

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

1	Application Area and Goals	1
1.1	Problem Statement	1
1.2	Goals	2
2	Data Selection	3
2.1	Structure and size of data	3
2.2	Basic data exploration	3
3	Preprocessing and Transformation	5
3.1	Preprocessing steps according to Python script	8
3.2	A list of problems we encountered	8
4	Data Mining	9
4.1	Algorithms	9
4.2	Three best performing algorithms	10
5	Interpretation / Evaluation	11

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 deep analysis and .

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

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

given data of already released movies and successful movies in the past, a model 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.

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

Chapter 2

Data Selection

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 consists out of several files in the .csv format containing specific info about movie casts, and external score. 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']
```

Untertitel bla bla

2.1 Structure and size of data

2.2 Basic data exploration

- In slides named: "structure and size of data"
- min. 1 Page
- Selection:

¹2017 Kaggle Inc

²Link to the dataset: <https://www.kaggle.com/rounakbanik/the-movies-dataset>

- What data is available?
 - What do I know about the provenance of the data?
 - What do I know about the quality of the data?
 - Wrong values, lot of null values
- Exploration
 - Get an initial understanding of the data
 - Calculate basic summarization statistics
 - Visualize the data
 - Identify data problems such as outliers, missing values, duplicate records
 - problem with number scales, data formats, truth of contents

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

Chapter 3

Preprocessing and Transformation

In total, two datasets were used to create features for testing the different classifiers: *movies_metadata.csv* and *credits.csv*. The dataset *movies_metadata.csv* contains 45,463 rows and 23 columns excluding the id-column. The dataset *credits.csv* contains 45,463 rows and 2 columns excluding the id-column. Based on the assumptions that budget and revenue were crucial numbers, the release year has an impact on those numbers and the genre, the production country such as production company, the spoken languages, the runtime and the fact whether a movie belongs to a collection or not are further important information, all others columns were dropped out from the *movies_metadata.csv*. **Nicht gerade wissenschaftlich hier** Figure 3.1 gives an overview which information was kept.

The reason why information, which could have been potentially interesting, had to be dropped, was mainly for time reasons. Preprocessing took about 70% of the timeperiod¹ of the whole project. Thus, the team was able to focus on preprocessing of mentioned columns. Still, chapter 5 provides a prospect, which steps were possible, if a larger timeframe was dedicated to this project.

After dropping out information, eleven columns remained. Combined with the two columns from the *credits.csv* dataset, thirteen columns were used as a basis to create features for finding the best performing classifiers.

In order to transform the data into a suitable representation for forecasting a movie's success, preprocessing was mandatory. All preprocessing steps included either the following:

- Merging of columns

¹Mainly due to the fact that heaps of problems arose from the dataset, which can be read in chapter 2.

adult	False
belongs_to_collection	{'id': 10194, 'name': 'Toy Story Collection', ...}
budget	300000000
genres	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...}
homepage	http://toystory.disney.com/toy-story
imdb_id	tt0114709
original_language	en
original_title	Toy Story
overview	Led by Woody, Andy's toys live happily in his ...
popularity	21.9469
poster_path	/rhIRbceoE9lR4veEXuwCC2wARtG.jpg
production_companies	[{'name': 'Pixar Animation Studios', 'id': 3}]
production_countries	[{'iso_3166_1': 'US', 'name': 'United States o...}
release_date	1995-10-30
revenue	373554033
runtime	81
spoken_languages	[{'iso_639_1': 'en', 'name': 'English'}]
status	Released
tagline	NaN
title	Toy Story
video	False
vote_average	7.7
vote_count	5415

Figure 3.1: Dropped columns of *movies_metadata.csv*. All kept columns are marked in yellow

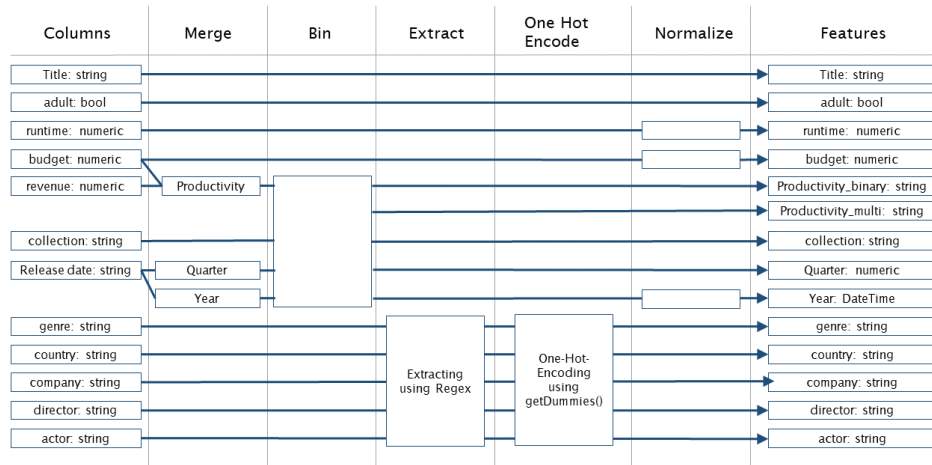


Figure 3.2: Features created during preprocessing

- Binning of features
- Extracting information out of columns
- One hot encoding
- Normalizing
- 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