Movie Revenue Predictor

•••

By
Naveen T - 1601CS55
Milan Jolly - 1601CS25

Introduction

Cast

Month

Budget

Year

Rating

Crew

Motivation

ORIGINAL MOTION PICTURE SOUNDTRACK
king arthur
LEGEND OF THE SWORD
MUSIC BY
DANIEL PEMBERTON

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

Importance to cast and crew

Inflation rate in the economy

Current Existing Models

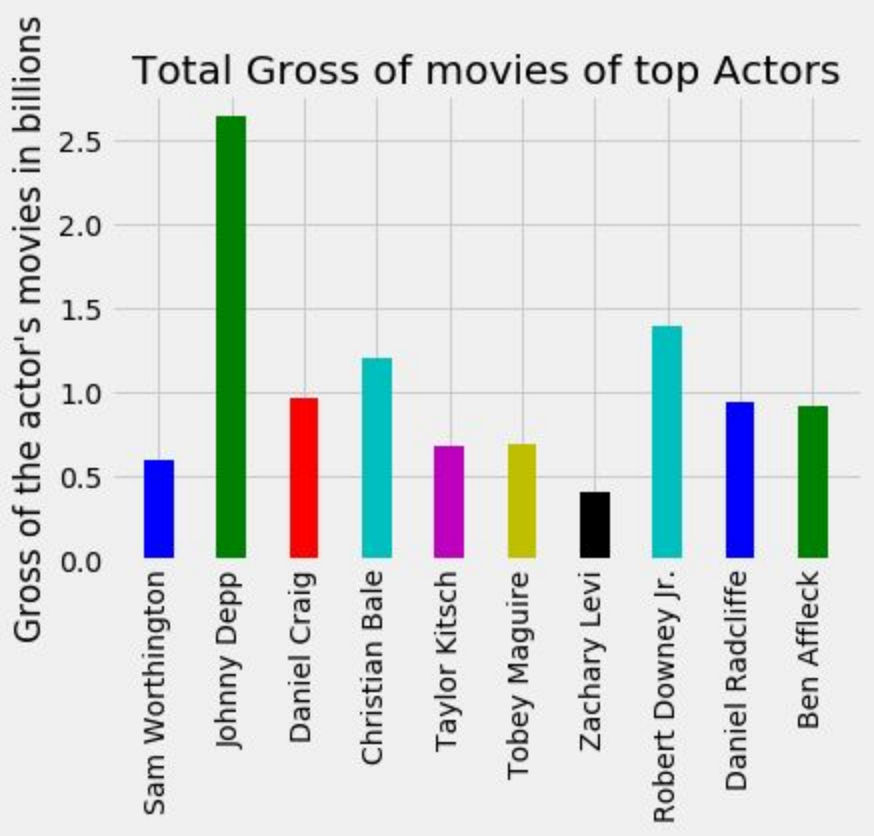
1. Using linear regression model and a dataset which contains the following features: Name, Year, Date, Genre, Wins, Awards, Nominations, Budget, International/National. The accuracy obtained with this implementation is roughly 68%.
2. Using Deep Neural Networks Model and IMDb dataset (Name, rating, 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%.

