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

List of Algorithms

List of Figures

List of Tables

Chapter 1

Application Area and Goals

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. *Do we need a source? koennte noch ausgefuehrt werden.*

The benefit 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, 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 Python. 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 tested, but also parameter tuning is applied with different performance measures [1].

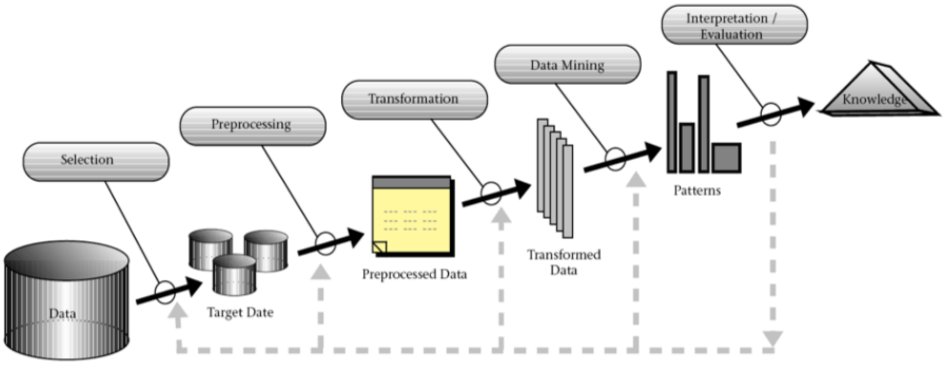
- Problem Statement and idea behind the project

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

- General introduction similar to Project outline

1.3 Theoretical framework

- keep it small
- roughly 1 Page



Chapter 2

Data Selection

- In slides named: ”structure and size of data”
- min. 1 Page
- Selection:
  - What data is available?
  - What do I know about the provenance of the data?
  - What do I know about the quality of the data?
- 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

Chapter 3

Preprocessing and Transformation

- 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 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

Chapter 5

Interpretation / Evaluation

- 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

Bibliography