Predict Movie Revenue With Movie Meta-Data

Lijun Sun
Computer Science Department, Tennessee Tech University

Objective

The aim of this paper is to derive a reasonably accurate movie revenue model with some pre-release movie attributes, such as budget, genre and director rating, without considering post-release attributes, like online-rating, vote count etc.

Introduction

Movie revenue is a key factor for measuring a movie's success. Movie meta-data contains a wealth of features that can be mined in order to determine what attributes correlate with their success. Literature review shows[1] [2] [3], there is little research about finding the correlation between movie data attributes and a movie's revenue. In this project, we obtain movie meta-data from The Movie Database(TMDB). We discuss the movie meta-data attributes with a focus on developing a movie revenue model for predicting movie profits as shown in Figure 1. The findings from the research can provide movie producers with information to make future movies profitable.

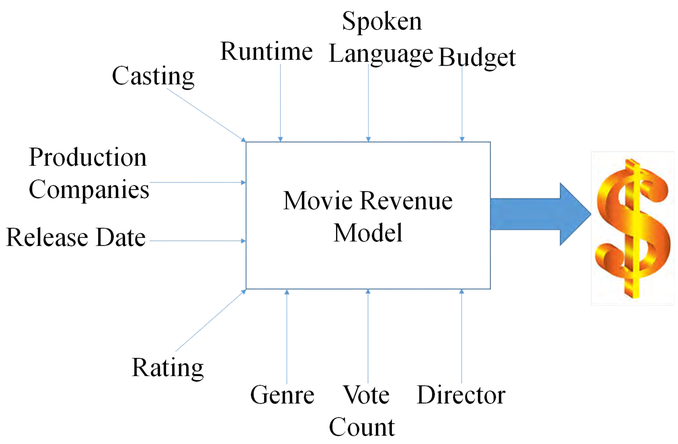

Figure 1 Movie Revenue Model Illustrate

Data Features Extraction

•Data Download

In order to obtain comprehensive information for developing the movie revenue model, three different movie data sets (represented as Dataset1, Dataset 2, Dataset 3) were downloaded from The Movie Database(TMDB).

•Data Mapping/Integration

Three data sets which contain different attributes need to be integrated as a complete dataset. Firstly, Dataset 2 is mapped to Dataset 3 with the common attribute "Director name" to get an new data set with attributes "Director" "Award" " Movie Title". Secondly, this new dataset is mapped to Dataset 1 which includes budget, revenue, and genre information, to get a final data set. The attributes of the final dataset are listed in Table 1.

•Data Cleaning/Filtering/Classification

1) During this procedure, a subset of movie with budget no less than 100K USD, revenue no less than 100K USD, and release year between 2000 to 2017 year is selected.
2) There are 21 different movie genres in the dataset. Each genre was assigned with a numeric ID [4].
3) Each director is rated based on the awards they received. The director rating varies from 0 (lowest) to 10 (highest).

TABLE 1 ATTRIBUTES OF FINAL DATA SET

Attributes	**Attributes**
Budget	Overview
Genres	Production Companies
Revenue	Production Countries
Title	Keywords
Director	Runtime
Award	Spoken Languages
Release Date	Vote Count
	Vote Average

Results

In this study, the movie revenue model was developed as a function of movie budge and director rating for each category of genre. The movie revenue model takes the following form.

$$\mathrm{Re}\,venue = k_0 + k_1 DR + k_2 Budget$$

where DR is the director rating; k_0 , k_1 and k_2 are coefficients that are obtained from data training. The data training results for the model for each genre are summarized in Table 2. Note that for some genres, there are insufficient data points for model training, and results were not available.

TABLE 2 DATA TRAINING RESULTS FOR MOVIE REVENUE MODEL

Genre ID	**Genres**	k_0	k_1	k_2	**R^2**
12	Adventure	-4.08E+07	2.61E+07	2.94	0.5291
14	Fantasy	-1.10E+07	-1.44E+06	2.78	0.569
16	Animation	1.42E+07	1.02E+07	3.26	0.3741
18	Drama	-5.19E+06	6.58E+06	2.22	0.4241
27	Horror	5.53E+07	3.97E+06	0.877	0.0616
28	Action	-7.04E+07	1.23E+07	3.48	0.56928
35	Comedy	1.81E+07	-2.35E+06	2.41	0.3608
53	Thriller	-4.58E+06	1.15E+07	2.09	0.4842
80	Crime	-3.16E+06	4.50E+06	1.86	0.4596
878	Sci_Fic	-5.75E+07	5.74E+05	4.14	0.5751
10749	Romance	3.95E+06	-2.23E+06	2.53	0.4252
10751	Family	4.15E+05	-4.13E+07	4.2	0.5422

Discussions

Figures 1 and 2 show the budget line fit plot and director line fit plot for the genre = 12, respectively. With such a simple model, the two-variable linear regression is able to achieve reasonably good accuracy with R^2 = 0.52911. This finding is consistent with accurate movie revenue models reported in the existing literatures [5]. It is more accurate than many other movie revenue models reported in the literatures. Considering the amount of efforts and the simple form of the proposed model in this paper, it is reasonable to claim the current model development is successful for genre 12.

