Quiz – Week 5 (5.50 – 8.35PM)

Q1. Give brief explanations for the questions below **(30 points)**

1. Importance of activation functions in neural networks and state 3 commonly used such activation functions
2. Explain “Bagging” ensemble method (Eg: used in Random Forests)

Q2. Predicting the Oscars **(70 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.
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

1. You are going to predict “Oscar_Best_Picture_won” feature; this will be your target variable. Remove all of the features which has the convention “Oscar_Best_XXX_won” except for the target variable “Oscar_Best_Picture_won”.

2. Convert the target variable’s type to a numerical type by doing the transformation, “Yes” = 1, “No” = 0.
3. Remove columns with high cardinality, i.e., for every column that has a unique value frequency of 70% or higher, remove them from the dataset.
4. Perform a time split and create a training dataset spanning the period 2000-2017 and a test dataset for the movies released in 2018 - use “year” feature for the data split
5. Create a tree-based model to predict the target “Oscar_Best_Picture_won”
6. Use the model to predict the test dataset and find the maximum predicted value

Optional: Go back to the initial dataset and find the movie in 2018 that is associated with the maximum predicted value.