Deep Learning

Evgenios Vlachos-2499
Vasileios Theiou-2685
Panagiotis Kounis-2540

17 June 2022

1 Methodologies

Firstly, we started our project by transforming the timestamp column into real time data. So, with the use of the given timestamp we were able to calculate the date and observe that the dataset contains data for 7 months. Hence, in order to find the distribution over time that were asked, we grouped the data by month and then by taking into consideration the customer id,country, city parameters we took the histogram of the the values count of the engagement level parameter. Same methodology was, also, used for the second task with quality of experience instead of engagement level. For finding the differences in the 2 distributions mentioned above, we performed the Kolmogorov-Smirnov test. This test indicates that if the p_value is greater than 0.05 we reject the null hypothsesis and the 2 distributions are equal. The engagement level duration was found by grouping data based on the viewer and then on the event. After that, we can take the first and last timestamp and calculate the duration that the viewer was in the event. Afterwards, we needed to iterate in each timestamp and subtract the i_+ 1 th for the i and multiple the result by the engagement. Lastly, we divided the result with the duration and found the engagement level duration. As for the correlations, we used Spearman's rank correlation because, it is not related with the distribution of the data and it is useful when the variables are measured on a scale that is at least ordinal. We dropped some columns that we believe that was not appropriate for our analysis: timestamp,date,Month,day_of_the event, because the do not contribute to our final prediction

2 PART I

2.1 Question 1

We decided to extract the above figure by depicting seven plots in the same figure in order to compare all the distributions of each month.We grouped by country id taking the engagement column from which we took the above figure with the seven histograms. We can see that the large amount of the records is observed with an engagement of 0.0-0.2 and from 0.9-1.0. Also, we can observe that these frequent values are found for the 9th month with the magenta color.

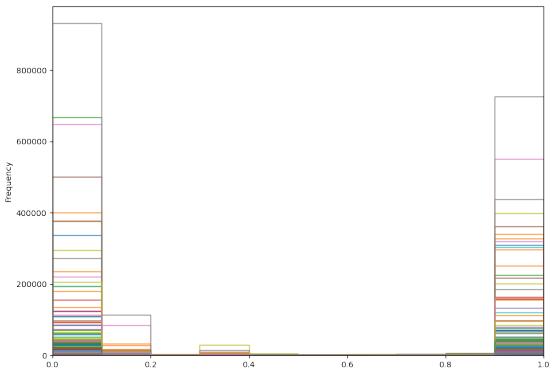

Figure 1: Engagement distribution over time based on country id

For the next two figures we concluded that they follow the same pattern as the first figure.

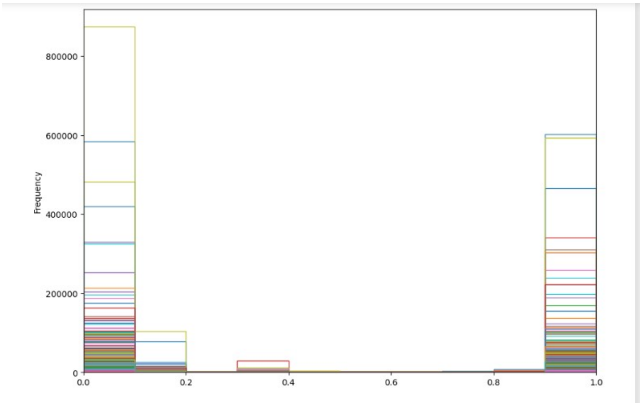

Figure 2: Engagement distribution over time based on customer id

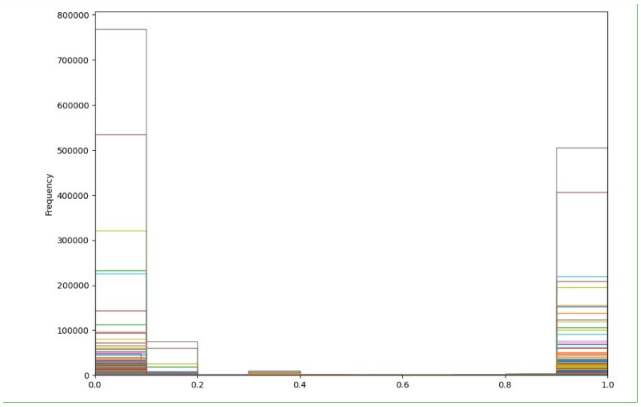

Figure 3: Engagement distribution over time based on city

2.2 Question 2

We decided to extract the above figure by depicting seven plots in the same figure in order to compare all the distributions of each month.We grouped by city id taking the qoe column from which we took the above figure with the seven histograms. We can see that the large amount of the records is observed with a qoe of 0.9-1.0.For the next two figures we concluded that they follow the same pattern as the first figure.

