International Journal of Modern Agriculture, Volume 10, No.1, 2021
ISSN: 2305-7246

Movie Profit Prediction

Dr.Charulatha B.S. [1],Abilash. R[2],Dr.VishnuPriya .A[3*],Divya peachi. M[4]

[1]Rajalakshmi Engineering College

[2]Jawahar Engineering College

[3]Madanapalle Institute of Technology & Science

[4]Chennai national college

[1]charu2303@yahoo.co.in,[2]abilashr.86@gmail.com,[3]a.vishnupriya2010@gmail.com,*

[4]divyapeachi99@gmail.com

Abstract

This paper produces a merit to today's generation though years passed people's love for movies and entertainment has not passed. Every year more than 200 films are released in Tamil film industry, sometimes this range reaches to 500 but the foremost circumstance that is to be inspected is the success rate of the movie. Every movie producer who invests money for the production of the movie is keen to know the conclusion of their investment, sometimes the result may surprise them, sometimes it may downhearted them. There is where our study plays an important role in early prediction of the conclusion that is our study propounds a decision support system using machine learning techniques. This experimentation avoids the investment risk by helping the investors to predict the success rate of the movie. We have accumulated different data from different movies to examine the profitability of the movie with its data and we have implemented different algorithms based on the collected data and have acquired a good accuracy with respective algorithms. This dissertation of early prediction is highly essential and favorable for the film industry. This precision of early prediction can be further improved and utilized by using larger data sets and features.

Keywords: *Classification, Ensemble Learning, Movie, Machine Learning, Prediction, Supervised Learning*

I. INTRODUCTION

The Indian cinema is tremendously popular. Every year thousands of movies are released in each language. Each film contains different themes, some may motivate us, some may disappoint us, some give us a lesson, some make us laugh, some make us feel the depth of the action, some may remind us of our sweet moments and sometimes it makes us even feel depressed by its sad endings. In short cinema has the capability to influence us both provincially and universally. But the authenticity is that though thousands of movies are released only some are successful. The term success in the film industry is extensively big which is wished by every individual who works in it. This is where the investors for whom cinema is a source of business come to chaos and confusion about their investment. So to resolve this chaos we have used our machine learning techniques in the name of movie profit prediction. A success of a movie is calculated in two ways either by its worldwide gross income or by its popularity. Some movies become popular but they don’t earn profit, whereas some movies earn profit but they are not popular and whereas there are some other movies which earn profit and also become popular. In our

cramming we are going to forecast the profitability of the movie before it reaches the box office using machine learning algorithms.

First of all let us know what is movie profit prediction? This indicates early prediction that is a decision support system where the results of the motion picture are forecasted before it is released on big screens. This type of early prediction advises the investors to know the success rate and profitability of the film by which they can have some confidence to invest in the particular motion picture.

Now let us know what machine learning is? Machine learning is an application of Artificial Intelligence whose algorithms are applied in many contradictory promising features. This system is closely related to statistical computing that emphasizes on making predictions. This is also cited as predictive analytics in disparate circumstances. This was prospered by Arthur Samuel in the year 1950. We have utilized this in our interpretation so as to predict the repercussion. We have collected disparatedata from 100 different movies of Tamil cinema industry and we have implemented seven different algorithms in the collected data to achieve a good accuracy. And so, the rest of the paper is organized by describing the collected data, its attributes, algorithms, experimental results and finally the conclusion of the interpretation.

II. DATASET DESCRIPTION

On appraising Tamil cinema industry we randomly espoused 100 different movies from the past 10 years with disparate themes and dissimilar profitability and then we have collected data according to the attributes by observing the nominated motion picture. Therefore the dataset consists of movies that were released in between 2010 to 2019 we have garnered the data by specifying unsystematic 10 disparate movies from each year. After the accumulation of data from surveying the nominated movie we also surveyed the movie trailer data from YouTube and then by consolidating both the assembled data we have predicted the gross earnings of that definite film. We have acquired 12,600 data sets after surveying 100 disparate movies. As the data sets are collected by our own selves personally the need of data cleaning and data processing is not mandatory. We possess 126 different attributes on the basis of Tamil film industry and accumulated the data according to it.

