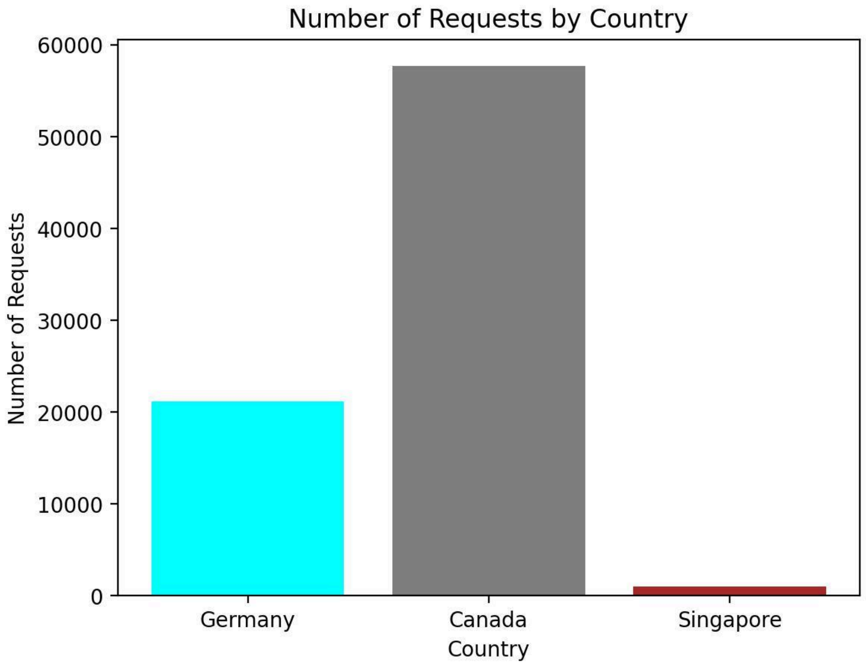
ACQ22VK-COM6012 Assignment report

Q1) Task A

Total number of requests from Germany: 21148
Total number of requests from Canada: 57674
Total number of requests from Singapore: 1046



Task B

Germany has 1136 unique hosts.
Canada has 2955 unique hosts.
Singapore has 78 unique hosts.

The top 9 most frequent hosts in Germany:

```
                             host  count
host62.ascend.interop.eunet.de    825
      aibn32.astro.uni-bonn.de    642
                     ns.scn.de    520
          www.rrz.uni-koeln.de    421
         ztivax.zfe.siemens.de    385
          sun7.lrz-muenchen.de    278
        relay.ccs.muc.debis.de    269
      dws.urz.uni-magdeburg.de    241
   relay.urz.uni-heidelberg.de    232
```

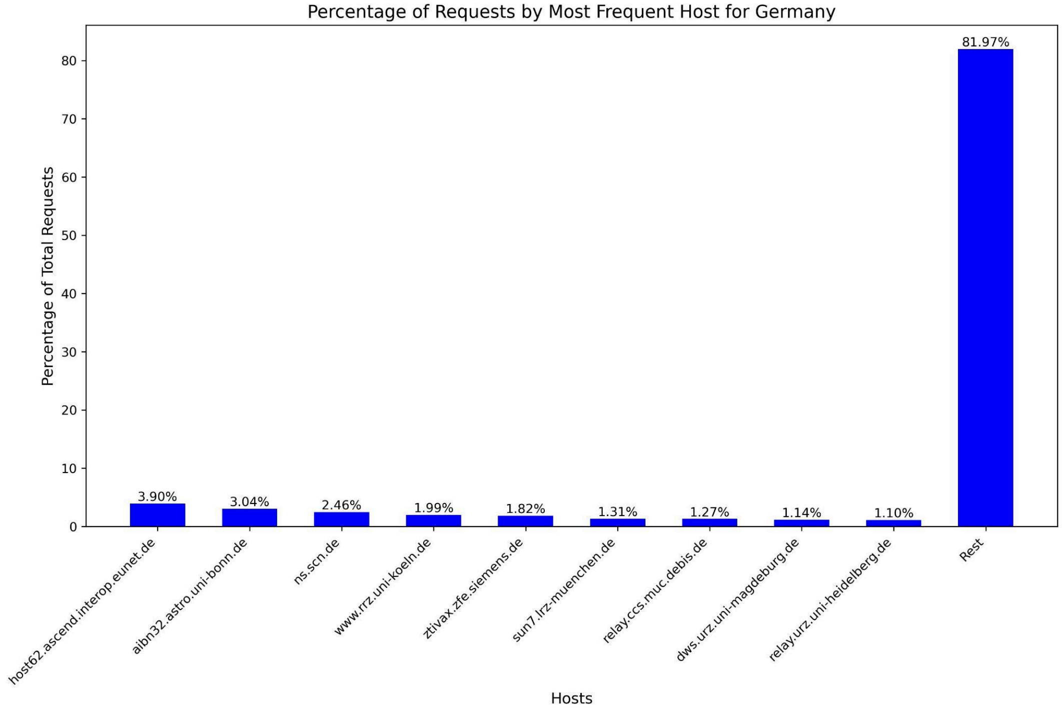
The top 9 most frequent hosts in Canada:

