**Movie Revenue Predictor**

Naveen T 1601CS55

Milan Jolly 1601CS25

**Overview:**

When we think about the realms of Science and Technology, it's hard to think of a domain that will change the world as much as machine learning and artificial intelligence will. I am excited that machine learning will be used as a tool, just as mathematics and also it will be an integral part in all fields of research that collect and analyze a large amount of data.

We wanted to contribute to this age of evolution, which merges the technology with human intelligence.To achieve our goal, a sound background in the field of machine learning is crucial.

There is a large amount of data related to the movies is available over the internet, because of that much data available, it is an interesting data set, which can be explored to understand and apply the theoretical knowledge of machine learning.

So the most interesting and complex machine learning problem is, predicting the revenue of the movie based on different features available on the dataset.

It’s so exciting for us to try our hands on different learning algorithms that we are to use in our project.

**Objective:**

According to a study, the movie industry in the United States generates revenue up to 10 billion dollars.But there is a great deal of uncertainty that the movie will be a hit or a flop. If we can predict the revenue of the movie based on the success rate of the crew and cast of the movie, it would save around millions for the production companies.

The questions that motivated us to do the project:

Can we predict the revenue of the movie before it is released?

Did the production companies find a successful formula to invest their money?

Which genre is more popular?

Can we classify the movie as animated or not using the job title of the crew members?

We will be mainly focussing on predicting the revenue of the movie.

**Current Status:**

There are a plethora of ways to implement the predictor relying on the nature of the dataset, the features under consideration and the model used to learn from the data.

To name a few which have already been realized:

1. Using linear regression model and a dataset which contains the following features: Name, Year, Date, Genre, Wins, Awards, Nominations, Budget, International/National. The accuracy obtained with this implementation is roughly 68%.
2. Using Deep Neural Networks Model and IMDb dataset (Name, rating, genre, budget, revenue), along with the movie poster, the accuracy obtained is 52%.
3. Using k-means clustering, genre separation, polynomial regression and linear regression, with the dataset being similar to the first one, the accuracy of this implementation is close to 42%.

**Novelty:**

Unlike most of the current implementations, we are giving high emphasis on the cast and crew for the movies as their performance and coordination is the thing that makes movies flourish. Also with the help of DNN, we are going to also give importance to the fact that what kind of combination of cast and crew grouping leads to the betterment of the movie.

Plus in accordance with the movies with year, we are also taking into the account the inflation rate for better training of the dataset.

Along with this we are considering the impact of a particular grouping of genres and establishing a relationship between the success of genres and the month on which the movie is released. For example Its common sense for Entertainment industry that there is an overall high demand for romantic movies in the month of February due to the Valentine’s Day.

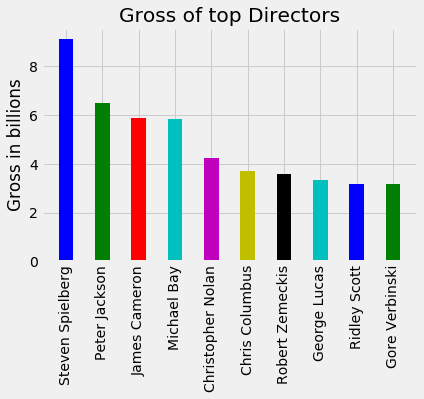
**Detailed progress so far :**

The data set for the project has been collected from kaggle. After exploring the data set, we planned to visualize the data in a better way to get started with the project. So, some data preprocessing was needed to be done. The dataset has few columns that are in JSON( javascript object notation) format, which is a lightweight data format for transferring the data. So we had to extract the required features and convert into list format which could be easily plugged into the learning algorithms.The features that we have in our dataset are the movie title, budget, genres, popularity, production companies, revenue, vote-average, vote-count, cast, and crew.

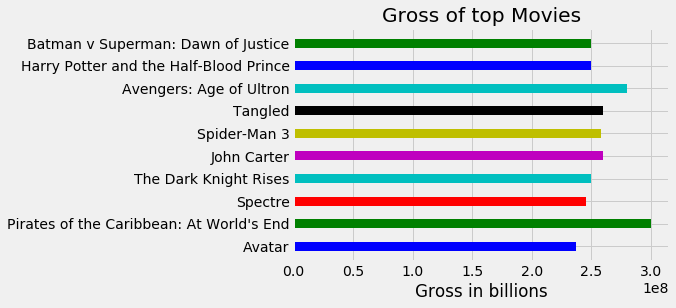
Along with the JSON format, there were problems with the data. Mainly the cast and the crew data were in string format while we needed them to be in a dictionary or list-like format. Conversion of cast and crew data to the dictionary like data was done in python with the use of pandas and numpy library.

After that to make bar charts, we used the now converted dict-like cast and crew data with the help of data present in different file related to gross and budget to plot some bar charts. But the data were distributed in 2 files so we had to do some extra data preprocessing there as well. The plots include:

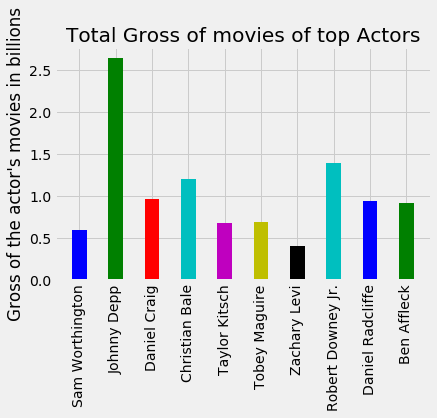
(i) The bar chart with Director on the x-axis and the height of the bars being the total gross of the movies directed by the respective Director for top 10 directors.



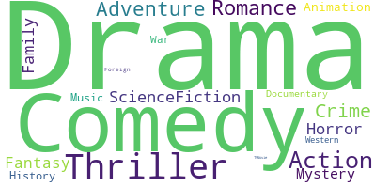
(ii) The bar chart with Movie name on the x-axis and the height of the bars being the total gross of the movie for top 10 movies.



(iii) The bar chart with main artist in the x-axis and the height of the bars being the total gross of the movies acted by the respective artist for top 10 artists.



**Word Cloud of Genre Representation :**



The above word cloud was generated from the data set which has around 4800 movies.We can observe from the above word cloud that Drama, Comedy, and Thriller are the most common genres in the movie industry.

**Plan of work with timeline :**

After the mid-semester examination, we have around 7 weeks of time to complete the remaining phase of the project, which will be to get the best model that fits the data.

So we will be fitting different models, like linear regression, SVM and then neural networks.

We will be resuming the work as soon as the mid-semester gets over.

Our Course lab dates :

1 / 3 / 2018 --- Understand the logistic regression and classification learning algorithms. And get to know the tools to use them.

8 / 3 / 2018 --- Build the first model based on logistic regression considering the movie revenue as a continuous variable. Try to improve the accuracy by selecting appropriate features.

15 / 3 / 2018 --- Then we move on to SVM ’s. Understand them better and build an SVM model and compare the accuracy that we got previously and try to improvise it.

22 / 3 / 2018 --- In this week we still continue to work with SVM and try different ensembling techniques to improve the accuracy.

29 / 3 / 2018 --- Then, we plan to use neural networks to try out more complex relations between the input features to get a better accuracy.

1 / 4 / 2018 --- This week will be dedicated to make some modifications to the previous build algorithms and improvise them.

We hope to stick to this timeline and gain more knowledge and apply them more effectively.