Movie Success Predictor

CMPE 255 Section 02

Dr. Gheorghi Guzun
Group 12
Nitsimran Singh SJSU ID: 015570425
Saikrishna Dosapati SJSU ID: 015950493
Sachith Gandham SJSU ID: 015929394
Faizali Mulla SJSU ID: 015256332
Github Link

Section 1

Introduction

The goal of our project is to predict a movie's success or failure by analyzing existing movie databases with its reviews and data related to the cast and crew and their movies in the past. Data mining techniques are applied to the movie's 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

Motivation

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 to 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.

Objective

Our objective is to predict the success of a movie with the help of existing data of movies, and their reviews. To do so 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.

Literature/Market review

- Since movies have been around for a long time, there was a 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.
- 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 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 movie's success. Unlike the approaches mentioned above, this work can be used to predict a movie's success even before its release.

Section 2

System Design & Implementation details

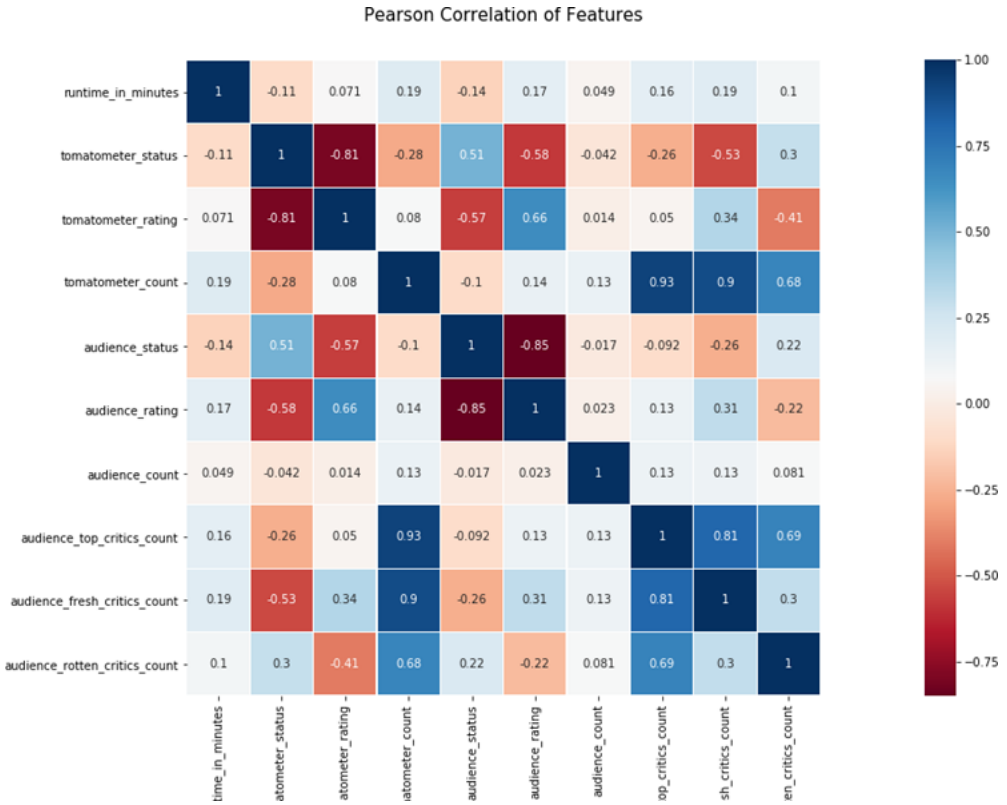
(1) Naive Bayes Classifier for sentiment analysis The sentiment analyzer we implement to categorize the movie reviews trains on a Naive Bayes Classifier. It is a probabilistic learning method based on Bayes theorem and follows a supervised learning approach. The goal of any probabilistic classifier is 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 a certain amount of skew. They are either left skewed or right skewed as seen in 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 the 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 Figure 2. Coming to the outlier 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.