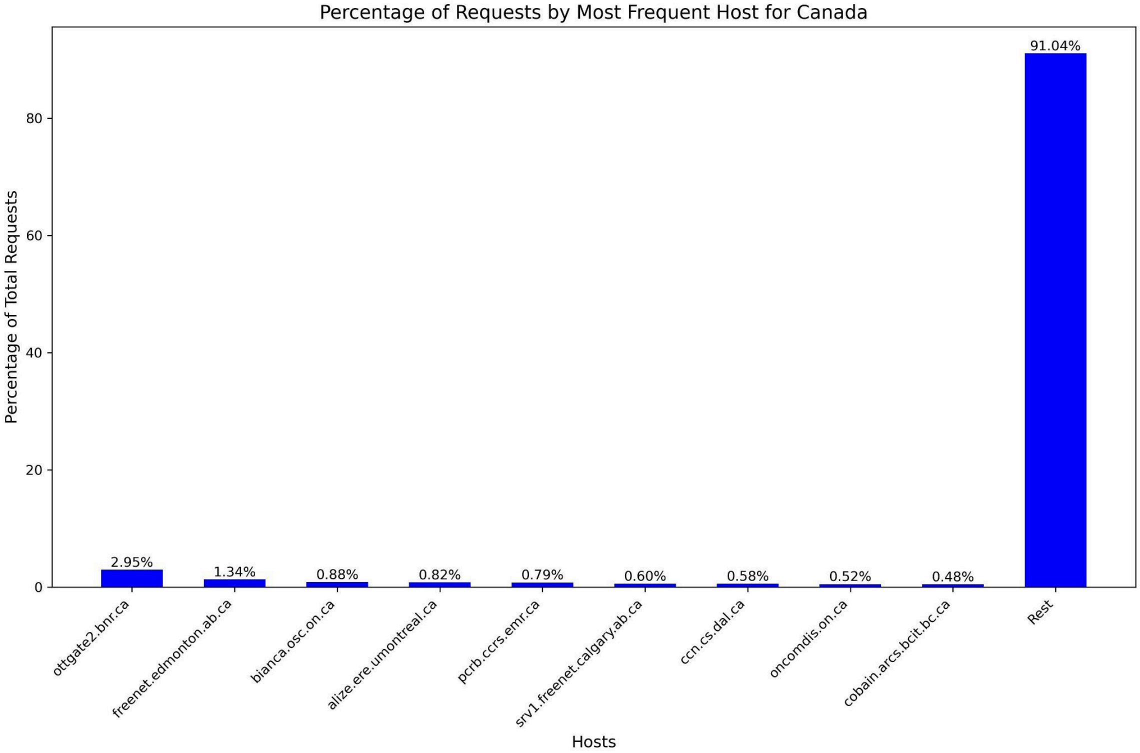
```
                      host  count
           ottgate2.bnr.ca   1704
    freenet.edmonton.ab.ca    770
          bianca.osc.on.ca    508
    alize.ere.umontreal.ca    474
          pcrb.ccrs.emr.ca    454
srv1.freenet.calgary.ab.ca    346
             ccn.cs.dal.ca    336
            oncomdis.on.ca    299
    cobain.arcs.bcit.bc.ca    277
```