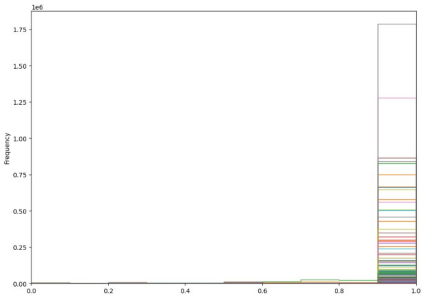

Figure 4: Viewers' QoE distribution over time based on country$_id$

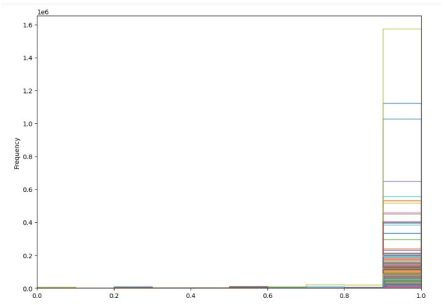

Figure 5: Quality of experience distribution over time based on customer id

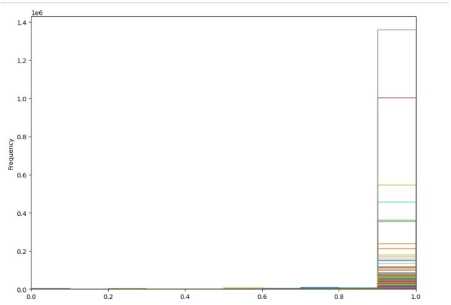

Figure 6: Quality of experience distribution over time based on city id

2.3 Question 3

In order to find if the two distributions differ from each other we performed the Kolmogorov-Smirnov test. This statistical test can understand if a given sample comes from a specific distribution and the chance that two distributions come from the same distribution. if p_value is greater than 0.05 we reject the null hypothsesis and the 2 distributions are equal. We performed this test in order to find the differences of the two distributions. We observe that the two distributions are equal. That makes sense because we believe that when someone has high engagement level it is highly likely that the viewer has at the same time high quality of experience and backwards. From the above figure, we observe that in many intervals the two distributions fall on top of each other.

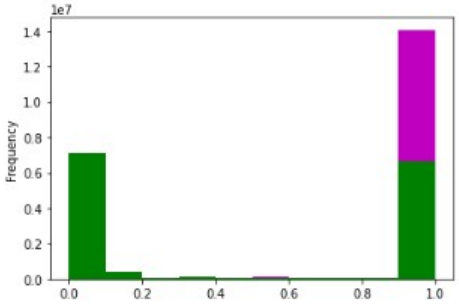

Figure 7: qoe and engagement different distributions

2.4 Question 4

First we grouped by viewer id and event id and sorted the records by date. We multiplied each engagement row of a specific viewer with the time that spent in the specific event. We repeated this process for all the viewers for all the events. And finally, we devided this sum with the total time that the viewer spent in the event. This value gave us the level of engagement duration for a specific viewer in a specific event. For extracting the figure, we followed the same process as before by taking a histogram for the engagement level duration column grouping by country id, city id, viewer type. In the last two figures first of all we observe many colors due to the fact that there are many categories. Moreover, we can understand that the large amount of the records are between 0 and 0.1 and from 0.8 to 1.0. So, we conclude that the majority of users in each city has a low engagement level duration. For the first figure we conclude that the engagement level duration is higher for viewers in the office than the users in their house due to the higher quality of internet.

