Data Science – Project Proposal

"Popular IMDB Movies characteristic"

Participants:

1. Name: Avi Barhom

   ID: 203572748

   Email: avibarhom1111@gmail.com

2. Name: Elad Cohen

   ID: 208931394

   Email: elad96c@gmail.com

Abstract:

"Similar to the abstract that will ultimately appear in your paper. It should be one paragraph long, for now perhaps only 5-10 lines. It should provide a high level summary of your project and outline your maingoals."

In this project, we will analyze the data of IMDB (The Internet Movie Database) -

Website which provides information related to Movies, TV-shows, and Video games. In our research we will focus at the information related to Films (Worldwide) .

Due to the amount of data gathered over the years, we are hoping to find significant insights related to movies. Our main research goal is:

- *" What are the most important attributes to predict gross income? "*

In addition, we found those topics interesting as well:

- *Which year was the best for movie production?"*
- *"Do user ratings similar to critics ratings?"*
- *"How long should a succesfull movie be?"*
- *"Which genres people rate higher?"*

Brief:

Dataset:

"What data sets do you plan to use? If you must do significant work to get the data or convert it into the proper format, then describe the process and approximate effortrequired."

The data we need for this project is found at the IMDB website.

Among the data in the website we can find for each movie: Film title, budget, gross income, release year, genre, ratings, critics, movie length, movie summary, actors, directors, and more…

In particular, The films we would gather information about are:

Movies with many ratings (1990 – Present) & Movies with high rating – Those can be filtered using IMDB search which allows sorting the highest rated or most voted movies first (for each year).

Unfortunately, IMDB doesn't support an official API.

Therefore, we will use Python with the BeautifulSoup Library to scalp information from the website.

To scrap IMDB, we will make a list of relevant movie pages, and then identify the related tags in the HTML code for the info we wish to extract.

Learning Tools:

"What learning tools do you plan to use (e.g., scikit, statsmodels, weka, other) and what methods (e.g., clustering, classification, regression, what methods in each,etc.)?"

The learning tool we chose to use is "scikit".

In specific, we will use (Linear) Regression for gross prediction,(RFC, SVM, NN) Classification for gross groups prediction, and (K-Means) Clustering for natural clusters of Top rated movies of all time.

Problem Analysis:

"How do you formulate the problem as a data scienceproblem (e.g., is it classification, association rule mining, etc.)? What exactly are you trying to predict (for prediction tasks) and how will you evaluate your results. How will you know if your results are good? What can you compare them to? It is critical that your problem iswell-defined."

For our problem we would need to analyze our Movies' rating data, Gross income data – and the relations between them and the other parameters on the dataset.

To recall, our main question was: *"What are the most important attributes to predict gross income?"*

Therefore, we'll use ML models to study the different attributes effect (such as budget, release year, genre, ratings, critics, movie length etc..) on the gross income parameter. By doing so, we hope to find the weights of the movie industry different factors on earnings.

As for the other topics mentioned earlier:

- *"Which year was the best for movie production?"*
- *"Do user rating similar to critics ratings?"*
- *"How long should a succesfull movie be?"*
- *"Which genres people rate higher?"*

We will use different plots and correlation matrixes in order to understand the different relations between each parameter (Year/Duration/Genre VS Rating & Earnings, User ratings VS Critics Ratings).

In order to evaluate the results, we will divide the data into 2 sections and use one of them for learning

And the other for evaluation.

The model will learn from different attributes each time so that we can see what attributes make our model more accurate.