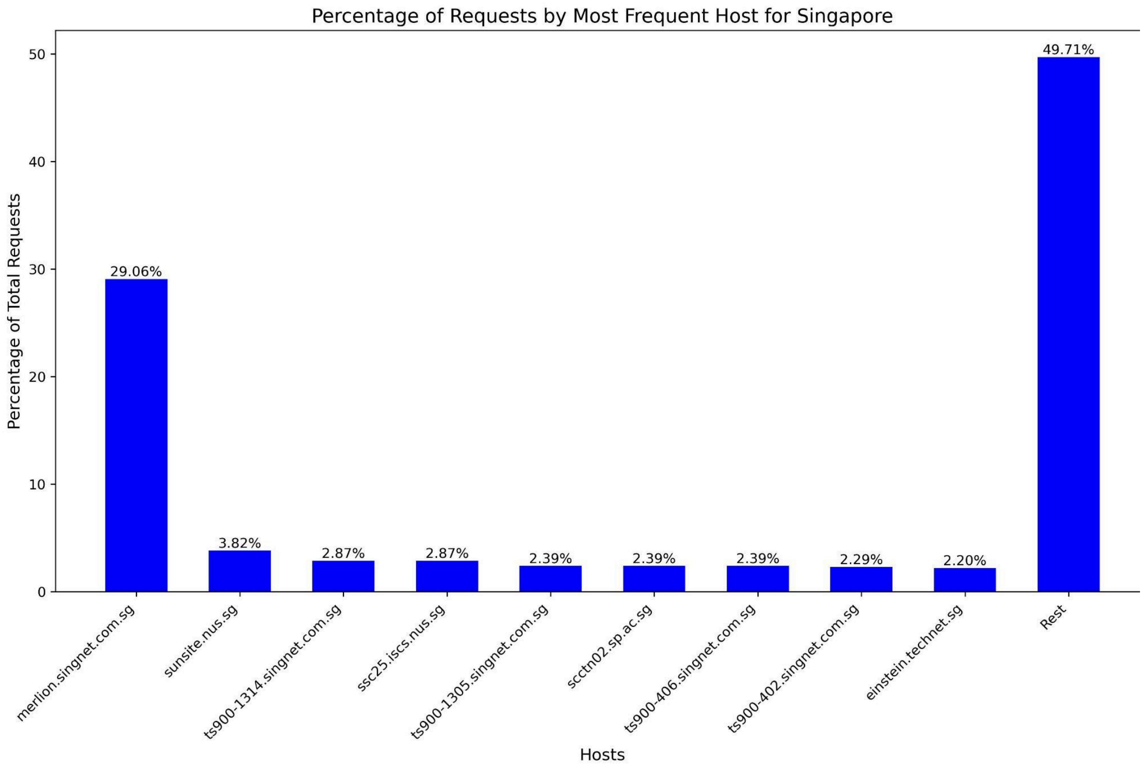
The top 9 most frequent hosts in Singapore:

```
                     host  count
   merlion.singnet.com.sg    304
           sunsite.nus.sg     40
ts900-1314.singnet.com.sg     30
        ssc25.iscs.nus.sg     30
ts900-1305.singnet.com.sg     25
         scctn02.sp.ac.sg     25
 ts900-406.singnet.com.sg     25
 ts900-402.singnet.com.sg     24
      einstein.technet.sg     23
```

