Project Proposal to Predict the Movie Popularity by Machine Learning

Assignment 3 of Machine Learning

Student Name: Yuhao Shi

Student Number: 13338239

1. Aim and objectives

The purpose of this project is to predict the popularity of a movie by building a machine learning model, based on the history movie records. This kind of analysis not only helps those movie companies to invest the correct movies and gain the considerable incomes, but also encourages the film screenwriters to write the better movie scripts, which can attract majority of the movie fans, so it could be a promotion for film industry. The aim of this project can also be divided into following 3 different objectives:

- Collecting sufficient movie records, including the movie name, type, actors, directors, the box-office of this movie, and so on from the latest movie databases to build the training dataset and testing set.
- Analyse the training set by multiple machine learning methods to develop the movie predictor classifiers.
- Validate the result by inputting the testing set into the developed models and comparing the result accuracy with those methods in order to decide the best movie predictor model.

2. Background

The reason why this project I choose to focus on film is the dramatically speed of the improving in film industry in worldwide, and for some countries, such as India, film industry is one of the mainstay industries, which brings much revenue to the state. Besides, as one of the entertainments for people’s leisure time, watching movies becomes more and more popular and is surpassing the Television as the top popular way of entertainment. According to the report of Motion Picture Association of American (MPAA), it clearly summarized the global trends of film industry and the increase of global box-office. For the global box-office, there are totally $41.1 billion in 2018, which is up one percent over 2017, although the rate of the increase looked small, the total amount increased amazingly, we can see from the histogram below that the box-office increases nearly $4.7 billion from 2014 to 2018. Considering its economic impact, more and more researchers and data scientists have conducted studies on the movie industry(Lee et al. 2018).

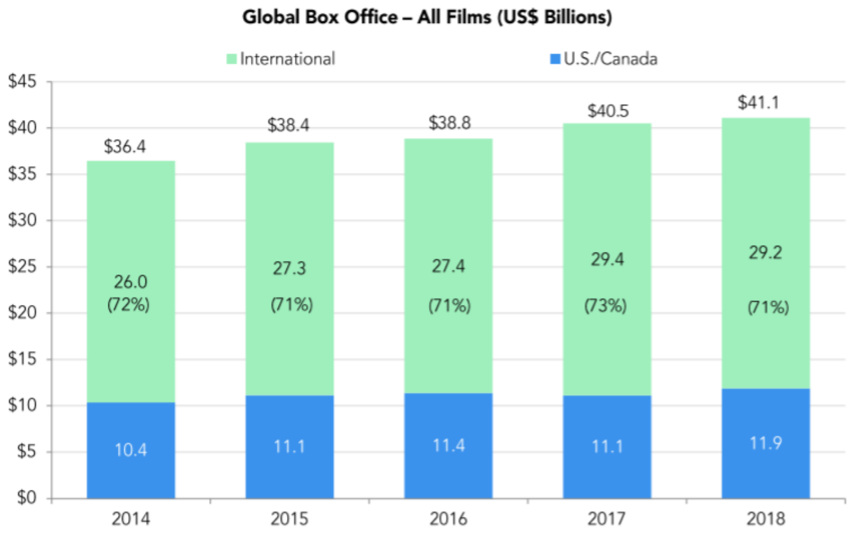


Figure 1: the global box-office of all films in 2018 (sourced from MPAA 2018 report) https://www.motionpictures.org/wp-content/uploads/2019/03/MPAA-THEME-Report-2018.pdf

As early as 1983, some researchers started to investigate and explore the factors that influence the performances of movie box-office. Litman (1983) has took the production cost, genre, release season and main actor's award story into consideration. After Litman's research, more and more experts tried to explore more factors which can affect movie success. Then, De Vany & Walls (1999) and Elberse (2007) mainly focused on how the star power can influence the performance of movies in the market. In recently years, with the arise of multiple social media, many researchers tried to gain the insight of the data from social media, like Mishne & Glance (2006) and Asur & Huberman (2010), they made their effort to predict the movie popularity by using the data from Twitter and web blog. Lee et al. (2018) believe that those researches have made a significant influence on the movie industry since they have provided some success guidelines to the producers and screenwriters, so that they can greatly reduce the risk of actual box-office does not meet the expected box-office, thus, they can reduce the failure of investment.