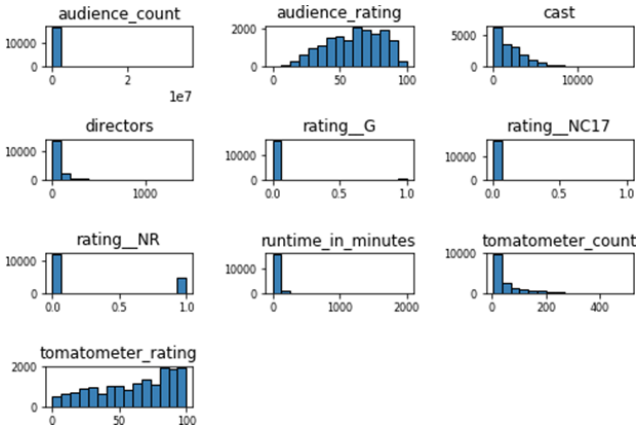


Figure 1: Skewness of Data Initially

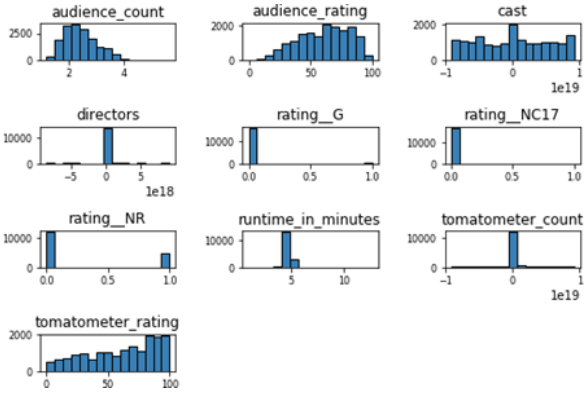


Figure 2: Skewness of Data After Applying Optimal Function

(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

Algorithm considered

(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 an ensemble algorithm. Ensemble algorithms are those which combine more than one algorithm of the same or different kind for classifying objects. For example, running prediction over Naive Bayes, SVM and Decision Tree and then taking a vote for final consideration of class for test object. Random forest classifier creates a set of decision trees from randomly selected subsets of the training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

(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 a 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.

(e) Support Vector Machine Support Vector Machine (SVM) is a supervised machine learning model that is used for classification. SVMs work by maximizing the margin between separating hyperplanes. 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.

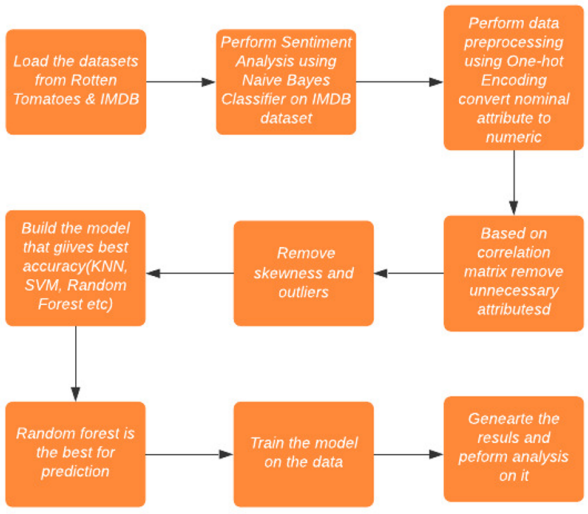
Algorithm selected (and why)

We selected Random Forest Algorithm, We selected Random Forest as it had the highest results for cross validation accuracy

Technologies & Tools used

The project will be implemented in python, we will thus be requiring Python 3.0, along with libraries numpy, Scikit-learn, Pandas. We will be using Kaggle in order to obtain the data set.

Design/architecture/data flow



Use cases

- It is wise to have a prediction before a monetary investment has been made, which is what we are trying to achieve.
- Movie success prediction is an important problem domain because it is an expensive task to create a movie.Movie success prediction can help prevent this problem by utilising data mining techniques to predict a movie's success ahead of its time. This will help the movie makers take a decision, based on its success prediction.
- The audience can also know the quality and success of the movie before actually spending money on watching it.
- The main use case would be filtering reviews to show to customers to provide the most helpful reviews possible by filtering them based on helpfulness, the number of people who voted, and the predicted sentiment.

Section 3
Dataset

(1) Rotten Tomatoes Data Set

(a) Rotten Tomatoes is one of the most popular film websites, which combines movie information, critic reviews and user 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 of 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).

Methodology

The movies data then will be preprocessed and we will be performing a KNN classifier to categorize the movies. The movie success to be predicted will be missing the rating attributes, which cannot be used, hence we will be using a 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 to the movies helps in improving the accuracy.

