Movie Reviews and Revenues

1. Introduction

Predicting box office for movies has been studied in economics, marketing and statistics. While predicting gross revenue may be hard since it largely depends on the quality of the movie, opening weekend revenue is more predictable for it is more influenced by the cast, budget, release time and some other forms of metadata.

In this project, we consider the problem of predicting a movie's opening weekend revenue using metadata about a movie—e.g., its genre, MPAA rating, and cast. We instigate the possibility of doing the prediction task solely based on the critique text data from different platforms like the New York Times, Variety. We also discuss how the results would be affected when text features are imposed. At last, we can perform some interesting analysis based on the trained mode such as identifying richer constructions that are good predictors.

2. Data

We use the data from http://www.cs.cmu.edu/~ark/movie$-data/, which is a data set gathering date for movies released in 2005–2009. For these movies, the metadata and a list of hyperlinks to movie reviews was collected from crawling Meta-Critic (www.metacritic.com).The metadata include the name of the movie, its production house, the set of genres it belongs to, the scriptwriter(s), the director(s), the country of origin, the primary actors and actresses starring in the movie, the release date, its MPAA rating, and its running time. The data set takes seven review websites that most frequently appeared in the review lists for movies at Metacritic, and obtained the text of the reviews by scraping the raw HTML. The sites chosen were the Austin Chronicle, the Boston Globe, the LA Times, Entertainment Weekly, the New York Times, Variety, and the Village Voice.

3. Predictive Task

3.1 Prediction and evaluation

In this project, we consider the problem of predicting the total revenue generated by a movie during its release weekend. We evaluate these predictions by using mean absolute error (MAE) and Pearson's correlation between the actual and predicted revenue.

3.2 Model

We use linear regression to directly predict the opening weekend gross earnings, denoted y, based on features x extracted from the movie metadata and/or the text of the reviews. That is, given an input feature vector $x \in R^p$, we predict an output $\hat{y} \in R$ using a linear model: $\hat{y} = \beta_0 + x^T\beta$.

To learn values for the parameters $\theta = \langle \beta_0, \beta \rangle$, the standard approach is to minimize the sum of squared errors for a training set containing n pairs $\langle x_i, y_i \rangle$ where $x_i \in R^p$ and $y_i \in R$ for $1 \leq i \leq n$:

$$\hat{\theta} = \underset{\theta=(\beta_0,\beta)}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} (y_i - (\beta_0 + x_i^T \beta))^2 + \lambda P(\beta)$$

A penalty term $P(\beta)$ is included in the objective for regularization. Classical solutions use an l_2 or l_1 norm, known respectively as ridge and lasso regression. The solution is a mixture of two, called the elastic net:

$$P(\beta) = \sum_{j=1}^{p} (\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|)$$

where $\alpha \in (0,1)$ determines the trade-off be- tween l_1 and l_2 regularization. We tune the α and λ parameters on our development set and select the model with the $\langle \alpha, \lambda \rangle$ combination that yields minimum MAE on the development set.

4. Features

We perform experiments to compare predictors only based on metadata, predictors only based on text, and predictors that use both kinds of information. The whole feature size including metadata and text features is 16600.

4.1 Metadata Features

Metadata features come directly from the dataset xml files. We considered seven types of metadata features, and evaluated their performance by adding them to our pool of features in the following order: whether the film is of U.S. origin, running time (in minutes), the logarithm of its budget, # opening screens, genre (e.g., Action, Comedy) and MPAA rating (e.g., G, PG, PG-13), whether the movie opened on a holiday weekend or in summer months, total count as well as the presence of individual Oscar-winning actors and directors and high-grossing actors. The best-performing feature set in terms of MAE turned out to be all the features. The metadata feature size is 117.

4.2 Text Features

The raw text data comes from the dataset xml files. It needs to be preprocessed before being used on the training task. We only included feature instances that occurred in at least five different movies' reviews. We used the common NLP technique bag of words with the vocabulary extracted from the training text data. We stem and downcase individual word component in all our features. With the use of python NLTK tools, we remove the stopping words and extract unigram, bigram and trigram text features. As a result, the vocabulary size is 16483.

5. Experiments and Results

We perform the predictions on three kinds of datasets (metadata only, text data only, both metadata and text data) using linear regression with two different kinds of loss functions (traditional mean squared error, loss function defined in 3.2). 1147 samples are in the trainset. 317 samples are in the validation set. 254 samples are in the test set. The MAE and Pearson correlation results on the test set are given as below.

	Self-defined loss function		MSE loss
	MAE($M)	Pearson correlation	MAE($M)
Metadata	6.4245 (α=0.6, λ=1)	0.69	7.37
Text data	8.6533 (α=0.9, λ=0.2)	0.72	8.68
Meta & Text	6.4231 (α=0.7, λ=0.3)	0.77	6.78

- The performance with the self-defined loss function, which is similar to elastic net, beats the performance of traditional simple linear regression.
- These all three datasets can all offer reasonable predictions on the opening week revenue since the MAE is low. The Pearson correlations between predictions and true values shows that the predictions are reasonable. The error is about 7 million dollars which is acceptable in box office prediction.
- Even though solely text data would not give the best performance. But it could be a good substitute for the prediction.
- Metadata and text data combined together yields the lowest loss and the highest the correlation. It further proves the possibility of higher accuracy when predicting movie box office when movie review data is also included.

6. Analysis and Findings

We also perform some analysis on the weighted model since the linear regression model is very easy to interpret. In movie box office scenarios, they can be interpreted as how much money a given factor could contribute.

	Feature	Weight ($M)
Metadata	Highest grossing actor Oscar winning actors Fantasy	1.72 0.027 0.037
Text data	Action Franchise Special Fun Good	2.764 0.84 0.67 0.53 0.31
Meta & Text	Highest grossing actor	1.67

- Highest grossing actors can indeed bring more value while Oscar winning actors may not have a big effect on the revenue.
- The sentiment and genre words in movie critics reviews have heavier weight. It makes sense since they do intuitively describe how the audience would feel about the movie.

7. Conclusion

We conclude that text features from pre-release reviews can substitute for and improve over a strong metadata-based first-weekend movie revenue prediction. We also proved the self-defined loss function can bring better performance of the model.