III. ATTRIBUTES DESCRIPTION

On the presumption of a motion picture we have created 126 different attributes. The generated attributes are based on the time duration, theme, location, costume, category, conclusion, time of release, popularity of actor, actress, director, writer and music director, no. of main roles and side roles, no. of roles involved in specific category and output. In accordance with this genre of attributes we have collected the data. And we have catalogued the accumulated data in the form of integers and designated the time in terms of minutes. The term depicted as output in the attribute engendered by us denotes whether the specific movie earned profit or loss and we have characterized it using the integer. Before the implementation of algorithms to the accumulated data we have fitted the data and then we have implemented the algorithms.

IV. IMPLEMENTATION OF ALGORITHMS

Machine learning is a notion which permits the machine to grasp through experience and examples. Though there are different types of algorithms we here operate supervised machine learning algorithms for the accumulated. The term characterized as supervised is interpreted as the machine learning task of educating a function that plots output to the input based on its example that is the supervised data embodies an input object and desired output object depending on its character. From supervised learning we have applied classification algorithms to the data. We have focused on classification. So, as our data consists of two outcomes we have made use of classification algorithms to the accumulated data.

From the list of classification algorithms we have used 7 disparate algorithms: Logistic Regression, Naive Bayes, Decision tree, Random Forest Classifier, Bagging Classifier, XGBoost Classifier and Voting Classifier.

4.1. Logistic Regression

Logistic regression is also expounded as a logit regression. It is a documentation model where in its basic form it utilizes logistic function to paradigm a binary dependent variable and many more intricate augmentations prevail. Logit regression appraises the parameters of logistic models. This is operated to allocate observations to a distinct set of classes. By utilizing logistic sigmoid function logistic regression transmutes its output to remit a probability value which can further be charted to distinct classes. Though logistic regression holds three conflicting types we have designated binary logistic regression for the data gathered for our work.

Now let us understand what is binary logistic regression? In a particular circumstance or case this is used to predict odds based on the predictors. The terminology odds here portray the probability value, that is the output of the data should consist of only two possible outcomes or the output is instance dissect by probability and can be exclaimed as non-case. As the data garnered by us has only two possible outcomes we have harnessed binary logistic regression for our exertion.

Now let us look at the mathematical formula of logistic regression:

They utilize sigmoid function to relay the linear regression into logit function. Here the entitled denomination logit function cites log of odds by utilizing logit function they calculate the necessary probability.

Linear regression equation,

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \quad (1)$$

here,

z = dependent variable

x_1, x_2, x_3 = independent variable

β_0 = intercept

β_1, β_2, β_3 = coefficients

k = number of observations

Sigmoid function,

$$p = \frac{1}{1-e^{-z}} \quad (2)$$

$$p = \frac{e^{z}}{e^{z}+1} \quad (3)$$

Now by adding sigmoid function to z value in linear regression equation,

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+1}} \quad (4)$$

Odds equation,

$$s = \frac{p}{1-p} \quad (5)$$

By substituting s value in p we get,

$$s = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} \quad (6)$$

Now by taking log on both sides,

$$Ln(S) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \quad (7)$$

And now by maximizing log likelihood we acquire mathematical formula of logistic regression.

$$\log\left[\frac{Y}{1-Y}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k \beta_k \quad (8)$$

Secured yield and accuracy by the implementation of logistic regression algorithm to the accumulated data is described in observation 1

Observation 1:

In observation 1 we have described the yield achieved by implementing a logistic regression classifier to the garnered data to predict the profitability of the motion picture. In our exertion we have considered four depths and endowed the accuracy for each depth. The four depths that we have pondered are 0,1,2 and 3. The graphical representation of the outcome on implementing a logistic regression algorithm to the garnered data is narrated below.
From the Figure 1 elucidation we can heed that the result on implementing a logistic regression algorithm to the collected data has acquired the same accuracy in all the four depths that is 0.62 and so the line in the graphical representation is straight.

