# Neuro-Predict:A Machine Learning Approach To Predict Success of Lollywood Movies

Muhammad Ijaz
*Department of Electrical Engineering*
*UET Lahore, New Campus*
Lahore, Pakistan
2017EE309@student.uet.edu.pk

Muhammad Talha Sial
*Department of Electrical Engineering*
*UET Lahore, New Campus*
Lahore, Pakistan
2017EE340@student.uet.edu.pk

***Abstract*—The cinema industry is growing at an unprecedented rate with hundreds of movies being released annually. Likewise, the capital investment being made in this industry is also topping the charts. The global box office passed $ 42 billion in 2019 and the forecast for 2020 is also expected set records. However, this industry, albeit lucrative, is riddled with risks. The movie industry faces its fair share of flop movies alongside its hit movies. In a scenario like this it helps to know whether a movie will be a hit or a flop prior to its release so the production houses can then accordingly plan their decisions. This paper aims to introduce an efficient approach to predict a movie's success prior to its release using several machine learning algorithms like Naive Bayes, Random Forest, KNN, Multi-Layer Perceptron, SVM and Decision Tree. Our study can be used as an insight by movie practitioners to predict the success of their movie before its release.**

KEYWORDS

Machine Learning, Movie Prediction, K-Nearest Neighbour, Random Forest, Artificial Intelligence, Bayesian Networks, Neural Networks, Data Mining.

## ABBREVIATIONS

- *NB(Naive Bayes)*
- *LR(Logistics Regression)*
- *RF(Random Forest)*
- *SVM(Support Vector Machine)*
- *MLP(Multi-Layer Perceptron)*
- *DT(Decision Tree)*
- *KNN(K-Nearest Neighbour)*
- *CV(Cross Validation)*

## I. INTRODUCTION

Lollywood is Pakistan's biggest movie industry, producing a fair share of movies annually, some of which are at an all time highest-grossing. However, due to several reasons less and less movies are susceptible of being a success in this present time. Absence of creative ideas, repetition of story lines and lack of equipment has all but contributed to the undoing of Lollywood Industry [1]. An industry which once was a booming success in the early years of its start is now an empty shell of its former glory, which has ultimately led the production houses to not dabble in this risky venture. The objective of this paper is to propose a machine learning model that can predict a movie's success or failure based on it winning a Lux Style Award. The model is trained on Star Cast of movies and by using an actor's/actresses's previous success ratio it can predict the movie's outcome. This approach is highly lucrative for directors as it eliminates the exhaustive and expensive approach of choosing the right star cast for their movies. With the expansion of movie industry, the availability of data on internet has similarly increased, making it possible to analyze the data and use it for prediction purposes. The proposed work extracts data about the movie's details from Omdb (open movie database Api website) using a web scraper. This creates a data set containing all the vital information needed such as release dates, star cast, director's name, composer, investment etc upon which our model can be trained. Although research on this topic has been going on for quite a while, Lollywood is still an industry which has remained unchartered on this area. It is hoped our study will be used as a proof of concept of the application of machine learning in Lollywood Industry and sway any likely investors into making the right decision

## II. LITERATURE REVIEW

Movie success prediction is a research induced topic with many experts and neural network scientists pouring their studies and hard work to develop a lasting model. In [2] authors collected their data from online sources using a web scrapper and selected a total of 3 features to make the prediction which are; Total number of screens used for a movie, IMDb rating and Music Rating (a feature exclusively in Bollywood). This study is unique in the sense of it using Music Rating as a feature. As every Bollywood movie contains a certain number of songs that have a direct impact on the movies rating the authors have used Logistic Regression as the solution which they claimed gave them better performance.

Some previous work [3], use IMDb dataset from kaggle.com to predict the IMDb score of a movie. This approach is quite data intensive as it uses 28 variables for prediction along with a time span ranging across 100 years. Their feature selection focuses more on the audience reaction rather than the movie's attribute. The key feature in their study is the number of audiences for any given movie. The authors propose a total of five models but claim Random Forest as the best classifier scoring upto 61% accuracy

In [4], authors have approached this problem in a different light. Apart from using a variety of classifiers to predict the movie's success, they have also tried to determine which feature holds the most predictive power in swaying the classifier to make a decision. Their features include Genre, Academy Awards, Directors and Leading Role. From their study it can be inferred which feature correlates best with another feature. This can significantly reduce the size of unnecessary features and make the data more compact to work with. Moreoever they have also tried to measure the accuracy of models at pre-production level to make the idea more persuasive to stakeholders

## III. Methods

### A. Data Details

The dataset we are using was collected from various online sources and is composed of a total of 219 movies spanning from 2002 to 2018 with a total of 533 actors and actresses. Our target variable is Awards and our possible predictors are Director Name, Producer Name, Release Date, Composer Name and Star Cast. However owing to noise in data and some empty values we have decided to use Star Cast as our only predictor by using the success ratio of the actors. Furthermore we have also taken into consideration the fact that some actors have only a few movies to their name and so we have excluded them from our dataset to ensure better accuracy

### B. Normalization

As our input variables are strings we have used a label encoder from sci-kit's pre-processing library to convert them into numericals. Furthermore we have scaled our data such that the distribution is now centred around 0, with a standard deviation of 1 by using StandardScaler from sci-kits library which assumes our data is normally distributed within each feature.

