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. Use data mining techniques to 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, 8 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 and their movies in the past. Data mining techniques are applied to the movies data set in order to extract patterns and identify trends that will help us in predicting a movie's success. Data mining is crucial in order to find hidden patterns and relationships among the attributes.

Movie success prediction is an important problem domain because it is an expensive task to create a movie whose success is totally dependent on various factors ranging across cast, crew, genre, run time, and audience opinion etc. A bad movie can turn out be a huge monetary loss to the investors involved in the movie. Movie success prediction can help prevent this problem ahead of its time. This will help the people making the movie make a decision whether or not to continue it, based on its success prediction. This success prediction has significant usage with the audience as well. The audience can know the quality and success of the movie before actually spending money on watching it. It is wise to have a prediction before a monetary investment has been made, which is what we are trying to achieve.

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 with several models (KNN, Naive Bayes, Random Forest, Logistic Regression and SVM) run on different combinations of the attributes to see how certain attributes affect the performance of the model.

1.2 Related Work

Since, movies have been around for a long time, there was good amount of work done in this domain of movie success prediction. Some of the earlier work ([1],[2],[3]) tried to predict the gross of a movie based on stochastic and regression models on the IMDB data set. They also categorised movies as a failure or success based on the revenue and applied binary classifications to the forecast. Revenue is not the only criteria that decides a movie's success. There are a number of other factors like cast and crew, genre, audience rating etc. that can have an impact on the success.

There were also some work done in 2009 [4], that tried to predict the success using news analysis. They have used the news data to make predictions. It was proved that news data was almost as good as IMDB data and the results were better when both the data sets were combined.

In 2015, there was a project about predicting the investment decisions about the movie [5]. Using historical data, this work helped the investors in the movie. However, not all the movies are profited through box office revenue.

There were some cases where some movies had profits from selling merchandise, or digital rights (Netflix, amazon etc). Some people also tried to predict the success based on the social media and the hype analysis. This was done by calculating the positive comments, likes on Facebook, Twitter tweets etc [6].

All of these approaches had predicted the success of the movie that already came out or the future predictions were based only on limited attributes that can not really influence a decision.

In our work, we have analysed various models with different combinations of attributes to find out which attributes really matter in predicting an upcoming movies success. Unlike the approaches mentioned above, this work can be used to predict a movies success even before its release.

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

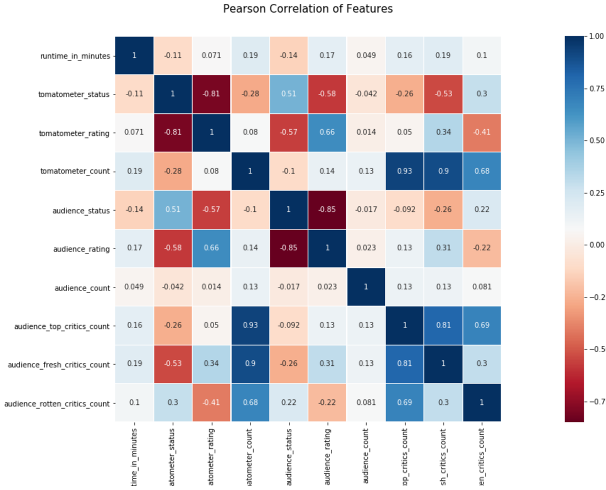
2 METHOD

2.1 Approach

(1) Naive Bayes Classifier for sentiment analysis
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 values 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 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.

(4) Removal of Skewness and Outliers
In the given data set, almost all the attributes are distributed with certain amount of skew. They are either left skewed are right skewed as seen in the Figure 1. Skewness is actually a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined. In order to remove the skewness, instead of directly applying cube root or log function on the data, We have done it individually attribute by attribute and we have also written two different functions to decide on how to remove the skewness. In the first function we have taken a variable i and looped it from 1 to 10. At each stage(i:1-10) we transformed the data with a power of i and then computed the skew. Then we took the value of i for which the skew was minimal. The second function is very similar to first, but instead of powering the data in this function we have rooted the data and got the value of i for which the skew was minimal. Eventually we have compared the results of both these functions and decided whether to apply power function or root function for each attribute in order to remove skewness. The results of which can be seen in the Figure 2.

Coming to outliers part, An outlier is basically a data point that is far away from other data points. In order to remove outliers we have implemented the Interquartile Range(IQR) approach. IQR is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 Q1. For each attribute we have calculated the 3rd quartile and 1st quartile, multiplied them with 1.5 and considered the data points that are present in between these two values and removed the data points are were out of this range.

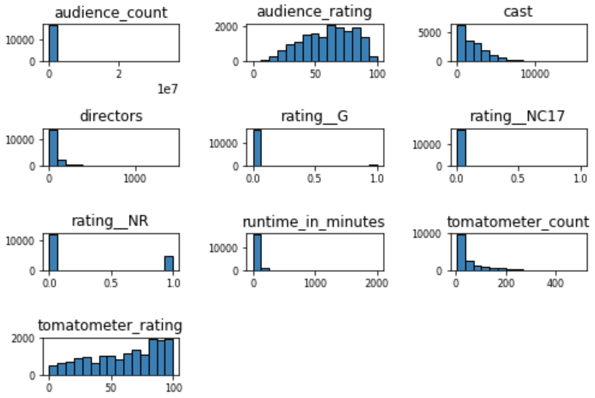


Figure 1: Skewness of Data Initially

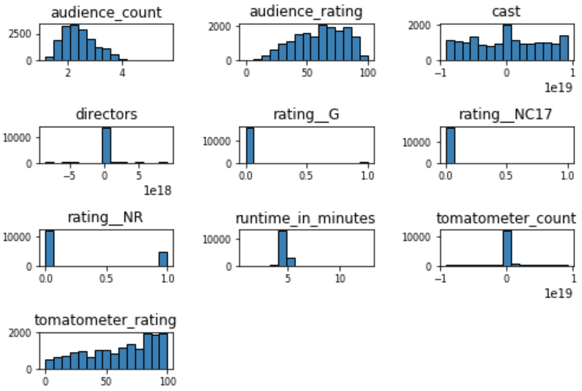


Figure 2: Skewness of Data After Applying Optimal Function

(5) Oversampling

After Removing the skewness and outliers, we observed that there was class imbalance in the final target table. One approach to address this imbalanced datasets is to over-sample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples don't add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE in short

(6) Machine Learning Models

(a) KNN

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data.

(b) Naive Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

(c) Random Forest

It is ensemble algorithm. Ensemble algorithms are those which combines more than one algorithms of same or different kind for classifying objects. For example, running prediction over Naive Bayes, SVM and Decision Tree and then taking vote for final consideration of class for test object. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. Irregular timberlands RF or arbitrary choice woodlands are a gathering learning strategy for characterization, relapse and different errands, that work by developing a huge number of choice trees at preparing time and yielding the class that is the method of the classes (in arrangement) or mean forecast (in relapse) of the individual trees. RF is an improvement over the decision tree algorithm as it corrects habit of over fitting in decision trees to their training set.

(d) Logistic Regression

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X. Contrary to popular belief, logistic regression is a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as "1". Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function. Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.

(e) Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning model that is used for classification. SVMs work by maximize the margin between separating hyper plane. In linear SVM the plane can be split by a line. For an example how the model could look like. For example, could the red values be answer A and the blue be answer B. If a new value would be introduced to the system and positioned on the red side, the model would predict the new value to be equal to answer A. If there are more answers possible a hyper plane is created to be able to split all the answers up in different areas. SVM are effective high dimensional, memory efficient, and versatile machine learning algorithms that work well with non-linear 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.

For removing skewness, we did not consider the traditional techniques such as applying Cube root , Square root or log on data. The reason we did not consider these is that by applying them we were not able to reduce the skew below standard value 0.8. Instead we have taken two functions to compute powers from 1 to 10 and roots from 1 to 10 on each attribute to compute the skewness at each stage and then decided which power or root value to be applied on the data in order to get minimal skew. For outlier detection and removal we had to choice between using either Z-score or IQR techniques and we decided on using the IQR approach as IQR can be used to identify outliers by defining limits on the sample values that are a factor k of the IQR below the 25th percentile or above the 75th percentile. The common value for the factor k is the value 1.5. A factor k of 3 or more can be used to identify values that are extreme outliers or "far outs" when described in the context of box and whisker plots.

