Predictions for Movie Success

Brooke Mosby and Michael Hansen

April 16, 2018

Case Study, Monday we review it all

Abstract

Cinema has become one of the highest profiting industries over the past century. The total box office revenue in North America alone amounted to $11.38 billion in 2016. With the possibility of great success, there is also a large risk of financial failure. This machine learning project is motivated by answering the question of what makes a movie successful. There are plenty of quantative data available for movies, such as the movies' budget, the release date, ratings etc., but in this analysis an attempt will be made to quantify movie information that is less measurable and then predict movie success.

1 Introduction

Research has been done to determine what aspects of a movie make it more successful; however, much of this research is contradictory. The research paper "Early Predictions of Movie Success: the Who, What, and When of Profitability"[1] states movies with a motion picture content rating 'R' will likely have lower profits, whereas the research paper "What Makes A Great Movie?"[2] states a motion picture content rating 'R' will have higher a box-office. Both papers analyzed thousands of movies, but came to opposite conclusions. Some variables used to predict movie success in these studies, included budget, motion picture content rating, and actor popularity.

Based on these previous models, the dataset used includes movie title length, run-time, motion picture content rating, director, genre, release date, actors, actor network scores, budget, opening weekend box-office revenue, and a list of other predictor variables. A network of actors was created to observe the impact certain popular actors and combinations of actors had on the success. Movie success will be determined by whether or not the movie turns a profit.

2 Data Scraped, Downloaded, Cleaned & Engineered

2.1 Dataset

A beginning dataset was downloaded from IMDb with 10,000 movies, each entry containing `Title`, `URL` (on IMDb), `IMDb Rating`, `Runtime` (in minutes), `Year`, `Genres`, `Num Votes`, `Release Date`, `Directors`. From this dataset, additional information on the `Budget`, `Gross`, `Opening Weekend`, `Actors_0-Actors_9`, `Oscar Nominations`, `Oscars Wins`, `Other Nominations`, `Other Wins`, `Meta Score`, and `Content Rating` was scraped and cleaned. The data points were collected from IMDb, which is a reputable source for information, according to their website,

> "We [IMDb] actively gather information from and verify items with studios and filmmakers"[3].

2.2 Cleaning Data

The data set was completed after gathering each data point, although the information is not clean or uniform. The first step to clean the data was to remove all commas across each column in the DataFrame. Removing commas made it easier to convert monetary amounts to integer values. Next, each date in the `Release Date` column was changed to a pandas date object, which simplified any calculations that rely on the release date of the movie. Each monetary value was converted into an integer in US dollar amounts. Each unique genre was made into a column with a true or false boolean for each movie entry. Using booleans instead of categorical data causes less problems when implementing regression.

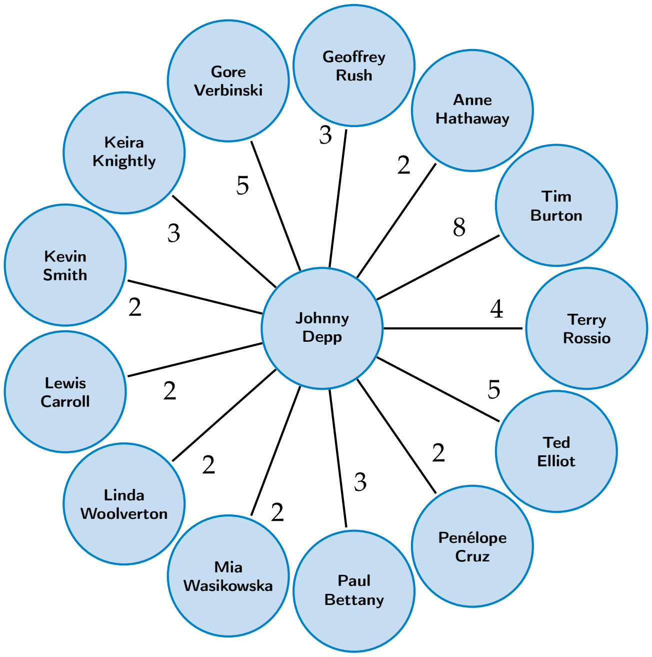
2.3 Feature Engineering

To resolve the disagreement in monetary amounts due to inflation, a dataset containing the CPI for each year from 1913 was used to adjust the values. The CPI (Consumer Price Index) describes the amount of purchasing power the average consumer has. For example, the purchasing power of 1 US dollar in January of 1913 is equivalent to approximately 24.69 US dollars in January of 2017. The length of the movie title was added, and a NetworkX graph of all actors was made. This was used to add a column describing the total edge weight for actors in each movie.

