

TDS 3301 DATA MINING

ASSIGNMENT PART 1

IMDB MOVIE ANALYSIS

|  |  |
| --- | --- |
| **GROUP MEMBERS** | **STUDENT ID** |
| LIYA SAFFURA | 1132702377 |
| AMIR RIDHWAN | 1132701767 |
| SURAYA IBRAHIM SHAH | 1151303737 |
| ILI FADHILAH AHMAD HIZZAD | 1151303720 |

Contents

Introduction 3

Data Dictionary 4

Insights 5

Data Mining Techniques 6

Data Quality Issues 7

References 8

Introduction

Movies have been the dominant form of popular art for the last 100 years. The motion-picture industry has been growing as many new movies are being released almost every single day.

The dataset that we have chosen is about movies given the name ‘IMDb 5000 Movies’.  It contains a total number of 5044 movies with 28 variables. The variables include “color”, “director\_name”, “num\_critic\_for\_reviews”, “duration”, “director\_facebook\_likes”, “actor\_3\_facebook\_likes”, “actor\_2\_name”, “actor\_1\_facebook\_likes”, “gross”, “genres”, “actor\_1\_name”, “movie\_title”,” num\_voted\_users”, “cast\_total\_facebook\_likes”, “actor\_3\_name”, “facenumber\_in\_poster”, “plot\_keywords”, “movie\_imdb\_link”, “num\_user\_for\_reviews”, “language”, “country”, “content\_rating”, “budget”, “title\_year”, “actor\_2\_facebook\_likes”, “imdb\_score”, “aspect\_ratio”, “movie\_facebook\_likes”. We downloaded this dataset from a Kaggler with the username chuansun76. He scraped the movies data from the IMDb website using a Python library called “scrapy”. IMDb, which stands for Internet Movie Database, is a well-known website that contains information of a huge number movies.

The reason why we chose this dataset is because we wanted to find out how we can determine a good movie before it is released without relying on reviews by other people or by intuition. We also want to find the latest trends in the movie industry.

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Color | The movie is in color or black and white |
| director\_name | Name of the movie director |
| num\_critic\_for\_reviews | The number of reviews of the movie given by critics |
| duration | Duration of the movie. |
| director\_facebook\_likes | Likes on the director’s facebook page. |
| actor\_3\_facebook\_likes | Likes on the actor3’s facebook page. |
| actor\_2\_name | Name of the second leading actor. |
| actor\_1\_facebook\_likes | Likes on the actor3’s facebook page. |
| Gross | The movie’s gross income. |
| genres | The movie’s genre. |
| actor\_1\_name | Name of the leading actor. |
| movie\_title | The title of the movie. |
| num\_voted\_users | Number of users who voted on the movie at the IMDb website. |
| cast\_total\_facebook\_likes | Total number of likes on all of the movie’s actors’ facebook pages. |
| actor\_3\_name | Name of the third leading actor. |
| facenumber\_in\_poster | Number of faces in the movie poster. |
| plot\_keywords | Keywords for the plot. |
| movie\_imdb\_link | The movie’s link in imdb. |
| num\_user\_for\_reviews | The number of reviews of the movie given by IMDb users. |
| language | Language used in the movie. |
| country | Origin of movie. |
| content\_rating | Suitability of a movie’s content to it’s audience. |
| budget | The budget cost for the movie. |
| title\_year | Year the movie was released. |
| actor\_2\_facebook\_likes | Likes on the actor3’s facebook page. |
| imdb\_score | Score given by IMDb users. Ranges from 0 to 10. |
| aspect\_ratio | Aspect ratio of the movie. |
| movie\_facebook\_likes | Likes on the movie’s facebook page. |

Data Dictionary

INSIGHTS

        After mining this dataset, there are a number of possible insights that may be beneficial. We are hoping to find some patterns by analysing this dataset that will lead us to these insights. Some of them include:

1. What kind of movies are good and what kinds are bad.
2. Movies with a certain genre has a higher rating than other genres.
3. Certain movie genre brings more profit than the other genres.
4. Facebook helps by making the actors and movies known and therefore increasing their profit.
5. The popularity of a movie is also affected by the total number of faces on it’s posters.
6. The profit of a movie correlates to the budget of a movie.
7. The number of total casts in a movie affects the popularity of the movie.

DATA MINING TECHNIQUES

The data mining technique that we would want to implement to this dataset is classification.  Classification method is suitable because it is easier to see a pattern of a known variable if it is put together under the same class. For example, movie that is based on the same genre may have same profit or imdb score. There are many kinds of classes that we can make from the dataset such as genre, number of likes on Facebook page, and IMDb Score.

Another technique that we may use is Association Rule Mining. We may use the plot keywords to find out which keywords usually go with one another and relate with the movie score to see if a movie with a group of keywords is a good movie.

DATA QUALITY ISSUES

We have discovered some data quality issues in this dataset that might cause some problems in our data mining tasks.

This dataset is not clean as it contains a lot of missing values as well as misspelling errors. Missing values come from unavailable data. For example, there are some directors who do not have a Facebook page. Hence, the value is set to 0.

It also contains a few columns that are not really important or beneficial in discovering patterns. The gross and budget column is inaccurate as all of them are not conform to inflation. Inflation rate changes from time to time and therefore causing a less accurate result.

There are also some outliers in this dataset. Other than that, the columns are messily arranged. Also, some data are incorrect while some are not up-to-date which may lead to inaccuracy of our analysis.

There are some columns in the dataset that we think are not relevant and necessary to our analysis. For example, “movie\_imdb\_link”.

References

Dataset - <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

Customer Price Index - https://www.minneapolisfed.org/community/teaching-aids/cpi-calculator-information/consumer-price-index-1800