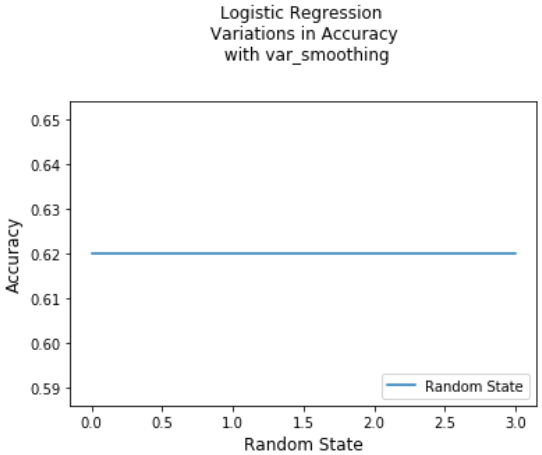

Figure 1: Graphical representation of the outcome on implementing Logistic regression along with its depths

4.2. Naive Bayes

Naive Bayes is also expounded as Idiot Bayes due to its calculation. To make their calculation tractable they have simplified the calculation of probabilities for each hypothesis. This classification algorithm is based on Bayes' theorem that this algorithm presumes independence among predictors. In elementary ways we can urge that this classifier presuppose that the presence of one feature is independent of the other feature in a class. Naive Bayes classifier is commonly exerted for very immense and sophisticated data sets.

Now let us look at the mathematical formula of naive Bayes:

This technique dispenses a way to calculate rear or posterior probability

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (9)$$

here,

P(c|x) = posterior probability

P(x|c) = probability of predictor

P(c) = prior probability

P(x) = predictor of prior probability

Secured yield and accuracy by the implementation of naive Bayes algorithm to the accumulated data is described in observation 2,

Observation 2:

In observation 2 we have described the yield achieved by implementing a Naive Bayes classifier to the garnered data to predict the profitability of the motion picture. In our exertion we have considered three depths and endowed the accuracy for each depth. The three depths that we have pondered are 7, 8 and 9. The graphical representation of the outcome on implementing a Naive Bayes algorithm to the garnered data is narrated below.

From the Figure 2elucidation we can heed that the result on implementing a Naive Bayes algorithm to the implemented data has acquired almost same accuracy in all the three depths that is 0.61, 0.63 and 0.65 in depth 7, 8 and 9 and so the line in the graphical representation has moved gradually according to the accuracy.

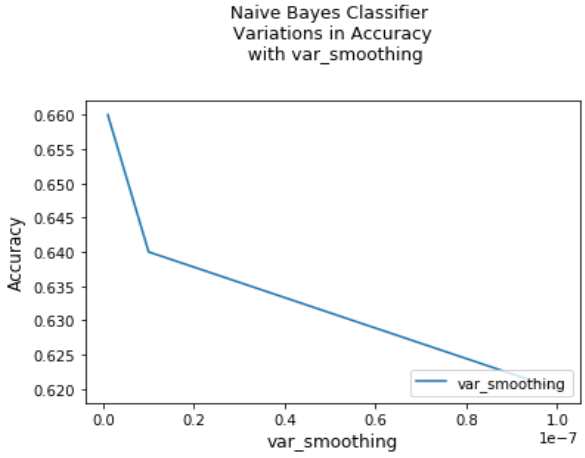

Figure 2: Graphical representation of the outcome on implementing Naïve bayes along with its depths

4.3. Decision Tree

Decision tree algorithm which is utilized as both classification and regression algorithm in the hoard data set. This erects a learning model to predict the value of the target variable. The values of the target variable or class is predicted by learning simple decision rules from the engendered learning model. The prediction of the target variable always starts from its root that is from the gathered data and then it compares the attributes of the accumulated data and the target data. On the basis of comparison accuracy is achieved here. Though decision tree holds two conflicting types we have designated categorical variable decision trees for the data gathered for our exertion.

Now let us understand what a categorical variable decision tree is. When a target variable is categorical then we make use of a categorical variable decision tree. In our exertion the target variable is categorical and so we have made use of categorical variable decision trees.

