Early Prediction of Movie Success using Machine Learning Models

D.M.L. Dissanayake
Department of Computing and Information Systems
Faculty of Applied Sciences
Wayamba University of Sri Lanka

V.G.T.N. Vidanagama
Department of Computing and Information Systems
Faculty of Applied Sciences
Wayamba University of Sri Lanka

ABSTRACT

The film industry is a multi-billion-dollar business that is spread all over the world. A very high number of films are being released every year. But only a few are successful and most are failures. If the success of a movie can be predicted and reduce the uncertainty at the early stages of the movie-making process, that will make a significant impact on the film industry because of the immense investments that are made. The success of a movie is based on several factors related to the past, present and future. By identifying the factors that are relevant to the success of a movie, it can be predicted accurately. Creating predictive models with the use of machine learning has become a trend in the recent past due to the availability of large volumes of data and high computational capabilities. Prediction models and currently available machine learning methods can be used to predict the success of a movie. This paper describes a novel approach using machine learning methods to predict the success of a movie in advance. In this paper, multiple regression and classification methods were used for training and testing the dataset and their performances were evaluated to identify the well-fitted model. The Support Vector Machine model showed a movie success prediction rate of 100% on the test data.

Keywords

Classification, Machine Learning, Movie, Prediction, Regression

1. INTRODUCTION

Movies are considered as one of the major mediums of entertainment in current society. A movie can be focused on a social issue, political issue, or a real-life story, etc... For some audiences, a movie is just a method of entertainment but for some audiences, it is a source of information. Thus, the movie has a significant influence on the public.

Even though a large number of movies are released per year, only a few of those movies have a higher success rate and make an impact on the public as well as the film industry. Making a quality movie is a complex process that consists of multiple stages. It is more than just writing a script, finding a cast, and filming the scenes. Each stage is a time-consuming phase with a fairly large cost attached to it. Therefore, making a film is an investment that has an immense risk for every film producer.

The producer and director can be identified as the main responsible parties in the process of making a film. One of their main goals is to make a successful movie. Since they carry the responsibility of making a successful movie it can cause pressure on them and that can affect the quality of the movie. Ultimately, the effort to make a movie successful can be the reason for its own failure.

The success of a movie is based on multiple factors. Budget, actors, script, locations, marketing strategies are a few of those. But out of all those factors, it is undeniable that every artistic and creative product must face the final judgment of the public. The public audience is a dynamic factor. Therefore, it is required to identify those factors correctly and the relation between those factors and the success of a movie. Then predicting the success of a movie is possible with the association of predictive analysis and machine learning techniques. Since the film industry is uncertain and rich, many researchers are carrying on the topic of movie success prediction. If an accurate model is built then the uncertainty in the film industry can be reduced and any financial loses can also be prevented.

2. LITERATURE REVIEW

The movie industry is a vast industry that involves a large number of investments. Despite the fact that the movie industry has large investments involved, the success or profitability of a certain investment is uncertain. Thus, an effective and reliable prediction model can be beneficial for investors and other industries linked to the movie industry. Numerous previous studies were performed on this topic, but the lack of satisfactory results has become a major drawback. The availability of multiple dynamic factors that have an impact on the success of a movie has made it hard to design an effective model to predict the success[1] . But in previous work by
[2]concluded that the total box-office receipts of a particular movie can be forecasted with very high accuracy after determining the first week of box office receipts. Their reasoning for that conclusion is demand for the box office receipts tends to tail off later after the first two weeks of the release of the movie. In this research work, several models have reached a higher accuracy of more than 90%. It has shown that higher accuracy can be achieved for the movie success prediction and hence satisfactory results can be achieved.

Many online platforms and websites provide a large amount of data regarding the movies. With the rapid growth of information sources, it can be a timely and possible task to design a good model for predicting the success of a movie. When building a predictive model online ratings and tags can be used as a good attribute regarding the audience approach for a certain movie. Because social media and online platform users have plenty of space to express their opinions regarding a movie and it is available for everyone[3]. This point is also proven in this current work. The selected 10 features for model creation have several features that are directly associated with the social media platforms.

Machine learning has become a popular and impressive technology in the recent past. It has been used with various fields and has performed excellently in the given conditions. Prediction can be introduced as a sub-topic in machine learning. Even though machine learning has the capability of addressing a variety of empirical questions it has some limitations. A major component in prediction in machine learning is data. Without proper data, it is impossible to get a quality and accurate result. After obtaining valid data, there are multiple steps to follow to build an effective model as the outcome[4].

After Initial data pre-processing steps both numerical and categorical attributes can be available for predictions. R and R studio can be used to convert categorical/nominal attributes to numeric values. In that process, it converts those categorical variables into binary features. If a variable has more than two possible values it will convert into n-binary features where n is the number of values. It is better to perform statistical tests in data before applying machine learning models even though machine learning techniques work on the principles of statistics[5].

With the availability of a large number of peer-reviewed journal articles published in the 2010 to 2020 decade regarding the prediction of the revenue of a movie and with the growing number of literature reviews on movie success prediction with the use of machine learning techniques, a systemic review will help to understand the research domain properly. A study has based on the following four questions has been performed.

1. What are machine learning techniques have been applied to predict the revenue of a movie and which are the most accurate techniques?
2. What are the data sources that commonly predict the revenue of a movie?
3. What features are used for movie revenue prediction and what is most important among them?
4. What are the evaluation metrics regularly used in movie revenue prediction?

With the use of the above questions, it is observed that cast, number of screens, and genre are the most used features in the revenue prediction articles. Regression, classification, and clustering approaches have mostly been used in these journal articles as machine learning models. Mean absolute percentage error, root mean square error, and average percentage hit rate is the most used evaluation matrices[6].

There have been several approaches used to predict the success of a movie. One approach is building a custom framework. Frameworks can have several steps in their architecture. The Movie Investor Assurance System (MIAS) system with the use of social network analysis and text mining techniques, extracts several types of features mainly including people is in the cast, what the movie is about, releasing date of the movie, and other hybrid features. Users of the MIAS framework can define their threshold value or a profitability metric according to their goals that are associated with the movie. Even though novel features that proposed had made a weighty contribution, experiment results show that the system has outperformed benchmark methods from a considerably large margin. The addition of a more realistic cast selection process and reduced computational complexity can use to improve the method of prescribing cast members[7].

Even [3] claims that a large amount of data available regarding movies[8]suggest a different suggestion regarding the Indian film industry. The Indian film industry called as the Bollywood industry is a major part of the cinema world. It is one of the largest movie industries with approximately 1500-2000 movies per year in 20 different languages. Even though the Bollywood industry is large in size data available on the Internet regarding it is limited. Since the data is limited, it is difficult to obtain data through websites. Therefore, web scraping, questioners have to use to obtain the data. A unique feature present in Bollywood movies is the music score of the movie. It could not be found in one place and therefore it has to collect in bits and pieces manually across the Internet. This will reduce the size of the dataset that has to use in the analysis.

In such instances as the Bollywood movie industry which currently does not have a large volume of data available on the Internet sentimental analysis can be effective. Predicting the success of a movie with the related attributes is a conventional method and those methods could give an idea about whether the movie is good or bad but not the box office collection. So social media contents with rich and custom information regarding viewers' preferences can have a major impact on a study in movie success prediction[9].

Another interesting approach is using a weighting schema for predictions. In these approaches, custom weights can be calculated and assigned to each attribute. Schemas with custom weighting are better at capturing the information and produce better results in most cases[10]. Even though the weighting approach produces better results it can be difficult to implement. Therefore, for this research work, feature selection methods were used.

Prediction of box-office success can be treated as a classification problem as well as a regression problem. Improvements for the prediction accuracy are proposed by Cinema Ensemble Model (CEM). New features like transmedia storytelling can make an impact and increase the accuracy of a model from past studies. Transmedia storytelling is delivering a single story across multiple media channels. It has been used and is currently used in both academia and industries[11]

3. METHODOLOGY

This research work followed various steps as shown in Fig. 1. The mentioned steps in Fig. 1 were followed in sequential procedure and a correct understanding of every step is required in advance. For this research, the dataset was obtained from the "Kaggle.com" website. Initially, it had both numeric and categorical columns and the dataset was not cleaned. Since the dataset is not cleaned it contained a considerable amount of noisy and null data. Therefore, data preprocessing methods had to be followed as the initial step.

3.1Data Preprocessing

Data preprocessing can be identified as the process that transforms raw data into a more usable and understandable format. It includes removing incorrect, incomplete, and inaccurate data as well as replacing missing values. Data preprocessing is required, as the use of raw data for analysis will result in inaccurate results. Preprocessed data can result in an accurate, complete, and consistent data analysis [12].