In order to overcome the problem of class imbalance on the target value, we have performed SMOTE Oversampling. Random sampling is the popular technique that is used in order to overcome the problem at hand. But the major drawback of Random undersampling is that it can discard useful data. As a result we chose SMOTE oversampling method. This method generates synthetic data based on the feature space similarities between existing minority instances. It does not discard important data by creating a synthetic instance. It finds the K-nearest neighbors of each minority instance, randomly selects one of them, and then calculate linear interpolations to produce a new minority instance in the neighborhood.

The biggest problem we were facing so far was the problem of class imbalance. As a result we have used oversampling. After this our next and final challenge was to build a model that gives the best accuracy in predicting the result of the movie. In order to do so, We have built several models based on different classifiers such as KNN, Naive Bayes, Random Forest, Logistic Regression and Support Vector Machines.While training each classifier we have first performed cross validation so that we can improve each model's performance and there by achieve better accuracy. The reason for testing out different classifiers was to see which classifier would give better performance in predicting the outcome of the test data, So that we can use that classifier as the ultimate classifier in predicting the outcome.

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

Our hypotheses entering the experiment were:

- **H.1** All models will not have same performance for the Prediction of Movie Success
- **H.2** There can be an influencing attribute in the success prediction
- **H.3** A genre can contain more successful movies
- **H.4** A movie rating can be a deciding factor for successful movies

3.3 Experimental Design

Procedure 1:
The performance of the machine learning model depends on various factors. One of which, for example is the way we divide the train-test split of the data(can play a crucial role in determining the performance). The size of data is many a times insufficient for the analysis. In this scenario, holding out the data might effect the analysis, but there are ways through which we can do the optimal utilization of the data in hand and overcome this. Approaches like cross validation to train the model multiple times can be helpful. To do the comparative analysis of the models, this test can be followed for multiple models thereby resulting in the proper usage of the data. Accuracy can be used as the metric in this regard to evaluate the model performance thereby helping us to prove our first hypothesis.

Procedure 2:
Attributes of movies data set which are meaningful, are nothing but the features of the data set, they play a major role in movie success prediction. These features can be used to answer many questions regarding the underlying factors that influence the success of the movie. This experiment helps us determining the influence of attributes which was previously not present in other studies, hence bringing novelty to the success predictions we are trying to make. The idea behind this thought is that there are few influencing attributes in this data set like the cast, directors, genre, movie rating etc. whose presence makes a lot of impact in the predictions and should be taken care while starting a production of a new movie. To detect which attribute is game changer, we can perform a series of test as follows:

1. Eliminate the cast attribute and build the model and analyse the performance of the model based on various evaluation metrics like accuracy, recall, f-score etc. for the best prediction model achieved.
2. Eliminate the directors attribute and build the model and analyse the predictions for the movie success prediction.
3. Eliminate the both these attributes and build the model and analyse the result as both these attributes were calculated using the existing data set.
4. Eliminate the other features like audience rating, tomatometer rating etc. on experimental basis and train the model to see their impact on the model performance and predictions.

By conducting these experiments, we can analyse the most influencing feature thereby helping us determine the results for the second hypothesis.

Procedures 3 & 4:
Few factors of a movie influence the audience to watch and make it a success. Some people prefer movies of Comedy genre while others may be interested in drama. Genre can be a driving factor many a times. On the other hand, the rating of the movie can also be a deciding factor. So, it is paramount that we need to invest time in analysing genre and rating features of the movie data so that we can get a wider choice of the public for the upcoming movies. To get a hang of this, we can perform a simple test on the dataset under consideration. We can filter the movies based on their success and record the the types of genres that are leading to the success of the movies along with keeping an eye on the rating. In this way we can use this data and predictions to determine whether the audience will prefer to watch a movie. These tests help us in proving the hypothesis three and four.

