**How to Make a Profitable Movie?**

*Nattapat Juthaprachakul, Rui Wang, Siyu Wu, Yihan Lan*

**Problem Definition**

Film industry has grown rapidly in recent years. More and more movies have been released. Some gains high box office profit while others do not make money or even lose money. Therefore, we are interested in the problem of “How to make a profitable movie?”; in other word, we are interested in finding out what important features are contributing to a profitable movie. Our goal is to find, preprocess, analyze, and visualize many available movie datasets in order to find important features related to movie profitability.

**Method and Experimentation**

**Step I: Data Collection**

In this project, our main datasets that we used are TMDB and UCI. The UCI datasets contain movie information from 1900 to 1999. It has two main files in form of HTML. The first file is “main.html” which contains movie name, id, year, and other related features. The second file is “cast.html” which contains movie id, title, actor, and director. Therefore, the first process that we need is to extract the dataset in HTML format and turn them into two CSV files. The first CSV file contains UCI movie id and year while another contains UCI movie id, title, actors, and directors. Later we joined these two CSV datasets on movie UCI movie id to get past movie information of the past hundred years. Since UCI movie id does not match TMDB movie id, movie title and year were used instead to combine UCI and TMDB dataset for further analysis.

For TMDB dataset, its cast data from 2000 to 2017 and its movie metadata from 1900 to 2017 are used. TMDB cast data contains TMDB movie id and the whole cast team members such as director and various actors. Considering the impact of celebrities on movies, only TMDB movie id, director name, and top five actor names are extracted. Movie metadata contains twenty-four movie features. TMDB movie id, genre, runtime, budget, revenue, title, release date, production companies, and production countries are selected as they are supposed to be related to movie profitability. Besides, movie ratings are also considered and TMDB rating data includes different user ratings. Movie average ratings are calculated and added into movie metadata. All data collection tasks were conducted in Pyspark which can be easily scaled for bigger movie dataset.

**Step II: Data Pre-processing for machine learning**

As of now, there are three different related movie datasets; namely, movies with ratings, past movie casts, and recent movie casts. For movies with rating, a variety of features include id, title, production company, country, genres, release date, budget, revenue, and rating. All of them are important and needed for later data analysis; however, some of them are needed for further data cleaning.

For some numeric movie features such as rating, budget, and revenue, there are many missing data. Instead of deleting all rows containing missing values, mean values are calculated and used to replace the missing values. For categorical features including company, country and genre, all missing values are deleted by a filtering method and only one value in each feature could be later used. Generally, a majority of movies are co-produced by a couple of different companies, filmed in many different countries, and categorized into various different genres. Because of this fact, for each categorical feature there are countless combinations. For the effectiveness of further analysis, only one value of each mentioned categorical feature will be used. Moreover, deleting duplicate rows of data with same movie id is also necessary.

For the cast datasets, by counting the numbers of all cast and director names in our datasets, we found out that there are too many different cast names. Therefore, the top 1000 frequent casts are selected instead by a following method. If a movie contains a selected frequent cast, the cast name is kept and the other will be dropped. However, if there is no cast matching with our 1000 frequent cast name list, one cast is randomly selected from all cast names. The three datasets are later joined together into one structured dataset.

The dataset was ready for the machine learning stage, but one problem still occurred. For Random Forest regression model, the maximum bins it could handle are 32. However, many features (including cast, director, company, and companies) have more than 32 unique values. To solve this problem, two methods are used which resulted in creating two datasets. The first way is to count and select the top 30 frequent unique values in cast, director, and etc. The frequent one is kept and all other unique values is renamed into ‘others’. The second way is to change all top 30 frequent unique values into ‘1’, and the rest values into ‘0’. All of the works mentioned were implemented by Pyspark framework (Dataframe and RDD) which can be easily scaled up for bigger movie datasets.

**Step III: Data Analysis by Machine learning and Feature Selection**

Two methods were implemented for feature selection. The first one is to calculate the correlation coefficients between features and movie profits and select the top 3 features that have the highest correlation coefficients. The other method is to use Random Forest regression model and select the feature importance coefficients. The model has been implemented three times; each time, the feature with the highest coefficient is selected and dropped before running another time. After selected three most important features, the relationships between each of them and profitability could be unraveled by using visualization tools. PysparkML was used for machine learning modelling and analysis.

**Step IV: Data Analysis and Visualization by Matplotlib and Tableau**

Pandas and Matplotlib were used for dataset understanding and visualization. The distribution of different movie genres, profit, and budget is shown in the Appendix. Additionally, the relationships of budget-profit, runtime-profit, and genre-profit are shown in the Appendix as well (Figure 2-8).

Additionally, in order to have a clear overview of how different countries perform in the movie industry, we used Tableau as our main visualization tool because it provides simple data manipulation methods and numerous rich visualization functions. First of all, to find which countries have a prosperous movie industry a hot map of global average movie revenue is formulated. Later, a deeper analysis of those countries was conducted to find out how their annual movie average profit changes overtime, which movie genres are more profitable than others, how runtime influences movie profitability, and which are their top 10 most profitable movies. Finally, the Tableau plots and figures are exhibited on a webpage (Figure 9).

**Result**

From correlation coefficients (Figure 1) and the feature importance coefficients (Table 1), the top three important movie features are budget, genres, and runtime respectively.

