# Movie Revenue Predictor

By Naveen - 1601CS55

Milan - 1601CS25

## Introduction :

The revenue of a movie is dependent on various factors like cast, crew, genre, budget, rating, year and month of release etc ,. There is no analytical formula for calculating the revenue of the movie because of its dependency on various factors. But we can predict the expected revenue of the movie by analyzing the revenue generated by previous movies. This prediction will be useful for the movie studio which will be producing the movie in deciding the expenses they will be spending for the movie. This prediction will also help the movie theaters to estimate the revenue they will be generating by screening a particular movie.

## Motivation :

The movie , “King Arthur: Legend of the Sword” had a production budget of $175 million dollars and loss of $150 million dollar.

The film was bestowed the dubious honour of "largest box office loss" by Guinness World Records and lead to the demise of production company Carolco Pictures and the blockbuster future for lead actress Davis.

If it was somehow known to production companies about the expected revenue of the movie they could have adjusted the expenses so as to gain maximum profit. They could use the Revenue prediction model to know about the market beforehand. So such a software is of dire need.

# Objective :

To design a machine learning model that is useful to the movie industries that spend billions of dollars. This model will predict the expected revenue of the movie that can be analyzed to draw conclusions that can save millions of dollars.

# 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 ANN, 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.

# Current existing models :

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%.

# Data collection and preprocessing:

The dataset was collected from kaggle website which had around 4800 Hollywood movies. The features that were associated with a particular movie were,

- Genres
- Budget
- Title
- Popularity
- Production Companies
- Release Date
- Vote_average
- Vote_count
- Revenue
- Cast
- Crew

The budget and the revenue in the original dataset were not adjusted towards inflation rates. So we collected the inflation rate in US from 1956 to 2017 as the budget and revenue were on US dollars. We took the inflation rate data from the website http://www.inflation.eu.

Then the genres, cast, crew and production_companies were in JSON format, which we converted into string and then to list. The genres were encoded and each genre were added as a new column and each cell was filled with either 0 or 1 ie) 1 if that movie belongs to that genre orelse 0.

Then for the production companies , each production company were assigned some score, by summing the revenues of the movie that production company occured divided by the number of occurrences. This method was best suiting for the model as it decreased the complexity of the problem by manyfold as there were 5016 different production companies in total.

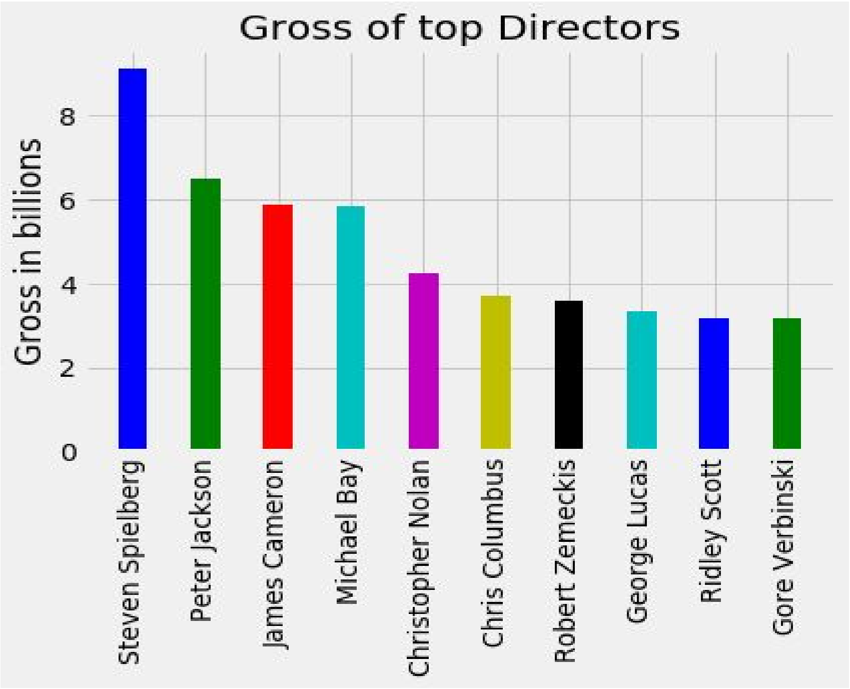
The cast in the movie dataset was arranged based on their credit score. So we took first five cast members from each movie . And assigned the a score to them based on the revenues and popularity of the movies they were in.Similarly from the crew we extracted editor, director and writer and assigned them similar scores done in cast.

From the date of release we extracted the month and year of release. Months were one-hot coded into 12 different columns.