4 RESULTS

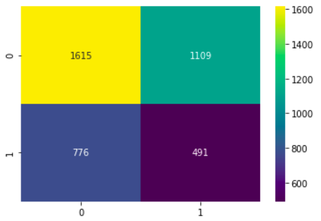
4.1 Results

(1) The following results obtained after pre processing of the data the models using 0.25 Test set size:

(a) KNN:
The accuracy obtained:

- Training Set: 0.7547344968436688
- Test Set: 0.5276872964169381

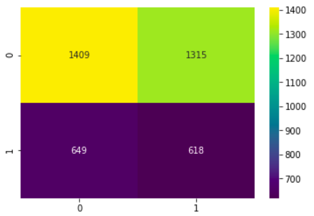
The confusion matrix:



(b) Naive Bayes:
The accuracy obtained:

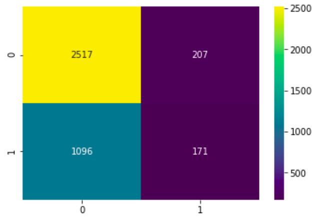
- Training Set: 0.5056937739819285
- Test Set: 0.5078927587070909

The confusion matrix:



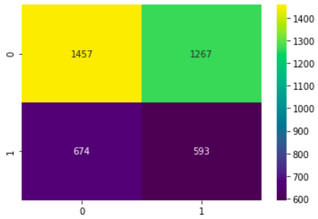
(c) Random Forest:
The accuracy obtained:
- Training Set: 1.0
- Test Set: 0.6735154096717615

The confusion matrix:



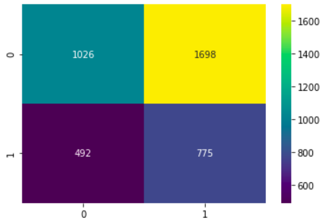
(d) Logistic Regression:
The accuracy obtained:
- Training Set: 0.503960886248298
- Test Set: 0.5136557253821098

The confusion matrix:



(e) SVM:
The accuracy obtained:
- Training Set: 0.51584354499931922
- Test Set: 0.45126534703081933

The confusion matrix:



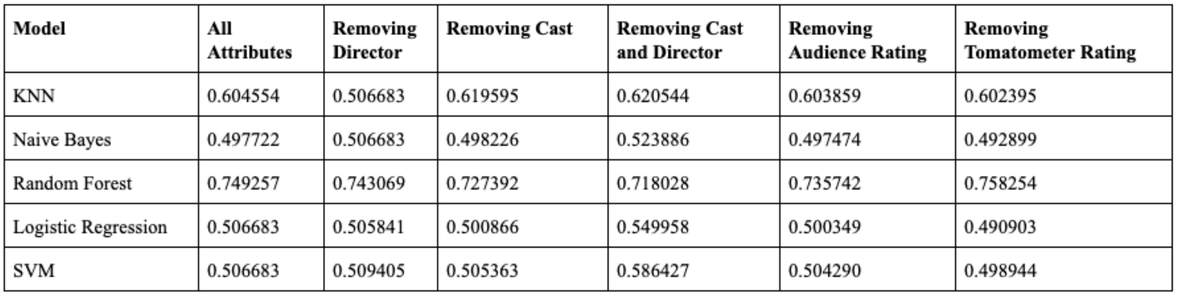
(2) The results of Cross Validation of the various models using 5 splits and 0.25 Test set size can be seen in the Figure 3.
(3) The results of Evaluation Metrics for Random Forest using Test set Size 0.25 can be seen in the Figure 4.
(4) The results of the genres feature values with the number for successful movies is as follows in Figure 5:
(5) The results of the rating feature values with the number for successful movies is as follows in Figure 6:

4.2 Discussion

(1) The results for the Test and Train accuracy as seen above in the results along with confusion matrices show that the Random Forest machine learning technique helps best to determine the Success prediction of the movie. We have also experimented the results with the cross validation test (Figure 3) to verify the results as this technique helps in increasing the data set size. The results help us prove our first hypothesis that various models can have different performance on the same data. Previous studies have shown various methods of movie success prediction mostly using stochastic processes only.
(2) Having found that the Random Forest is the best model for this dataset, we have conducted the tests to find the most influential attribute that alters the performance of this model in lieu of the movie success prediction. From the Figure 4, we can see that we have experimented with the model by considering all the attributes and also by removing the Director, Cast, Director and Cast, audience rating and tomato-meter rating individually. We have used the evaluation metrics like Accuracy, Precision, Recall, Specificity and F-Score to analyse the impact. From the results obtained, we can state that the performance of the model was influenced more when we removed the Audience Rating and Tomato-meter Rating attributes. Hence, this result shows that some attributes are the most influential attributes than others proving our second hypothesis.
(3) According to the Figure 5, we can see that the count for the movies which were successful were most likely of the Drama genre. This analysis can be used in consideration in the future to help predict the success of a movie. Analysis similar to this were not performed earlier and can greatly help the movie makers to consider. This also proves our third hypothesis.
(4) In the opinion to the Figure 6 which includes the count for the successful movies in various ratings, we can conclude that the audience prefer to go to the movies that are R rated. Might be that the movies are preferred by the families. This experiment validates the fourth hypothesis.

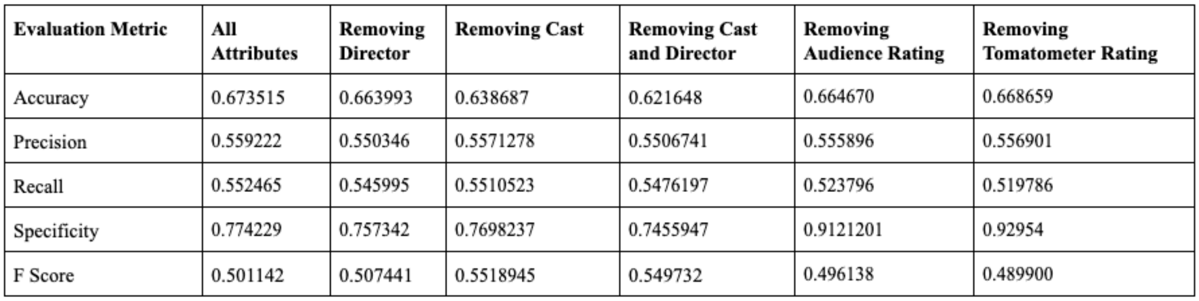
4.3 Conclusion

The rotten tomatoes data set is an interesting data set which includes data of movies along with the reviews. The movie reviews helped us in performing the sentiment analysis using Naive Bayes Classifier. We have given more importance to the movie reviews provided by the Top Critics for the movies using the attribute 'Topic Critic'.
After building the five different machine learning models we found out that the Random Forest represents the movie success prediction more accurately. We have also conducted experiments to determine the influencing attributes and found that the cast and director features highly influence the prediction of the success. The results for highest successful movies was present in Drama Genre and R Rated movies.



Model	All Attributes	Removing Director	Removing Cast	Removing Cast and Director	Removing Audience Rating	Removing Tomatometer Rating
KNN	0.604554	0.506683	0.619595	0.620544	0.603859	0.602395
Naive Bayes	0.497722	0.506683	0.498226	0.523886	0.497474	0.492899
Random Forest	0.749257	0.743069	0.727392	0.718028	0.735742	0.758254
Logistic Regression	0.506683	0.505841	0.500866	0.549958	0.500349	0.490903
SVM	0.506683	0.509405	0.505363	0.586427	0.504290	0.498944

Figure 3: Cross Validation Accuracy Results



Evaluation Metric	All Attributes	Removing Director	Removing Cast	Removing Cast and Director	Removing Audience Rating	Removing Tomatometer Rating
Accuracy	0.673515	0.663993	0.638687	0.621648	0.664670	0.668659
Precision	0.559222	0.550346	0.5571278	0.5506741	0.555896	0.556901
Recall	0.552465	0.545995	0.5510523	0.5476197	0.523796	0.519786
Specificity	0.774229	0.757342	0.7698237	0.7455947	0.9121201	0.92954
F Score	0.501142	0.507441	0.5518945	0.549732	0.496138	0.489900

