Box office prediction

Jørgen Vikan Eidsvåg, Erik Fiksdal, Ulrik Fagerberg, Preben Brenne Johannessen

ML Gruppe 7

17.11.22

DESCRIBE THE PROBLEM

SCOPE

For this assignment, one of the suggested Kaggle Competition: ‘Box Office Prediction’ was selected as project. For this project we wanted to see if we could make a good estimation of the revenue for a given movie, given all the details around said movie, including details pertaining to post-release.

This Project was executed, primarily as an experiment, and has limited application in the real word.

A more realistic and applicable version of this project would utilize additional data to make a strong estimate of where resources should be focused in the developing of a movie. Things such as: how much is spent on advertising, what genres the movie encompasses, intended release window. Companies that produce movies would have an interest of seeing how profitable a movie could be, given key factors. Estimating a revenue before a movie is made is a hard ask though, as there are many factors, some of them hard to predict in and of themselves. One of the factors that matters could be who the cast is, who the director is, or if the movie is based on an already popular IP. And things might change unexpectedly: actors or directors might fall out of favour, the movie genre might already be saturated in the industry, real world events could make a movie extra relevant. Being experimental and mostly a proof of concept for machine learning in general, our project did not take these factors into consideration, but made predictions on the most indicative data we had at hand.

One of the key factors used in our model was the popularity of a movie. This ‘popularity’ is a number from TMDB ranging from 0 to 338 with no real explanation of how it has been made but it carries a significant correlation to revenue.

A scenario where we would like to predict the revenue of a movie, would be before its creation, yet popularity is a value that we will only get to generate some time after its release.

With this is mind, we need to recognize that this machine learning-model carries no genuine business impact, and only works as a proof of concept. In retrospect this is something we might have avoided using.

DATA

All data used in this project is pre-processed data from TMDB, which is imported in csv files. The data is consistent and with correct labelling before being imported. The data consisted of numbers, Json objects and string objects. The training data consists of 3000 movies, where these consists of corresponding labels. And the test data consists of 4811 movies, with the same labels as the training data except for the revenue column, in which we were tasked to create the estimate for this. For the most part, the data were consistent with what we expected, but in some cases the values were missing or misleading, so we had to create or edit these values to something that would more represent a more correct estimate. For example, there were some movies that had incorrect release years, but after closer inspection these were labelled as being released in 2034 instead of 1934. Some movies were also missing runtime. And by visualising the correlation between budget and the other labels we could see that there was a correlation between them, so we decided to give a runtime of the mean of the other movies. Some of the movies had a budget of 0, this would drastically skew the estimate of the revenue, therefore we decided to give these movies a budget of the mean of the other films. This would in turn give a more realistic estimate. Considering all the data that was in the datasets, we quickly saw that most of the labels would be useless for our purposes. Labels like the movie's homepage and the poster path were discarded for this model, and a lot of these labels were missing a more in-depth model might possibly use more of the labels, but for our model, we only decided to use the labels that gave us the best correlation values. These were the labels runtime, budget, popularity and two new labels we created. These two labels were crew_amount and cast_amount, from the correlation matrix, we could see that these were the labels with the biggest impact on the revenue. We can also note that some of the labels that we did not use possibly could be implemented in the future, because we assume that they could help improve the result, but as a proof of concept this model will suffice. Especially the label indicating the production companies would be interesting to research more into, it could also be interesting to investigate the impact of a A-list celebrity’s impact on the revenue.

MODELING

We tested 6 different models, and viewed their mean squared error to estimate the baseline results of each model. From this we could visualise that the mean squared error of Random Forest Regressor gave us the lowest result, I.e., lower error, more accurate results.

From there we were able to fine tune this model. To achieve this, we used Randomized Search CV to find and apply the best hyperparameters. From this we found our first model. We visualised the results in a bar chart and saw that our model was slightly lower than the actual result. From this we can argue that since a few movies will have a much bigger that average revenue compared to the model we trained, and the fact that movies have a bigger revenue today than they had 100 years ago, this model works to give a moderate estimate to the revenue of a movie.

We also investigated the feature importance's of the model and saw that only budget and popularity had a big impact on the revenue, 68% and 17.5% respectively, we could make a new model with only these features and see if this would make an impact on the revenue. We could also investigate the impact of the release year or some other of the unused labels or edit the budget to compensate for inflation.(if that is not currently done to the values.)

DEPLOYMENT / GOING FORWARD

The newest version of the model as well as the code that generates it, lies on GitHub. However, the newest version will not automatically be deployed from GitHub, because we don’t have CI testing implemented.

For now, the model has been deployed via Heroku. this is publicly available, and anyone can make use of it. We will need to make sure our intended users are the only ones to have access to the site in the future.

In theory, if the model were to be taken into production it would need to be changed so it does not base itself on the “popularity” metric in any way. This data is obviously only available after the movie is already made. It is strongly suggested to re-train the model a couple of times a year with up-to-date data to avoid any performance deterioration. This is on the assumption that batch learning remains the preferred approach.

Moving forward a few interesting areas could be explored, and potentially improve the model if desired. It is not much of a stretch to assume a correlation between genre and revenue. Broadly speaking, an action comedy is likely to generate higher revenues than a documentary could hope to. Likewise, it is likely worthwhile taking a closer look at the relations between the film production companies reputation and the revenue a film is likely to generate. Finally, we would like to suggest doing more research into the release year of a movie. We were not able to conclude much around its correlation with revenue and might even suspect that the results we observe in this area are not entirely correct, but it is an area that would be interesting to look further into.

REFERENCES

List sources you’ve used during the planning of the project. The list of references should indicate the feasibility of your project.

https://www.themoviedb.org/

https://www.kaggle.com/