Prediction of the Box Office in Taipei

Aji Setyoko M10715805 M10715805@mail.ntust.edu.tw

陳勁宇 M10801K03 M10801K03@mail.ntust. edu.tw

賴伯昇 M10801003 M10801003@mail.ntust.edu.tw

Abstract

Firm industry plays an important role in Taiwanese market, and so does box office prediction. Related works shows that classification model and the information from the social media improve the accuracy of the box office prediction. The study adopted the combination of the information of the movie dataset from the Taiwanese government and the popularity information from IMDb to do the classification prediction and the regression for the box office localized in Taipei. The result shows that the Taipei box office can be predicted by 90% accurate in the classification model and by 88% accurate in the regression model.

1. INTRODUCTION

Since the trend of globalization, the firm market has grown rapidly and steadily in nowadays, and it further attracts investors to invest more and more sources in the film industry, Walls and McKenzie (2012b). According to the report from Box Office Mojo (2020), in 2019, the box office hits a new record by $USD 42.1 billion in global revenue, and also in the same year, Taiwan, with the 56th world rank population, is firmly ranked 17th in the global world box office by acquiring $NTD 10.78 billion. By this, Taiwan has become one of the countries that plays an essential role in the film industry. Consequently, by having Taiwan as a representative country to study about the film features and the fruitfulness of film business strategy is worthwhile.

However, the success of the film industry is hard to be determined. One of the example case is a film named 'Cats,' which released in 2019. 'Cats' is a very famous musical adaptation film. Before this film was released, this film had a very high reputation. Not only having an excellent response on the original movie, but it also has an outstanding cast. Yet, the film lost about $USD 100 million in total. The film just sold $NTD 17.8 million in the Taiwan box office and $USD 38 million for world box office achievement. By taking into account of this problem, box office prediction is becoming a widespread issue. Not only to provide the theory and technique support, box office prediction also gives the film company or related investor some indicators that how much budget they are going to spend on or which decisions they should make.

The previous research topic which addressed box office prediction had done in many different ways. In Saraee, White, and Eccleston (2004), Zhang and Skiena (2009), a popular film database named the Internet Movie Database (IMDb) is employed and show convincing results. In other case, Nagamma et al. (2015) revealed that the accuracy of the prediction would be improved by using the class label, and the classification accuracy will also be improved by using the clustering method as the complementary step. Then, Apala et al. (2013) also shows that by using the information from the social media will help to construct the prediction model. Another interesting work is also stated by Lee (2008), which shows that the origin of the film become one of the critical aspect for box office prediction.

Following the factor which already stated above, the study aims to build a box office prediction model localized in Taipei, based on the original information from the movie and popularity information from the social media by using the clustering, classification, and regression. The study will focus on two experiment of Taipei box office prediction. The first one consists of the combination of K-Means and random forest, by using the clustering technique to cluster and label the movie first, then do the classification prediction. The second one adopts random forest regression to predict the real number of the box office directly.

In the end, in line with the contribution of this research, the model will predict if the film would be popular or acceptable for the Taiwanese market in advance. We believe that our result could be used for the related industry to do the decision making about film business opportunities.

In our experiments, we use the data from Taiwan Movies Company. We also collect the IMDB data as supplementary data. In total, we have a 1300 movies record from all movies in Taipei between 2015-2019. The data separated randomly into training and testing with 70% and 30%, respectively.

In the following section, we will describe the details of the the data collection, then followed by experiments setting, then result and discussion, and the last is the conclusion of our experiments.

2. DATA COLLECTION

The dataset from the ministry of culture in Taiwan is used in this study, which provides the number of the movie with the box office revenue for each film. The dataset also provides the category of each film, however, this feature is not entirely available for all film titles. We then gathered data from IMDB to get an additional feature for the learning model. The features from the IMDB database consists of the budget of the film production, the adulty of the film, the category of the film, and the popularity of three primary casters, producer, director, and also the screenplay.