Data preprocessing consists of four main steps. Those steps are,

1) Data cleaning
2) Data integration
3) Data reduction
4) Data transformation

In this work, the duplicate rows and rows with duplicate values were removed from the dataset to reduce the noise of the dataset. If not, that raw data can cause underfitting or overfitting problems. From the initial twenty-eight columns, some did not have an impact on the predictions and some cannot be used for the predictions. Therefore, those columns were removed before moving further. Removing those columns causes dimensionality reduction as well.

3.2 Feature Selection

Arguably the features (independent parameters) are the main factor of any machine learning model because they make a large impact on the model. Identification and selection of parameters have to be performed more precisely so that it enables the most accurate model to be built.

Feature selection is the process that identifies the most relevant attributes to the target variable from the number of input variables available. Feature selection is used because all the input variables in the dataset are rarely useful to build a model. There are two main feature selection techniques which are, Supervised and Unsupervised. The supervised method can be further divided into a wrapper, filter and intrinsic[13].

This paper uses multiple feature selection processes which are the Chi-square method, F-Test method, feature importance method and Recursive Feature Elimination (RFE). The reason for using more than one feature selection method is that it allows for a selection of highly correlated parameters while ensuring the correctness of the selection. Results from all earlier mentioned feature extraction methods were considered before selecting the final features that are correlated with the target variable.

The Chi-square method is a statistical test that is based on a hypothesis. This test is used for categorical attributes available in the dataset. It assumes that the observed frequency of a categorical variable matches the expected frequencies for the categorical variable. This method is used to calculate the chi-square value between each categorical attribute and the target attribute and then determine the attributes based on the chi-square values[14].

Analysis of variance (ANOVA) is a statistical hypothesis testing method that is used to determine whether two or more attributes are from the same distribution or not. ANOVA is mainly used when there are numerical variables as well as categorical variables present in the distribution[15].

Even though the ANOVA method is also calculating a score for each independent variable as same as the chi-square method, the calculating methodology is different from each other. In this paper’s implementation, it was decided to select the best ten parameters.

The feature importance method will also select the features by assigning a score calculated especially to each input attribute. That score denotes the importance of each feature concerning the target variable. Since feature importance score can indicate which features may be most relevant to the target variable, it can provide an insight into the dataset[16]. The result for the feature importance method is shown in Fig. 1 and all the selected ten features have higher scores and those scores are almost equal. Hence, it can be said that the selected features have almost equal impact on the dependent variable individually.

The recursive feature elimination is a method that fits a model and removes the weakest features from the available features. It will stop its recursion once the pre-specified number of features is reached. In this method, features are ranked either by the model's coefficient attributes or feature importance attribute[17].The feature importance method works in the backward selection of the predictors. Initially, all the attributes except the target variable will be used to build the model and then the importance score is computed. After computing the importance score, predictors with the least important score are removed and the model is re-built. Then again, the feature importance of the available predictors will be calculated. This process will iterate until the remaining features are equal to the pre-specified limit[18].For the recursive elimination method, this paper uses the logistic regression and a thousand iterations were set as the number of maximum iterations.

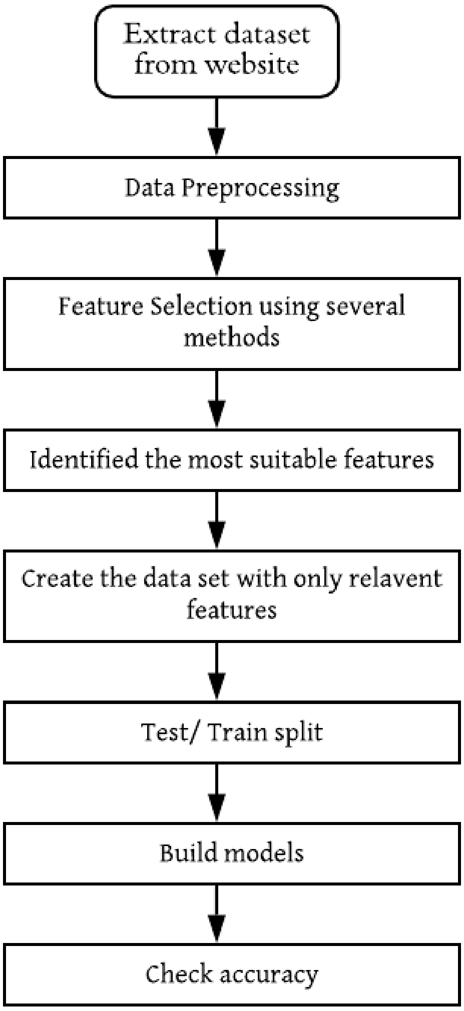
Extract dataset from website
Data Preprocessing
Feature Selection using several methods
Identified the most suitable features
Create the data set with only relavent features
Test/ Train split
Build models
Check accuracy