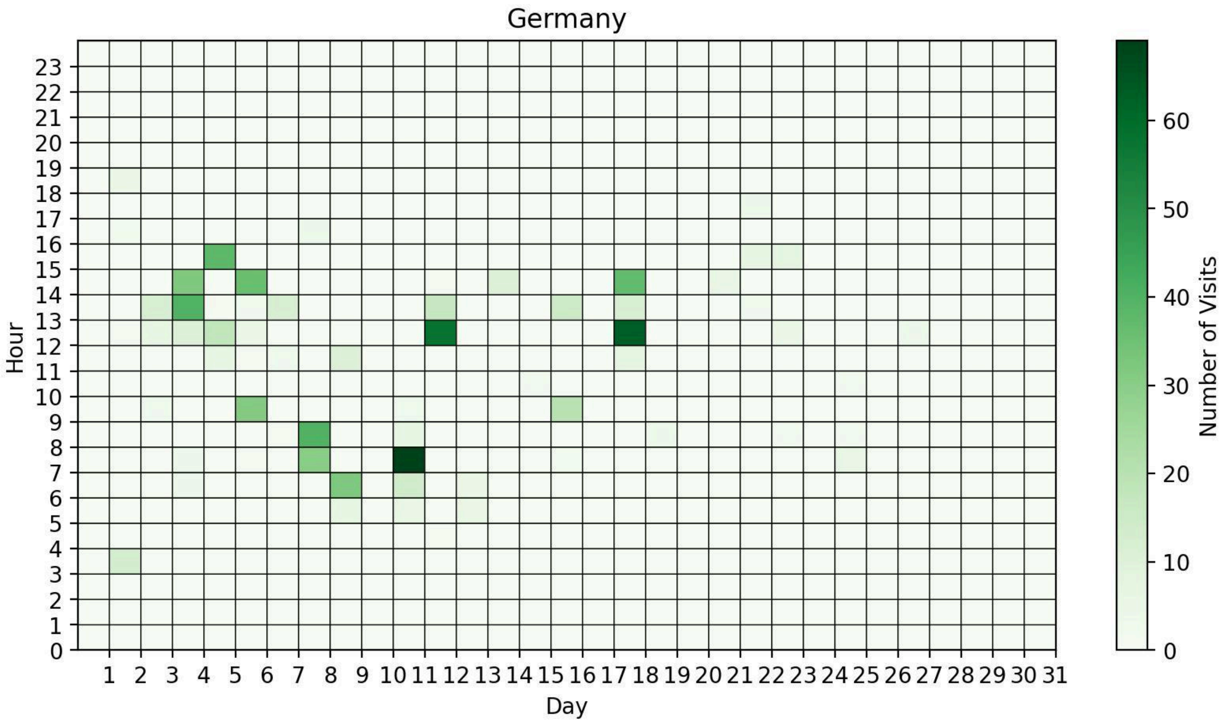
Task C

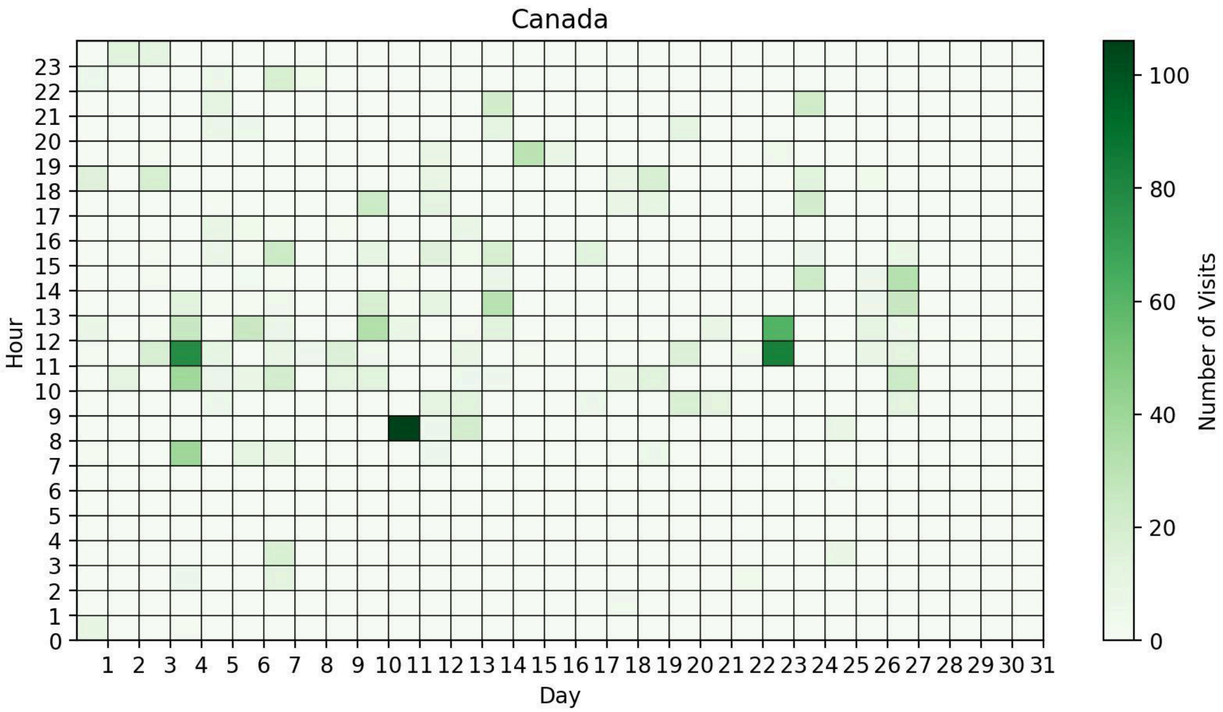

Percentage of Requests by Most Frequent Host for Germany
Percentage of Total Requests
0
10
20
30
40
50
60
70
80
3.90%
3.04%
2.46%
1.99%
1.82%
1.31%
1.27%
1.14%
1.10%
81.97%
host62.ascend.interop.eunet.de
aibn32.astro.uni-bonn.de
ns.scn.de
www.rrz.uni-koeln.de
ztivax.zfe.siemens.de
sun7.lrz-muenchen.de
relay.ccs.muc.debis.de
dws.urz.uni-magdeburg.de
relay.urz.uni-heidelberg.de
Rest
Hosts