2.3.1 Actor Network

Using the Python library NetworkX, a network of actors was made to help determine the success of a movie. Each node in the network is an actor that has appeared in a movie, and that actor node has an edge to another actor node if they appeared in a movie together.

The weights on each edge correspond to the amount of movies the two actors appeared in together. In the example below, edges with Johnny Depp having weight 1 were ignored. The other nodes and edges demonstrate how the graph was assembled, actor to actor with the edge weight being the number of movies they appear in together according to the dataset used.



2.4 Final Dataset

The most significant variables in the final dataset are `Title`, `Length of Title`, `Content Rating`, `Runtime (mins)`, `IMDb Rating`, `Genres`, `Content Rating`, `Meta Score`, `Oscar Nominations`, `Oscar Wins`, `Other Nominations`, `Other Wins`, `Director`, `Release Date`, `Budget`, `Budget Adjusted`, `Opening Weekend`, `Opening Weekend Adjusted`, `Gross`, `Gross Adjusted`, `Profit`, `Profit Adjusted`, `Profit Bool`, `Directors Prev Total Profit`, `Actor_0-Actor_9` (the top ten actors in each movie), and `Actor Weights`. A separate NetworkX object holds the actor nodes and their connections.

3 Methods

Various supervised machine learning models are used to predict certain characteristics of the movie that determine a movie's success. The variables that our models attempt to predict are the movie's IMDb rating, Metacritic Score, the number of votes it recieves on IMDb, the number of Oscar nominations, the number of Oscar wins, the number of other award nominations, the number of other award wins, the number of total award nominations, the number of total award wins, gross, the profit,the gross adjusted for inflation, the profit adjusted for inflation, and finally whether or not a movie turned a profit. These varaibles will hereon be referred to as the movies' independent variables. For each model the data is split into a training and testing set to determine how accurate each method proves.

3.1 PCA

Principal Component Analysis (PCA) is used to reduce the dimensionality of the large dataset while retaining important information. For each Machine Learning model, PCA is used to reduce dimensionality to 10, 20, and 50 components, and then compared to find the best dimensionality. Because our dataset contains over 100 columns, PCA is essential for our models to avoid extremely costly computations and produce more accurate results.

3.2 Linear Regression Model

The simple supervised machine learning model, linear regression, attempts to predict the movies' dependent variables. This regression method attempts to predict these variables with a linear approach by modelling the relationship between the scalar and categorical independent variables and the dependent variables, that we have previously determined. Because this is simple to implement and does not contain costly computations, it is a great starting model for our data.

3.2.1 Linear Regression Model Analysis

Regarding the results of this model,there are very few things that linear regression can predict accurately. It is important to note that the best estimator was PCA with 50 components, which is the largest number of components used in the predictions. There are some variables that seem to have potential for a large increase in accuracy The first notable varaible is the IMDb rating, with has an error percentage of about 10% with potential to be increased. The Oscar nominations and wins are predicted extremely accurately from linear regression alone, and the boolean for whether or not a movie returns a profit is correct more than 70% of the time, which is perhaps one of the most important variables movie producers will look for, and one that will hopefully have better accuracy with different models. Other variables do poorly when predicted with linear regression.

3.3 Ridge Regression Model

The ridge regression model is essentially the same as the linear regression model; however, it implements a regularization term to prevent overfitting, and potentially prove more accurate than linear regression.

3.3.1 Ridge Regression Model Analysis

Ridge Regression results perfom as poorly as the linear regression model, with very few notable differences. The best estimator was still 50 PCA components.