The combined data from Taiwan Movie dataset and IMDb dataset consists of seven main features, the detailed information of them will be described in the following paragraphs. The one the example of the films is shown as Table 1.

Table 1 Example of default data

Movie name Eng	The SpongeBob Movie: Sponge Out of Water
Produce country	USA
Movie length	93 分
Language	English
Audio effect	SDDS | Datasat | Dolby Digital
Budget	74000000
Caster 0	3.641
Caster 1	1.67
Caster 2	2.703
Director	0.98
Producer	1.678
Screenplay	4.138
Category	Animation, Fantasy, Family, Comedy, Adventure

2.1 Budget

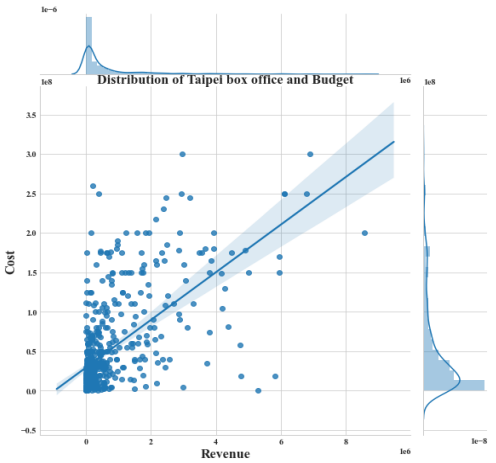

Figure 1 Distibutio of Taipei box office and Budget.

Movie feature is one of the features which obtained from the IMDB database. As follow in Chen and Cheng (2008), the quality of product is likely depend on the budget. We brought this idea to become one of the features to predict the selling revenue of a movie film.

The distribution of the movie budget and the box office is shown as Figure 1. Most of the movie budget is between $USD 10 millions to 60 millions, and as the budget incearses, the box office will increase as well.

2.2 Taipei Box Office Revenue

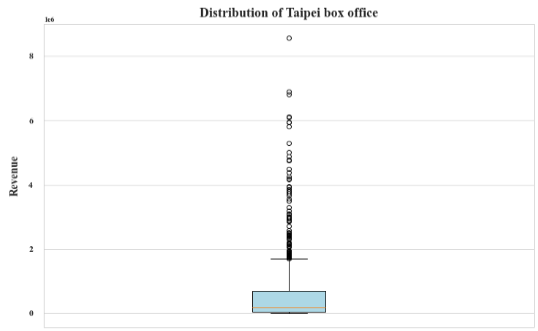

Figure 2 The visualization of Taipei Box Office

Taipei Box Office is an important feature which used to validate our prediction model. This feature also represents the response of the market for each movie. Followed by the Taipei movie market stated in the introduction, this feature could represent the world response for a film.

Figure 2 is a boxplot which shows the distribution of the Taipei box office. The statistics indicates that the range of 25% of data is significantly more extensive than the other 75%, and the box office mainly gathers in the range between $USD 40.6k to 699k.

2.3 Category

As follow in Walls and McKenzie (2012a), category movie or genre movie has an impact on the attention of the people. We then combine features from Taipei Movies and the IMDB dataset. We need to combine the data to overcome the unprovide data from the Taipei Movie dataset.

In Figure 3, the distribution of the genres for the movies shows that for the last four years, Drama genres is the most produced movie, followed by Comedy and Thriller.

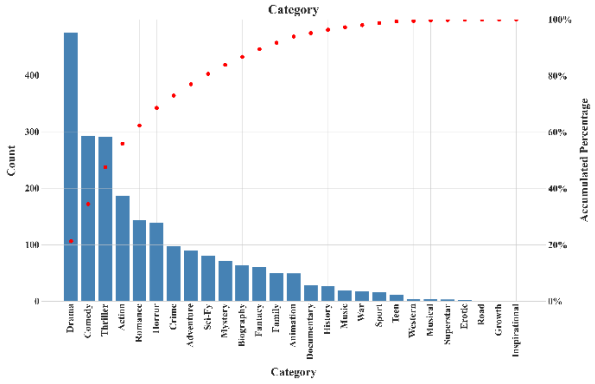

Figure 3 Distribution movie film genres

2.4 Director, Screnplay, and Actor Popularity

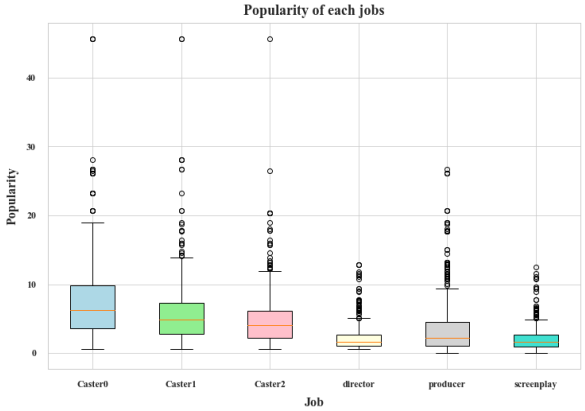

Figure 4 Caster popularity distribution

Apala et al. (2013) shows the usefulness of follower crew (Caster, Producer, Director, and Screenplay) in social media. By this, we employ a similar feature, popularity, which obtained from the IMDB dataset. The popularity features in the dataset represents the popularity of the crew in front of the IMDB user forum.

The distribution of the data shown in Figure 4. All of the results indicates that the range of 25% of data is significantly different than the other 75%, especially in the casters, the range of the distributions are quite broad.

2.5 Produce Country

As stated above in Lee (2008), the source of the movie production has an impact on the income. So, to take into

account of this problem, we employ the source of the movie production feature into the model.

In Figure 5, the result shows that the USA become the country that produces the most movies. Statistics also shows that the second most movie production country is just 25% compared to the USA. On the other hand, other countries have a normal distribution compared to each other in the number of movie production.

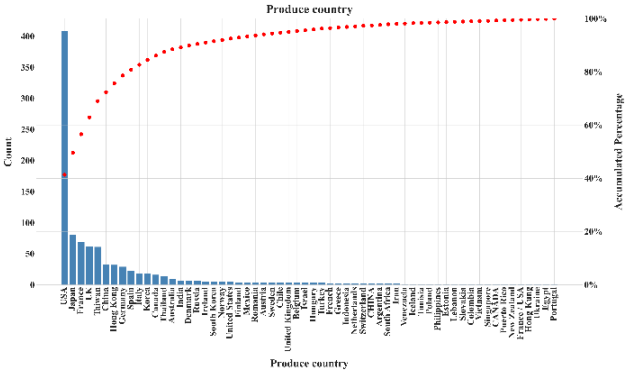

Figure 5 Origin movie production distribution

2.6 Language

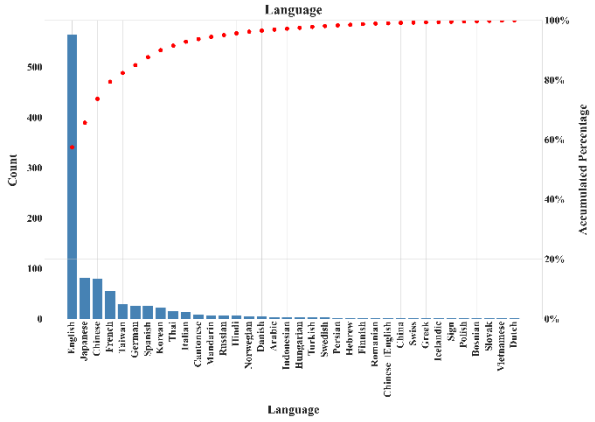

Figure 6 Movie language distribution

To increase the leverage of the origin features, language feature is employed to build the movie revenue prediction.

In Figure 6, the result shows distribution of the language in the films. Clearly, English is dominant compared to others. Then, followed by the Japan and the Chinese. Here, we see that Japan and Chinese get on the position because the database crawled in Taiwan.