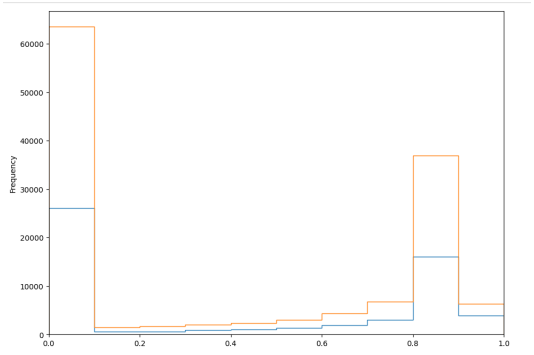

Figure 8: engagement level duration over viewer type

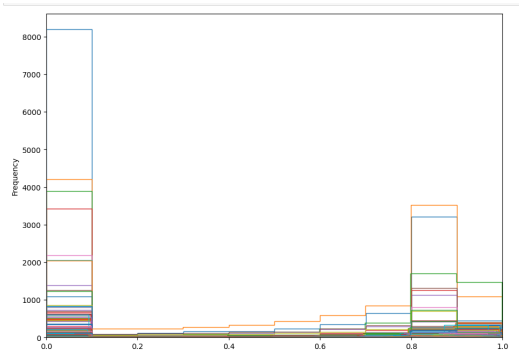

Figure 9: engagement level duration over city

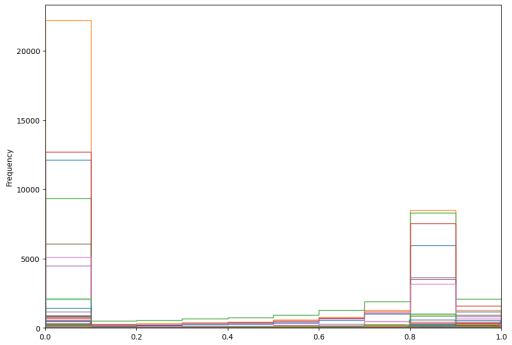

Figure 10: engagement level duration over country

2.5 Question 5

We again applied the Kolmogorov-Smirnov test which indicated that the two distributions are not equal. We can observe in the diagram plot that the two distributions differ because the distribution based on the country id and customer id has high standard deviation and low mean value, while on the other hand the distribution based on the city id and customer id has very low standard deviation and very high mean value.

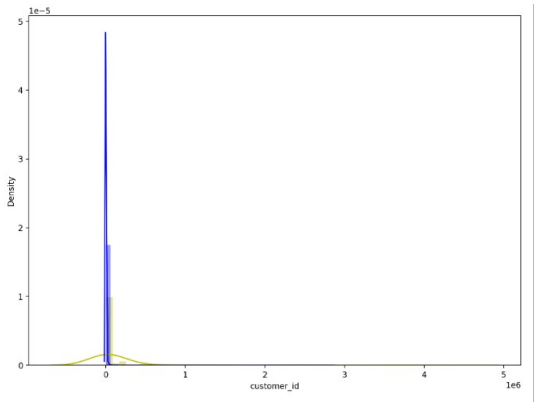

Figure 11: Countries-Cities that follow different distributions per customer

2.6 Question 6

Correlation is a metric to evaluate the strength of association between two features and the direction of their relationship. A correlation close to 1 indicates strong positive relationship between the two variables and as one increases, the other also increases at the pace of the other. On the other hand, correlation close to -1 indicates strong negative relationship and as the one variable increases, the other decreases at the pace of the other. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. And finally, a correlation very close to zero indicates no relationship between the variables. In this part of our analysis we had to find the correlations between the data points of section 2. There were 3 possible ways to find those correlations using Pearson's, Kendall's, Spearman's approaches. Each approach corresponds to specific reasons that are related with the structure of the data points. Pearson's approach measures the relationship between linearly related variables which are normally distributed. Furthermore, kendall's rank correlation, is a metric that measures the strength of dependency between two features and also, Spearman's rank correlation, measures the degree of association between two variables. In our case the most appropriate correlation matrix to understand our data is the Spearman's rank correlation because, it does not carry any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal. These methods were applied only for measuring correlations between numerical features. For measuring the correlation between numerical and categorical and categorical with categorical features we applied two different techniques. For the first case, we applied ANOVA test in 95% cnfidence interval to see if the mean of our numeric variable changes with different values of the categorical variable. That doesn't give us a correlation, but it tells us if there's a relationship, and we found that the only features that are not correlated was customer_id and buffer_ms, viewer_type and buffer_ms. For the second case we performed chi squared test

in 95% confidence interval. If we assume that two variables are independent, then the values of the contingency table for these variables should be distributed uniformly. And then we check how far away from uniform the actual values are. We found that all variables are correlated.