Secured yield and accuracy by the implementation of decision tree algorithm to the accumulated data is described in observation 3,

Observation 3:

In observation 3 we have described the yield achieved by implementing a decision tree classifier to the garnered data to predict the profitability of the motion picture. In our exertion we have considered nine different depths and endowed the accuracy for each depth. The nine depths that we have pondered are 1, 3, 5, 7, 10, 15, 20, 25 and 30. The graphical representation of the outcome on implementing a decision tree algorithm to the

From the Figure 3elucidation we can heed that the result on implementing a Naive Bayes algorithm to the collected data has acquired same accuracy for seven depths that is 0.55 as average accuracy in depth 5, 7, 10, 15, 20, 25 and 30 whereas different accuracy for the rest two depths that is 0.69 and 0.6 as average accuracy for the depths 1 and 3. And so the line in the graphical representation has moved gradually according to the accuracy.

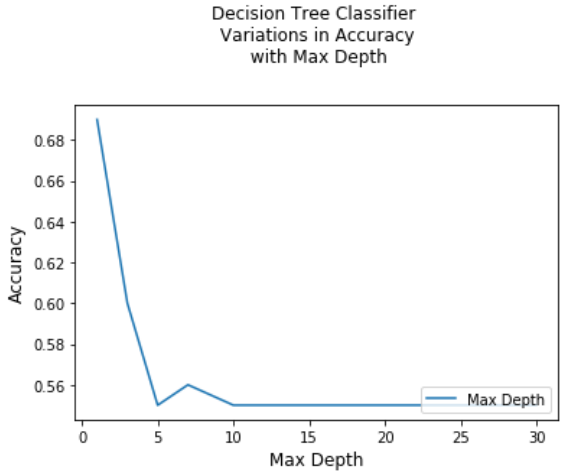

Figure3: Graphical representation of the outcome on implementing Decision tree along with its depths

4.4. RANDOM FOREST

Random forest is a supervised algorithm which is utilized for both classification and regression data sets. Though it is used for both the datasets this algorithm is preferably used for classification. Cluster of decision trees leads to a random forest algorithm as we can acquire it by the name of this classifier. This algorithm forges decision tree in the data specimen and then it acquires the prediction from each data specimen. And after this process it then finally determines the best denouement by voting. This has a major advantage when compared to a decision tree, It reduces the issue of overfitting as it is an ensemble method.

Secured yield and accuracy by the implementation of random forest algorithm to the accumulated data is described in observation 4,

Observation 4:

In observation 4 we have described the yield achieved by implementing a random forest classifier to the garnered data to predict the profitability of the motion picture. In our exertion we have considered six trees and four different depths for which we have endowed the accuracy under this observation. The six trees that we have pondered are 3, 5, 7, 10, 20, and 25 and the four different depths that we have considered are 3, 5, 9 and 12. The graphical representation of the outcome on implementing a random forest algorithm to the garnered data is narrated in Figure 4.

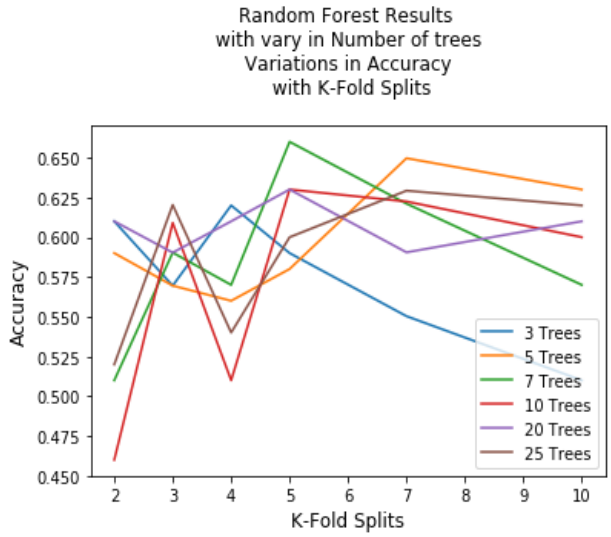

Figure 4: Graphical representation of the outcome on implementing Random forest along with its trees