Graphs

Figure 3: Cross Validation Accuracy Results

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 4: Random Forest Metric 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 5: Genre Frequency

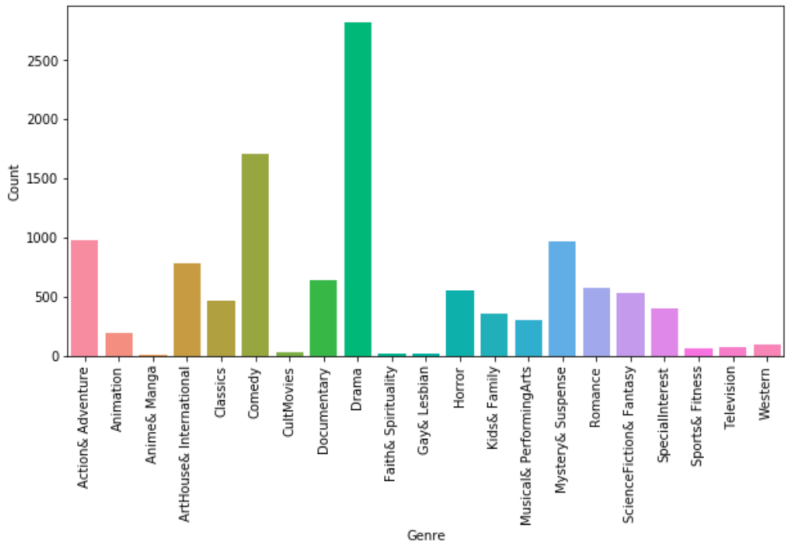
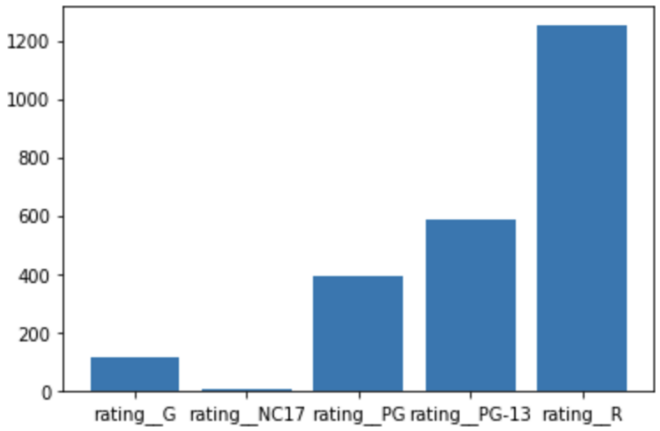


Figure 6: Rating Frequency



Analysis of results

Results:

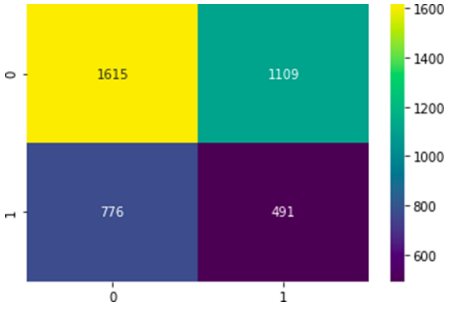
(1) The following results obtained after pre processing of the data the models using 0.25 Test set size:

(a) KNN: The accuracy obtained:

• Training Set: 0.7547344968436688

• Test Set: 0.5276872964169381

The confusion matrix:

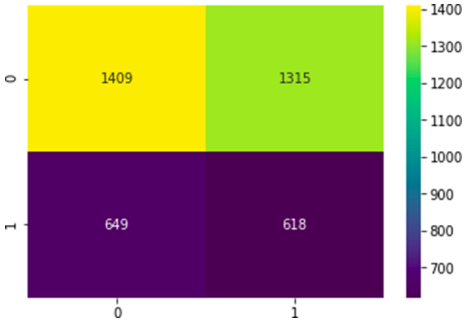


(b) Naive Bayes: The accuracy obtained:

• Training Set: 0.5056937739819285

• Test Set: 0.5078927587070909

The confusion matrix:

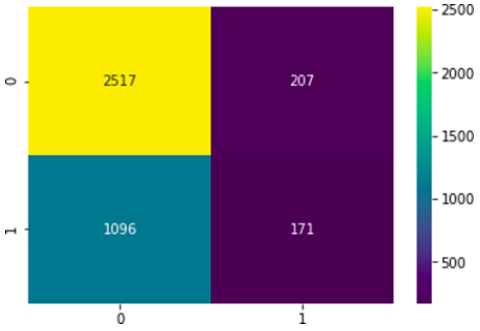


(c) Random Forest: The accuracy obtained:

• Training Set: 1.0

• Test Set: 0.67351540967176151

The confusion matrix:

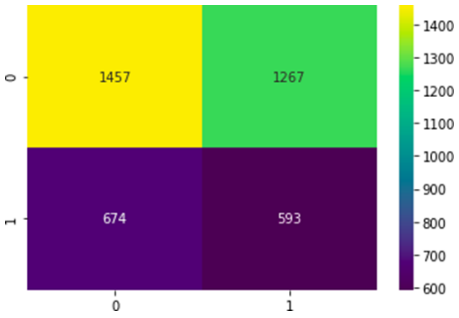


(d) Logistic Regression: The accuracy obtained:

• Training Set: 0.503960886248298

• Test Set: 0.5136557253821098

The confusion matrix:

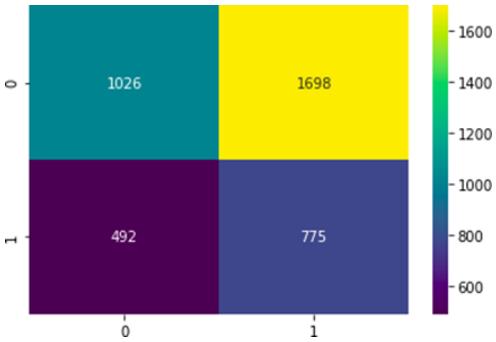


(e) SVM: The accuracy obtained:

• Training Set: 0.5158435449931922

• 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 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

Section 4

4.1 Decisions Made:

- Initially we used the Naive Bayes classifier to 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.

4.2 Difficulties Faced:

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 thereby 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.

4.3 Things that worked

- We have taken a step further in inculcating the novelty by considering what model and what underlying influencing factors are impacting the movie success, factors like cast and crew, genre, audience rating etc
- We have also gathered results like which genre is most liked or how a rating of a movie can impact any success
- There were some experiments with analysing news data to make predictions. It was proved that news data was almost as good as IMDB data.
- Work on how to predict the success based on the social media hype by calculating the positivity in the comments, number of likes was also done.

4.4 Things that didn't work well

- Some of the earlier work categorised movies as a failure or success based on the revenue. Apart from revenue, there are other factors like cast and crew, genre, audience rating etc. that have an impact on the success.
- Processing the movie poster attribute to estimate the attractiveness on the audience through techniques like image processing could not be accomplished.
- Conglomeration of the data regarding the movie from the social media like Facebook, Twitter, Youtube could also be tried to fetch better accuracy and results.

4.5 Conclusion

- The rotten tomatoes data set is an interesting data set which in- cludes 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 fea- tures highly influence the prediction of the success. The results for highest successful movies was present in Drama Genre and R Rated movies.

- 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 movies 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 by using techniques like Web Scraping.

Section 5

Tasks and assignments of team members to different components

- Loading data and Preprocessing: Saikrishna Dosapati, Sachith Gandham
- Training Model and Validation model: Nitsimran Singh, Faizali Mulla
- Data Analysis: Sachith Gandham, Nitsimran Singh
- Writing report: Nitsimran Singh, Saikrishna Dosapati
- Creating presentation: Sachith Gandham, Faizali Mulla

Everyone finished the assigned tasks and no one changed the role.

Resources

[1] J.S.SimonoffandI.R.Sparrow,"PredictingMovieGrosses:WinnersandLosers, 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.SawhneyandJ.Eliashberg,"AParsimoniousModelforForecastingGross Box-Office Revenues of Motion Pictures," Marketing Science, vol. 15, no. 2, pp. 113–131, 1996.

[4] W.ZhangandS.Skiena."Improvingmoviegrosspredictionthroughnewsanalysis". In Web Intelligence, pages 301-304, 2009.

[5] MichaelT.LashandKangZhao,"EarlyPredictionsofMovieSuccess:theWho, 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.