**ALY6040 90248 Data Mining Applications SEC 01 Summer 2023 CPS [BOS-D-HY]**

**Module 5 Assignment — Individual Quiz**

A black and white logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by**

Zihan Ma [ma.zihan1@northeastern.edu](mailto:ma.zihan1@northeastern.edu)

**Instructor**

 Prof. Kasun Samarasinghe

**Date**

08/10/2023

**Individual Quiz**

**Q1. Loss function behavior (40 points)**

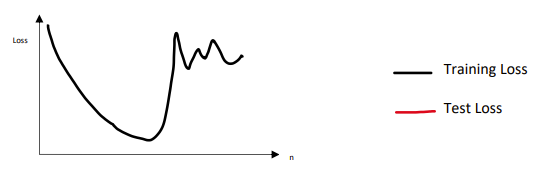
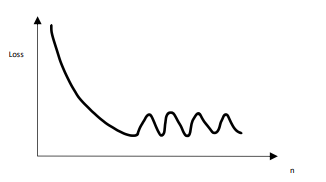
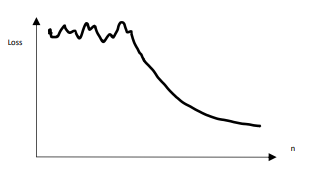
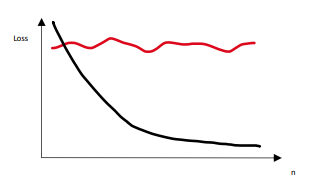
Following are different instances of how a loss function could behave during training and

validation for an ML model.

The x-axis represents the training iterations, and the y-axis represents the training/ test loss

function.

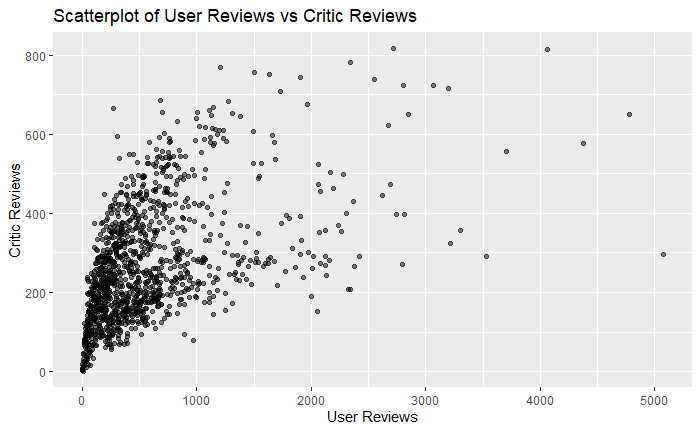
For each instance, please explain 1 reason for this behavior and 1 method to fix it.

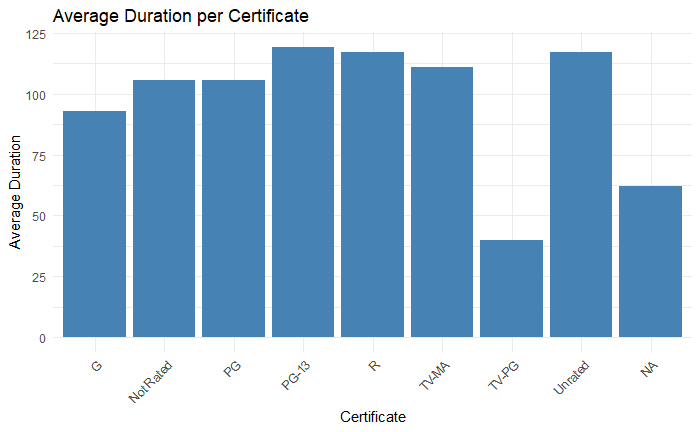
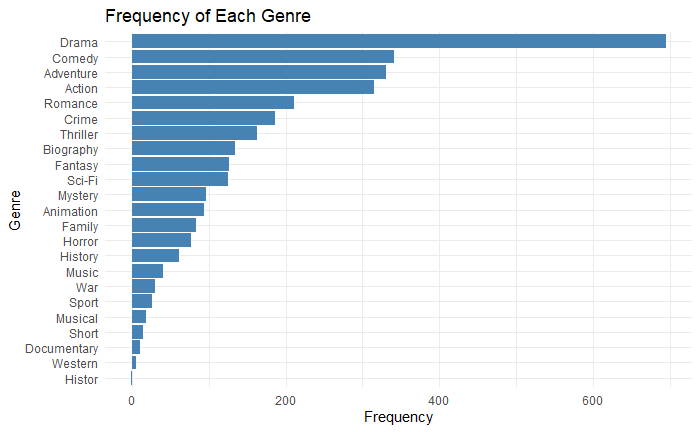
1.   
   The problem could be about the data. In the original data or your data engineering process, some data didn’t normalize correctly or accidentally got deleted would cause this. Try adding a step between suspicious steps to replace 0 or nan data with the average number if there is one. Then remove half of them until the graph changes back.
2.   
   It seems the model refuses to converge. Maybe there are too many restrictions being added to the model. It’s better to start over or reduce the complexity and learning rate until it goes back to normal.
3.   
   It seems the mistake happens on the data being pushed into the model. Consider regrouping the data to flatten the curve. The data should have a consistent pattern for machine learning.
4.   
   It seems the activation function it’s not working efficiently, or the learn is too low for the model to start learning. Try increase the learning rate, then try using other activation function.
5.   
   The model is overfitting. It should be because the data used for training is so little that cause the model learned nothing from it, or there are way too many layers in the network. If it’s not the case, the testing loss should at least move a little at the start. Use a dataset with more data to test the model. If it still not solve, simplify the model and test again.