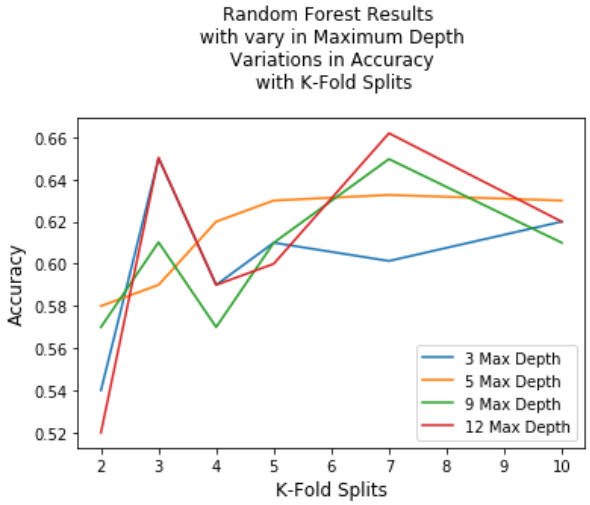

Figure 5: Graphical representation of the outcome on implementing random forest along with its depths

From the Figure5 elucidation we can heed that the result on implementing a random forest algorithm to the collected data has acquired almost the same accuracy in all six trees that is 0.57, 0.59, 0.58, 0.57, 0.60 and 0.58 as average accuracy in trees 3, 5, 7, 10, 20, and 25 and in case of depths we have acquired same accuracy in three depths that is 0.60 as average accuracy for depths 3, 9 and 12 whereas different accuracy in the remaining depth that is 0.61 as average accuracy in depth 5. And so the line in the graphical representation of the pondered trees and depths have moved gradually according to the accuracy.

4.5. Bagging Classifier

Bagging classifier is a supervised classifier predictive technique. This is a group of meta-estimators. To form a final prediction at first they fit to the base classifier on an arbitrary subset of the primal dataset and after this process they jumble their solitary predictions. This ensemble method is also interpreted as bootstrap aggregation method. Now, we know that the ensemble method is expounded as a technique that fuses the predictions from numerous machine learning algorithms to yield forbye accurate predictions. This type of meta-estimators are utilised to reduce the variance of the other algorithms that have large variance by inaugurating randomization appraised in its construction task and then it makes a group that is ensemble out of it. Secured yield and accuracy by the implementation of bagging classifier algorithm to the accumulated data is described in observation 5,

Observation 5:

In observation 5 we have described the yield achieved by implementing a bagging classifier to the garnered data to predict the profitability of the motion picture. In our exertion we have considered four trees and endowed the accuracy for each tree. The four trees that we have

pondered are 5, 10, 15 and 20. The graphical representation of the outcome on implementing a bagging classifier algorithm to the garnered data is narrated below.

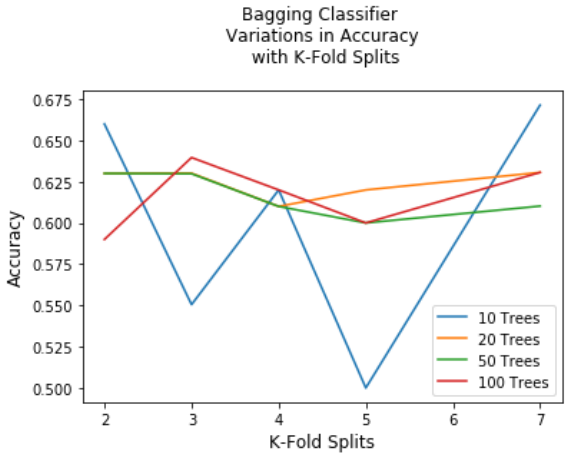

Figure 6: Graphical representation of the outcome on implementing Bagging classifier along with its trees

From the above elucidation we can heed that the result on implementing a bagging classifier algorithm to the collected data has acquired same accuracy in two trees that is 0.61 as average accuracy in trees 15 and 20 whereas different accuracy for the rest two trees that is 0.60 and 0.62 as average accuracy in trees 5 and 10. And so the line in the graphical representation has moved gradually according to the accuracy.

4.6. XGBoost Classifier

XGBoost classifier is a decision tree rooted machine learning algorithm that is ensemble. They make use of gradient boosting as a framework. For the production of large scale problems these classifiers are battle tested. They are delineated for their speed and good performance which is their major advantage. To make use of it first we need to download and install this software library to our system and so it is expounded as an open source software library. This algorithm has acquired the most favorable among the competitors as this yields the scalable and portable reverberation.

