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 usually means high upfront investment costs and the success of a movie cannot be guaranteed. That is why producing a movie is a risk for all investors. Since a lot of factors influence the success of a movie (e.g. main actors, storyline, setting of the movie, etc.) it is a challenge to determine upfront whether it is going to be lucrative. In order to solve this problem, Data Mining is a good technique. Therefore, the main research of this project 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 metadata with 24 different features (e.g. budget, release date, genre and the classifier revenue) for about 45,000 movies.

In addition to the movie metadata, the cast and the crew of all contained 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 of a movie will be based upon the generated revenue. 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.

After selecting and preprocessing the data, the dataset will be split into two stratified datasets. One dataset containing 70% of the data rows will be used to train the classifiers and tune hyperparameters (training set). The other data set containing 30% of the data rows will evaluate the project at the end (test set).

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

Since the metadata of the movies already contain 24 features and not all of them might be relevant, irrelevant information will be discarded. In order to find out the most important features, try and error of different classifiers and set of features will be applied. Movies with missing features will be filtered out beforehand. Naive guessing of the majority class will provide a baseline to be exceeded by each classifier in question.

Classifiers like Naive Bayes, k-NN, Decision Trees and Random forest will be considered in this project. For all classifiers expecting hyperparameters, hyperparameter tuning will be performed using grid search. Examples of targeted hyperparameters are the number of neighbours and distance measures in k-NN and the depths and split measures of Decision Trees. More hyperparameters will be considered during the project work.

It is planned to compare the average F1-score of the different classifiers under a stratified ten times cross-validation on the training dataset and draw the resulting ROC curves to show the performance of each classifier in a diagram.

4 Measurement of Success

Success of a classifier can be measured in mulitple ways. This project tries to predict the success of a movie as good as possible. Therefore the main focus is placed on the F1-score of the developed classifier. Finding and computing other performance measurements designed for multiclass classifiers will be part of the project work.

5 Expected Results

The goal of this project work is to create an application, which provides the functionality to enter all known metadata of a new movie, in order to predict the range of revenue (and therefore the success) this movie will generate. 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.