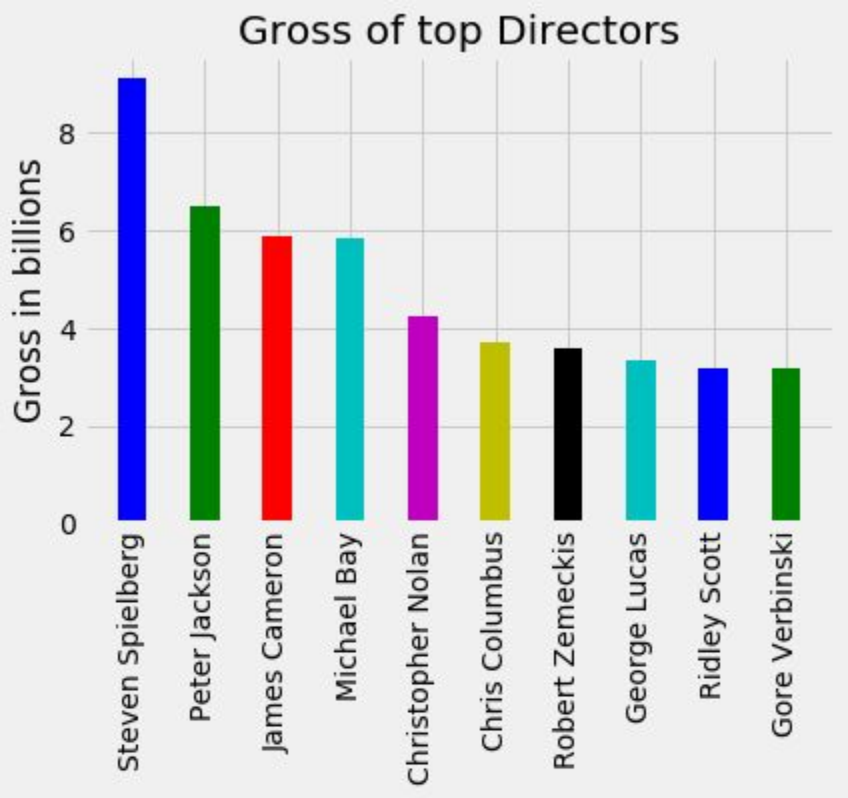
Data Collection and Data Preprocessing

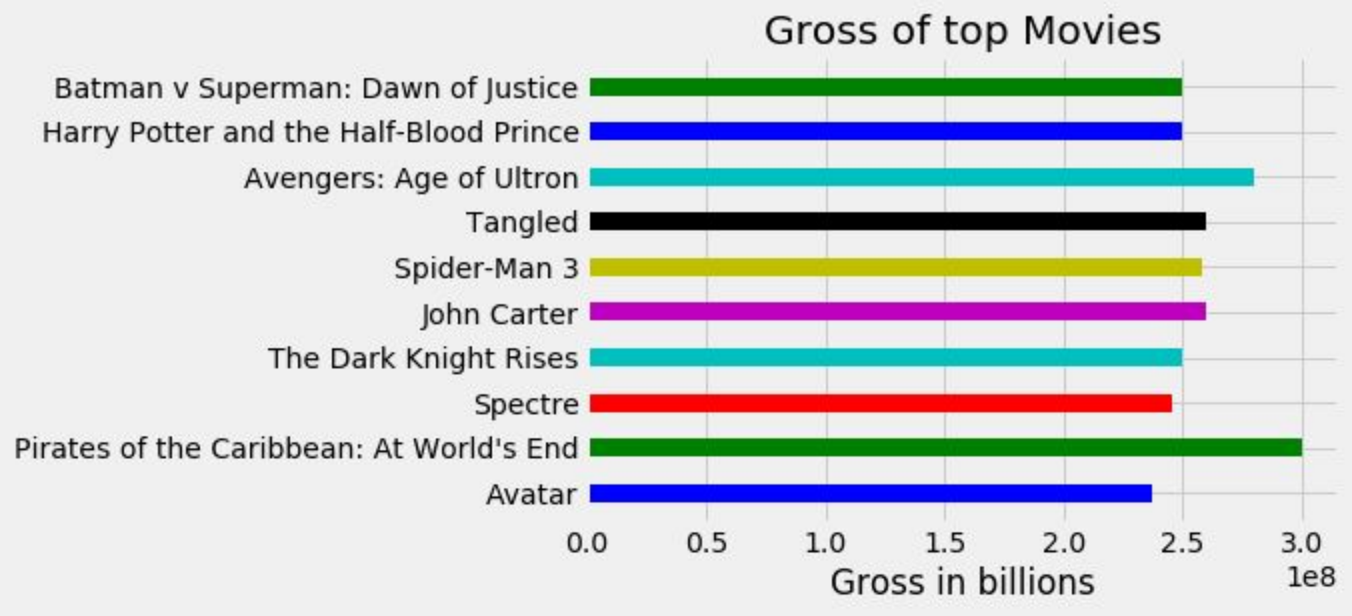
1. Inflation Rates: To account for the inflation rates over the years from 1956 to 2017, we have used the data of inflation rates to reduce every year's money value to that of 1956.
2. JSON to dict format: The data related to cast and crew for each movie were in data format and they were converted to dictionary format.
3. Score Allocation to production companies, cast and crew members.
4. Month of release added to the dataset.