Secured yield and accuracy by the implementation of XGBoost algorithm to the accumulated data is described in observation 6,

Observation 6:

In observation 6 we have described the yield achieved by implementing a XGBoost classifier to the garnered data to predict the profitability of the motion picture. In our exertion we have considered four trees and endowed the accuracy for each tree. The four trees that we have pondered are 10, 20, 50 and 100. The graphical representation of the outcome on implementing a XGBoost classifier algorithm to the garnered data is narrated in Figure 7.

From Figure 7elucidation we can heed that the result on implementing XGBoost classifier to the collected data is almost the same accuracy in all the four different trees that is 0.62, 0.59, 0.60 and 0.58 as average accuracy in trees 10, 20, 50 and 100. And so the line in the graphical representation has moved gradually according to the accuracy.

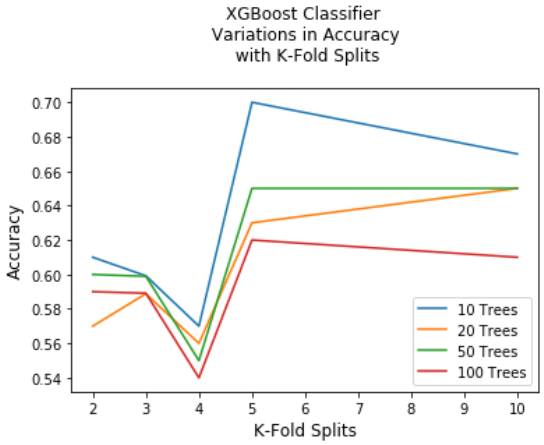

Figure 7: Graphical representation of the outcome on implementing XGBoost classifier along with its trees

4.7. Voting classifier

Voting classifier is a classification predictive technique that associates innumerable models and yields an output based on highest probability. Instead of generating a new model for each algorithm to yield the accuracy we have designed a single model which is trained by these models and yields the output according to their majority of voting. There are two distinct types in voting classifiers: in those two types we have made use of a hard voting classifier for our exertion.

Now let us understand what a hard voting classifier is. In this classifier the output is predicted according to the votes. When we have the output with two possible outcomes then the outcome with the majority votes is reported as the final predictions. We have made use of this classifier in our exertion and have predicted the output utilising it.

Secured yield and accuracy by the implementation of voting classifier algorithm to the accumulated data is described in observation 7,

Observation 7:

In observation 7 we have described the yield achieved by implementing a voting classifier to the garnered data to predict the profitability of the motion picture. In this observation we have not pondered any trees or depths. The graphical representation of the outcome on implementing a voting classifier algorithm to the garnered data is narrated in Figure 8

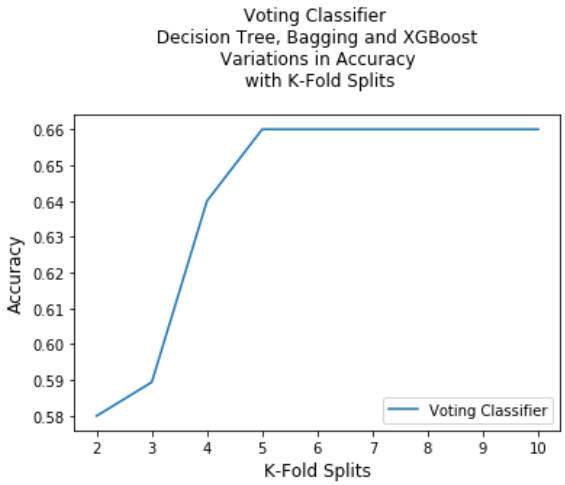

Figure 8: Graphical representation of the outcome on implementing Voting classifier

From the Figure 8 elucidation we can heed that the result on implementing a voting classifier algorithm to the collected data has acquired an average accuracy as 0.62. And so the line in the graphical representation has moved gradually according to the accuracy.

