Predicting a movie's box office and analysing it's key features

Karen Saksakulm, Silver Spitsõn

Business understanding

Business goals

As the movie industry is rather big and complex, it’s quite hard to hit a jackpot, so we believe that it would be important to analyse a variety of movies to find the key features that affect the movie’s box office. Our purpose is to predict the gross income by given attributes. We also find it important to uncover the main features affecting the movie gross. The business goal is to be able to combine as many positively affecting elements when making a new movie so it could do it’s best. We’d like to discover these features so that in five years there would be no movies from our company that would lose money or earn less than twice the budget.

Assessing the situation

We are using data from Kaggle, which consists of a movie list taken from IMDb. The available online resources for us to learn more about the movie industry are IMDb and Rotten Tomatoes. As the data we are using is free and available for everybody, it doesn’t force us to being very careful of others seeing it. Requirements wise, we do not have a very long time to complete this project and for us to consider this project finished, we should see at least a 80% prediction accuracy and a reasonably extensive analysis of the features.

A rather big risk is that after a lot of time and work put into this project we won’t reach the expected outcome and we will not be able to guarantee that our movies will be profitable. One of the contingencies is that one of our computers could break down during work without having been able to push our changes to the repository. One solution is just to push changes to the repository every time a change is made.

Terminology:

Budget – the price of making a movie

Company – the movie’s production company, who paid for making the movie and who earns the profit movie makes

Gross, box office – the total amount of revenue the movie generated, without having deducted any fees

Net profit – the actual profit after working expenses not included in the calculation of gross profit have been paid

Director – the producer of the movie, who stages the scenes

Genre – a category of artistic composition, characterized by similarities in form, style, or subject matter

Rating- a film's suitability for certain audiences based on its content

- R – restricted, under 17 requires accompanying parent or adult guardian, contains some adult material
- PG – parental guidance suggested, some material may not be suitable for children
- PG-13 – parents strongly cautioned, some material may be inappropriate for children under 13
- NC-17 – no children allowed, or more specifically no one 17 or under allowed
- G – general audiences, all ages admitted

Runtime – movie length in minutes

Released – movie released date

Score – IMDb users rating

Star – main actor, actress of the movie

Votes – number of user votes in IMDb

Year – year of the release

Writer – writer of the movie

We don’t have any costs related to this project. Cost-benefit plan is not necessary for us.

Defining data-mining criteria

The deliverables of the data-mining in our project are the predicted gross of movies and main features affecting the gross size. We will use different algorithms to find the most accurate one to predict the box offices. After the predicting and analysing our goal is to make a presentation so everyone else would understand our process to the final solution. Our technical criteria to support the outcome is to have a prediction accuracy of at least 80%. Our goal is to find two main criterias that affect the gross income the most.

Data understanding

Gathering data

For addressing the data-mining goals it is important to have a selection of attributes to predict the gross, so we could find two of them affecting the gross the most. We need a csv file to extract the data and use it to make predictions. The data is available for us and is already in csv format. We will be using a movie industry dataset from Kaggle (https://www.kaggle.com/danielgrijalvas/movies). The information has been gathered from IMDb.

We successfully imported the data into our Jupyter notebook file, and confirmed the integrity of it.

Describing data

The movie industry dataset includes 15 attributes for 6820 movies. The given attributes are budget, company, country, director, genre, gross, name, rating, released date, runtime, score, votes, star, writer, and year.

As we have a year column and a release date column, it seems reasonable to convert the date column to just day and month and the year would be separately.

We except the director and country might have the biggest effect on gross as we had a look on the data. It also seems logical that the score and star would affect the success of a movie.

Exploring data

There is some shortage in the data. Some of the movies are unrated, but this can still be used for analysis, for example, how much a rating even matters for a movie’s performance. There are few single ratings which will be mostly useless.

Some movies are missing the budget value, so it’s set to 0.0. It would probably be a good idea to replace this with NaN for clarity and easier handling during data manipulation.

Some movies have very few votes, so we might have to take that into consideration when working with the user score of a movie.

Verifying data quality

About 30% of the movies don’t have a budget so this could mean that we cannot use the budget field in our predictions or analysis. A possible solution is to try and gather these numbers from another source. If this turns out to be unsuccessful and we’re still left with too many NaN values, then we might have to consider dropping the feature completely.

Project plan

Id	Task	Member	Expected time
1	Business understanding	Karen, Silver	4 hours
2	Data understanding	Karen, Silver	4 hours
3	Data preparation	Karen	2 hours
4	Gathering new budget values	Silver	2 hours
5	Assessing a model to use	Silver	1 hour
6	Predicting gross values	Silver	3 hours
7	Finding accuracies and repeating the last task if needed	Silver	3 hours
8	Finding and analyzing the main attributes affecting the gross	Silver	5 hours
9	Making a model for movies to not make a loss	Silver	5 hours
10	Analyzing the highest-grossing movies and comparing them to the model made	Karen	6 hours
11	Making plots for presentation	Karen	2 hours
12	Making the presentation	Karen	5 hours
13	Presenting our project	Karen, Silver	3 hours