Predicting Movie Gross

Dara Shorten

1 Introduction

The movie industry is a multi-billion dollar industry with studios expecting to make profit on movies they produce. In this project I try to predict the gross a movie will make using various machine learning algorithms. Using data gathered from IMDB data-set I hope to predict the gross that film makes, using features of previously recorded movies. For instance Avatar has a gross of 300+ million however the budget of these films also have a major influence on the gross generated by the film. The data set includes features such as actor Facebook likes, genre and and content-rating. These features will be used to help predict the gross of an unseen movie. While predicting exactly what a movie will make is extremely difficult I hope to predict the gross of a film within a certain range. Using multiclass classification I hope to the gross generated within a certain bucket. These bucket will range from 0K to 1m, 1m - 10m, 10m-100m and 100m-1b. To acomplish this i split the data set into an 80:20 ratio. Eighty per cent of the data will be used as a training set while the remaining 20 per cent will be used to test the overall acuracy of the algorithms.

2 Related Research

Some research has been conducted in this field of machine learning. A paper written by Jason van der Merwe and Bridge Eimon set out to predict the total gross a movie will make. Eimon and van der Merwe also used a bucketing system to group predictions of gross. This technique is preferred as trying to predict the exact amount of movie could is extremely difficult if not near impossible. Eimon and Merwe used a Logistic regression on their training set and received an accuracy level of 52 percent.

Dara Shorten
CIT, Bishopstown Co.Cork, e-mail: dara.shorten@mycit.ie

Another study conducted by Rui Xue, Yanlin Chen from Standford also set out to tackle this problem of predicting the gross generated by movie. In their research they used both SVM and Naive Bayes to try and predict the outcome. Using feature selection they conducted that the most important features of a movies gross include Director, actors and genre to name a few. The use of SVM allowed for the expansion of multi-class classification of the algorithm. Using SVM and multi-class classification their error rate on their training data came to 20 per cent, while on the test data an error rate of 23.58 per cent. These are very high levels of accuracy. Going forward in my research I will look to mirror the results received using SVM multi-class classification.

3 Algorithm/ Model Data

3.1 Dataset

The dataset I have chosen is an IMDB dataset retrieved from Kaggle. This dataset contains details on 5000 movies. The dataset is an extremely detailed dataset with 28 unique features. Certain features include actors names, gross of the film, budget and numerous other features.

3.2 Feature Selection

First problem of the data set is that some of the data has no impact on the overall accuracy of the algorithm. Features such as IMDB link, number of critic reviews, These features will be dropped from the out set as they have no impact on the final result. The reason critic reviews and IMDB score is being dropped is because a movie about to be released will have no data when the movie is initially released. From looking at the data we may also drop features such as Actor Facebook likes. This feature can be deleted as another field called cast total Facebook likes includes the total number of Facebook likes for each of the actors. This feature will be used to assess how famous actors and directors. On inspection of the data I found that not all films used to same currency for their gross. For instance the film ”The Host” according to the dataset cost 12 billion dollars to make, however the case is it cost 12 billion won which is roughly 10 million dollars. On the constraint of time I will only use movies created in USA because the currency is all in dollars.

3.3 Mapping values

Some of the values in the data set must also be given numerical values in order to be used by the algorithm. To assign certain features a numerical value the use of one-hot encoding will be used. The reason I am using one-hot encoding is such that one variable isn't seen as greater then another by the algorithm. For instance if the genre Comedy was given a numerical value of 15 and the genre Action was assigned a numerical value of 3 then depending on the algorithm Comedy may be seen as larger the Action when in fact this is not the case.

3.4 Gross Bucketing

As stated previously trying to predict the exact gross a movie could make would be extremely difficult. To combat this problem I have assigned gross buckets to categorise gross amounts. These ranges are from 0-100k, 100K-1M, 1M-10M, 10M-100M, 100M+. Each of these ranges are assigned a value from 1-5. 1 being in the top range and 5 being the lowest bucket.

3.5 Algorithm

For this project I will run multiple algorithms to test the prediction. The various algorithms are Decision tree, KNearest Neighbour, Random forest classifier and Naive Bayes. The reason I am using multiple classifier techniques is to give me a range of results. This range of results helps pick the most accurate classifier to use.

4 Empirical Evaluation

4.1 Results

Non Normalised

Algorithm	Results (percentage)
Decision Tree	43.25
SVM	62.35
KNearest Neighbour	31.05
Random Forest	57.08
Naive Bayes	52.47

As seen from the results above we see that SVM has the most accurate result of all the algorithms. With a prediction level of 62 per cent based on the model fed into the algorithm I believe that this is a reasonably accurate predication. I believe that SVM performs best here because the gross of each have been separated in different buckets based on gross. SVM can be expanded to deal with multiclass classification and that is why I believe it has the highest accuracy in comparison to the other algorithms.

Normalised Data

Algorithm	Results (percentage)
Decision Tree	43.46
SVM	62.35
KNearest Neighbour	31.50
Random Forest	55.87
Naive Bayes	52.47

After the data has been normalised we see little change in the results. Again we see that SVM has the highest accuracy amongst all the algorithms.

4.2 *Feature Impact*

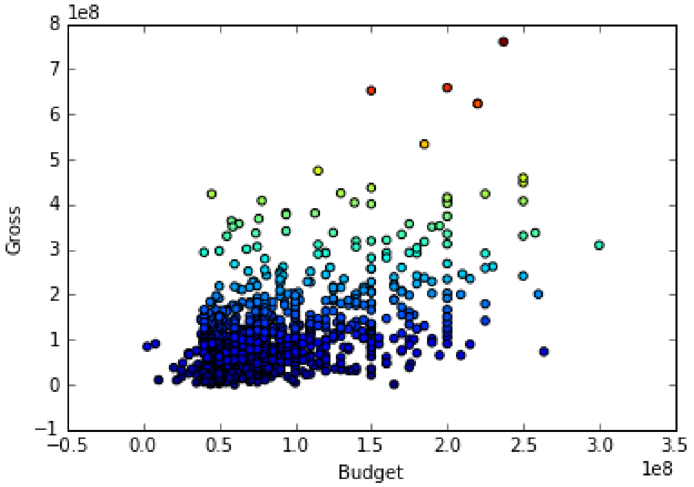

Fig. 1 All Amazon regions working

In the figure above we see that as the budget increases so does the gross of the movie. There is a general upward graph to indicate that budget has an impact on gross.

5 Conclusion and Future Work

5.1 Conclusion

Conclusion is that it is possible to predict on how much gross a movie can generate given enough data. It is clear that certain features have a major impact on the overall gross of a film. Using SVM we get a pretty reasonable prediction rate. Although I didn't receive the same level of accuracy as Rui Xue and Yanlin Chen believe that if given extra features such as studio and production company then the accuracy may increase.

5.2 Future Work

For future work will entail gathering information about studios that produced the films. I believe that studio that produce these films may have an impact on the prediction result.