3.4 Random Forest Regression Model

The random forest regression model is a type of additive model that makes predictions by combining decisions from a sequence of base models. A random forest is made by growing multiple binary trees. At each node the data is split into two children nodes, a decision made based on the residual sum of squares, or RSS. For random forests using regression, the predicted value at a node is the average response variable for all observations in the node.

3.4.1 Random Forest Regression Model Analysis

The Random Forest Regressor outperforms both other models in determining whether or not a movie turns a profit, with 76% accuracy, and predicting the Oscar wins and Oscar Nominations. However, IMDb rating and the Meta Score rating both did worse with the random forest regressor, although they are arguably the least important variables when predicting movie success.

3.5 Catboost Regression Model

There exist several implementations of random forest boosters, among these are GBM, XGBoost, LightGBM, and Catboost. Catboost seems to outperform the other implementations, even by using only its default parameters. Not only is it more accurate, but also faster than the other methods, making it an ideal machine learning model to implement on our dataset. In Boosting each tree is grown using information from previously learned trees, then each tree is fitted on a modified version of the original data set. Because the random forest regression gave the best results, the next natural step was to examine how Catboost performed on our dataset.

3.5.1 Catboost Regression Model Analysis

Catboost lives up to the hype and out-performs all of the models so far, with the exception of predicting the Oscar nominations a movie receives. Catboost is able to predict whether or not a movie will turn a profit with almost 80% accuracy.

3.6 Neural Network Classification Model

Neural networks are one of the biggest advancements in machine learning in the 21st century. A neural network is based on a collection of connected nodes, or neurons. In nerual network implementations, the output of each node is calculated by a non-linear loss function based on the sum of its inputs. The edges in the network typically have a weight that adjusts as learning proceeds. The weight increases or decreases based on the sum of the inputs. Neurons may have a threshold for the sum of inputs, before they are activated. Typically, neurons are organized in layers, where different layers may perform different kinds of transformations on their inputs. In our implementation, the neural network was costly to compute, so only the profit boolean was predicted with the neural network.

3.6.1 Neural Network Classification Model Analysis

The nerual network is unable to classify whether or not a movie turned a profit, as well and the random forest regressor and the catboost regressor are, with an accuracy of about 75%.

4 Case Study

5 Conclusion

Results	Linear	Ridge	Random Forest	Catboost	Neural Net
IMDb Rating Average Percent Error:	10.9439%	10.9444%	11.5794%	10.1450%	N/A
Meta Score Average Percent Error:	22.2333%	22.2320%	21.2724%	21.8721%	N/A
Oscar Nominations Average Error:	0.3688	0.3687	0.3422	0.3839	N/A
Oscar Wins Average Error:	0.2755	0.2754	0.2365	0.2264	N/A
Profit Bool Accuracy:	72.1221%	72.0720%	76.5765%	79.0790%	74.9299%

After analyzing the results of all the machine learning algorithms implemented, it becomes apparent that determining the success of a movie is very difficult. Some characteristics of movies, such as word of mouth appeal, are very difficult to quantify. Including an actor network score into our models improved the accuracy of our predictions, however there are still many other variables to account for that detract from the accuracy.

It is worth noting that every regressor model failed to accuractely estimate the profit a movie turned, because this value was extremely unpredicatable. Because Catboost was so fast and inherently more accurate, it enabled us to fine tune more than other models, to produce the most accurate predictions. Overall Catboost performed the best.

6 References

[1] T., Michael, et al. “Early Predictions of Movie Success: the Who, What, and When of Profitability.” [1506.05382] Early Predictions of Movie Success: the Who, What, and When of Profitability, 29 Jan. 2016, arxiv.org/abs/1506.05382.

[2] University of California - Davis. "What Makes A Great Movie?." ScienceDaily. ScienceDaily, 16 August 2007. <www.sciencedaily.com/releases/2007/08/070815135034.htm>.

[3] “Where Does the Information on IMDb Come from?” IMDb, IMDb.com, help.imdb.com/article/imdb/general-information/where-does-the-information-on-imdb-come-from/GGD7NGF5X3ECFKNN?ref_=helpart_nav_22#.