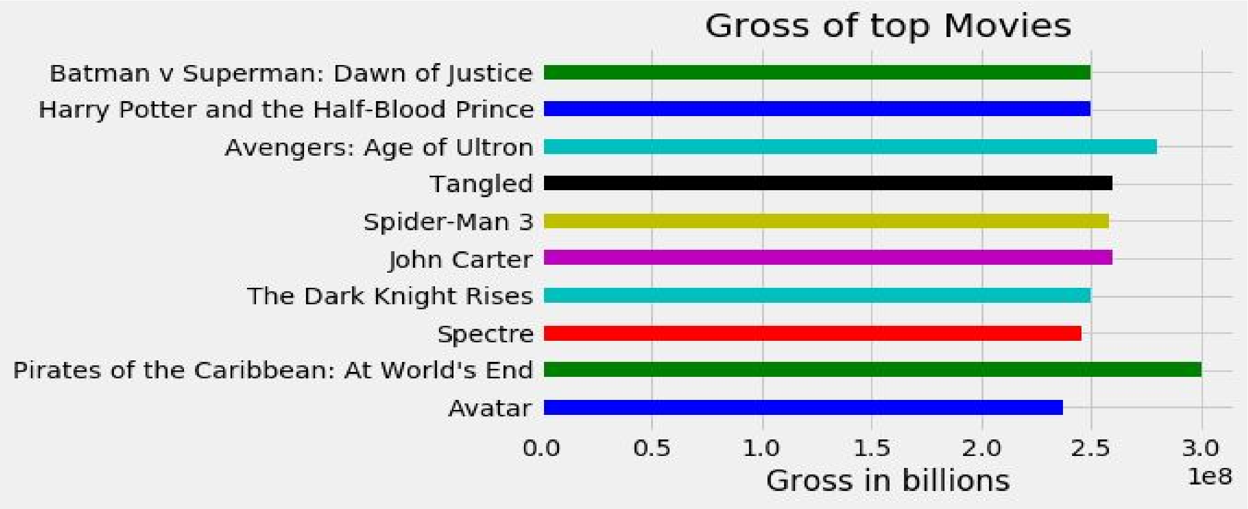
## Data Visualization:

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:

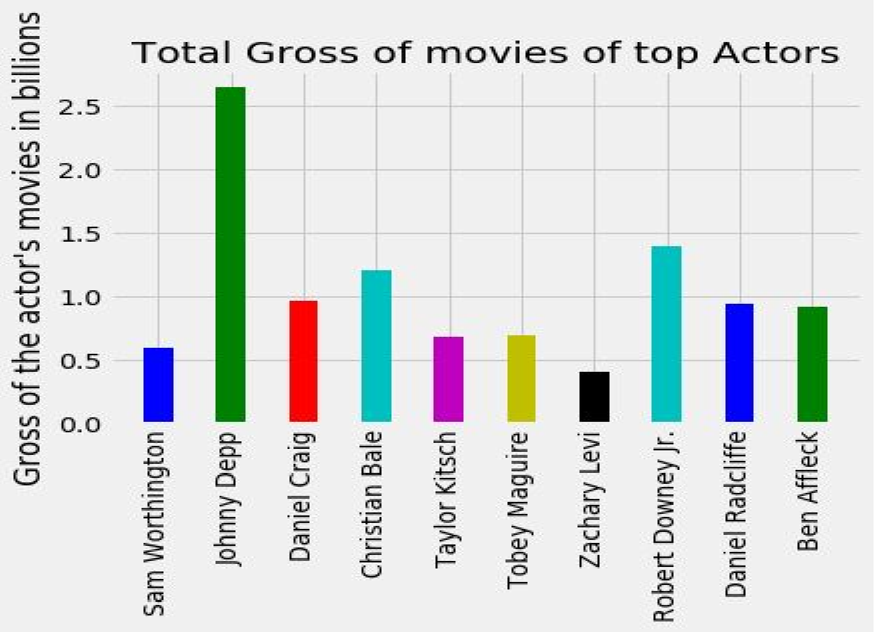
(i) The bar chart with Director on the x-axis and the height of the bars being the total gross of the movies directed by the respective Director for top 10 directors.