Figure 4: Random Forest Metric Results

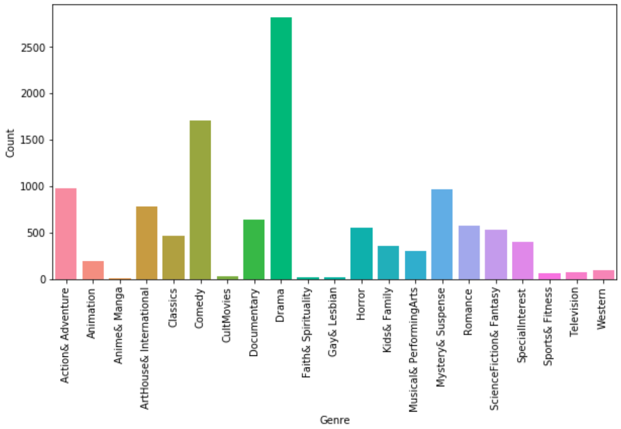


Figure 5: Genre Frequency

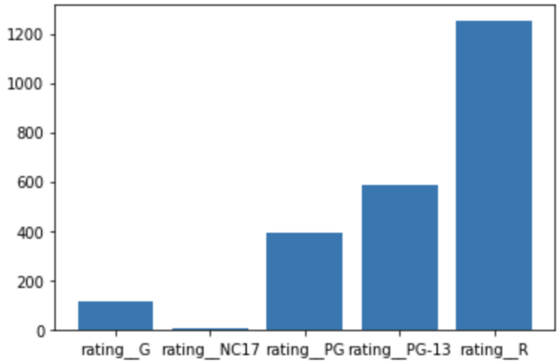


Figure 6: Rating Frequency

These are some of the results which were not present in the previous studies and these results can be helpful to predict the success rate of movie as it is of utmost importance since billions of dollars are invested in the making of each of these movies every year.

In the future, we would like to accumulate and increase the number of movies, reviews and features in the dataset, through various approaches and techniques like Web Scraping etc.

Sometimes, movie posters can also attract the audience and drive them to watch the movie. Our dataset also includes an attribute called 'Movie poster url'. In future we would like to consider and process this data attribute using approaches like Image recognition etc. and determine its influence on the audience.

The data from social media sources such as Facebook likes (we were not able to include this due to high occurrence of missing data for directors in the data set), Twitter, and YouTube content of movie can also be included in the success prediction of movie. These can be done using conglomeration of various data sets from different sources.

5 VIRTUAL MEETING SCHEDULE

5.1 Past Meetings

We have conducted the Virtual Meeting using Zoom in past for the following timings and all the meetings were attended by all team members.:

(1) March 20^{th} - (4:00pm - 6:00 pm)
(2) March 27^{th} - (4:00pm - 6:00 pm)
(3) April 3^{rd} - (4:00pm - 6:00 pm)
(4) April 10^{th} - (4:00pm - 6:00 pm)
(5) April 17^{th} - (4:00pm - 6:00 pm)
(6) April 22^{th} - (4:00pm - 6:00 pm)

GITHUB

We have used the online collaboration platform. Please find the source code in the following link:
https://github.ncsu.edu/stallur2/Movie-Success-Predictor
The data for the Rotten Tomatoes Movie review is more than 100MB hence it has been compressed, please unzip the file before execution of the code.

REFERENCES

[1] J. S. Simonoff and I. R. Sparrow, "Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers," Chance, vol. 13, no. 3, pp. 15–24, 2000.
[2] A. Chen, "Forecasting gross revenues at the movie box office," Working paper, University of Washington, Seattle, WA, June, 2002
[3] M. S. Sawhney and J. Eliashberg, "A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures," Marketing Science, vol. 15, no. 2, pp. 113–131, 1996.
[4] W. Zhang and S. Skiena. "Improving movie gross prediction through news analysis". In Web Intelligence, pages 301-304, 2009.
[5] Michael T. Lash and Kang Zhao, "Early Predictions of Movie Success: the Who, What, and When of Profitability", June 2015.
[6] J. Duan, X. Ding, and T. Liu, "A Gaussian Copula Regression Model for Movie Box-office Revenue Prediction with Social Media," Communications in Computer and Information Science Social Media Processing, pp. 28–37, 2015.
[7] 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
[8] 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).
[9] 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).
[10] Cary D. Butler,Eric Jackson,Li-jing Chang,Mahzabin Akhter,Md Atiqur Rahman,Predicting Movie Success Using Machine Learning Algorithms(LACCEI 2017).