Movie Success Predictor

Sai Shruthi Madhuri Kara
NC State University
Raleigh 27606
skara2@ncsu.edu

Sruthi Talluri
NC State University
Raleigh 27606
stallur2@ncsu.edu

Sameer Thumallapally
NC State University
Raleigh 27606
vthumma@ncsu.edu

Sai Santosh Balusu
NC State University
Raleigh 27606
sbalusu@ncsu.edu

ABSTRACT

To predict a success of a movie with the help of existing data of movies, and their reviews. Create and analyze the predictions and the useful data in determining the success.

KEYWORDS

datasets, data mining, sentiment analysis, text tagging, naive bayes, exploratory data analysis, one hot encoding, multi label binarization, feature engineering

ACM Reference Format:
Sai Shruthi Madhuri Kara, Sruthi Talluri, Sameer Thumallapally, and Sai Santosh Balusu. 2020. Movie Success Predictor. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND BACKGROUND

1.1 Problem Statement

The goal of our project is to predict a movie's success or failure by analyzing existing movies database with its reviews and data related to the cast and crew associated and their movies in the past. We will be taking both the data regarding the movie such as cast, crew, rating, genre etc. and also reviews associated with the movie given by critics. We will then compute the accuracy of our predictions and try various modifications by removing few attributes to see the impact of them.

1.2 Related Work

An algorithm will be designed to perform the analysis on the reviews attribute of the movies dataset to discretize the content into a binary attribute(positive or negative), and to obtain a generalized opinion associated with the movie it is then stored as a new attribute added to the movies data. The movies data then will be pre processed and a KNN classifier will be performed to categorize the movies. The latest movies to be predicted will be missing the rating attributes, which can not be used, hence we will be using regression model to predict an approximate value.
The idea of the project is to determine how well the model performs in case a predicted rating value for the movie is used instead of the true value, and also determine how addition of review attribute, or others to the movies helps in improving the accuracy.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 METHOD

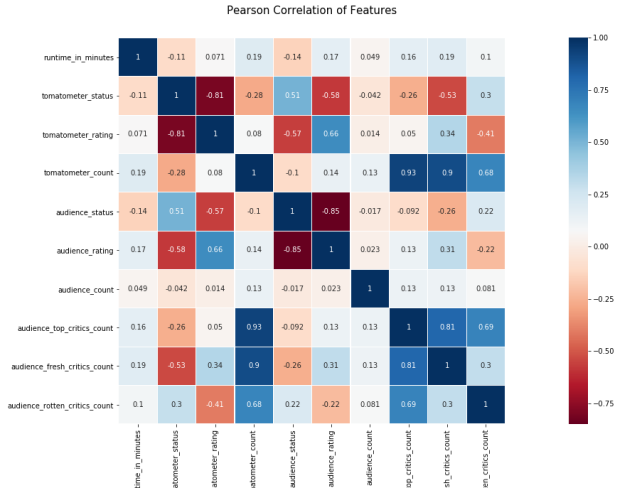
2.1 Approach

(1) Naive Bayes Classifier
The sentiment analyzer we implement to categorize the movie reviews trains on a Naive Bayes Classifier. It is a probabilistic learning method based on Bayes theorem and follows a supervised learning approach. The goal of any probabilistic classifier is to analyze all the features and all the classes and to determine the probability of the features occurring in each class, and to return the most likely class.
In our application, the features correspond to the words and the class corresponds to positive or negative. Unlike other classification models, Naive Bayes requires very little training. In sentiment analysis, when a new input is given to the trained model, it simply analyzes the probability of every word in the input (review) and picks the corresponding class based on the cumulative probability of all the words in the input.

(2) One-hot Encoding and Multi Label Binarization
In the data we had a few categorical attributes such as Genre and Rating. We had to encode the rating attribute into numerical values as it was categorical data and not all the models can work accurately on categorical data. One idea was to import Label encoder library and encode the vales but the problem with label encoding is that it assumes higher the categorical value, better the category and in the given scenario rating attribute is not an ordinal one. So we used another approach called One-hot encoding on attribute such as Rating and Genre.
One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For the Genre attribute we have used MultiLabelBinarizer module as it was a multi valued attribute. MultiLabelBinarizer takes an enumerable list and turns it into columns with binary values

that represent the list. So for the genre attribute we have first formatted it to be in the form of a list rather than string, transformed it into a series of columns with binary values and finally ordered them alphabetically.

(3) Removal of Attributes
We have used the method of determining the correlation between attributes. To gather the data of the inessential attributes and remove them. The Correlation matrix of the data is as follows:



From the above correlation matrix between attributes of the movie data, it is evident that the attributes tomato-meter status and audience status are highly correlated with the data attribute tomato-meter rating and audience rating respectively. As a result we have removed the attributes tomato-meter status and audience status from our data frame. This has helped us in reducing the dimensions in our data.

2.2 Rationale

The idea behind the using the Naive bayes classifier is to be easily categorize the reviews(each movie had an average around 56 reviews) for each movie as it takes linear time to train the classifier. Another important factor was to be able to combine both the data of the movies with its reviews as a single dataframe. Hence using the classifier algorithm and determining the scores for each movie with the consideration of Top Critic attribute in the reviews was important.
The One hot encoding of few of the attributes helped us in converting the data that can easily be processed into the models for the movie success or failure predictions.
The correlation matrix also helped us in reducing the data dimensions successfully.
Another important task was to convert the data regarding the cast, and directors for each. movie into numerical or categorical values to be processed into the model. Here, we have tried various approaches and finally using the data about the number of movies each of them were involved in we converted them into meaningful numerical values for the attributes.
Hence using the above approaches we were able to combine the data of movies and their corresponding reviews and pre-process the data attributes to be suitable input to the model.

3 EXPERIMENT

3.1 Data Sets

(1) Rotten Tomatoes Data Set
(a) Rotten Tomatoes is one of the most popular film websites, which combines movie information, critic reviews and users reviews.
(b) Movies are divided in three categories according to the critics reviews and in two categories according to the users reviews:
Critics reviews categories:
Certified fresh: at least 75% of critics reviews are positive and 5 reviews come from top critics
Fresh: at least 60% of the critics reviews are positive
Rotten: less than 60% of the critics reviews are positive
Users reviews categories:
Upright: at least 60% of the users reviews are positive
Spilled: less than 60% of the users reviews are positive
All the records have been scraped as of 07/11/2019.
(c) The movie dataset includes 16,638 movies with attributes such as movie description, critic consensus, rating, genre, cast, and all the Rotten Tomatoes scores. The critics reviews dataset includes 930,942 reviews from critics with attributes such as critic publication, critic icon, and review content.
(d) Data has been scraped from the publicly available website https://www.rottentomatoes.com.

(2) IMDB 5K Movie Facebook Likes Data Set
(a) IMDB data set containing 5K movie data which includes 28 attributes related to the movie information. The attributes which are under consideration for the experiment of this project are actors and their Facebook likes and directors along with the Facebook likes.
(b) Data has been scraped from the publicly available, a large data set of informal movie reviews from the Internet Movie Database (IMDB).

(3) IMDB Dataset of 50K Movie Reviews
(a) IMDB dataset having 50K movie reviews for natural language processing or Text analysis. This is a dataset for binary sentiment classification containing substantially more data than previous benchmark data-sets.
(b) The data consists a set of 25,000 highly polar movie reviews for training and 25,000 for testing. So, predict the number of positive and negative reviews using either classification or deep learning algorithms.
(c) Data has been scraped from the publicly available, a large dataset of informal movie reviews from the Internet Movie Database (IMDB).

3.2 Hypothesis

The following are the questions that are investigated before the experiments are run:

(1) Cast - Directors Attributes
   (a) How do we convert the nominal attributes like cast and directors into numeric format to use in the model for generating precise results?
   (b) How can we tackle the count for the cast members as for each movie there can be supporting cast apart from the lead actors?
   (c) For some movies there can be more than one director working. How to handle and convert this kind of format?

(2) Sentiment Analysis
   (a) How do we tag each various number of reviews for each movie, depending upon the sentiment expressed in them?
   (b) The reviews for each movie vary in number and a common trend in them is to be identified and attached to each movie. What is the approach that can be followed for the same?
   (c) Some of movie reviews are provided by the Top critics, how do we take this factor into consideration while accessing each review for the movie?
   (d) How should we train the system to be able to analyse the movie reviews?

3.3 Experimental Design

(1) Cast-Directors Attribute
The questions for these attributes circle around the crucial fact that the attributes cast and directors are non-numeric or nominal attributes. These cannot be used directly for the model training as it will not be supported and produce superior results. Some movies can have a count of cast as 2 while the others can have a count as high as of 20. This creates a huge range and variation. The same is the case with the director attribute for the movies.
The format of these attributes is a continuous string of names separated by commas. There has to be a processing phase to be able to separate or enumerate these values.
The following are the approaches that we have experimented with, for pre-processing of these type of attributes:

   (a) Approach 1:
   We have found a supporting data set(IMDB 5K movie Facebook likes data set). Our main idea was to find the popularity of the actors based on the Facebook likes. We have analysed and compared that data set with our data set. The supporting data set consists of three columns for the actors namely, actor1_name,actor2_name,actor3_name along with the Facebook likes of each of the actors namely actor1_facebook_likes, actor2_facebook_likes, actor3_facebook_likes.
   We have created a dataframe and merged all the actors into a single column mapping them with their Facebook likes. Then we have summed up the Facebook likes of all the people in the cast in the main dataset using this supporting dataframe and replaced the non-numeric data with this count. This way we get the popularity factor of the cast. The same process was repeated for the director attribute as well.

   (b) Approach 2:
   Another approach we have experimented with is that we would find the count of movies that the cast have done and the directors have directed. This way we will know the popularity of the cast/ director. This information is maintained in a dataframe. For each cast/director in the corresponding attribute of a movie, we have replaced with sum of frequency of the all cast/director multiplied by a factor of 10. This way the entire column is processed.

(2) Sentiment Analysis
The movie reviews for each movie are to be processed and attached to the movie data, and be formatted to be an input to the model for further analysis.
The format of data consists of all the movie reviews with a movie link an unique identifier for each movie. The reviews data initially had to be grouped for each movie as a processing phase.
The following is the approach we have experimented with to generate the results for the movie review data:

Approach:
We have found a supporting data set(IMDB Dataset of 50K Movie Reviews). We used the data containing the Reviews with a positive and negative tag to train the Naive Bayes classifier. The data helped the classifier to analyze the reviews sentiment and was useful as it also was done for the reviews which would contain similar words and sentence formations.
We then used the classifier to classify our data into positive and negative reviews, we later used the attribute of Top Critic for the reviews to give a weighted sum to the total value. The positive value increases the score by 1 , negative decreases the score by 1, but if it contains Top Critic both increase and decrease is done by 2. We determined the value of the total score as positive or negative.

4 RESULTS

4.1 Partial Results

(1) Cast-Directors Attribute
   (a) Approach 1:
   The entire data set consists of 16,638 movies. We have decided that if a particular cast or director of the movie has Facebook likes are available, then we use them. Otherwise, we assume that he is not popular and set his Facebook likes to zero. To check whether this approach is fruitful or not we need to analyse how many cast or director attribute rows are set to zero. By following this approach of the cast or director popularity using Facebook likes, the results are as follows:
   Cast : 3427
   Director:9374

We can see from the above that this experiment was not fruitful as we have expected. It has lead to more number of zero values which would create many number of missing values in the data. So, we have moved onto another approach to solve the problem of categorising the data attributes of cast and directors.

(b) Approach 2:
This approach was successful as we could gauge the popularity of the cast and directors with the idea that the more number of movies you are cast or the more number movies you direct, you are known to more people. The number of movies determines the popularity of the cast or director. There were no zero values in this approach strengthening the analysis.
We therefore replaced the attribute values containing the cast and directors of the movies with the value of the count of movies they were present in.

(2) Sentiment Analysis
The Naive Bayes classifier approach using the existing movie reviews was very useful. We could then tag all the movie reviews and assign the scores by a factor of 1 or 2 for each positive and negative review, and the Top Critic attribute of the reviews.
The Scores of each movie was classified as positive or negative and movie_link an unique identifier for each was appended. The new attribute was then combined with the existing movie data frame using the unique identifier movie_link.

5 PROPOSED WORK

5.1 Design of Future Experiments

Currently, we have completed the data collection and merging of the data where the sentiment analysis has been performed on almost a million of reviews. We have also performed feature engineering thereby selecting the attributes which are meaningful to the analysis, and finally considering them as essential features. The data set has been pre processed and formed in a full fledged manner.
The following are the activities currently planned, which will form the future milestone of the project:

(1) We would be performing an extensive data pre-processing on the complete data set to make sure we are transforming the data in such a manner that we present the underlying problems to predictive model to ensure greater accuracy.
The steps for which include:
(a) The detection and handling of the outliers using approaches like modified Z-Score method, IQR method etc. based on the feasibility of the data.
(b) The handling of the missing values using approaches like linear regression or other novel approaches based on the exploratory data analysis, we will be performing and the experimenting like finding a supporting data etc.
(c) We will perform the exploratory data analysis to analyse the skewness of the data using different techniques like log transformation, square root transformation or Box-Cox Transform based on the results we will obtain applying each of these approaches.
(d) Further scaling of data plays a crucial role for the performance of the Machine Learning algorithms we will be applying. We will be performing the normalization of the data using methods like min-max scaling and will ensure that the data is ready to be fed into the predictive models as an input.

(2) Once, data transformation is completed, we will be doing the movie success prediction using the test data by experimenting with various models. We are doing the research and analysis with the goal of finding the answers to the following thoughts in our mind:
- Which features or attributes are more essential or predictive than others in predicting whether a movie is a success or a failure?
- Can we find any suitable state-of-art algorithms which we can modify according to the project needs?
- Can we improve the traditional algorithms like the Random Forest etc. so that we can improve the performance of the algorithms which will in turn cater to the accuracy of the movie prediction?

(3) The following are some of the novel approaches in our mind. Going further with the experiments, we will dig deep into these thoughts.
As of now, we have planned to explore deep into finding the suitable state of art algorithms. Some of the algorithms which each of our team member will be exploring and try to enhance it according to our needs of the project of predicting movie success or failure are as follows:
- Decision Trees
- Random Forests
- SVM + Kernel
- Ensemble Methods

We will also be performing the Hyper Parameter tuning to further improve the accuracy on the algorithms we will be exploring and implementing.

5.2 Plan of activities

We have currently planned the work division among each of the team member as follows:

- Each team member will be performing the exploratory data analysis and contributing to the data processing steps as mentioned above like the detection of outliers, handling skewness of data, data normalization etc.
- Each member of the team will be exploring the algorithms stated above keeping in mind the set of questions mentioned and try to find the solutions for those questions thereby introducing the novelty to the project.

6 VIRTUAL MEETING SCHEDULE

6.1 Past Meetings

We have conducted the Virtual Meeting using Zoom in past for the following timings:

(1) March 20^{th} - (4:00pm - 6:00 pm)
(2) March 27^{th} - (4:00pm - 6:00 pm)

6.2 Scheduled Meetings

We will be conducting Virtual Meetings using Zoom according to the following Schedule:

(1) April 3^{rd} - (4:00pm - 6:00 pm)
(2) April 10^{th} - (4:00pm - 6:00 pm)
(3) April 17^{th} - (4:00pm - 6:00 pm)
(4) April 22^{th} - (4:00pm - 6:00 pm)

REFERENCES

[1] Javaria Ahmad, Prakash Duraisamy, Amr Yousef, Bill Buckles. Movie Success Prediction Using Data Mining, 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), PP 1- 4.

[2] K Meenakshi, G Maragatham, Neha Agarwal, Ishitha Ghosh. A Data mining Technique for Analyzing and Predicting the success of Movie,2018 J. Phys.: Conf. Ser. 1000 012100

[3] Rijul Dhir, Anand Raj. Movie Success Prediction using Machine Learning Algorithms and their Comparison,2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC).

[4] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts, Learning Word Vectors for Sentiment,The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

[5] Cary D. Butler,Eric Jackson,Li-jing Chang,Mahzabin Akhter,Md Atiqur Rahman,Predicting Movie Success Using Machine Learning Algorithms(LACCEI 2017).