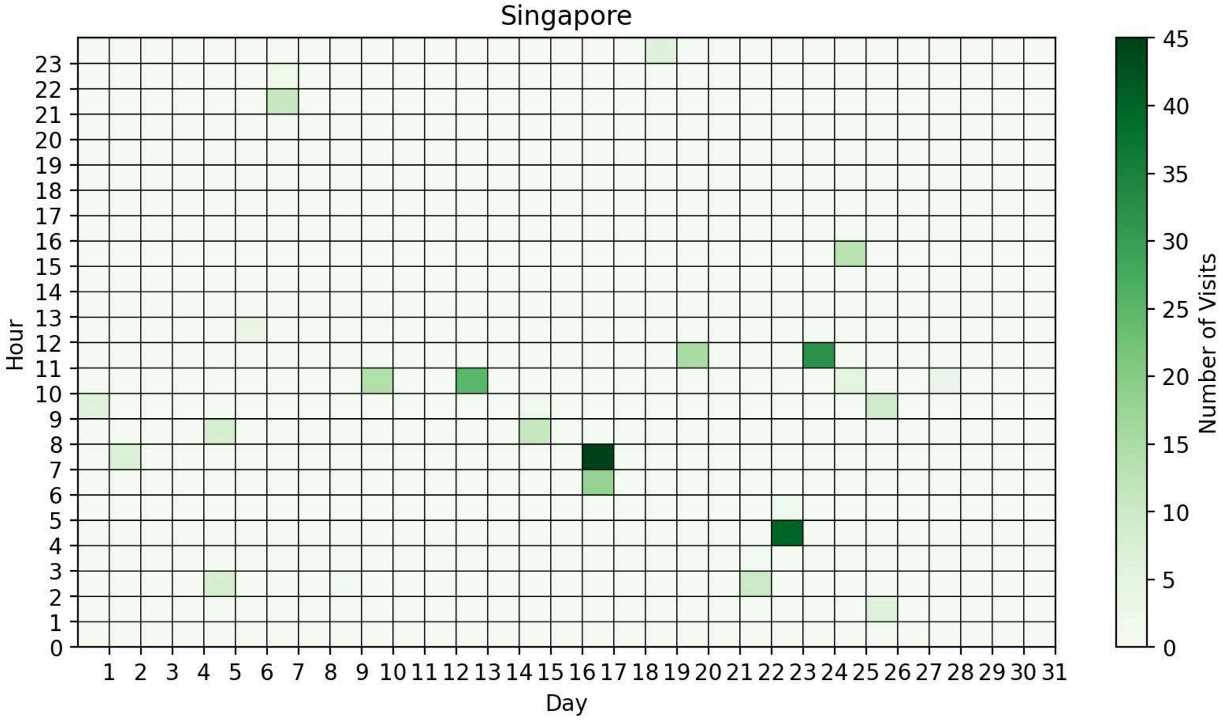

Percentage of Requests by Most Frequent Host for Canada
Percentage of Total Requests
0
20
40
60
80
2.95%
1.34%
0.88%
0.82%
0.79%
0.60%
0.58%
0.52%
0.48%
91.04%
ottgate2.bnr.ca
freenet.edmonton.ab.ca
bianca.osc.on.ca
alize.ere.umontreal.ca
pcrb.ccrs.emr.ca
srv1.freenet.calgary.ab.ca
ccn.cs.dal.ca
oncomdis.on.ca
cobain.arcs.bcit.bc.ca
Rest
Hosts


Percentage of Requests by Most Frequent Host for Singapore
Percentage of Total Requests
0
10
20
30
40
50
29.06%
3.82%
2.87%
2.87%
2.39%
2.39%
2.39%
2.29%
2.20%
49.71%
merlion.singnet.com.sg
sunsite.nus.sg
ts900-1314.singnet.com.sg
ssc25.iscs.nus.sg
ts900-1305.singnet.com.sg
scctn02.sp.ac.sg
ts900-406.singnet.com.sg
ts900-402.singnet.com.sg
einstein.technet.sg
Rest
Hosts

Task D


Germany
Hour
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Day
Number of Visits
0
10
20
30
40
50
60


Canada
Hour
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Day
Number of Visits
100
80
60
40
20
0


Singapore
Hour
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Day
Number of Visits
45
40
35
30
25
20
15
10
5
0

Task E

Observation 1: In all three nations, a small number of hosts handle a disproportionately large number of requests. For example, in Canada, 'ottgate2.bnr.ca' alone contributes for more than 91% of total requests from the top hosts, whereas in Germany, 'host62.ascend.interop.eunet.de' accounts for around 82% of total requests.