3. METHOD

There are three main steps in our experiments, which consist of Data Collection, Classification Experiments, and Regression Experiments. The flow for these experiments shown in Figure 7. The first experiment is Data Collection, which indicates by a gray background. This part consists of Data Crawler and Data Preprocessing. The second part is Classification experiments, indicate by the blue background. Classification experiments flows consist of Clustering, Labelling, Random Forest Classification, and Label Predicting. The third part is Regression experiments, which only consist of Random Forest Regression and Predict Taipei Box Office.

3.1 Data Preprocessing

The result of Data Crawler are features shown as Table 1 mentioned in the paragraph 2. Data Preprocessing uses to combine the Taiwan Company Dataset and the IMDB dataset. In detail, data preprocessing are consists of data filtering, data encoding, and data normalizing.

Data cleaning uses to drop the data from the rows if the features not complete. The default features that must be provided for each film are Produce Popularity, Screenplay Popularity, Director Popularity, three Caster Popularity, Budget, Length of the Film, and Category of the film.

Data encoding uses to encode features into appropriate forms. In our work, we only need to encode the Category feature to become a One-Hot Encoding format.

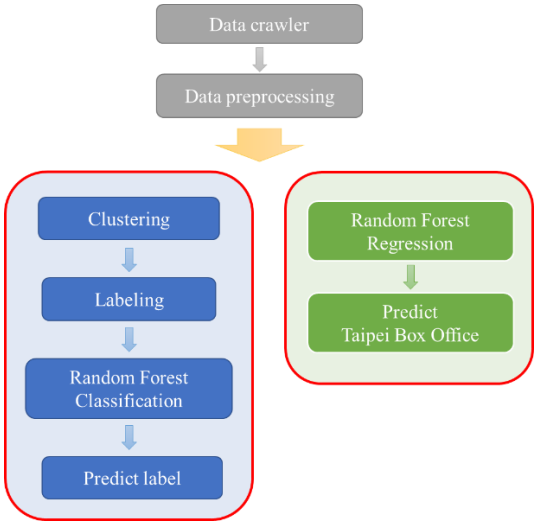

Figure 7 Algorithm Flows

Data normalizing employs to make every feature have the same distribution. We do a 0-1 normalization on all popularity features, budget, and length of the film. The formula of this normalization is shown in equation 1.

$$z_{i\_norm} = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad \text{.....} \quad \text{equation 1}$$

3.2 Classification

There are two main steps in classification experiments. First is the labeling process, and the second is generating a predictive model to the labeled data.

For the labeling process, we use the K-Means algorithm the generate the label for each group of movies. The features which fed to the K-Means is the default features.Acoording to the employs the elbow method, the number of class label is chosen to be 5.

The predictive model built by using the labeled data which already separated into training and testing. We employ a random forest classification algorithm to be the model. The result of this experiment are presented in the following section.

3.3 Regression

In regression experiments, we directly employ the normalized data to build the model. Random Forest algorithm is employed to solve this experiment. The model also tested to predict the performance of the model. The result of this section shown in Section 4.

4. RESULT AND DISCUSSION

In Figure 8, the result shows the visualization of the

labeling results in the properties of Budget and Taipei box office. This result is obtained by using K-Means. As shown in the figure, there are five colors which represent the different classes.

From the analytic statistic performance, we make a ratio performance for every class. This analysis shows that the most profitable cluster is Group 3. Then for the most unprofitable is Group 5.

