Mining for The Perfect Movie

Student Project Data Mining HWS17
Team 6

Starring

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. A lot of factors influence if a movie is going to be successful (e.g. main actors, storyline, setting of the movie...). These are too many factors, to easily tell the success of a movie. In order to solve this problem, Data Mining is necessary. The main research is focused on building a solution, which will help stakeholders to predict the success of a movie:

- Will be a movie good or will it be a flop? (based on revenue)
- What influences a good movie?

2 Data Usage

To predict the revenue of a movie, a classifier has to be built based on a big set of movies. Therefore a big focus will be placed on the Movie dataset from Kaggle. 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, also the cast and the crew of all the movies are going to be taken into account. During the project, it is going to 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 use to predict in which range the revenue of the new movie will be. 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.
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. We will work with classifiers like KNN, Naive Bayes, Decision Trees and Random forest.

On all the classifier some hyperparameter tuning will be performed, to get the best 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 ten times cross-validation on the dataset and draw a Roc-curve.

4 Measurement of Success

After finding the best classifier on the dataset with all the parameters set we will run a train and test split on the data. The train data won't contain any test data. We will compute the accuracy, recall, precision and the f1 score of the prediction of the test data and 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, make predictions as a market researcher or getting an overview as producer or director of a new movie.