To sum up, those existing studies about the prediction of movie’ s popularity are not based on the current situation. Firstly, the databases they used as the training data are old, because the people’s purchase power in recently years is entirely different from the years before 2010. Secondly, the features those studies chosen are not enough, due to the development of computing power and more and more efficient methods appears these years, considering as much as possible features related to movie box-office could pose a positive influence on the prediction. Finally, the previously researches implemented different machine learning algorithms, such as regression analysis, regression tree, neural networks and support vector machine, so it is very necessary to make a comparison between those algorithms in order to choose the best model to predict the movie popularity more accuracy.

3. Research project

3.1 Significance

This project helps movie companies, producers and screenwriters to reduce the risk of loss after the film is released. If they refuse doing such analysis, there may cause a huge financial loss. For example, in 2013, a Korean movie, called Mr. Go, expected to attract at least 5 million people in Korean to go to the movies, because its record-breaking produce cost and investment, which reached nearly 20 million US dollars. However, there were only 1.5 million attendance until the movie quitted the market. Thus, the analysis of the project can provide some evidences for movie companies and investors, so that they can re-consider the production cost and the expected incomes to avoid the huge loss.

3.2 Innovations

The innovations in this project can be divided into three aspects. Firstly, the database this project used is the latest, the movies in the database are all published from 2017 to 2019. Considering that there are thousands of movies released every year, so we just choose some representative datasets. For example, the top 300 of the box-office in the three years must be used in the database, and for the movies which rating below 4.0 (full mark is 10.0), we will decide to randomly retrieve some data, we can set the number as a parameter to examine the best one. Then for the other movies in the middle of the rank, we can randomly get some data. For the data source, we will crawl data from different authoritative websites, like Internet Movie Database (IMDb), WorldwideBoxoffice, Rotten Tomatoes. Secondly, we will consider the related features as much as possible, which mentioned in the previously studies, what we will do is to combine those features. Thirdly, we will use the comparison methods for our project in order to find the best model to predict the movie popularity.

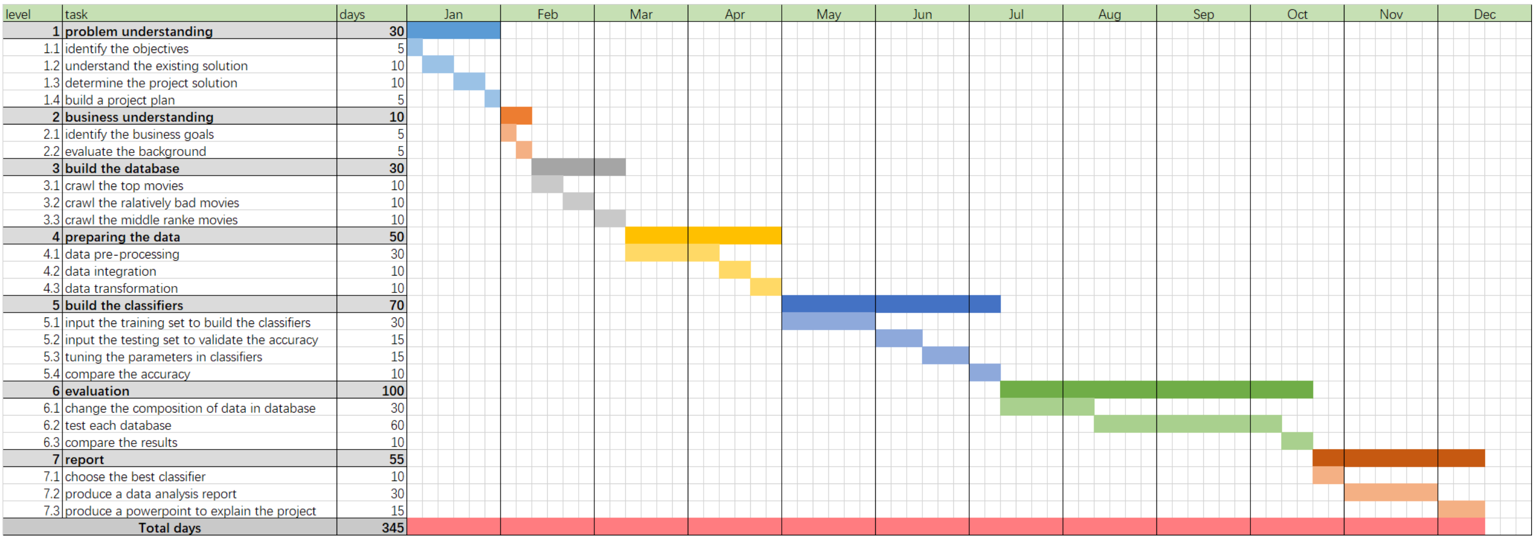
3.3 Outline

In order to build a model which can predict the popularity of the movies, we need to achieve the three objectives which mentioned in the first part.