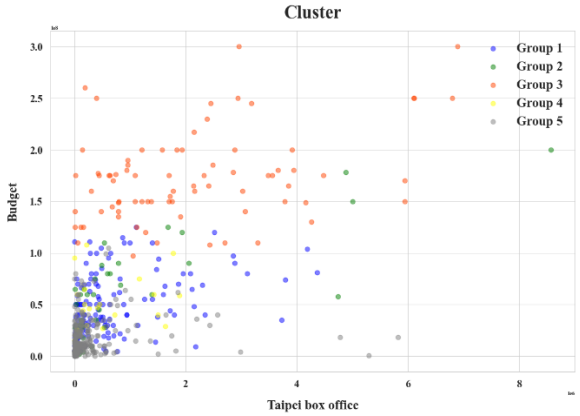

Figure 8 Clustering with K-Means result

The target for the classification model is the label from the clustering process. The result of the classification process with a random forest algorithm shown in Table 2 and the confusion matrix of prediction is shown in Table 3.

Table 2 Classification performance with 5 cluster

Cluster	Precision	Recall	F1	Support
1	0.86	0.91	0.89	47
2	0.68	1.00	0.81	13
3	0.95	0.87	0.91	23
4	0.40	1.00	0.57	2
5	1.00	0.89	0.94	88
Accuracy			0.90	173

In Table 2, the result shows that F1-Score averagely reach 90%, which means our model has excellent performance to predict movie in belongs to each cluster. Moreover, the F1 score for group 3 is 0.91, which means the model has 91% accuracy in predicting the film which is profitable.

Table 3 shows the confusion matrix of the prediction, and the result shows that the model has excellent performance for the 5 groups especially in the group 5.

Table 3 Confusion matrix of classification model

True Prediction	1	2	3	4	5
1	36	1	2	2	0
2	0	10	1	2	0
3	0	0	21	0	0
4	0	0	0	3	0
5	10	0	0	0	85
	1	**2**	**3**	**4**	**5**

True Label

The result of the regression model to predict the number of the Taipei the box office under 4 K-Fold validationo is shown in Table 4. Here we can see that our model has a very good performance by achieving 88% R^2 value. This number means that our model can predict the value of Box Office in Taipei by 88% accuracy in average, even if the 4-Fold validation is adopted.

Table 4 K-Fold validation for regression model

Fold	1	2	3	4	Average
Score(R2)	0.82	0.89	0.91	0.90	0.88

5. CONCLUSSION

In this study, we show that the combination of the movie data from the government and the popularity information from the IMDB can be used to predict the characteristic of the movies and predict the box office value in Taipei.

In advance, our proposed model can solve the prediction problems by having accuracy 88% for regression model and 90% for classification model. The information provides the invester or the movie related industry some indicaters to do the decision making about film business opportunities or reference to do the adjustment. In the future, the prediction performance might be increased by employing more data by employing another dataset.

6. REFFERENCE

Apala, K. R., M. Jose, S. Motnam, C. Chan, K. J. Liszka, and F. de Gregorio. 2013. "Prediction of movies box office performance using social media." In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, 1209-14.

Chen, Jinchuan, and Reynold Cheng. 2008. "Quality-aware probing of uncertain data with resource constraints." In *International Conference on Scientific and Statistical Database Management*, 491-508. Springer.

Lee, Francis L. F. 2008. 'Hollywood movies in East Asia: examining cultural discount and performance predictability at the box office', *Asian Journal of Communication*, 18: 117-36.

Mojo, Box Office. 2020. '2019 Worldwide Box Office'. https://www.boxofficemojo.com/year/world/2019/

Nagamma, P, HR Pruthvi, KK Nisha, and NH Shwetha. 2015. "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction." In *International Conference on Computing, Communication & Automation*, 933-37. IEEE.

Saraee, Mohamad, S White, and J Eccleston. 2004. 'A data mining approach to analysis and prediction of movie ratings', *WIT Transactions On Information And Communication Technologies*, 33.

Walls, W David, and Jordi McKenzie. 2012a. 'The changing role of Hollywood in the global movie market', *Journal of Media Economics*, 25: 198-219.

Walls, W David, and Jordi %J Journal of Media Economics McKenzie. 2012b. 'The changing role of Hollywood in the global movie market', 25: 198-219.

Zhang, Wenbin, and Steven Skiena. 2009. "Improving movie gross prediction through news analysis." In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 301-04. IEEE.