**Q2. Predicting the Oscars (60 points)**Description:

Please download the Oscar\_2000\_2018.csv dataset provided. This dataset amounts to a total of 1,235 movies from 2000 to 2018, where each film has 100+ features including: It sports 20 categorical, 56 numeric, 42 items, and 1 DateTime field totaling 119 fields giving you plenty of details about various aspects of the past nominees and winners. The dataset is organized such that each record represents a unique movie identified by the field movie\_id. The first 17 fields have to do with the metadata associated with each movie e.g., release\_date, genre, synopsis, duration, metascore.

Tasks: Part 1: EDA

1. Using a scatterplot or a pair plot show the relationship between features “user\_reviews” and “critic\_reviews”. Find the Pearson's correlation coefficient(r) between the 2 features.   
   
2. Plot the average “duration” per “certificate” feature. In other words, x-axis would be “certificate” and the y-axes would be the average duration.   
   A screenshot of a computer code

   Description automatically generated
3. Plot a histogram for the “genre” feature. Note that the field “genre” needs to be split first to find the frequency for each individual genre type; “Comedy”, “Romance”, “Action” etc. (Hint: Functions like “strsplit” in R or “split” in Python can be used)   
     
   

Part 2: Model Building

Call:

rpart(formula = Oscar\_Best\_Picture\_won ~ rate + awards\_wins +

Golden\_Globes\_won, data = train\_data)

n= 1183

CP nsplit rel error xerror xstd

1 0.24408610 0 1.0000000 1.0021883 0.2307674

2 0.05976529 1 0.7559139 0.9273356 0.1928155

3 0.05626739 2 0.6961486 0.9425111 0.1802626

4 0.01485802 3 0.6398812 0.9805908 0.1823363

5 0.01000000 4 0.6250232 0.9611623 0.1815757

Variable importance

awards\_wins Golden\_Globes\_won rate

76 16 8

Node number 1: 1183 observations, complexity param=0.2440861

mean=0.01521555, MSE=0.01498404

left son=2 (1161 obs) right son=3 (22 obs)

Primary splits:

awards\_wins < 25.5 to the left, improve=0.24408610, (0 missing)

Golden\_Globes\_won < 1.5 to the left, improve=0.11826430, (0 missing)

rate < 7.95 to the left, improve=0.05043795, (0 missing)

Surrogate splits:

Golden\_Globes\_won < 3.5 to the left, agree=0.984, adj=0.136, (0 split)

rate < 8.75 to the left, agree=0.983, adj=0.091, (0 split)

Node number 2: 1161 observations, complexity param=0.05976529

mean=0.006890612, MSE=0.006843131

left son=4 (1117 obs) right son=5 (44 obs)

Primary splits:

awards\_wins < 14.5 to the left, improve=0.13334470, (0 missing)

Golden\_Globes\_won < 1.5 to the left, improve=0.07273917, (0 missing)

rate < 8.05 to the left, improve=0.03248562, (0 missing)

Surrogate splits:

Golden\_Globes\_won < 2.5 to the left, agree=0.972, adj=0.250, (0 split)

rate < 8.65 to the left, agree=0.964, adj=0.045, (0 split)

Node number 3: 22 observations, complexity param=0.05626739

mean=0.4545455, MSE=0.2479339

left son=6 (7 obs) right son=7 (15 obs)

Primary splits:

awards\_wins < 38 to the right, improve=0.18285710, (0 missing)

rate < 7.95 to the left, improve=0.04102564, (0 missing)

Golden\_Globes\_won < 1.5 to the left, improve=0.01000000, (0 missing)

Surrogate splits:

Golden\_Globes\_won < 2.5 to the right, agree=0.818, adj=0.429, (0 split)

Node number 4: 1117 observations

mean=0.0008952551, MSE=0.0008944537

Node number 5: 44 observations, complexity param=0.01485802

mean=0.1590909, MSE=0.133781

left son=10 (29 obs) right son=11 (15 obs)

Primary splits:

rate < 8.05 to the left, improve=0.04474327, (0 missing)

awards\_wins < 19.5 to the right, improve=0.03985108, (0 missing)

Golden\_Globes\_won < 1.5 to the left, improve=0.01042471, (0 missing)

Surrogate splits:

awards\_wins < 24.5 to the left, agree=0.727, adj=0.200, (0 split)

Golden\_Globes\_won < 3.5 to the left, agree=0.705, adj=0.133, (0 split)

Node number 6: 7 observations

mean=0.1428571, MSE=0.122449

Node number 7: 15 observations

mean=0.6, MSE=0.24

Node number 10: 29 observations

mean=0.1034483, MSE=0.09274673

Node number 11: 15 observations

mean=0.2666667, MSE=0.1955556  
  
A screenshot of a computer screen

Description automatically generated