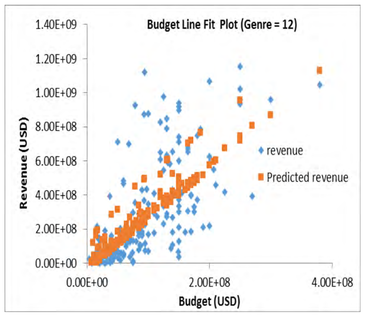

Figure 1 Budget Line Fit Plot for Genre = 12.

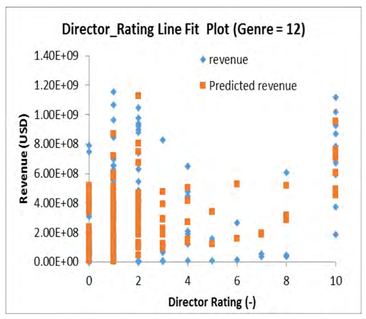

Figure 2 Director Rating Line Fit Plot for Genre = 12.

Figures 3 and 4 show the budget line fit plot and director line fit plot for the genre = 14, respectively. It can be seen from these two figures that the movie revenue model can achieve reasonably accurate prediction with R^2 = 0.5690. Due to space limitation, the linear regression results for other genres were not shown in the poster.

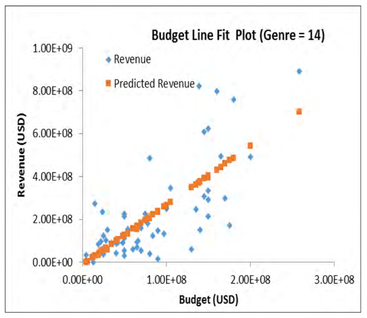

Figure 3 Budget Line Fit Plot for Genre = 14.

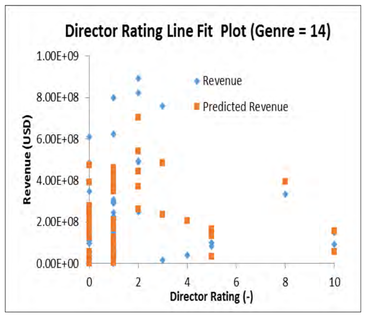

Figure 4 Director Rating Line Fit Plot for Genre = 14.

Conclusion and Future Work

- This paper presents a movie revenue model using a multivariate linear regression method. The model predicts a movie revenue based on the pre-release attributes: budget, genre, director ratings. Each revenue model was computed based on every genre.
- The result shows movies with genres like Action, Family, Fantasy and Adventure, can be more precisely predicted than movies in other genres.
- The model is least suitable for predicting movie revenue for horror-type movies.

For future work, the model can be further improved by the following work:

- Add more features/attributes (include star power, or post-release attributes, such as movie rating) into the model for prediction.
- Comparing the model structure with other methods, such as non-linear regression or other classification methods.

References

[1] Kabinsingha, S., Chindasorn, S., & Chantrapornchai, C. (2012). Movie rating approach and application based on data mining. International Journal of Engineering and Innovative Technology (IJEIT) Volume, 2.
[2]Lim, Y. J., & Teh, Y. W. (2007, August). Variational Bayesian approach to movie rating prediction. In Proceedings of KDD cup and workshop (Vol. 7, pp. 15-21).
[3]Dooms, S., De Pessemier, T., & Martens, L. (2013, October). Movietweetings: a movie rating dataset collected from twitter. In Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys (Vol. 2013, p. 43).
[4]Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications, 30(2), 243-254.
[5]Kaimann, D., & Pannicke, J. (2015). Movie success in a genre specific contest: Evidence from the US film industry (No. 98). Ilmenau Economics Discussion Papers.

Acknowledgement

The author would like to acknowledge Dr. Eberle from Department of Computer Science at Tennessee Tech University for providing support for this research.