Fig 1:Basic steps involved in this research

With the use of results from the above-mentioned methods, the final features were selected. They were,

1) Num_critic_for_reviews
2) Director_facebook_likes
3) Actor_3_facebook_likes
4) Actor_1_fb_like
5) num_voted_users
6) Cast_total_fb_like
7) Num_users_for_review
8) Budget
9) Actor_2_fb_likes
10) Movie_facebook_likes

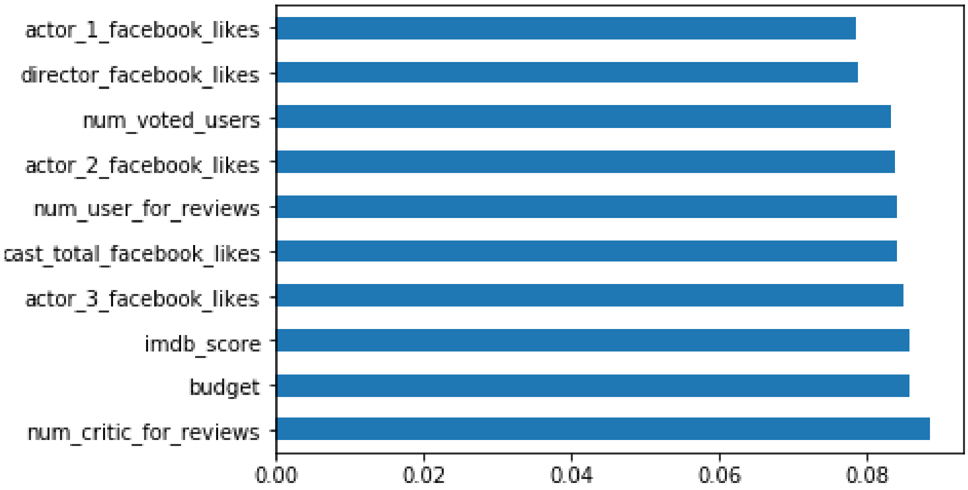

Fig 1: Selected features with their score from the feature importance method

The correlation heatmap was used to check the validity of the features selected by the feature selection methods according to their correlations. A correlation heat map is a two-dimensional heatmap that shows the correlation between two attributes. The color of the cell indicates the correlation level and it is ideal for analysis purposes. If the two features have either strongly negative (-1) or strongly positive (+1) correlations those predictor variables can be considered as features for training the models.

3.3 Building regression models and classification models

A new data set was created that only has the above-selected features as independent parameters. The new dataset was separated into two parts as the dependent variable and the independent variables. In the process of building a regression model "gross" was considered as the dependent variable. For the classification problem, a new column of "Success" was added as the dependent variable. It contains only binary values (0,1). Value 1 was assigned to the success column if the budget of the movie is less than the gross. Otherwise, the value 0. With the completion of separating dependent and independent variables, a test-train split was performed. 20% of the total data set was allocated to the test set and the remaining 80% to the training set.

Regression analysis quantifies the relationship between the independent variable and dependent variable. Multiple linear regression has two or more variables as independent variables. Polynomial regression algorithm models the relationship between independent and dependent variables as a polynomial of the n^{th} degree. Support vector regression and support vector machine methods are supervised algorithms and they output the relationship of dependent and independent variables in a hyperplane which separates the classes well.

Classification analysis identifies and assigns data that has the same properties into the same categories. In logistic regression classification, logistic function when modeling the dependent variable. K- nearest algorithm finds the distance between data points and assigns the data points that have nearly the same distance into the same class. Regression tree analysis in both regression and classifications problems, will divide the dataset into smaller subsets and create the tree structure while splitting the dataset into subsets.

Random forest algorithm builds multiple decision trees and merge them to achieve accurate results. The naive Bayes algorithms use the Bayes theorem for classification.

3.4 Evaluating built models

Models we build should be able to perform well and give accurate predictions with the test data. Therefore, we always need to ensure that our model is accurate and reliable with its predictions. So that we can apply it for predictions without any doubts. To ensure the accuracy and performance are high, we use performance evaluation matrices to assess the performance of our models. Matrices are required in the performance evaluation to quantify the performances. The metrics are depending on the task of the model. Those tasks can be classification, ranking, regression, clustering, etc..[19]. R^2 Score, mean square error and root mean square error can be identified as one of the major estimators to evaluate a regressor. These factors will describe how the model fits and give an estimation about the relationship between the dependent variable and independent variables.

