The future success of movies

Project Work Team 6

presented by
Team Member Names Here

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

October 2017

Chapter 1

Outline

1.1 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 can not be guaranteed. Thatswhy 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 neccessary. The main research is focused on building a solution, which will help steakholders to predict the success of a movie:

- Will be a movie good or will it be a flop?

- Predict revenue of a movie

- What influences a good movie?

1.2 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 Data Set 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 a necessary.

- Movie Data from Kaggle

- https://www.kaggle.com/rounakbanik/the-movies-dataset

- IMDB movie data

- Box office data if neccessary

1.3 Methodology

The measurement of success will be based upon the created revenue of the movie. Since the revenue is a continiuous attribute it will be binned into 10 bins/classes. The concept of binning allows to later on 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, unrelevant information is going to be discard. In order to find out the most important features, an approach of different classifier and set of features will be applied. Movies with missing features will be filtered out beforehand.

For the best prediction, different classifier will be used on the data set. Since Naive Bayes gives good results, despite the assumption of independence of the features, Naive Bayes will be taken as Baseline to compare the different classifier. We will work with classifier 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 with different depths and split techniques will be applied (Entropy, Gini). To compare the results of the different classifier, it is planned to run a 10 times cross validiation on the dataset and draw a Roc-curve.

- Predict revenue / visitors for new movies in the cinema.

- Preprocessing and encoding

- Feature selections

- Binning revenue in different "classes"

- Train test split and cross validation

- Different classifiers

  - Knn
  - Naive Bayes (as baseline)

- **–** Decisions tree if possible
- **–** Random forest

- Support vector machines

- Parameter optimization

1.4 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.

- Different metrics

1.5 Expected Results

The goal of this project work is to create an application, which gives the 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.

- Put in a new movie with some data you have about it

- Get a prediction of the movie success

- Multilabel classifier