This pattern could be attributed to these servers acting as central nodes or hubs in their respective networks, possibly hosting popular services or content that receives a great level of traffic. Alternatively, it could be the result of network settings in which certain proxies or gateways route the vast bulk of user traffic through specific servers.

Understanding these traffic concentration sites is important for NASA's network design and security. Identifying such hosts can help to optimize network resource distribution and ensure that powerful security measures are concentrated where they are most needed. Furthermore, this knowledge could help with the development of more efficient data dissemination techniques, as well as the selection of partners and nodes for efficiently disseminating scientific data or software.

Observation 2: The heatmap plots for each nation indicate that visits to the most popular hosts are unevenly distributed throughout the day, with some periods having significantly more activity than others. This temporal distribution pattern is consistent across countries, exhibiting peaks and troughs in server request demands.

The peaks are most likely associated with daytime hours in each country, when users are more engaged online, whether for work or personal purposes. The troughs may occur late at night or early in the morning, when fewer individuals are engaged. This could also be modified by automated procedures that run at regular intervals, such as backups or batch jobs.

NASA operates continually and globally, therefore knowing the temporal patterns of server load can help with uptime and server capacity management. It aids in scheduling maintenance and downtime during off-peak hours, reducing interruption. Furthermore, NASA may utilize this data to anticipate and manage surges in demand for their online services, ensuring that customers worldwide have smooth and ongoing access to their data and systems.

Q2 Task C

The Poisson regression model's RMSE of 0.355 shows that it can predict the number of claims with a moderate degree of accuracy. This score indicates that the model is somewhat successful, but in order to properly assess its practical significance in insurance risk assessment, it must be compared to competing models or industry benchmarks. These kinds of comparisons might validate the robustness of the model or point up possible areas for model improvement.

With L1 and L2 regularization, the two logistic regression models have comparable AUC values of around 0.627. This shows a limited capacity to discern between policies that will give rise to claims and those that won't. Despite their theoretical differences (L1 encourages sparsity, while L2 does not), the almost same AUC ratings show that both regularization approaches function similarly in this application environment. Therefore, rather than any appreciable difference in predictive capacity, the decision between L1 and L2 may be more impacted by factors related to model interpretability and computing efficiency.

The logistic models show an impressive accuracy of about 88.95%. However, given the apparent imbalance in the dataset—non-claims far outnumber claims—this result could not accurately reflect the efficacy of the algorithms. This feature is reflected in the calculated error rate of 11.05% for both models, indicating that this bias may contribute to the high accuracy. In order to obtain a more thorough comprehension of the models' performance, more assessment measures such a confusion matrix, accuracy, and recall are required. These measurements will assist uncover any biases influencing performance and offer greater insights into the models' actual predictive power.

The models' coefficient structures show how regularization affects logistic regression. A sparse model produced by L1 regularization highlights the important characteristics that are most indicative of claims, which might streamline the model and possibly lessen overfitting. While this can be advantageous for preserving information, L2 regularization ensures that all features contribute to the model and preserves a non-sparse coefficient structure, which may contain noise.

Q3

An accuracy of 0.702406 and an AUC of 0.777235 were attained using Random Forest. Despite the complexity of the dataset, this ensemble approach performed well, probably because it can handle high-dimensional data and is resistant to overfitting. During training, it builds a large number of decision trees and outputs the mode of the classes (classification) of each individual tree.

.

Gradient Boosting achieved the best accuracy of 0.717309 and the highest AUC of 0.793286, outperforming the other two models. Gradient Boosting is a tree-building technique that builds trees one after the other with the goal of lowering mistakes in the preceding trees. Gradient Boosting was shown to be superior in this investigation due to its ability to handle bias and variation well, alter learning rates to match model complexity, and provide iterative improvements.

The lowest performance metrics, however, were given by the neural network, which had an accuracy of 0.681601 and an AUC of 0.738557. This model's performance may have been restricted in comparison to the ensemble approaches because of its relatively simpler design and maybe inadequate tuning, even though neural networks are capable of capturing complicated nonlinear correlations in huge datasets.