### C. Sampling

We have experimented with various sampling methods in order to glean maximum accuracy. Among those are Train Test Split, Stratified K_Fold, Stratified Shuffle Split and Hold Out Method. Within Stratified K_Fold we tested each model individually by changing the value of K from K=2, K=10 and K=20 as shown in Table 1.1. In k-fold the entire dataset is randomly divided into folds of equal size . The model is then trained and tested K number of times. This approach ensures that every entry of the dataset is likely to be included in training and testing dataset. Our split ratio divides 80% of data for training and reserves 20% for testing.

### D. Feature Selection

In past studies several models have been made that use a variety of features for prediction, ranging from Social Media Reach, Sentiment Analysis and IMDb score to Movie Budget, Director Rank and No. of Screens used. Neverthless, for the sake of brevity we are only using the Star Cast as our input feature and we have included the Award as our target variable. Our model will provide the users with the choice of choosing three actors/actresses as the Star Cast for a movie and then predict the likely outcome of the movie being a success (Award=1) or a failure (Award=0).

### E. Classification Methods

*1) Gaussian Naïve Bayes:* Naïve bayes comes under supervised learning method and is the most basic and widely used algorithm for most of the machine learning approaches and comprises of the Bayes Rule (Equation 1). It is generally applied on the set of continuous data and as its name suggests, it follows the gaussian-normal distribution. Its theory lies on the assumption that every feature is independent of each other given the value of class variable.

$$P(\theta|\mathbf{D}) = P(\theta)\frac{P(\mathbf{D}|\theta)}{P(\mathbf{D})} \quad (1)$$

*2) KNN(K Nearest Neighbour):* K nearest neighbour is a basic non-parametric algorithm of machine learning and falls in the category of supervised learning. Most of the uses of K nearest neighbour lie in the pattern recognition and data mining side. It is basically used for the prediction of target variable by creating an imaginary boundary to classify the data. When new data points come in, the algorithm will try to predict that to the nearest of the boundary line.

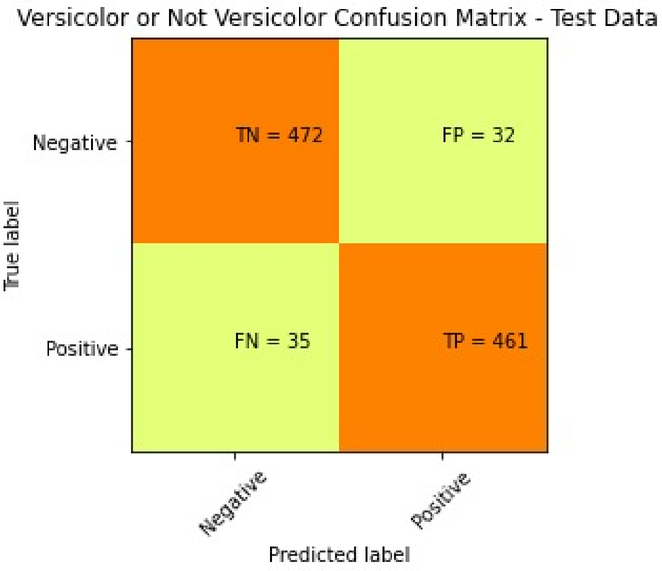

Fig. 1. KNN Confusion Matrix.

*3) Decision Tree:* Decision tree is the machine learning algorithm that uses the tree structure for solving a classification problem. It divides all the data into smaller subsets and a decision tree is then formulated containing the decision as well as the leaf nodes.

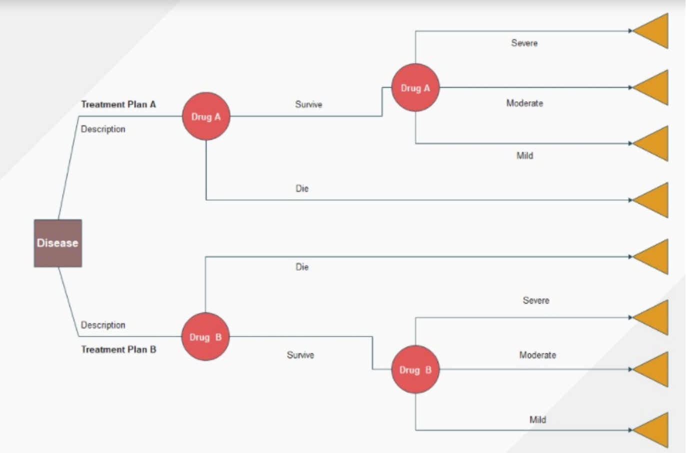

Fig. 2. Decision Tree.

