterns in the data without greatly simplifying the model. This leads to a key difference which is that L1 regularisation can lead to sparser models by removing features entirely, while L2 regularisation tends to produce denser models by retaining all features with diminished weights.

- **Observation 3: Accuracy and Regularization**
  The accuracy of the logistic regression models with L1 and L2 regularisation is almost the same, with L1 achieving 94.8917% and L2 slightly higher at 94.8927%. This slight difference suggests that for this particular dataset, choosing between L1 and L2 regularisation doesn't greatly affect the model's accuracy. However, the type of regularisation might influence other performance aspects, like how easily the results can be interpreted and how well the model can adapt to new, unseen data, depending on the most influential features and the data's underlying characteristics.

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

TASK-A: Using Pipelines and Cross-Validation for Hyperparameter Tuning

The approach starts with carefully preparing and sampling the dataset to ensure class balance, which is critical for effective model training and validation. By taking a 1% random sample from the whole dataset and guaranteeing an equitable distribution between classes, the technique prevents against any model learning biases. This stage is especially important in classification tasks such as discovering the Higgs boson particle, because imbalances across classes can drastically alter learning outcomes.

For the hyperparameter tuning:

1. **The Random Forest Classifier** was customised using three hyperparameters**: numTrees (**number of trees in the forest**), maxDepth (**maximum depth of each tree**), and maxBins (**maximum number of bins used for feature splitting**),** each having three options to explore different configurations**.**
2. **Gradient Boosting Trees Classifier** explored **maxIter** (maximum number of iterations), **maxDepth**, and **stepSize** (learning rate) to find the optimal hyperparameters.
3. **Multilayer Perceptron Classifier** tested different neural network structures by adjusting **layers**, the number of **maxIter** (iterations for convergence), and **blockSize** (size of the block of input data for each iteration).

These models were tested in a cross-validation setup with a 5-fold method to assess each hyperparameter combination's performance on the sampled data, ensuring that the findings are robust and resistant to overfitting. This phase successfully discovers the optimal configuration for each model using the AUC(Area Under the ROC Curve) metric. The same train-test split in the sampled data was used throughout these evaluations to guarantee consistency in performance comparisons.

TASK-B: Model Training on the Full Train Dataset and Evaluation on the full Test Dataset

After identifying the best hyperparameters from the sampled dataset, the models are then retrained using these optimised settings on the full dataset. The models are evaluated using the same training and test data splits, maintaining consistency in the testing process. This ensures a reliable comparison of performance across the models, highlighting how each algorithm adapts and performs with the full dataset.

Comparison of performances of the three algorithms:

The best hyperparameters obtained after the cross-validation process for hyperparameter tuning are given below:

Best Hyperparameters for each Algorithm		
Algorithm	Parameter	Value
Random Forest	numTrees	10
	maxDepth	10
	maxBins	50
Gradient Boosting Trees	maxIter	30
	maxDepth	5
	stepSize	0.2
Neural Network	layers	[28, 5, 5, 2]
	maxIter	100
	blockSize	20

The evaluation results obtained after transforming the fitted model on both the Sampled Test data and the full Test data are provided below:

Evaluation Results			
Algorithm	**Data Type**	**AUC**	**Accuracy**
Random Forest	Sampled Test	0.7755	0.7019
Gradient Boosting Trees	Sampled Test	0.7906	0.7139
Neural Network	Sampled Test	0.7425	0.6832
Random Forest	Full Test	0.7778	0.7047
Gradient Boosting Trees	Full Test	0.7961	0.7194
Neural Network	Full Test	0.7505	0.688

Analysis:

- **Gradient Boosting Trees** consistently outperforms the other two models on both sampled and full test data, achieving the best AUC and accuracy. This shows that GBTs are very effective at dealing with the complexities of the data used, potentially due to the sequential error correction inherent in boosting approaches**,** where each subsequent learner focuses on the errors made by the previous ones, attempting to correct these errors in the next round of training.
- **Random Forest** also performs well, with a modest improvement in both AUC and accuracy when comparing sampled data to the complete dataset. This improvement suggests that the Random Forest model may benefit from more data, which could provide a more diversified set of instances for better generalisation.
- In both test datasets, the **Neural Network**, particularly a shallow one, lags behind the tree-based algorithms in terms of AUC and accuracy. Although performance improves slightly when as we move from sampled to full test data, the Neural Network continues to fall behind tree-based models. This lower performance could be attributed to the Neural Network's shallow architecture, which may be insufficiently deep or complicated to effectively gather and model the complexities present in the data.

Conclusion: The Gradient Boosting Trees model has the best scalability and overall effectiveness in this situation, followed by the Random Forest and then the Neural Network. The increase in performance metrics from sampled full test data for all models indicates that larger datasets benefit each model, most likely due to improved generalisation from more comprehensive training. However, the degree of improvement and overall performance vary greatly, with GBTs providing the highest benefit and the most consistent performance across both datasets. This comparison highlights the significance of model selection based on both the nature of the data and the computational resources available, as well as the potential benefits of adopting ensemble approaches such as Gradient Boosting Trees in complex classification tasks.

Question 4: Movie Recommendation and Cluster Analysis

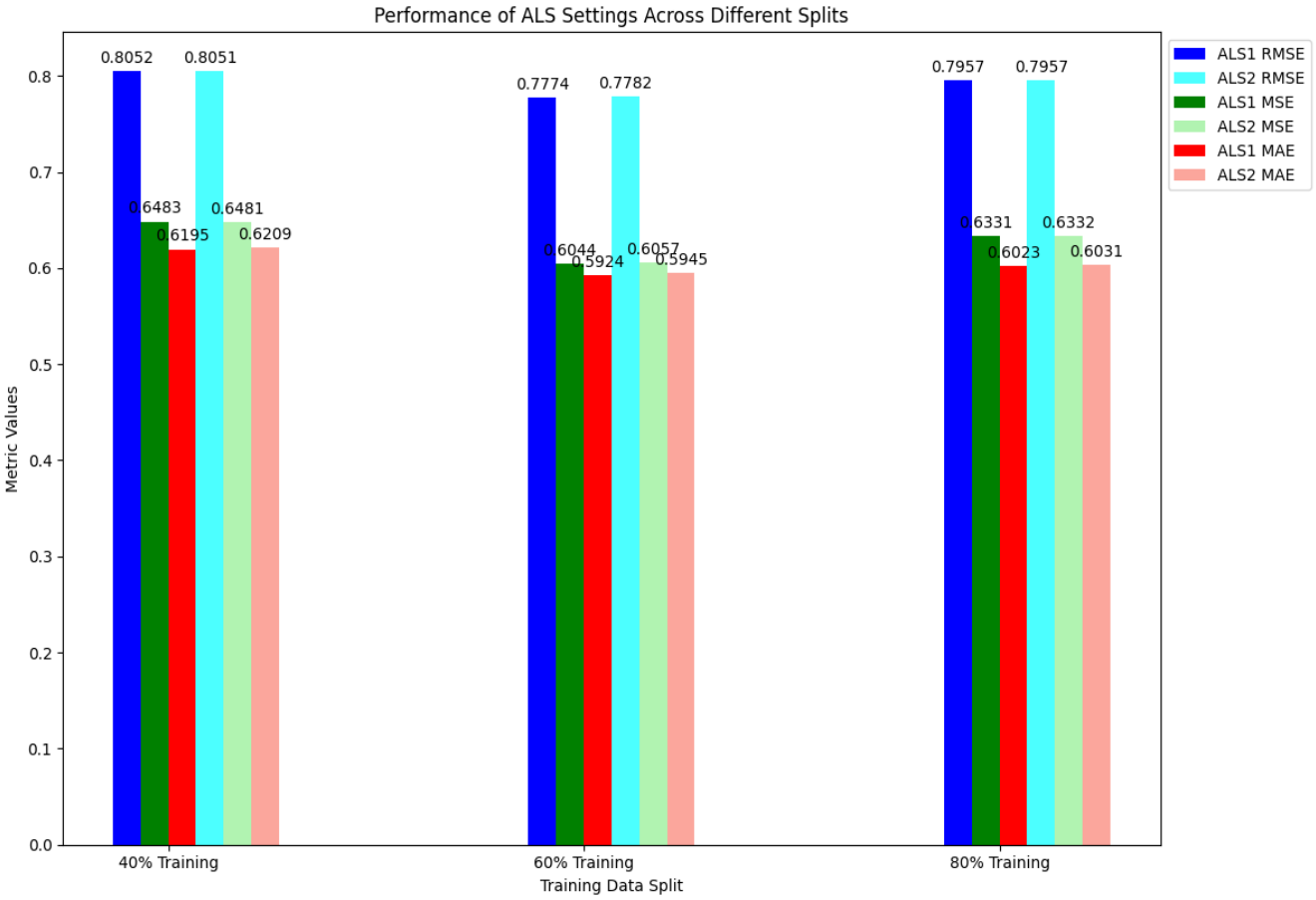
TASK-A: Time-split Recommendation

The ALS (Alternating Least Squares) algorithm looks at the existing data about how users have interacted with different items—like movies, products, or articles—and notices where there's missing information, such as a user not having rated a movie yet. The algorithm then makes predictions about these missing interactions. For example, if a user liked similar movies, the algorithm might predict they'll also like this one. These predictions help the system recommend new items to users based on what they seem to like or dislike from their past actions.

Using the python code, different metric values(Root Mean Square Error, Mean Square Error and Mean Absolute Error) were computed for two settings of ALS systems. The values are presented below in tabular format:

ALS Settings Metric Values				
Setting	**Split**	**RMSE**	**MSE**	**MAE**
ALS Setting 1	40%	0.8052	0.6483	0.6195
ALS Setting 1	60%	0.7774	0.6044	0.5924
ALS Setting 1	80%	0.7957	0.6331	0.6023
ALS Setting 2	40%	0.8051	0.6481	0.6209
ALS Setting 2	60%	0.7782	0.6057	0.5945
ALS Setting 2	80%	0.7957	0.6332	0.6031

These values have been visualised in bar plots, as shown below:

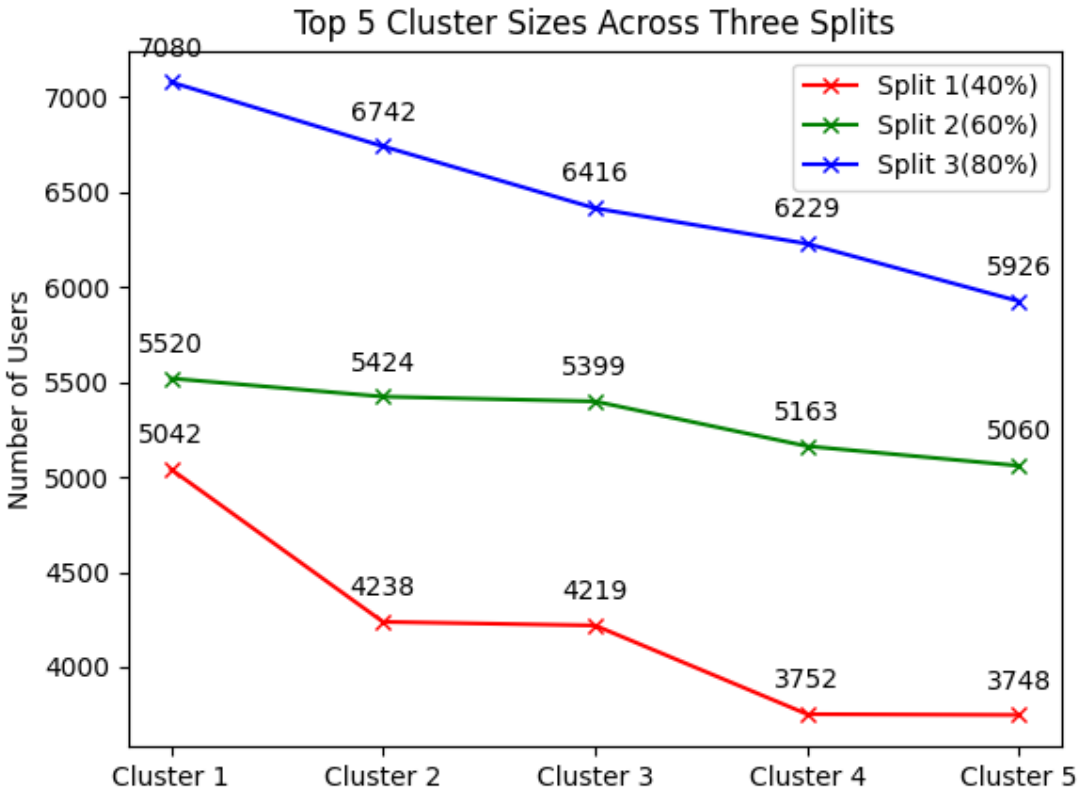


TASK-B: User Analysis

For each of the three time splits, K-means clustering with K=25, was used to clusters all the users based on user factors as learned by the second setting of the ALS system, and the top five largest user clusters were determined. The size of (i.e. the number of users in) each of the top five clusters are tabulated below:

Top 5 Cluster Sizes Across Three Splits					
Split	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Split 1(40%)	5042	4238	4219	3752	3748
Split 2(60%)	5520	5424	5399	5163	5060
Split 3(80%)	7080	6742	6416	6229	5926

The visualised form of the sizes of the top 5 clusters for each of the 3 time splits is presented below:



The top movies, that is, the movies having an average rating greater than or equal to 4 (>=4) are determined and named as top_movies. The data from 'movies.csv' was used to find the genres for all of the top_movies. The top_movies are grouped based on genres and the top 10 genres for each time split were determined and are reported below:

Top Genres for Split 1	
Genre	**Count**
Drama	57
Comedy	27
Thriller	19
Crime	15
Romance	14
Action	11
Adventure	9
Mystery	8
War	8
Horror	6

Top Genres for Split 2:	
Genre	**Count**
Drama	405
Comedy	173
Romance	122
Thriller	85
Adventure	84
Crime	84
Documentary	72
Action	69
War	66
Mystery	42

Top Genres for Split 3:	
Genre	**Count**
Drama	1394
Comedy	542
Romance	367
Crime	263
Thriller	257
Documentary	245
War	157
Adventure	156
Action	142
Mystery	123

TASK-C:

Observation 1: Optimal Data Split for Enhanced Recommendation Accuracy

- **Observation:** In both ALS settings examined, the 60% data split consistently outperformed the 40% and 80% splits in terms of performance measures (RMSE, MSE, and MAE).
