Mining for The Perfect Movie

Student Project Data Mining HWS17
Team 6

Presented by

Steffen Jung
Adrian Kochsiek
Martin Koller
Marvin Messenzehl
Daniel Szymkowiak

Submitted to the
Data and Web Science Group
Prof. Dr. Heiko Paulheim
University of Mannheim

October - December 2017

1 Problem Statement

In today's world, each movie production comes with a lot of uncertainty for all stakeholders involved. The production of a movie is usually very expensive and the success of a movie cannot be guaranteed. That is why producing a movie is a big risk for all investors. Since a lot of factores influence the success of a movie (e.g. main actors, storyline, setting of the movie...) it is not easy to determine upfront whether it is going to be lucrative. In order to solve this problem, Data Mining is a good technique. Therefore, main research focuses on building a solution, which will help stakeholders to predict the success of a movie:

- Will the movie be popular or will it be a flop? (based on revenue)
- Which factors contribute to a good movie?

2 Data Usage

To predict the revenue of a movie, a classifier has to be built based on a large set of movies. Therefore, the main focus will be placed on the movie dataset from Kaggle[1]. This dataset contains the metadata with 24 different features (e.g. budget, release date, genre and the classifier revenue) for about 45,000 movies.

Additional to the movie metadata, the cast and the crew of all the movies will be taken into account. During the project, it will be evaluated if other data sources are necessary. Other data sources might contain IMDb movie data or official box office data with visitor numbers and specific revenues.

3 Methodology

The measurement of success will be based upon the created revenue of the movie. Since the revenue is a continuous attribute it will be binned into several bins/classes. The concept of binning allows to predict the expected range of revenue for a new movie.

Since the metadata of the movies already contains 24 features and not all of them are relevant, irrelevant information is going to be discarded. In order to find out the most important features, an approach of different classifiers and set of features will be applied. Movies with missing features will be filtered out beforehand.

[1]The Movies Dataset by Rounak Banik. Can be found under `https://www.kaggle.com/rounakbanik/the-movies-dataset`

For the best prediction, different classifiers will be used in the dataset. Since Naive Bayes gives good results, despite the assumption of independence of the features, Naive Bayes will be taken as a baseline to compare the different classifiers.

Fruthermore classifiers like KNN, Decision Trees and Random forest will be used in this project work. For all classifiers, hyperparameter tuning will be performed, to receive the best possible results. Additionally, KNN with k from 1 to 10 and decision trees with different depths and split techniques will be applied. To compare the results of the different classifiers, it is planned to run several success measurement techniques under ten times cross-validation on the training dataset and draw a ROC curve.

4 Measurement of Success

After finding the best classifier on the dataset with all the parameters set, a train and test split will be performed on the data. The train data won't contain any test data. Factors computed are: the accuracy, recall, precision and the F1 score of the prediction of the test data. A high F1-score will show a high success.

5 Expected Results

The goal of this project work is to create an application, which gives the option to enter all known metadata of a new movie, in order to predict the range of revenue (and therefore the success) this movie will create. For such an application several fields of use are conceivable. This ranges from looking for hot new movies as a fan, making predictions as a market researcher or getting an overview as producer or director of a new movie.