Correlations between numerical and categorical features: all correlated except from: customer id and buffer_ms with P-Value for Anova: 0.396 and viewer_type and buffer_ms with P-Value for Anova: 0.129

	qoe	engagement	buffer_ms
qoe	1.000000	0.007316	-0.581653
engagement	0.007316	1.000000	-0.002949
buffer_ms	-0.581653	-0.002949	1.000000

Figure 12: spearman's correlation for numerical features

From the below diagram we can understand that there is not high correlation between the variables. The only correlation that is noteworthy is the negative correlation between buffer ms and qoe. That makes sense because the quality of experience of a viewer depends on the quality of the internet. So as the buffer ms increases(low quality of internet) the qoe decreases and as the buffer ms decreases the qoe increases.

2.7 Question 7

i)Number of viewers during the event: We calculated how many viewers attended each event and added on column in our dataset which indicates that calculation. ii) Day of the event: We turned the timestamp column to weekday and added the column on our dataset. iii) Duration of the event: We sorted the dates of each event and we took the last record minus the first in order to find the duration of the event. iv) Countries: It is a column of our dataset. v) Viewers' retention: We grouped by each event and by each viewer id in order to found all the records first and last row for the date column to subtract them. In that way we can find how much time did each user participated in every event.

To find the correlation of the engagement column with all these 5 columns, we performed spearman's rank correlation for the numerical variables, and ANOVA test with 95% confidence interval to find the correlation between engagement and the categorical variable Country and Day of the event. We found that both categorical variables are correlated with the engagement variable. From the above correlation matrix we do not observe any strong correlation between engagement and the other features.

	engagement	event_viewers	event_duration	retention_time
engagement	1.000000	-0.018376	0.110944	0.060266
event_viewers	-0.018376	1.000000	0.474774	0.201270
event_duration	0.110944	0.474774	1.000000	0.438897
retention_time	0.060266	0.201270	0.438897	1.000000

Figure 13: Correlations between viewer engagement and the following factors: i) Number of viewers during the event, ii) Day of the event, iii) Duration of the event, iv) Countries, v) Viewers' retention (how much time did each viewer participated in the event)

3 PART III

3.1 Data selection

Firstly, we need to take into account the correlations of the features. After performing the anova and the chi square tests and observing the spearman's correlation matrix we can see that there are not any highly correlated features that need to be dropped. So, it was up to our critical thinking and trial and error to find out which features should be selected as input to the models. Thus, we concluded that the input of the model should consist of 3 features, which are relates to the network's status. These features are: engagement, quality of experience and buffer ms. The 2 first contain information from the viewer's side, while buffer ms shows the duration of a player's buffering to play the next video fragment in milliseconds. Features like country id, viewer type etc were used for statistical information, about which we will discuss later.

3.2 Model Architecture

The goal of this project is to detect network anomalies. Our perspective of the project is that since there were not mentioned any specific network anomalies(targets), we need unsupervised methods and therefore, we considered the problem as One Class classification. Anomalies are a lot less frequent than the normal status and so, we needed a model that could detect which points of the whole data stand out of the others, namely the outliers. Hence, we used the Isolation Forest algorithm, which gives us the opportunity of outliers selection.

3.3 Loss Function

The Isolation Forest algorithm is not trained based on a loss function, but instead uses a dedicated algorithm just like decision trees.

3.4 Data Training

Due to the fact that anomalies are very few compared to the normal data and that we do not have a specific target, we decided to train our model in the whole dataset. If we had a specific anomaly we could train the model on the normal data only, but we believe that the model will learn the patterns of normal performance because of their frequence and will be able to detect outliers. More specifically, we split the data into 80% for train and 20% for test because on 100% we got memory error. After the training was finished, we use the model to make predictions on both training and test datasets. The predictions outcome is -1 for anomalies and 1 for regular data. Also, we calculate the scores of the decision function used for predictions. These scores value from -1 to 0 and indicate how sure the model was for the prediction it made. We then plotted a histogram of these scores and based on that we decided to set a threshold on -0.02. We can observe that most of the predictions with a lower score than -0.02(stronger predictions) follow some patterns, like having high buffer ms values and median or small qoe and engagement. So, we can define the anomaly based on the above observation and now have labeled data to evaluate the model.

For the parameters of the model, we tried different values and through trial and error we concluded that the best parameters are the default ones except the contamination. Contamination indicates the percentage of outliers in the data. This, we set this parameter to 0.01.