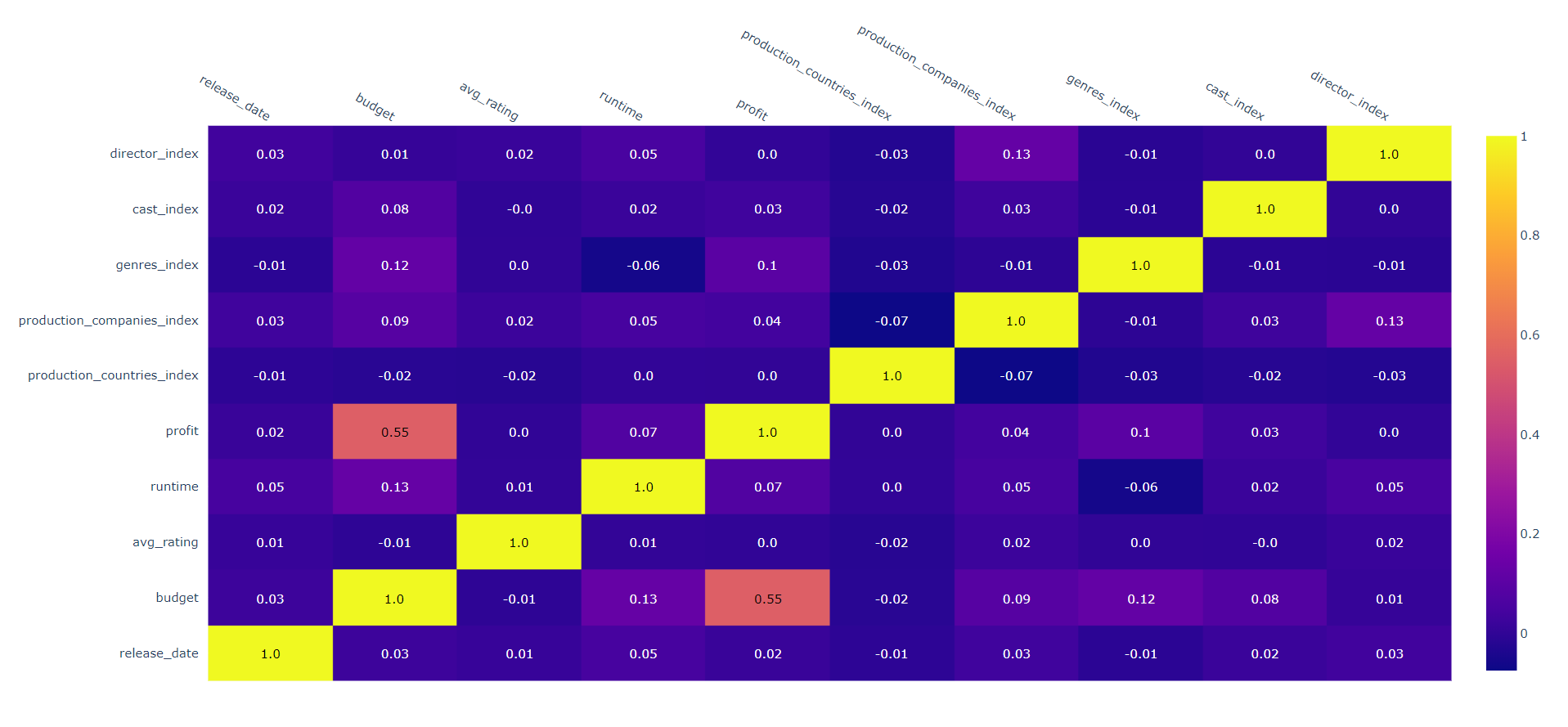


Figure 1. Correlation Coefficients

|  |  |  |  |
| --- | --- | --- | --- |
| The 1st time | Budget(0.68) | Runtime(0.10) | Genres(0.09) |
| The 2nd time | Genres(0.33) | Runtime(0.32) | Rating(0.13) |
| The 3rd time | Runtime(0.41) | Release date(0.17) | Company(0.08) |

Table 1. Feature Importance Coefficients

The analysis above show that the top 3 profitable countries in film industry are America, Britain and Germany. The profitable genres in America are animation and adventure. In Britain, genres in family and adventure are more profitable. Besides, action and adventure are profitable in Germany. It can also be concluded that budget is directly proportional to profit that a profitable movie is supposed to have enough budget; genres is also an important factor in profit, and the profitable genres are adventure, action, thriller, romance and comedy; besides, runtime of a profitable movie cannot be too short or too long that 90 - 160 minutes runtime is a good choice.

Finally, web is built to visualize the result. The web which is country-oriented shows the budget, genres, and profit of the most developed countries in the film industry.

**Challenges and Future Works**

Throughout the project, we have encountered three main challenges. The first challenge is in the data collection step as the UCI movie dataset is in HTML format. The process of extracting data from HTML and converted them into Dataframe is demanding and a problem that we never encountered in class. To solve this problem, “urlopen()” function from “urllib.request” module was used to fetch the webpages. Later, we utilized “findall()” function to find all tables on the page and form a Dataframe. The second problem is about dataset itself. As the distribution of different movie genre is shown in the Figure 5, we can see that there are four genres that dominate our dataset namely: Drama, Thriller, Comedy, and Romance. In the future, we might find and use more balanced dataset for our analysis. The last problem is about missing value in the dataset such as missing movie budget and ratings. Therefore, in the future, If we have an access to bigger, more updated, and complete movie dataset, our analysis could be substantially improved.

**Project Summary**

* Getting the data: Acquiring/gathering/downloading. (score=2)
* ETL: Extract-Transform-Load work and cleaning the data set. (score=4)
* Problem: Work on defining problem itself and motivation for the analysis. (score=2)
* Algorithmic work: Work on the algorithms needed to work with the data, including integrating data mining and machine learning techniques. (score=2)
* Bigness/parallelization: Efficiency of the analysis on a cluster, and scalability to larger data sets. (score=1)
* UI: User interface to the results, possibly including web or data exploration frontends. (score=4)
* Visualization: Visualization of analysis results. (score=3)
* Technologies: New technologies learned as part of doing the project. (score=2)

**Appendix**

The plots in appendix indicate relationships between budget and profitability, genres and profitability, and runtime and profitability.

|  |  |
| --- | --- |
|  |  |
| Figure 2. Movie Budgets | Figure 3. Movie Profits |
|  |  |
| Figure 4. Relationship Between Genres and Profits | Figure 5. Number of Movies in Each Genre |
|  |  |
| Figure 6. Regression: Budget-Profit | Figure 7. Scatter Plot: Runtime-Profit |

|  |
| --- |
|  |
| Figure 8. Relationship Between Budget-Profit and Genres |
|  |
| Figure 9. Web Page for Visualization |