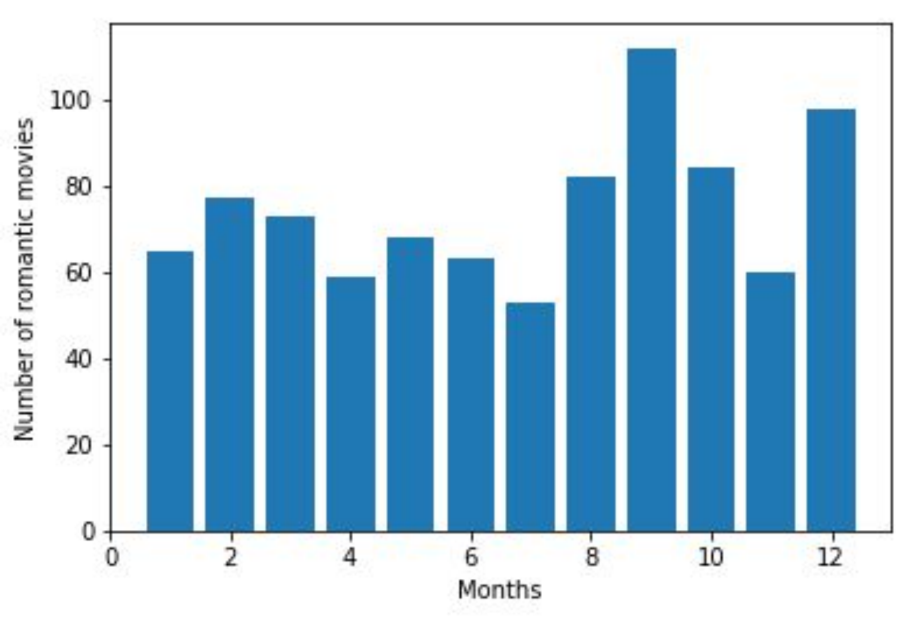
Data Visualization

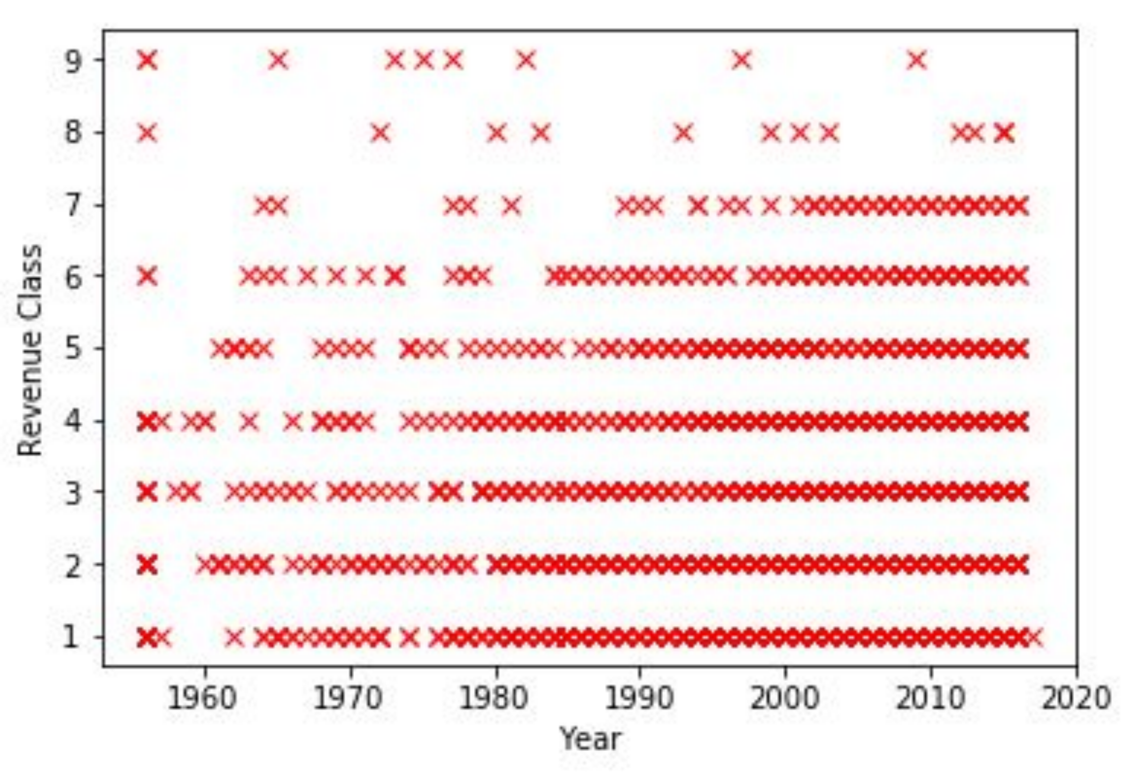
Some plots for visualizing the data are as follows:


Total Gross of movies of top Actors
Gross of the actor's movies in billions
0.0
0.5
1.0
1.5
2.0
2.5
Sam Worthington
Johnny Depp
Daniel Craig
Christian Bale
Taylor Kitsch
Tobey Maguire
Zachary Levi
Robert Downey Jr.
Daniel Radcliffe
Ben Affleck


Gross of top Directors
Gross in billions
0
2
4
6
8
Steven Spielberg
Peter Jackson
James Cameron
Michael Bay
Christopher Nolan
Chris Columbus
Robert Zemeckis
George Lucas
Ridley Scott
Gore Verbinski


Gross of top Movies
Batman v Superman: Dawn of Justice
Harry Potter and the Half-Blood Prince
Avengers: Age of Ultron
Tangled
Spider-Man 3
John Carter
The Dark Knight Rises
Spectre
Pirates of the Caribbean: At World's End
Avatar
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e8
Gross in billions


Number of romantic movies
0
20
40
60
80
100
0
2
4
6
8
10
12
Months


9
8
7
6
5
4
3
2
1
Revenue Class
1960
1970
1980
1990
2000
2010
2020
Year

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.

Revenue Class	Range in Millions
1	Less than 1
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

Reasons for discretization of the data

- 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).

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.

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

Support Vector Machines(SVMs)

“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.

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%.

Artificial 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.

In our Neural Network model we got a test set accuracy of 73.98% and training set accuracy of 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.