3.5 Evaluation

We passed again the whole dataset for prediction in order to detect outliers. Since we now have labeled data we are able to evaluate the model based on the classification report. More specifically, our idea is to capture all the anomalous point in the system. So its better to identify few points which might be normal as anomalous(false positives) ,but not to miss out catching an anomaly(true negative). Thus, we are interested in having high recall and low false negative rate. We can see form the classification report bellow that the recall is 100%. This fact indicates that our model caught almost all the negative cases. Specificity is

```
array([[   90476,       82],
       [  115503, 20409309]], dtype=int64)
```

Figure 14: Confusion matrix

the proportion of true negatives that are correctly predicted by the model and is measured as tn / (tn+fp). Hence, we can now calculate specificity which is 43.92%. This show that from all the negative predictions the 43.92% were true negative. For further evaluation, we also checked the false negative rate which was 0.0004%.

3.6 Difficulties and different approaches

One of the first problems in our analysis was there were not an predefined anomalies to search for so we had to train the whole dataset and based on its predictions and our critical thinking to define the anomalies. Furthermore, throughout our model analysis we tried many different models to test for outliers/anomalies. One of the first problems we faced was that due to the large amount of records in the dataset some of these algorithms were not able to perform, like One-Class-SVM. We, then, tried different models such as Local Outlier Factor, Elliptic Envelope, but Isolation Forest turned out to have the best recall and false negative rate.

3.7 Main Observations

Firstly, we needed to define which records are anomalies. This was done after training the model and observing the results. Since the most of the predicted outlier consisted of high buffer ms values and low or medium quality of experience and engagement values, we concluded that most anomalies have buffer ¿ 200, qoe ¡ 0.6 and enagement¡0.6. Also, with the algorithm results we were able to perform some statistical information to explain our results. Hence, we made some bar plots to check which countries, cities , events, customers, viewers and viewer type contained the most anomalies. The results of the top 20 features with most anomalies are shown on the pictures bellow:

country_id	anomalies_count
1.0	22596
0.0	14293
4.0	7135
2.0	7046
3.0	6614
12.0	4332
25.0	4016
7.0	3875
13.0	1956
5.0	1901
18.0	1842
6.0	1422
8.0	1365
16.0	1259
23.0	981
17.0	880
14.0	836
24.0	833
9.0	743
10.0	553

Figure 15: Countries Anomalies

event_id	anomalies_count
12.0	7562
6.0	6819
18.0	3595
29.0	3161
2.0	2993
0.0	2257
10.0	1704
39.0	1574
3.0	1510
158.0	1472
44.0	1180
82.0	1063
182.0	1054
156.0	1052
201.0	1028
11.0	923
66.0	895
1.0	892
42.0	817
163.0	814

Figure 16: Event Anomalies

customer_id	anomalies_count
4.0	13391
6.0	11455
1.0	8443
0.0	6444
2.0	6350
7.0	5219
15.0	4443
14.0	4360
8.0	3845
17.0	3528
16.0	3517
21.0	2831
5.0	2298
9.0	1898
12.0	1880
3.0	1567
19.0	1546
22.0	1426
11.0	1059
18.0	805

Figure 17: Customer Anomalies

viewer_type	anomalies_count
WFO	64512
WFH	26046

Figure 18: Viewer type Anomalies

viewer_id	anomalies_count
51734.0	145
55689.0	135
5485.0	130
77645.0	116
285.0	110
30804.0	93
40803.0	90
14403.0	87
75760.0	86
67862.0	85
13616.0	85
6222.0	84
49785.0	82
3416.0	82
11981.0	81
83114.0	81
26704.0	74
15942.0	74
24232.0	72
84619.0	70

Figure 19: Viewer type Anomalies

Last but not least, we can see that from the top 20 viewers with anomalies, they are all from countries contained in the list with top 20 countries with most anomalies. The above observation is helpful as we can conclude that in these areas there might be connection problems and slower internet speed. On a smaller scale, the same conclusion can be applied for cities. Furthermore, most people with network anomalies are working from office which is probably happening due to the fact that there are a lot of people connected in the same network. Also, there are few matching cases such as the above in customers top 20. These observations can help customers find out about anomalies on their events which might be due to buffer ms, engagement or quality of experience, so that they improve the content of their events to motivate viewers to participate more. Lastly, we can see that most of the customers that contain anomalies in their events are based on some of the top 20 countries with network anomalies, which might cause the problems.