R^2 score is also called the coefficient of determination. It is a statistical measure that uses by regression models to measure the amount of variation of the dependent variable that can be predicted by independent variables. In simple terms, the R^2 score illustrates how well the data fit the regression model.

$$R^2 \text{ score} = 1 - SS_{res}/ SS_{tot} \text{----------(1)}$$

Where;

$SS_{res}$is the sum of squares of residual errors
SS_{tot} is the total sum of errors

R^2 score can take any value between 0 and 1. The value of the R^2 score can be referred to as the percentage of the changeability of the dependent variable that can be explained by the model[13]. Mean Square Error (MSE) can be defined as the mean of the square of the difference between real and estimated values.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{n}(actual\ value - predicted\ value)^2 \text{----------(2)}$$

Where MSE is the mean square error, and N is the total number of rows in data.

Root Mean Square Error (RMSE) can be defined as the standard deviation of the errors that occur while predicting the values. It is the same as the root mean square error but the root value of the error is considered while calculating.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(actual\ value - predicted\ \ value\ )^2}{N}} \text{-------------------(3)}$$

Where RMSE is the root mean square error and, N is the total number of rows in the dataset.

It is a fact that no machine learning algorithm is perfect when it applies to real-world scenarios because of the uncertainty of real-life scenarios. Therefore, there can be misclassifications. The classification made by the binary classification model can be divided into four major parts. Those are;

1) True positive
2) True Negative
3) False Positive
4) False Negative

True positive is when the model correctly predicts the positive class. That means both predicted value and actual value are positive. True negative is predicting the negative class correctly while actual value is in the negative class. False-positive is the wrong prediction of the positive class while the actual class is in the negative class. False-negative is wrongly predicting the negative class while the actual class is positive. False-positives and false-negatives occur due to misclassifications. Since there is a misclassification, it confuses classifying. This confusion of classifying the data can be described using a 2×2 matrix called a confusion matrix[20].With the use of the confusion matrix, it is possible to calculate accuracy, precision, and recall. Precision is also called the positive predictive value. It can be defined as the ratio of correct positive predictions to the total positive predictions. Simply, precision means what proportion of positive predictions was actually correct.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \text{------------------------(4)}$$

The recall is the ratio of correct positive predictions to the total real positive values. That is what proportion of actual positives were correctly identified.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \text{------------------------------(5)}$$

Accuracy is the ratio between correctly predicted values and the total number of predictions. In terms of the confusion matrix, it can be defined as follows.

$$\text{Accuracy} = \frac{True\ Positive + True\ Neagative}{True\ Positive + True\ Negative + False\ Positive + False\ Neagtive} \text{--------(6)}$$

The Receiver Operating Characteristic (ROC) curve is used to ensure the calculated values of accuracy.ROC curve illustrates how the classification model performs at all classification thresholds. It graphically represents the relationship between two parameters, true positive rate (recall) and false-positive rate. The area under the ROC curve (AUC) is a key factor of the ROC curve. Higher the AUC better the classifier. A perfect classifier will have an AUC of 1[21].

4. RESULTS AND DISCUSSION

4.1 Results

In this research paper, five regression models and seven classification models were built and relevant data required to measure the accuracy and fitting of the model was reported. Relevant ROC curves were also drawn and coefficient matrices were also generated for further inquiry of performances. While creating models their accuracy was also measured. R^2 score, mean square error, and root mean square error values were calculated for each regressor model that was built as tabularized in table 1.

To measure the accuracy of each classification model, a confusion matrix was calculated. With the use of a confusion matrix and ROC curve, it is easy to ensure the fidelity of the models proposed here. Table 2 illustrates the accuracy, precision, and recall observed for every classification model that is proposed in this paper.

4.2 Discussion

As per Table 1, all the regression models reported a relatively low R^2 score. The lowest R^2 score value was 0.3366 and which was from the polynomial regression model and the highest R^2 score was 0.6822 which was recorded by the random forest regressor. R^2 score demonstrates how well the model is fit. The ideal R^2 score is 1 and if the R^2 score is closer to 1, the model is said to be the closest fit. Support Vector Regression (SVR) regressor had a negative R^2 score value. That means the SVR regressor that we created was not fitted to the trend in the dataset. Since SVR recorded a negative R^2 score value another SVR model was created with the scaled parameters. But it also resulted in a negative value. Hence SVR model can be mentioned as out of trend from the dataset.