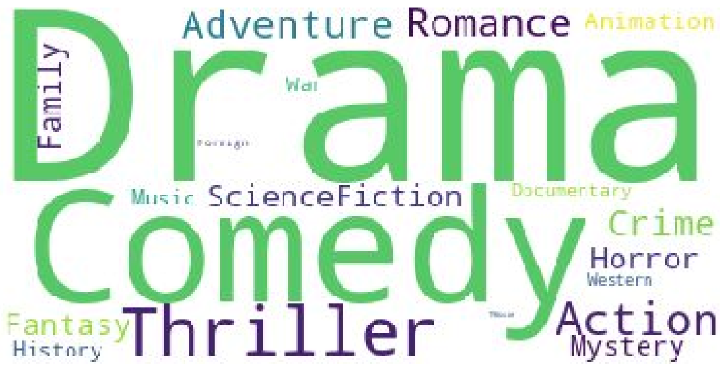
(ii) The bar chart with Movie name on the x-axis and the height of the bars being the total gross of the movie for top 10 movies.



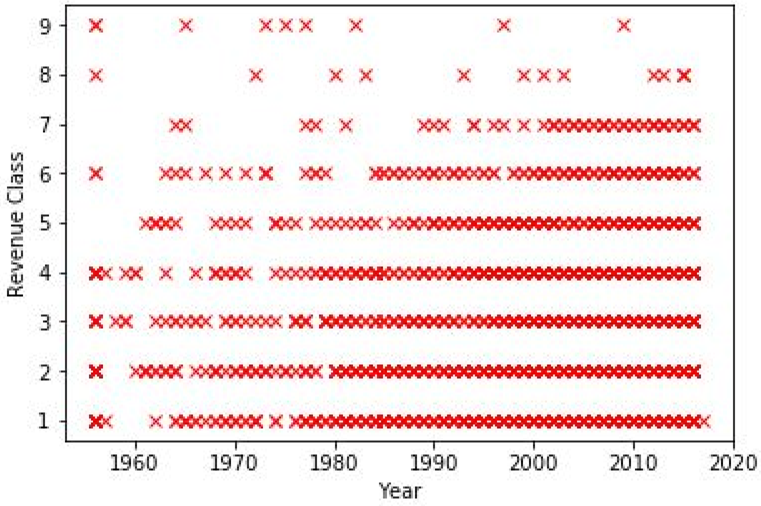
(iii) The bar chart with main artist in the x-axis and the height of the bars being the total gross of the movies acted by the respective artist for top 10 artists.



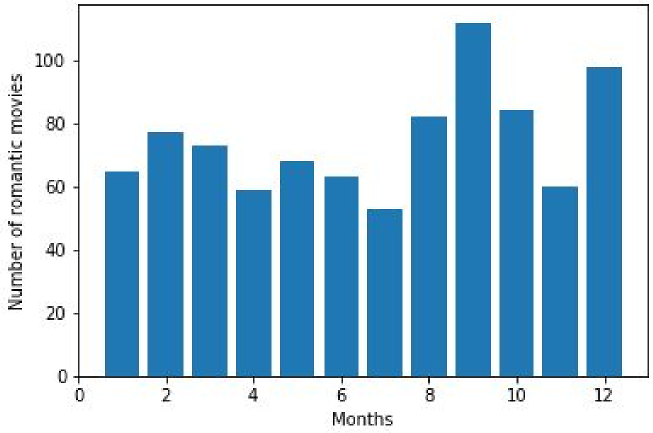
( iv ) The below word cloud was generated from the data set which has around 4800 movies.We can observe from the above word cloud that Drama, Comedy, and Thriller are the most common genres in the movie industry.



(v) The below is a plot which gives us the info and some relation about the movie revenue class and year where each ‘x’ is a movie.



(vi) The below shows us the relation between romantic movies and month of a year. Here we have used number of romantic movies as our y-axis while month of the release as x-axis.



## Discretization of the Revenues:

A movie can be classified into any one of the nine categories ranging from “flop” to a “blockbuster” based on the revenue.

| Class Number | Range in Millions |
|---|---|
| 1 | Less than 1 ( flop ) |
| 2 | Less than 10 greater than 1 |
| 3 | Less than 20 greater than 10 |
| 4 | Less than 40 greater than 20 |
| 5 | Less than 65 greater than 40 |
| 6 | Less than 100 greater than 65 |
| 7 | Less than 150 greater than 100 |
| 8 | Less than 200 greater than 150 |
| 9 | Greater than 200 ( blockbuster ) |

This range for each class is decided by comparing the success rate of the movie and referenced from a research paper written by Ramesh Shardha.

The reasons for choosing discrete values over continuous values are that,

- discrete values are closer to a knowledge-level representation (Simon, 1981)
- data can be reduced and simplified through discretization (Liu et al., 2002),
- for both users and experts, discrete features are easier to understand, use, and explain (Liu et al., 2002); and finally,
- discretization makes many learning algorithms faster and more accurate (Dougherty, Kohavi, & Sahami, 1995).