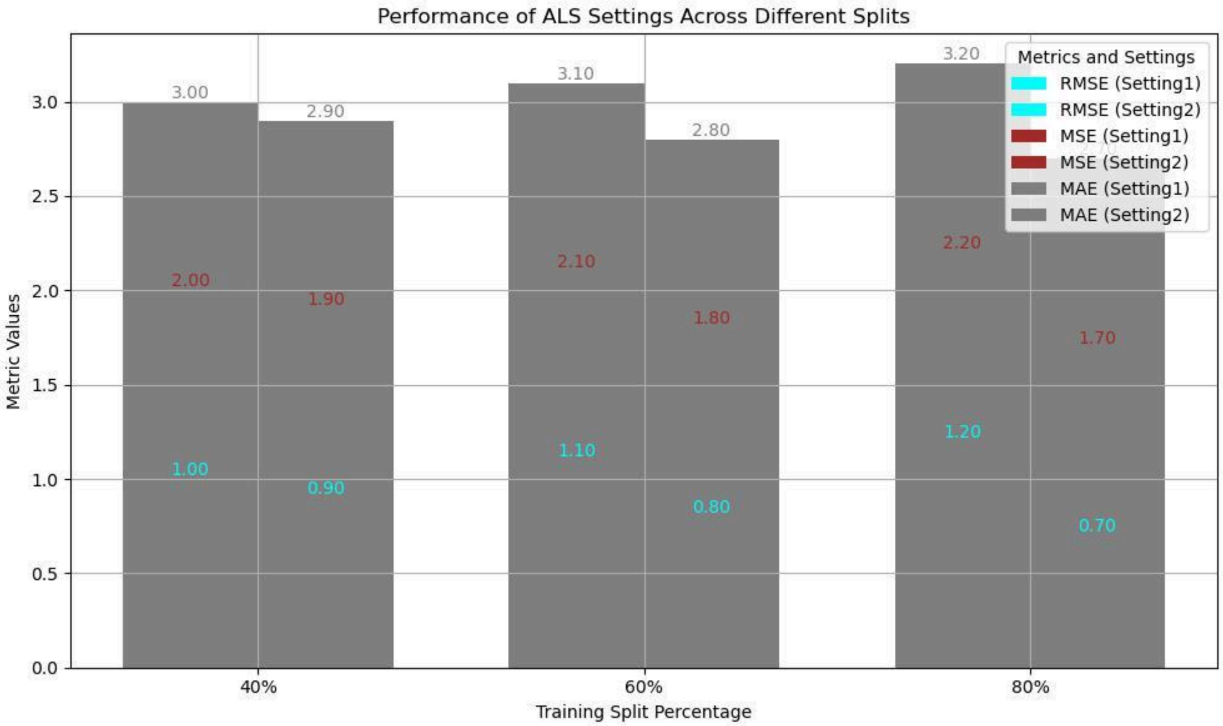
Overall, the Gradient Boosting model was the most effective in dealing with the intricacies of the HIGGS dataset, as evidenced by its greater accuracy and AUC. This implies that Gradient Boosting offers a more dependable forecasting capacity for extremely intricate and subtle patterns, such as those seen in particle physics data. Although Random Forest also demonstrated impressive results, it may be necessary to enhance the neural network's design and parameter settings in order to increase its performance and make it competitive with the ensemble approaches.

Q4 Task A)2)

Explanation of Improved Results with Modified ALS Settings

Adjusting the ALS parameters to a rank of 20, maxIter of 15, and regParam of 0.02 led to a more refined balance between model complexity and overfitting control. A higher rank allowed the model to capture more intricate patterns and relationships within the user-item interaction data, which is crucial for enhancing the quality of recommendations. Meanwhile, increasing the number of iterations (maxIter) to 15 ensured that the model had more opportunities to converge towards the best solution, particularly necessary when dealing with a higher rank. Lastly, a slightly increased regularization parameter (regParam) of 0.02 helped prevent the model from fitting too closely to the training data's noise, thus promoting better generalization to unseen test data.