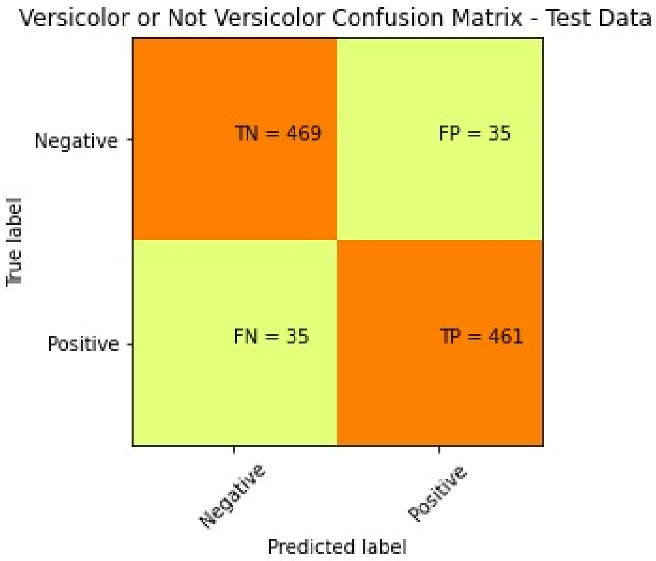

Fig. 3. Decision Tree Confusion Matrix.

*4) Random Forest:* Random forest is an ensemble (The algorithm that takes account of many machine learning algorithm at once) learning method widely used for classification as well as the regression type problems. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It as actually built over the decision tree classifier in the sense that it takes into consideration the over-fitting of training set.

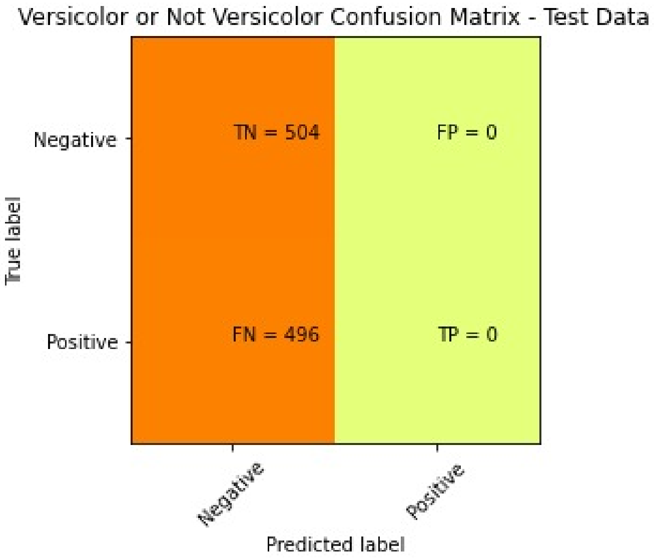

Fig. 4. Random Forest Confusion Matrix.

*5) SVM(Support Vector Machine):* SVM is supervised machine learning model that is castoff for the classification problems. Its principle lies in exploiting the margin amongst separating hyper plane. SVM are very effective high dimensional, memory efficient, and handy machine learning algorithms that work fine with a set of non-linear data.

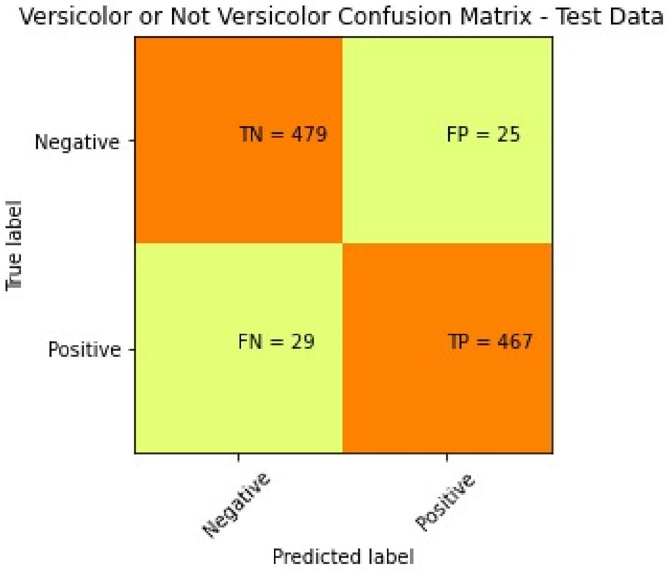

Fig. 5. SVM Confusion Matriz.

