DATA SCIENCE FOR BUSINESS ANALYTICS

IMDB MOVIE RATING PREDICTION

Shixin Li

Kashyap Uppuluri

Hao

Rohit Wankhede

Instructor: Prof. Brian D’Alessandro

Introduction:

Movies have always formed an integral part of recreation and culture amongst human beings. The worldwide market for movies is huge and is a major source of revenue and income as well. According to recent estimates, the global box office returns for movies released all over the world is around $ 39.4 billion.Close to 5000 movies were released the world over. Most movies have a relatively high value of production cost, and hence it is imperative to come up with ways and means to maximize the revenue for every movie released. One way of doing this is to try and gauge public sentiment about a movie before its actual release. This sort of feedback can help producers to get a sense of the kind of response their movie will get. Before the advent of the Information age this sort of feedback was gained through conventional wisdom and empirical results. In today’s times there is plenty of scope for the use of data science to solve such problems.

This paper details the analysis of the IMDB dataset, a dataset which contains records of movies from all over the world along with information associated with every movie such as title, genre, gross income, language, country of origin. Based on votes provided by users, a rating is assigned to every movie.

The goal of this paper is to identify the correlation between the various attributes of a movie with the movie rating. For example, is there any correlation between the budget of a movie with its rating i.e. does a high budget movie have a higher rating than a low budget movie. Also whether movies from a particular genre are more popular than others. This can then be used to predict the rating of a movie, if it were to be listed on the IMDB website.

IMDB

Launched in 1990, as of November 2015 IMDB has approximately 3.5 million titles (including episodes) and 6.8 million personalities in its database. There are close to 64 million registered users on IMDB.IMDB registered users can rate every movie in the website (on a scale of 1 to 10). Users can rate a movie how many ever times they want, but each rating will overwrite the previous one.

Data

The dataset for the paper has been taken from the IMDB website. The data was scattered over different files. One of the initial data engineering tasks was to collate the data in one file.

The dataset consists of various attributes (around 49) for every movie. Some of these features are redundant or beyond the scope of this paper, hence the first task was to select certain relevant features of the many attributes available.

The following information is available related to every movie: actor ,actresses, countries , complete-cast , cinematographers, costume designers , budget , votes , genre , keywords , language, locations, ratings , running – time, release – date etc.

As the information is scattered over different files, on identifying the relevant and feasible features the first task is to form a single data frame consisting of the movie titles and the relevant features.

Some of the features such as alternate-versions, biographies, crazy-credits, goofs are being dropped on the basis that they are in the form of full text and will require different sort of natural language processing techniques which are beyond the scope of this project.

Some features such as name, title, costume designers etc are being dropped as features on the basis that it is obvious that they have no contribution towards the rating of a movie.

After a substantive study of the relevance and feasibility of implementation the features selected are:

1. No of Votes
2. Release Year
3. Budget
4. Gross Income
5. Genre
6. Country

The final data frame consists of around 33,000 rows along with their respective features.

From the selected features, some required further engineering in order to make them suitable to be processed by one of the prediction models.

For e.g. The genre feature, for every movie was in the text format. In order to make this feature usable by a model, this feature was converted into a numeric value.

i.e. For every record a ‘1’ or ‘0’ was assigned on the basis of the genre of the movie.

For the ‘Budget’ column, as the values for different movies from various countries was in different currency formats. All values were converted to USD ($) for the sake of standardization.

For the Release year, as the values for the rest of the columns such as genre and country are binary , the data was normalized such that its value lies between 0-1.A larger value indicates that the movie is more new whilst , a smaller movie indicates that the movie is relatively older.

For e.g. For a movie with release year 2001 , normalized value = 0.87 and for a movie with release year of 1924 = 0.27

For country of origin as well, the data has been normalized such that movies with country ‘USA’ have been given the value, while other movies are given the value zero for the ‘ Country’ column.

Dataframe after engineering:

Feature Importance

In order to get a sense of the importance of features which regards to predicting the rating of a movie, we can plot a graph for feature importances.

These feature importances are assigned on the basis that which features increase entropy to the maximum and contribute to maximizing information content.

Modelling

The following models are implemented and compared on the basis of various metrics discussed in the further sections.

1. Decision Tree
2. SVM
3. Naïve Bayes
4. k – Nearest Neighbors
5. Logistic Regression

Model Selection

Results

Deployment

Future Scope

References