V. OBSERVATION RESULT

The above narrated inventory is about the catalogue of algorithms that are implemented to the data horned by us and their observation. So from the above observation we can regard that the implementation of classification algorithms to the garnered data to predict the profitability of the motion picture has acquired and yielded a good accuracy. By implementing classification algorithms we have acquired the following accuracy, for logistic regression algorithm we have acquired 62% as average accuracy, for naive bayes algorithm we have acquired 65% as average accuracy, for decision tree algorithm we have acquired 69%, for random forest algorithm we have acquired 61% as average accuracy, for bagging classifier algorithm we have acquired 62% as average accuracy, for XGBoost algorithm we have acquired 62% as average accuracy and for voting classifier algorithm we have acquired 62% as average accuracy.

VI. CONCLUSION

The aspiration of this exertion is to predict the profitability of the motion picture. We have acquired a good accuracy on the implementation of the classification algorithm to the horned dataset. But a success of a movie is not only related to the features of the movie whereas it is also related to the audience who plays a major role. That means our exertion is not the direct success but whereas it is a one way by which we can predict the success of the movie whereas the remaining lives in the audience's hand who play vital roles not only in case of cinema but also in distinct cases like politics and economics. When the economic stability is low, audiences will avoid to watch movies on screen from where the profitability of the movie begins to reduce this situation is barely confronted. And so we conclude that the profitability of a movie can be

predicted using our exertion but whereas it is not the final prediction the success rate also depends on the audience who reviews it.

REFERENCES

1. https://ieeexplore.ieee.org/document/8703320
2. Danussvar Jayanthi Narendran, Abilash R and Charulatha B.S. Exploration of Classification Algorithms for Divorce Prediction. In Proceedings of international Conference on Recent Trends in Machine learning, loT, Smart Cities and Applications, Springer 2020
3. Abilash R and Charulatha B.S. Early Detection of Diabetes from Daily Routine Activities: Predictive Modeling Based on Machine Learning Techniques. In Proceedings of International Conference on Big-Data and Cloud Computing, Springer 2019
4. https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3
5. https://towardsdatascience.com/what-makes-a-successful-film-predicting-a-films-revenue-and-user-rating-with-machine-learning-e2d1b42365e7
6. https://www.diva-portal.org/smash/get/diva2:1106715/FULLTEXT01.pdf
7. http://dspace.bracu.ac.bd/xmlui/bitstream/handle/10361/9015/13301028%2C13301019_CSE.pdf?sequence=1&isAllowed=y
8. http://www.laccei.org/LACCEI2017-BocaRaton/student_Papers/SP499.pdf
9. https://www.researchgate.net/publication/322138608_A_Machine_Learning_Approach_to_Predict_Movie_Box-Office_Success
10. https://www.researchgate.net/publication/332826643_Movie_Success_Prediction_using_Machine_Learning_Algorithms_and_their_Comparison
11. https://link.springer.com/article/10.1007/s11042-019-08546-5?shared-article-renderer
12. https://syslog.co.in/wp-content/uploads/2019/11/Movie-Success-Prediction-using-Machine-Learning-Algorithms-and-their-Comparison.pdf
13. https://www.statisticssolutions.com/what-is-logistic-regression/
14. https://en.wikipedia.org/wiki/Logistic_regressio
15. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
16. https://www.quora.com/What-is-the-math-behind-logistic-regression
17. https://www.geeksforgeeks.org/understanding-logistic-regression/
18. https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis
19. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
20. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
21. https://www.geeksforgeeks.org/naive-bayes-classifiers/
22. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
23. https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4
24. https://en.wikipedia.org/wiki/Decision_tree_learning
25. https://www.geeksforgeeks.org/decision-tree-introduction-example/

26. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
27. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
28. https://builtin.com/data-science/random-forest-algorithm
29. https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674
30. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm
31. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html#:~:text=A%20Bagging%20classifier.,to%20form%20a%20final%20prediction.
32. https://www.geeksforgeeks.org/ml-bagging-classifier/
33. https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/
34. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d
35. https://data-flair.training/blogs/xgboost-algorithm/
36. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/
37. https://www.geeksforgeeks.org/ml-voting-classifier-using-sklearn/
38. https://medium.com/@sanchitamangale12/voting-classifier-1be10db6d7a5
39. http://rasbt.github.io/mlxtend/user_guide/classifier/EnsembleVoteClassifier/
40. https://stackabuse.com/ensemble-voting-classification-in-python-with-scikit-learn/