*6) Logistics Regression:* Logistics regression is one of the most popular algorithm that can be used to predict a binary outcome and one or multiple continuous or categorical predictor variables. It is easier to train and implement as compared to other methods. Logistic regression works well for cases where the dataset can be separated using a linear function: A linear dataset is the one whose classes can be separated by drawing a straight line between them.

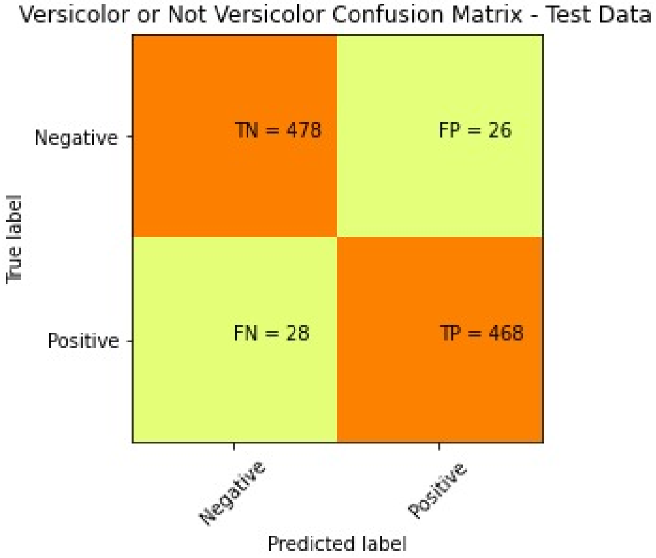

Fig. 6. Logistics Regression Confusion Matrix.

The results of each classifier with its corresponding error estimator is shown in Table 1.

## IV. Results and Discussion

Table 1 is presenting the accuracy comparison of various classifiers experimented with different estimators. The table clearly shows that Multi-Layer Perceptron is the best algorithm giving an accuracy of 95.2%. Furthermore, it can also be inferred that by using K-Fold Cross Validation, the performance of almost all the classifiers increases drastically. Random Forest performed the worst, giving an accuracy of 23.6%. MLP is a deep learning supervised technique useful in machine translation, speech recognition and image recognition. As its name suggests, it contains an input layer, an output layer and a hidden layer. The user gives input features to the input layer and it and then it uses feed-forward procedure to propagate each result to the next layer thereby increasing the accuracy at every instance. It is used as a precursor for complex neural networks

| | NB | LR | RF | SVM | MLP | DT | KNN |
|---|---|---|---|---|---|---|---|
| **Test Train Split** | 70.45% | 72.72% | 72.72% | 40.9% | 36.36% | 29.45% | 31.81% |
| **N_Fold CV_5** | 71.2% | 94.2% | 50.4% | 94.2% | 95.1% | 93.3% | 93.1% |
| **N=10** | 93.7% | 73% | 50.4% | 94.6% | 95.2% | 93% | 93.1% |
| **Hold Out** | 76.36% | 39% | 23.63% | 40% | 38.18% | 37.27% | 35.45% |

Table 1: Evaluation of classifiers

| Error Estimators | Percentage Error |
|---|---|
| **Test Train Split** | 49.37% |
| **N_Fold CV_5** | 15.5% |
| **N=10** | 15.28% |
| **Hold Out** | 58.58% |

Table 2: Errors

## V. Conclusion

Our works and findings pave way for the application of Machine Learning Techniques in successfully predicting the success of a movie before its release. Based off the results we got it was found that neural network algorithms performed much better than Bayesian models in some instances while Bayesian outperformed ensemble algorithms at all instances. Amongst all the other estimators N-Fold Cross Validaton seemed to outperform every other estimator used. Our model used Lux Style Award as the basis of movie being a hit or a flop and also as a target variable while Star Cast was used as input variable. The Star Cast seemed to hold considerable explanatory power in predicting the movie's success and failure. We would like to continue our research further by adding more features like movie genre, budget, release data, movie rating and increasing the dataset to include movies from before 2000. Additionally, we would like to test our model on unsupervised learning techniques like CLARA-Clustering and comparing their performance to our supervised learning ones.

## References

[1] Dr Erum Hafeez Aslam, "Lollywood-Pakistani Cinema Through a Traditional Lense,", June, 2015.

[2] Garima Verma, Hemraj Verma, "Predicting Bollywood Movies Success Using Machine Learning Technique,", Feb.1, 2019.

[3] Rijul Dhir, Rijul Dhir, "Movie Success Prediction using Machine Learning Algorithms and their Comparison,", Nov.7, 2018.

[4] Salman Masih, Imran Ihsan, "Using Academy Awards to Predict Success of Bollywood Movies using Machine Learning,", Feb, 2019.