So we chose to divide the revenue into 9 classes after referring to these research papers.

Final step of the data preprocessing was to remove the unwanted features. So we used the variance as the parameter to do this. We set the threshold to be 0.8 , which means that if a particular column has a repeating number more than 80% then that column is deleted from the dataset.
Now the data is ready to be fit in various models of ML.

## Logistic Regression :

Logistic regression sometimes called the logistic model or logit model, analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve.

The logistic curve is represented as,

$$logit(\ y\ )\ =\ ln\left(\frac{p}{1-p}\right)\ =\ \alpha\ +\ \beta\chi$$

This logistic regression equation represents for a two dimensional case.
Extending this to multiple predictors,

$$logit(\ y\ )\ =\ ln\left(\frac{p}{1-p}\right)\ =\ \alpha\ +\ \eta_1\chi\ +\eta_2\chi_2 + \dots\dots\dots + \eta_n\chi_n$$

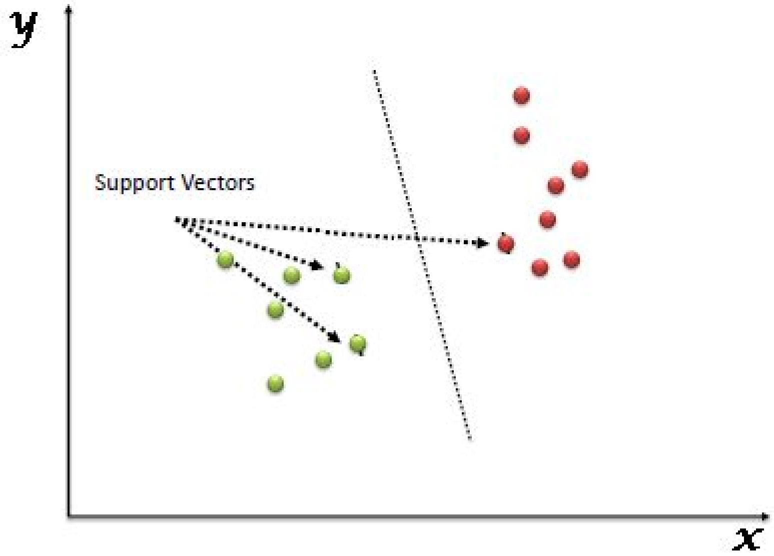
Therefore probability is determined as,

$$p\ =\ \left(\frac{1}{1+e^{-(\alpha+\eta_1\chi+\eta_2\chi_2+\dots\dots\dots+\eta_n\chi_n)}}\right)$$

In our Logistic regression model we got a test set accuracy of 61.49% and training set accuracy of 64.16%.

## Support Vector Machines( SVM ) :

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However,  it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiate the two classes very well (look at the below snapshot).



SVM uses kernel trick to handle classes which are not separable. In this SVM model we used Radial Basis Function ( rbf ) kernel as it can take care of classes which are grouped as clusters, more efficiently.

In our SVM model we got a test set accuracy of 62.85% and training set accuracy of 73.14%.

## Neural Networks :

Artificial neural networks (ANNs) are biologically inspired computer programs designed to simulate the way in which the human brain processes information. ANNs gather their knowledge by detecting the patterns and relationships in data and learn (or are trained) through experience, not from programming.
Since the revenue of the movie has many predictors the artificial neural network is the best fit as it can find complex relations among the independent variables.

In our model we used Keras library to implement the artificial neural network in more efficient way.We used three hidden layers and in total five including the input and output layer.
The activation function used in the hidden layers use ReLU function and the final output layer uses sigmoid function.
In our Neural Network model we got a test set accuracy of 73.98% and training set accuracy of 80.21%.

| S.no | Model | Test Set accuracy | Training set accuracy |
|---|---|---|---|
| 1 | Logistic regression | 61.49% | 64.16% |
| 2 | SVM | 62.85% | 73.14% |
| 3 | Neural Network | 73.98% | 80.21% |

# Conclusion :

The results show that Neural network employed has a pin point accuracy of 73.98% . Compared to the limited number of previous movie revenue predictors , the prediction accuracies are significantly better.Compared to the logistic regression and SVM models used , neural network performed significantly better.The result of this project shows the power of Neural network in solving complex prediction problems.This model can be used in the initial forecasting phase by the production companies in deciding the budget and other expenses to be spent on the movie.

# References :

1. Predicting box-office success of motion pictures with neural networks (Ramesh Sharda*, Dursun Delen)
2. Predicting Movie Revenue (Nikhil Apte, Mats Forssell, Anahita Sidhwa)