- Firstly, we need to build the dataset, which can be regarded as the most time-consuming part but highly essential. As we explained, we will combine the top movies, relatively bad movies and middle rank movies together as the dataset, but the number of movies in each part need to treat as a parameter to examine the one who has the best performance.
- Then, we need to divide some data from the whole database as training data and testing data and input the training data into different algorithms.
- After training the data, we will get different classifiers, so input the testing data to validate the accuracy. After the process of tuning parameters, we will get many accuracy results with different number of data in dataset, parameters. What we need to do is to choose one to three relatively better classifiers as our final model, which can be used for the future prediction.

3.4 Gantt chart

In this project, the process includes problem understanding, business understanding, build database, preparing the data, build the classifiers, evaluation and produce the report. Besides, this project should be completed within one year, so the Gantt Chart below shows the whole project will finish at most 345 days.



level	task	days
1	**problem understanding**	**30**
1.1	identify the objectives	5
1.2	understand the existing solution	10
1.3	determine the project solution	10
1.4	build a project plan	5
2	**business understanding**	**10**
2.1	identify the business goals	5
2.2	evaluate the background	5
3	**build the database**	**30**
3.1	crawl the top movies	10
3.2	crawl the ralatively bad movies	10
3.3	crawl the middle ranke movies	10
4	**preparing the data**	**50**
4.1	data pre-processing	30
4.2	data integration	10
4.3	data transformation	10
5	**build the classifiers**	**70**
5.1	input the training set to build the classifiers	30
5.2	input the testing set to validate the accuracy	15
5.3	tuning the parameters in classifiers	15
5.4	compare the accuracy	10
6	**evaluation**	**100**
6.1	change the composition of data in database	30
6.2	test each database	60
6.3	compare the results	10
7	**report**	**55**
7.1	choose the best classifier	10
7.2	produce a data analysis report	30
7.3	produce a powerpoint to explain the project	15
	Total days	**345**

Figure 2: the timeline of this project

3.5 Expected outcomes

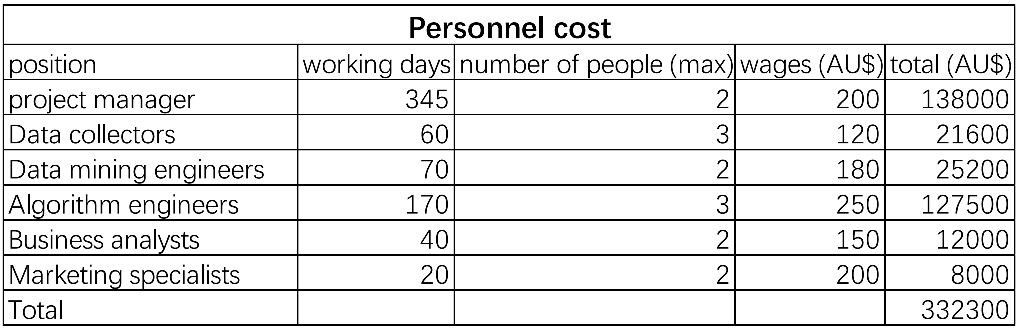
The purpose of this project is to predict the movie popularity, so the results we gain from the classifiers we built are the accuracy of the models. Due to the results will pose an essential influence on people's investment behaviours, so the ideal outcome of this project should be the higher accuracy of the prediction. The higher accuracy the model produces, the less risks of loss that movie companies will face.

4. Budget

In this project, the budget is mainly composed of the flowing parts.

- Personnel cost

The personnel cost will be regarded as the main part of all the cost, and very different position works for different periods and paid by different standards of wages, the estimated maximum cost of personnel shows below.