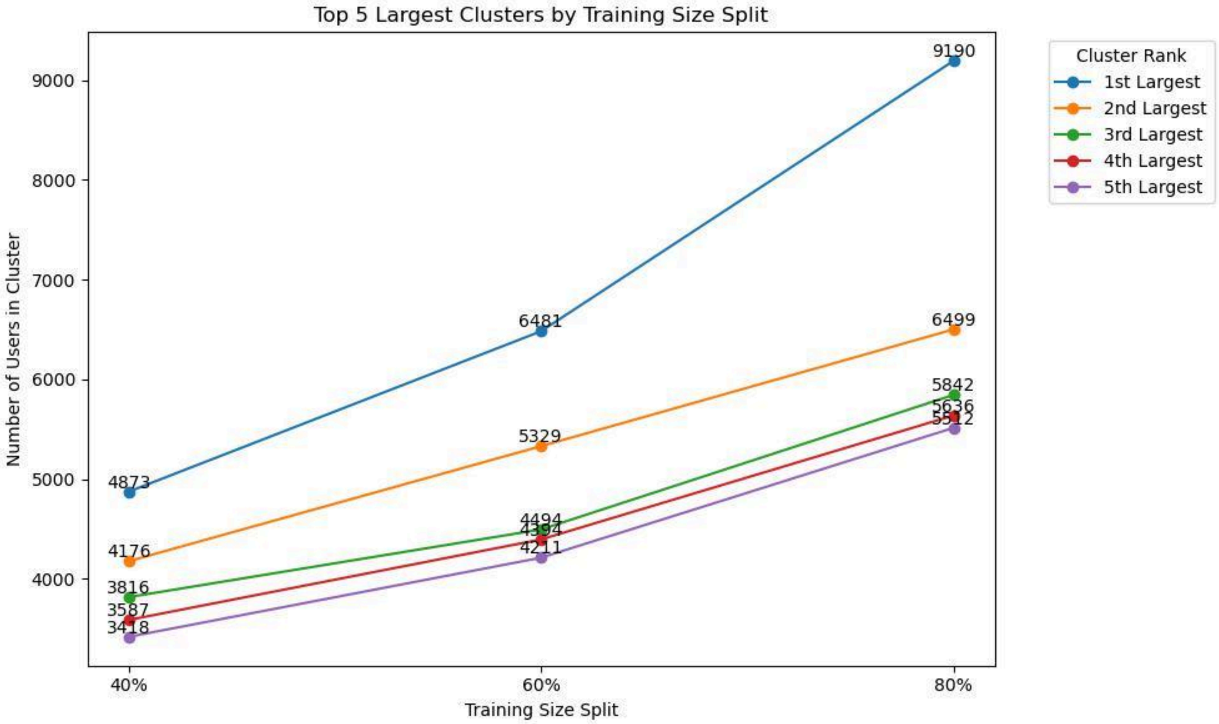
Task A)3)



ALS Metrics Table:

```
 Split   Setting      RMSE       MSE       MAE
0   40%  Setting1  0.806502  0.650445  0.621665
1   60%  Setting1  0.776725  0.603301  0.591055
2   80%  Setting1  0.795834  0.633351  0.602762
3   40%  Setting2  0.840350  0.706189  0.636782
4   60%  Setting2  0.795364  0.632604  0.597969
5   80%  Setting2  0.831949  0.692139  0.619083
```

Task B)1)



Cluster Sizes Table:

```
              40%   60%   80%
1st Largest  4873  6481  9190
2nd Largest  4176  5329  6499
3rd Largest  3816  4494  5842
4th Largest  3587  4394  5636
5th Largest  3418  4211  5512
```

Task B)2)

Top Genres from High Rated Movies (40% Split):

```
Drama        1639
Comedy       1212
Thriller      629
Romance       628
Action        513
Adventure     405
Crime         373
Sci-Fi        275
Children      258
Horror        229
```

Top Genres from High Rated Movies (60% Split):

```
Drama        3888
Comedy       2764
Romance      1352
Thriller     1297
Action       1074
Crime         879
Adventure     789
Horror        692
Sci-Fi        582
Children      433

----------------------------------------------------------------------
Top Genres from High Rated Movies (80% Split):
Drama        5242
Comedy       3509
Romance      1770
Thriller     1746
Action       1432
Crime        1198
Adventure    1044
Horror        929
Sci-Fi        759
Fantasy       580
```

Task C

Observation 1:

The ALS model metrics improve as the training data increases from 40% to 60% but slightly decline from 60% to 80%.
This pattern likely arises because while more training data generally provides more learning opportunities, excessively large data might include noise or less relevant information which could detract from the model's ability to generalize well.
This finding is significant for Netflix since it indicates the ideal quantity of training data to balance generalization and learning. Finding the ideal data size for a recommendation system aids in its optimization, making it more reliable and accurate while also enhancing user experience by recommending more pertinent information.

Observation 2:

When more training data is utilized, user clusters get bigger, with the biggest clusters growing at the fastest rate.

This expansion may be explained by the fact that as more data becomes available, the model gets more adept at capturing and organizing increasingly unique user preferences, enabling more precise and detailed grouping.

Netflix may make judgments about user segmentation techniques and content customization by taking into account how user clusters grow as more data becomes available. By more closely aligning with viewer preferences, it allows for more focused marketing and content suggestion tactics that can improve user happiness and engagement.