The mean squared error and root mean squared error has resulted in higher values. The reason for those higher values is that the parameters were not scaled before fitting them to the model. Also, outliers can generate a higher error value.

One of the major results observed was that the decision tree regressor and random forest regressor models have resulted as overfitting models. That can be due to the high noise in the data set and less training data. To avoid overfitting, tree pruning can be used with the decision tree regressor and model parameter tuning can be used with the random forest method.

All the classification models that we built performed well except for the Kernel Support Vector Machine (KSVM) model. A major highlight was that the SVM model resulted in 100% accuracy and 1.0 precision and recall value. That means out of all the regression models and classification models SVM classifier has the best result and is very well fitted to the dataset. The naïve Bayes classification model had 71.54% accuracy. Generally, it can be considered as a good accuracy, but in comparison to the accuracy of other classification models, it is a slightly low value for accuracy. Out of all seven classifiers kernel SVM was identified as the overfitting model since it records lower values of accuracy, precision, recall for the test set than the training set.

Table 1. Results of regression models

	Multiple Linear Regression	Polynomial Regression	SVR			Decision Tree Regression	Random Forest Regression
			With Feature Scaling	Without Feature Scaling			
R^2 Score	0.4656	0.3336	-0.5317	-0.1033		0.4907	0.6822
Mean Square Error	320183	324668	917728	537545		305126	1903958
Root Mean Square Error	56584745.50	56979725.788	95798150.071	73317506.968		55238247.27	43634374.40

Table 2. Results of classification models

	Logistic Regression	K-Nearest Neighbour	SVM	Kernel SVM	Naïve Byes	Decision Tree	Random Forest
Accuracy (%)	92.48	98.52	100.0	53.95	71.54	97.85	94.09
Precision	0.9621	0.9777	1.0	0.5371	0.8163	0.9775	0.9682
Recall	0.8944	0.9949	1.0	1.0	0.6030	0.9824	0.9195

Out of the seven classification models built, the SVM model had the highest accuracy. Because it did not have any false positive or false negative productions as shown in Fig. 3. But the kernel SVM model has a low accuracy level which is 53.95%. That accuracy level is due to the almost same number of predictions of false-positive and true positive values as shown in Fig. 4. Even though accuracy is average in kernel SVM it has a recall of 1. It implies that kernel SVM returns a large number of predictions but those predictions are not highly accurate.

Higher precision and recall values have been recorded for all the classification models except for the kernel SVM model. Higher precision and recall values emphasize that the models can identify the relevant data and it is correct almost all the time. ROC curves were used to ensure the performance demonstration by the accuracy, precision, and recall values.

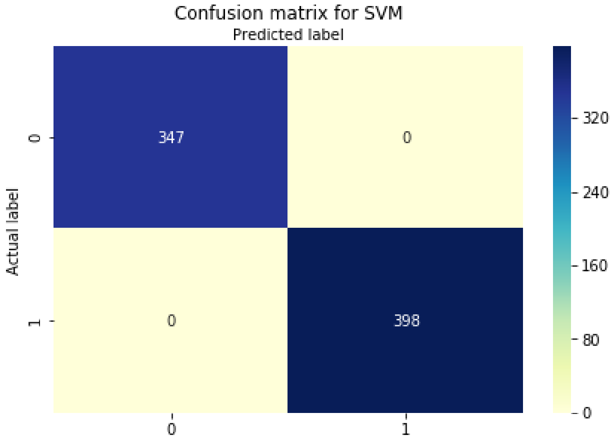

Fig 3: Confusion matrix of SVM

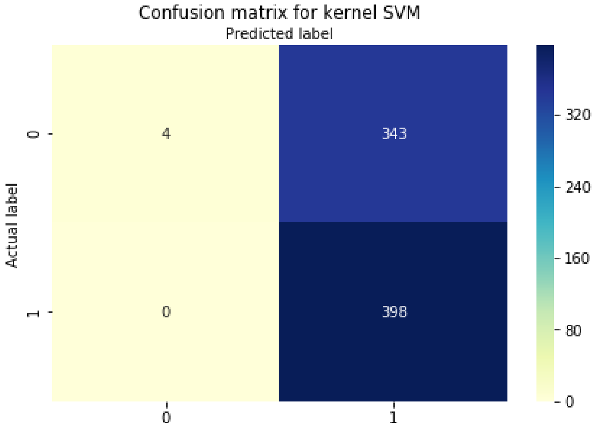

Fig 4: Confusion matrix for kernel SVM

ROC curve assured the fact that SVM classifier has the highest accuracy and kernel SVM and naïve Bayes classifiers are slightly underperforming concerning the other classification models. The 100% accuracy suggested by the confusion matrix is confirmed by the ROC curve graph. As shown in Fig. 5 area under curve value is 1 for the SVM classifier.

