Predicting the rating of a movie based on its duration, country, language, and director

Jordanović Aleksandra
Radojičić Filip

1. Motivation

Movies are one of the most famous means of entertainment. The widespread use of The Internet has led to large volumes of data related to movies being available online. In addition to good entertainment, today, the film is also a means of global business. Nowadays, in addition to quality, popularity is also very important, as well as a good rating of the film. So, this app is developed with the purpose to help to predict the success of the film depending on the duration, language, country, and director.

2. Research questions

The app should be able to predict the rating of a movie based on its duration, country, language, and director. A set of data from the IMBD site is available at the following link https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+ratings.csv. There are many pieces of information in this dataset, it contains various information about 85.855 films, but it is necessary to use data from the following columns::
duration - duration of a film,
country - country of origin of a film,
language - a language of a film,
director - director of a film.
In column weighted average is data that the app should predict.

3. Related work

We didn't find an article that tries to solve exactly the same problem, but there is one that solves a similar. In this article [1], authors solved predicting the genre and rating of a movie based on its synopsis. They use datasets from seven different websites. When it comes to preprocessing, movies that were missing information were removed. Also, they have split the dataset in proportion 80:20 for training and testing, respectively. They solve the problem by using CNN. But also, they performed numerous experiments using deep learning models and compare them with other solutions that solve the same problem using one of the popular traditional approaches such as SVMs and Random Forests. In this discussion, they say they concluded that deep learning based methods are much better than traditional.

4. Methodology

To solve this problem we use Random Forest Classifier. Also, we use RandomSearchCV to optimize parameters. Using RandomSearchCV we choose random parameters and evaluated them. After that, we fine-tune "best" parameters from the previous step using GridSearch.

When we talk about preprocessing since there is not much data with missing values, we are decided to remove them. Also, we use LabelEncoding to encode the values.

5. Discussion

First of all, we randomly split the initial dataset from Kaggle into two wholes. 20% of the dataset, i.e. 16 978 data we save in CSV file for the test, and the rest 80% we save in "Trains.csv".

That 80%, namely 68 877 data we randomly split for train and validation data set in proportion 70:30.

The first "best" parameters for RandomForest that we got by using RandomizeSearchCV in 10 iterations is*: 'n_estimators': 1000, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': 80, 'bootstrap': True*. The result that we get on the validation set by using RandomForest with these parameters was 0.30951462961150797 value of micro f1 metrics.

After that, we increased the number of iterations, and in 100 iterations we get the next parameters: *'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': 90, 'bootstrap': True*

With these parameters, we get a 0.34063998434289067 value of micro f1 metrics.

Finally, in 220 iteration we get a 0.35688423524806734 value of micro f1 metrics, with best parameters: *'n_estimators': 1000, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': 10, 'bootstrap': True.*

The next step was fine-tuning using GridSearch. With GridSearch we get that the best parameters for this model is: *'bootstrap': True, 'max_depth': 9, 'max_features': 'auto', 'min_samples_leaf': 5, 'min_samples_split': 4, 'n_estimators': 999*

So, the final result on the validation set is 0.3574713768470496 value of micro f1 metrics.

Using this final model, we predicted the test dataset and get a 0.35730457578646324 value of micro f1 metrics.

6. References

[1] Battu, V., Batchu, V., Gangula, R. R. R., Dakannagari, M. M. K. R., & Mamidi, R. (2018). Predicting the Genre and Rating of a Movie Based on its Synopsis. In Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation.