Mining the Success for Movies

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

Contents

Chapter 1

Application Area and Goals

This paper represents a documentation for the data mining project *"Mining the Success for Movies"*[1]. The structure of this paper follows the classical data mining process. Chapter 1 provides an overview of the problem the project is based on and is complemented by the goals and objectives of this project. Afterwards, chapter 2 deals with the structure and size of the data. Here, a closer look will be taken at the dataset at hand. Questions that had to be answered were for example which information were provided in the original dataset or which problems were identified concerning for instance outliers or missing values. Upon that, chapter 3 explains which preprocessing steps had to be taken in order to cleanse the dataset to prepare it for the data mining step and model learning. Chapter 4 describes which data mining techniques regarding algorithms and parameters were used to learn an expedient model in respect of the goals set in Chapter 1. Finally, chapter 5 closes this paper by describing which insights could be won for the problem at hand. Here, a critical reflection is delineated how the model could be improved further in order to provide even more precise results.

1.1 Problem Statement

Already well before new movies are being produced, every stakeholder certainly is interested in the monetary success of the given movie. In order to predict the success costly methods are being applied, such as market investigations or analyses.

The benefit which Data Mining brings to the analysis of large datasets, can also be transferred onto the stated problem of predicting a movie's success. Based on given data of already released movies and successful movies in the past, a model

[1] Information in this paper refers to the (sample) dataset and python scripts handed in for a classification problem

is being learned which shall be applied on upcoming or planned movies. In order to learn and apply the model various pieces of information are taken into account. Just a few are budget, revenues, runtime, genre and information on the release. Information on the dataset and on all preprocessing methods which were applied will be provided in chapter 2.

1.2 Goals

The goal of this project is to learn a model which will predict how successful a not yet released movie will be. This is done by using common data mining techniques in the Python programming language. As a main objective the question *"Based on revenue, will the movie be popular or will it be a flop?"* shall be answered for all possible combinations of information on a new movie as precisely as possible. In order to be as precisely as possible, not only different algorithms are being tested, but also parameter tuning is being applied with different performance measures [2].

[2] Further information on applied techniques and evaluation methods is provided in chapter 4

Chapter 2

Data Selection

2.1 Structure and size of data

The selected dataset onto which a classification model shall be learned is provided by Kaggle [1]. It is named *The movies Dataset*[2] and contains metadata on approximately 45,000 movies in its raw format. It is provided and updated by Rounik Banik. The complete dataset consits out of several files in the *.csv* format containing specific info about movie casts, metadata, and external scores. For the outcome of this project the central file, used for further preprocessing is named *movies-metadata.csv*. This csv-file holds 24 columns in total, which can be expected in the graphic below.

```
['adult', 'belongs_to_collection', 'budget', 'genres', 'homepage', 'id',
'imdb_id', 'original_language', 'original_title', 'overview', 'popularity',
'poster_path', 'production_companies', 'production_countries',
'release_date', 'revenue', 'runtime', 'spoken_languages', 'status',
'tagline', 'title', 'video', 'vote_average', 'vote_count']
```

Figure 2.1: Columns of the file *movies-metadata.csv*

The structure of the individual attributes is very different. In addition to boolean values, strings and numeric float values (e.g *budget* or *runtime*), many attributes contain longer texts (e.g *overview*), arrays or even a list of JSON objects (e.g.*production-countries*)These different formats must be taken into account in the later preprocessing step and need to be processed individually, so that a well-functioning classification model can be worked out.

[1] 2017 Kaggle Inc

[2] Link to the dataset: https://www.kaggle.com/rounakbanik/the-movies-dataset

2.2 Basic data exploration

If you take a closer look at the data it can be noticed that the overall quality varies significantly. This can be attributed to the fact that the record is maintained by a community and is not provided by a larger organization or company. Therefore, a demanding quality assurance process is difficult to realize. One quality issue and special case is for example the encoding of string attributes of Asian movies and the Chinese and Japanese character set. Another aspect of importance is the absence of a lot of values. Especially for older movies (before 1960) there are just a few information available. This lack of information needs to be taken into account for the further steps. The most critical effect has the absence of values to numerical attributes like *revenue* or *budget*, which are also critical for the success of a classification model with respect to financial interests. Here around 34000 records are containing a zero or nothing in either the revenue or the budget column, what decimates the dataset heavily.

In addition to that, there is no information about the currencies of financial attributes. Complicated because later computations will rely heavily on revenue and budget data. Average budget value, average revenue (without zero). Correlation between revenue and budget. The most common genre is... . Average number of actors per movie.

Back to revenue and budget take into account that this is data from past 60 years. A lot of things changed in movie economics (prices, consumer Verhalten, globalization). Introduced new column, productivity value A lot of values are scaled wrongly. Little example with 2 movies. 2000000 in revenue is just stated as 2, don't know the scale of data. Will be a big problem

Main issue quality of data. Further explained in preprocessing...

Chapter 3

Preprocessing and Transformation

In order to pick out columns that have a significant impact on forecasting a movie's success, the following assumptions concerning the importance of information in the metadata were made:

- The budget and revenue are crucial numbers.
- The release year has an impact on numbers such as budget and revenue.
- The genre, the production company and the runtime are important information.
- Transform data into a representation that is suitable for the chosen data mining methods
  - number of dimensions
  - scales of attributes (nominal, ordinal, numeric)
  - amount of data (determines hardware requirements)
- Methods
  - Aggregation, sampling
  - Dimensionality reduction / feature subset selection
  - Attribute transformation / text to term vector
  - Discretization and binarization
- Good data preparation is key to producing valid and reliable models

- Data preparation estimated to take 70-80% of the time and effort of a data mining project!

3.1 Preprocessing steps according to Python script

3.2 A list of problems we encountered

1. **list further problems we had and solved!**
2. Prod. Comp.: Same prod. company named differently -¿ using Regex to solve (Steffen)
3. dataset: 5 datasets have duplicates

Chapter 4

Data Mining

- Input: Preprocessed Data
- Output: Model / Patterns

1. Apply data mining method
2. Evaluate resulting model / patterns (using P, R, F1, not accuracy)
3. Iterate:
    - Experiment with different parameter settings
    - Experiment with different alternative methods Improve preprocessing and feature generation Combine different methods

4.1 Algorithms

- **Random Forest**
- **Decision Tree**
- KNN
- Bayes
- NeuralNet
- svc

4.2 Three best performing algorithms

- Pick best three algos
- GridSearch
- Why does each classifier perform how it performs (unausgeglichene Klassen, ...)?

Chapter 5

Interpretation / Evaluation

Maybe also prospect?

- Output of Data Mining
  - Patterns
  - Models
- In the end, we want to derive value from that, e.g.,
  - gain knowledge
  - make better decisions
  - increase revenue