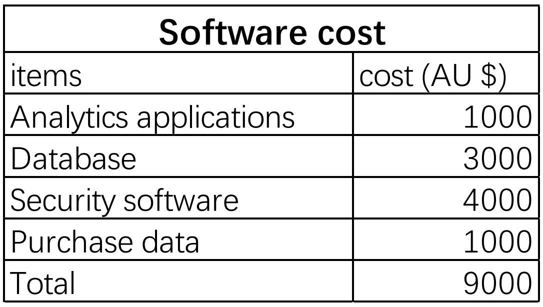


Personnel cost				
position	working days	number of people (max)	wages (AU$)	total (AU$)
project manager	345	2	200	138000
Data collectors	60	3	120	21600
Data mining engineers	70	2	180	25200
Algorithm engineers	170	3	250	127500
Business analysts	40	2	150	12000
Marketing specialists	20	2	200	8000
Total				332300

Figure 3: table of personnel cost

- Software cost

For the software we need in this project, we may use some professional data analysis tools, so such tools may require us to pay for them. Besides that, we will also need a new database to storage the movie data which are the latest data we may purchase. Consider this project is highly related to the interests of film companies, so we must buy some security software to protect the data lost. The final cost shows below.

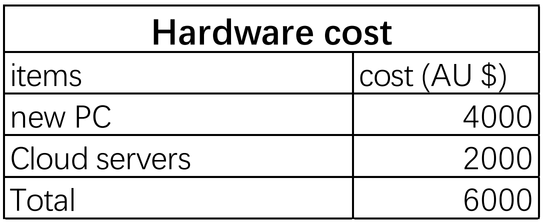


Software cost	
items	cost (AU $)
Analytics applications	1000
Database	3000
Security software	4000
Purchase data	1000
Total	9000

Figure 4: table of software cost

- Hardware cost

Talking to the hardware, the first thing we need is the computers, which can available for all the employees work at the same time. And we may need a cloud server to store all the related documents and files and it is convenience for us to share documents.



Hardware cost	
items	cost (AU $)
new PC	4000
Cloud servers	2000
Total	6000

Figure 5: table of hardware cost

5. Personnel

This section lists the employees required for complete this project and explains the responsibilities of each position.

5.1 Project manager

This position can be assigned to one or two people, the project manager is responsible for the entire process of this project, which can be regarded as the leader of the team. They must keep monitoring the other people's work to ensure each part can be finished timely and without any mistake.

5.2 Data collectors

They are responsible for the database building, what they need to do is to collect the data from different movie websites and store in the database after cleaning the data.

5.3 Data mining engineers

The data mining engineers have to finish the data processing part, they must gain the insight view of the data and visualize the data to ensure their teammates could understand the meaning of the data.

5.4 Algorithm engineers

They have to build the classifiers to train data and validate data, in order to make work more efficient, there could be three or four people in this position, and each of them is responsible for one algorithm.

5.5 Business analysts

The people in this position may cannot understand the programming and algorithms, but they perform well in business fields, so they can analysis the results and diagrams produced by engineers based on the business background. Finally, they need to produce a report about the project.

5.6 Marketing Specialists

They are people who directly facing the investors and clients, so it must finish the power point section and explains to the investors and clients.

References

Asur, S. & Huberman, B.A. 2010, 'Predicting the future with social media', IEEE Computer Society, pp. 492-9.

De Vany, A. & Walls, W.D.J.J.o.c.e. 1999, 'Uncertainty in the movie industry: Does star power reduce the terror of the box office?', vol. 23, no. 4, pp. 285-318.

Elberse, A.J.J.o.m. 2007, 'The power of stars: Do star actors drive the success of movies?', vol. 71, no. 4, pp. 102-20.

Lee, K., Park, J., Kim, I. & Choi, Y.J.I.S.F. 2018, 'Predicting movie success with machine learning techniques: ways to improve accuracy', vol. 20, no. 3, pp. 577-88.

Litman, B.R.J.T.J.o.P.C. 1983, 'Predicting success of theatrical movies: An empirical study', vol. 16, no. 4, pp. 159-75.

Mishne, G. & Glance, N.S. 2006, 'Predicting movie sales from blogger sentiment', pp. 155-8.

Video pitch

https://www.youtube.com/watch?v=tSkC67LYiNg&feature=youtu.be