D10: KAGGLE - Movie Ratings

Authors: Kevin Kliimask, Jens Jäger, Taavi Eistre

Repository: https://github.com/kevinklmsk/ids2022

Dataset: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

Business understanding (Task 2)

Background

The global film industry is huge: it generates over $100 billion in revenue annually. Critically acclaimed movies usually tend to turn a larger profit than lower rated movies, so tools that help to determine what makes a movie successful are very much of value to the industry.

Business goals

A machine learning model that could accurately predict the audience rating of a movie would be very sought after by the movie industry. Past experiences and wisdom could be used to replicate successful movies. The model could predict the rating of the movie even before it leaves the drawing board. A movie that would be predicted to not be popular would not have to go into production at all, which would save the producers from a potential loss.

Business success criteria

For the model to be successful, it will have to accurately predict the outcomes of movie prototypes, where keywords, actors, genres etc. are the only information that exist of the movie. The model will have to take into account many features of a movie, even features that may seem unrelated to the success of the movie at first glance. The model must also predict movies with combinations of features that it may not have seen before, as many movies try to be as original as they can.

Inventory of resources

Our team consists of 3 coursemates who all have some experience regarding data science. We are all attending the course “Introduction to Data Science” and some of us have had experience with web scraping and handling the scraped data. In addition, we are proficient in using Python for data science related tasks. Furthermore, we possess a dataset that contains metadata on over 45,000 movies and has 26 million movie ratings from over 270,000 users.

Requirements, assumptions and constraints

The project’s deadline is 12th of December, which means there will not be enough time to finish a fully optimized model. Rather, we will finish a working proof of concept that shows our business proposal is feasible. Our dataset has a public domain license, meaning that we will not have to legally request access to the data. As our model will also take into account actors and directors, we would have to acquaint ourselves with data protection laws if we wished to commercialize our product.

Risks and contingencies

Our biggest concern is the time limit of the project. If we run into hurdles in our work, we will have to either obtain additional qualifications to better solve the problem, or we will have to simplify our solution. Which resolution we choose will depend on our available time. Work will be divided between our team members to better manage our time.

Terminology

Machine learning model - a program that can find patterns or make decisions from a previously unseen dataset.

Costs and benefits

There will be no costs as we will be doing free labor. Benefits of such a project would be huge for film production firms, given the fact that the model would be accurate in its predictions.

Data understanding (Task 3)

Outline data requirements

To address the data mining goals, the data must include movie ratings and metadata about the movies. The metadata should include features such as the cast, the production companies, the production countries, the genres, popularity, language and budget. The wider the metadata, the more features our model can take into account when predicting movie ratings. Other than that, there are no hard requirements for the data.

Verify data availability

After looking at the given dataset about the metadata, we could see that the majority of the data did exist. We did spot however that the budget column had a lot of zero values, which we were hoping to use in the selection criteria. Subsequently, we decided not to use budget in our model. We also saw that a lot of our data was in JSON format, which we will need to reformat. There were also quite a lot of movies without a production company or a production country, so we decided to cut these movies out.

Define selection criteria

For this model we will be using ‘movies_metadata.csv’ and ‘keywords.csv’ files with these specific columns:

- Keywords
- Genres (will be split up using pandas dummies or similar function)
- Original_language (will be split up using pandas dummies or similar function)
- Production_companies (will be split up using pandas dummies or similar function)
- Production_countries (will be split up using pandas dummies or similar function)
- Runtime
- Vote_average (our ratings)

Describing data

Our dataset contains information of 45431 unique movies. The description of our data:

- Keywords - json objects that show the keywords associated with the given movie
- Genres - json objects that show what genres fit the given movie
- Original_language - string value that shows the original language used in the given movie
- Production_companies - json objects that show the production companies that were involved in the making of the given movie
- Production_countries - json objects that show the countries in which the given movie was made
- Runtime - float value that shows the total runtime of the movie (in minutes)
- Vote_average - float value in the range of 0 to 10 that shows the rating of the given movie

Exploring data

- Keywords - There are 19953 different keywords. The top 3 keywords are “woman director” (3115), “independent film” (1930), “murder” (1308).
- Genres - The top 3 genres are “Thriller”, “Comedy” and “Drama”. There are 20 different genres.
- Original_language - The most popular language is English, with 71% of entries being in English, other languages make up 29% of the movies.
- Production_companies - The production company with the most movies was Warner Bros. (about 2% or 1250 movies), with Metro-Goldwyn-Mayer (about 2%) and Paramount Pictures (about 2%) following. There were a total of 23537 production companies.
- Production_countries - The United States of America had the most movies (about 45%), next was the United Kingdom (about 9%) and France (about 8%). Total there were 156 production countries.

- Runtime - 9% of the movies have a runtime under 60 minutes, 79% of the movies have a runtime between 60 and 120 minutes, the rest of the movies have a runtime larger than 120 minutes.
- Vote_average - The mean of ratings was around 5.6 with a minimum and maximum rating of 0 and 10.

Verifying data quality

We discovered that almost 3000 movies have a rating of 0, which potentially means that they are missing a rating. This suggests that we can’t use those rows for our data. We also found that there were a lot of production companies (23537), which would make our model really slow due to over 20000 new dummy features. This would also suggest not to use production companies in our model.

Project plan (Task 4)

We will proceed our work with the following tasks:

- Research machine learning models and determine models to try to use for our project. All of our members will spend about 5 hours each on this task.
- Polish and revise all of the skills and knowledge that we have obtained thus far during the course. All of our members will spend about 4 hours each on this task.
- Analyze the data. Try to find out how each feature might influence the movie's rating, so we can determine which features need to be taken more into account than others. Jens and Taavi will spend a total of 8 hours on this task.
- Split the data in two for training and testing. Develop the learning models. All of our members will spend 5 hours each on this task.
- Test the models on the selected testing data to figure out the accuracy of the model. All of our members will spend 5 hours each on this task.
- Create plots and point out of the most interesting findings while working on the task. Kevin will spend 4 hours on this task.
- Sum up all of the steps and create a project poster for the poster session. All of our members will spend 2 hours each on this task.

We will use the following Python data science packages for our work: Pandas, Numpy, Matplotlib, Scikit, Sklearn, Tensorflow. If the need arises, we will find other packages to support our work. We will try using machine learning and potentially also neural networks for predicting ratings with keywords. We will be using Github to submit our work.