Also, the nearly 50% accuracy of the kernel SVM model is confirmed by the ROC curve since it has nearly 0.5 value for the area under curve value as shown in Fig. 6.

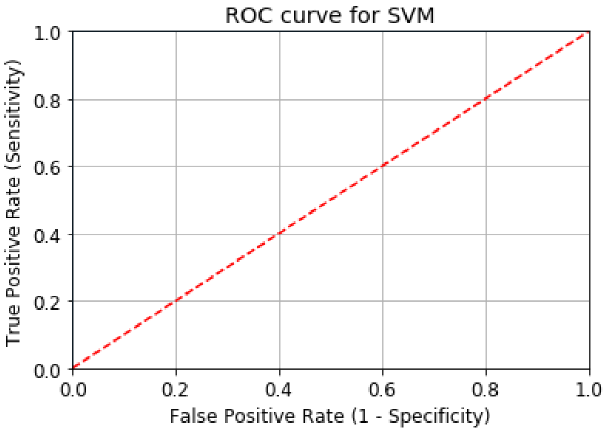

Fig 5: ROC curve of SVM that verify the 100% accuracy

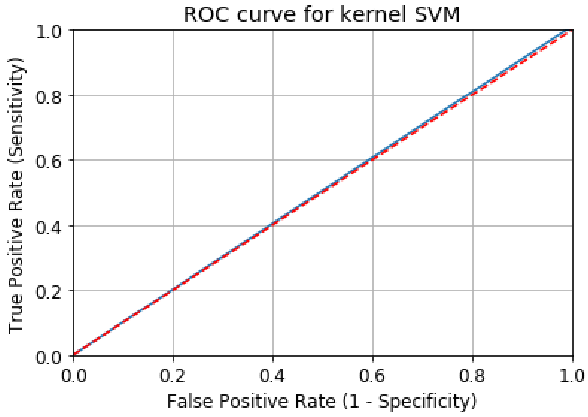

Fig 6: ROC curve of kernel SVM that verifies the nearly 50% accuracy

5. CONCLUSION AND FUTURE WORKS

5.1 Conclusion

The purpose of this research was to use machine learning models to predict the success of a movie in advance. The related research works have not used both classification and regression models simultaneously. But in this research, both regression models and classification models were used. Hence it provides an option regarding whether to treat movie success prediction as a classification problem or regression problem in future works. The initial dataset which was downloaded from the "Keggale.com" website has 5024 rows. But with the feature selection process, 25% of rows were removed.

As per[22]data preprocessing has to be at higher levels to achieve high accurate models. This paper was able to achieve high accuracy for almost all the proposed classification models. Impact achieved here with a maximum accuracy of 01 for the SVM classifier. This implies that the approach used here for data pre-processing is valid and productive.

As regression models, the paper built multiple linear, polynomial, SVR, decision tree, and random forest regressors. But these regressors have resulted in poor performance. Two of them were overfitted. But the classification models that were built performed with higher accuracy. Except for the kernel SVM and naïve Bayes classifiers, while all other classifiers resulted in accuracy higher than 90%. Even though naïve Bayes is lacking inaccuracy, it has an accuracy of more than 70%.

After observing all the results and values related to models, it was concluded that the used data set still has some noise data even after the data preprocessing. Hence the overfitting occurred. One of the major factors identified from the selected features is that social media, Facebook in this case has a major contribution and impact on the success of a movie. Finally, it can be concluded that the data values were well fitted to the created classifiers rather than the created regressors. Since the SVM classifier resulted in the highest values for accuracy, precision and recall which is 1.0, that can be defined as the best model for this research problem.

5.2 Future works

This research work only focused on a model that can be used for prediction. Proposed models depend on feature selection methods to identify the most suitable features. But it is possible to use advanced statistical analysis to identify the features that have a high impact on a success of a movie. Also, the data set was obtained from a website. Instead of downloading a dataset with the use of a method such as data scraping, conducting surveys is also encouraged. With the emerging fact that social media has a major influence on the film-making industry analyzing social media concerning the film industry can result in interesting findings.

6. ACKNOWLEDGMENTS

I would like to express my deepest appreciation for all the immense support given by the all-academic staff at the Department of Computing and Information Systems, all my colleagues, and my family members.