- **Possible Causes:** This optimal performance at the 60% split is most likely owing to a well-balanced proportion of data that is large enough to capture a wide range of user behaviours and interactions but not so extensive that it introduces too much noise or complexity into the model. The 40% split may be insufficient, missing key information required to effectively understand user preferences, whereas the 80% split may incorporate diminishing returns on additional data, which may include irrelevant or redundant information that adds to, or even negatively impacts predictive accuracy.
- **Utility to a Movie Website:** For Netflix, using the 60% split to train their movie recommendation system involves employing a dataset that is optimally sized for maximum prediction accuracy while avoiding the downsides of overfitting and heavy processing demands. This optimal use of data not only improves the quality and relevance of movie recommendations, but it also increases operational efficiency by reducing unnecessary computing costs. It enables Netflix to provide a highly responsive and user-centric service, boosting user engagement and happiness with more accurate and personalised content recommendations.

Observation 2: Genre Popularity Trends in Content Streaming Platforms

- **Observation:** Drama is consistently the most popular genre throughout all time splits, followed by Comedy, but the count of each genre grows significantly as more data is added in each split.
- **Possible Causes:** Drama and comedy are broad and popular genres that appeal to a large audience, so they are more likely to be rated by users. The significant increase in counts with larger time splits implies that when more user data is acquired, the intrinsic popularity of these genres becomes more obvious, indicating a genuine preference among a broader audience.
- **Utility to a Movie Website:** This observation is critical for streaming services such as Netflix since it identifies which genres have broad popularity. Prioritising these genres allows Netflix to ensure that their content library meets the preferences of the majority, potentially increasing client retention and happiness. Furthermore, these trends can help Netflix plan marketing tactics and create new content, with an emphasis on genres that have been shown to draw the most views.