7. REFERENCES

[1] I. S. Ahmad, A. A. Bakar, M. R. Yaakub and S. H. Muhammad, "A Survey on Machine Learning Techniques in Movie Revenue Prediction," SN Computer Science, p. 235, 2020.

[2] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," Expert Systems with Applications, vol. 30 , no. 2, p. 243–254, 2006.

[3] S. Wei, X. Zheng, D. Chen and C. Chen, "A hybrid approach for movie recommendation via tags and ratings," Electronic Commerce Research and Applications, vol. 18, pp. 83 - 94, 2016.

[4] X. (. Wang, J. H. (. Ryoo, N. Bendle and P. K. Kopalle, "The role of machine learning analytics and metrics in retailing research," Journal of Retailing, 2020.

[5] S. M. R. Abidi, Y. Xu, J. Ni, X. Wang and W. Zhang, "Popularity prediction of movies: from statistical," Multimedia Tools and Applications, pp. 47-48, 2020.

[6] I. S. Ahmad, A. A. Bakar and M. R. Yaakub, "Movie Revenue Prediction Based on Purchase Intention Mining," Information Processing and Management, vol. 57, no. 5, 2020.

[7] M. T. Lash and K. Zaho, "Early Prediction of Movie Success: The Who, What and When of Profitability," Journal of Management Information Systems, vol. 33, no. 3, pp. 874 - 952, 2016.

[8] H. Verma and G. Verma, "Prediction Model for Bollywood Movie Success: A Comparative Analysis of Performance of Supervised Machine Learning Algorithms," The Review of Socionetwork Strategies, 2019.

[9] S. Mundra, A. Dhingra, A. Kapur and D. Joshi, "Prediction of a Movie's Success Using," in Information and Communication Technology for Intelligent Systems, Singapore, 2019.

[10] R. Parimi and D. Caragea, "Pre-release Box-Office Success Prediction for Motion Pictures," Machine Learning and Data Mining in Pattern Recognition, vol. 7988, pp. 571-585, 2013.

[11] K. Lee, J. Park, I. Kim and Y. Choi, "Predicting movie success with machine learning techniques: ways to improve accuracy," Information Systems Frontiers, vol. 20, pp. 577-588, 2018.

[12] S. Anunaya, "Data Preprocessing in Data Mining -A Hands On Guide," 10 08 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/. [Accessed 12 08 2021].

[13] J. Browniee, "How to Perform Feature Selection With Numerical Input Data," 18 08 2020. [Online]. Available: https://machinelearningmastery.com/feature-selection-with-numerical-input-data/. [Accessed 21 05 2021].

[14] J. Brownlee, "A Gentle Introduction to the Chi-Squared Test for Machine Learning," 31 10 2019. [Online]. Available: https://machinelearningmastery.com/chi-squared-test-for-machine-learning/. [Accessed 21 05 2021].

[15] J. Brownlee, "How to Choose a Feature Selection Method For Machine Learning," 27 11 2019. [Online]. Available: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/. [Accessed 19 03 2021].

[16] J. Browniee, "How to Calculate Feature Importance With Python," 20 08 2020. [Online]. Available: https://machinelearningmastery.com/calculate-feature-importance-with-python/. [Accessed 21 02 2021].

[17] "Recursive Feature Elimination," [Online]. Available: https://www.scikit-yb.org/en/latest/api/model_selection/rfecv.html. [Accessed 09 07 2021].

[18] M. Kuhn and K. Johnson, "Feature Engineering and Selection: A Practical Approach for Predictive Models," 21 06 2019. [Online]. Available: https://bookdown.org/max/FES/. [Accessed 09 07 2021].

[19] S. Mutuvi, "Introduction to Machine Learning Model Evaluation," 16 04 2019. [Online]. Available: https://heartbeat.fritz.ai/introduction-to-machine-learning-model-evaluation-fa859e1b2d7f. [Accessed 21 05 2021].

[20] V. Jayaswal, "Performance Metrics: Confusion matrix, Precision, Recall, and F1 Score," 14 09 2020. [Online]. Available: https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262. [Accessed 07 08 2021].

[21] V. Jain, "Idiot's Guide to Precision, Recall, and Confusion Matrix," [Online]. Available: https://www.kdnuggets.com/2020/01/guide-precision-recall-confusion-matrix.html. [Accessed 08 08 2021].

[22] N. Quader, M. O. Gani and D. Chak, "Performance Evaluation of Seven Machine Learning Classification Techniques for Movie Box Office Success Prediction," in 3rd International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 2017.