International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206481

Predicting the IMDB rating by using EDA and machine learning Algorithms

Prayas Dixit*, Shahzeb Hussain, Gaurav Singh
Infosys Limited, Pune, Maharashtra, India

Article Info
Volume 6, Issue 4
Page Number: 441-446
Publication Issue :
July-August-2020

Article History
Accepted : 10 Aug 2020
Published : 16 Aug 2020

ABSTRACT

The film industry is not only the center of Entertainment but also a huge source of employment and business. Well, famous actors and directors can ensure the publicity of a movie but can't promise a good IMDB score. We have collected the data available online about these Hollywood movies and their IMDB ratings to create our dataset. After getting the dataset we have incorporated various exploratory analysis techniques and then applied various machine learning algorithms to predict the IMDB rating. Finally, identified the best-fit algorithm which gives the most accurate prediction.

Keywords : Exploratory Data, Regression, Regression, Multiclass Classification.

I. INTRODUCTION

Movies are the best way to refresh our minds and a really good source to gain practical knowledge sometimes. So, when a viewer decides to see a movie, he or she should want to get as much information about the movie beforehand as possible. So where can they get this information? Well, movie trailers only reach the people who go to the theater or watch television regularly. Some movies start their advertising campaign a year or months in advance to guarantee the success of their movie. But these trailers can't tell the consumer about the quality of the movie. This is where the IMDB movie ratings come in the picture.

There are various platforms such as rotten tomatoes, IMDB, and few others which provide ratings and reviews from the people. Based on that people can identify which movie they should watch. Because of the honest reviews, these platforms are getting famous.

This paper proposes a model that is enough to predict the IMDB rating of a movie based on various factors such as duration, budget, genre, language, and various others, which helps the user to decide whether to see a movie or not. We have used the historical data of movies to predict the success of upcoming movies. The goal is to identify the IMDB ratings with maximum accuracy.

The paper is organized as follows:
This is the first section which is Introduction.
Section II describes the initial stage which is data collecting and preprocessing. Section III includes the algorithm which is applied to the dataset. Section IV portrays the results.

Section V is the final section in which we have concluded our paper.

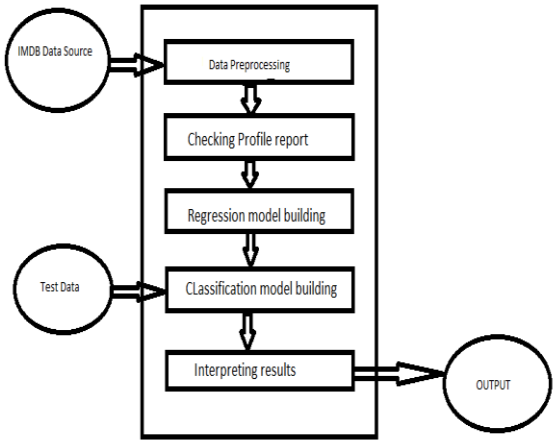

Fig. 1: General Flow

II. DATA PREPROCESSING

There are 28 variables for 5043 movies, spanning across 100 years in 66 countries in the data. Also, there are 2399 unique director names and thousands of actors/actresses. There are 27 variables which are the possible predictors and "imdb_score" is the response variable.

There are various steps followed in this process of data cleaning and are as follows.

A) Missing Value Treatment

Handling the missing value is one of the most important and common tasks which needs to be performed meticulously. There are various reasons due to which data goes missing.
Some of the techniques which are used to handle missing data are mentioned below.

1. **Deletion** - This is the best method in most of the cases unless the nature of missing data is 'Missing completely at random'.

   i. Pairwise Deletion - In this case, the analysis is done on the variables present in the table, and only the missing observations are ignored.

   ii. Listwise Deletion - In this case, rows that contain the missing values are deleted.

   There is always some loss of information in both cases and listwise deletion suffers the maximum information loss when compared.

2. **Imputation** - It also has some methods which are used to perform imputation.

   i. Average Technique - Calculation and then imputing the missing values with mean, median, and mode are the most popular technique depending upon the type of dataset.

   ii. Predictive Techniques - A predictive method can be also used to impute missing values which assumes the nature of missing values. There are various statistical methods like regression techniques and machine learning methods like SVM.

B) Outlier Analysis

Outliers are nothing but the extreme values that differ from the rest of our data i.e it diverges from the overall pattern of the sample. Outliers are of two types, first one is Univariate which can find out when analyzing the distribution of values in single feature space. Another one is Multivariate and can be found in an n-dimensional space and we need to train a model to find this type of outlier because it will be very difficult to find out just by using the human brain. There are some most common causes for outliers, few of them are mentioned below.

i.Data processing errors.
ii.Data entry errors.
iii.Measurement errors.
iv.Sampling errors.

C) Converting Categorical variables to numerical variables

Sometimes the data set will contain categorical variables and stored as text values that represent various traits. In our dataset, there were categorical values too such as color, genres, and few others. It is always challenging how to use this kind of data in the analysis. Various algorithms do not require any conversion and support categorical values but many others require the conversion. Here, we have applied various tools of python such as Pandas and Scikit-learn, and with their help, we converted the required categorical data into suitable numerical data.

III. MODEL BUILDING

Here in this section, we will both type of model building for prediction i.e Regression and classification.

A) REGRESSION MODEL BUILDING

- Here the first step that we have performed is the splitting of the dataset. One part is the training data which is used to train the model and the other one is the testing data which is used to test the performance. We have used the 'train_test_split' to split the data in an 80:20 ratio i.e. 80% of the data will be used for training the model while 20% will be used for testing the model that is built out of it. Below is the snippet of code that we have used for the splitting of the data.

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,rand
om_state=123)
```

- The second step that is also an important step to increase the efficiency and speed of the model building process is Feature elimination. In our approach, we have performed Recursive feature elimination. It is a method that fits a model and removes the weakest features until reaches a count that is specified initially.

This technique builds a model on the entire set of predictors in testing data and computes an importance score for each predictor. After the computation those predictors get removed that have the least score among others, the model is re-built and importance scores are computed again. After these two steps, our process of modeling begins.

i) Simple linear Regression

Linear regression models are used to predict the relationship between two variables. Various variables are used to predict the target variable and are known as independent variables and the target variable is also known as the dependent variable. This model works by fitting a line to observed data. Most models such as logistic and nonlinear regression use a curved line but linear regression uses a straight line. Below is the formula for simple linear regression.

$$Y=\beta o + \beta 1X + \varepsilon$$

- B_0 is the intercept, the predicted value of y when the value of x is 0.
- On any given value of the independent variable (x), y is the predicted value of the dependent variable.
- B1 is the coefficient of regression.
- e is the error of the estimate, or how much variation there is in our estimate of the regression coefficient.
- x is the independent variable.

ii) Support Vector Machines with Linear, Polynomial, RBF Kernels

SVM is a supervised machine learning which can be used for both Classification as well as Regression. It takes a longer time to process than other algorithms. It follows the idea of fitting a hyperplane between the feature and tries to separate them into different domains.

The points nearest to the hyperplane known as the "support vector points" and the distance of these vectors from the hyperplane are called the "margins".

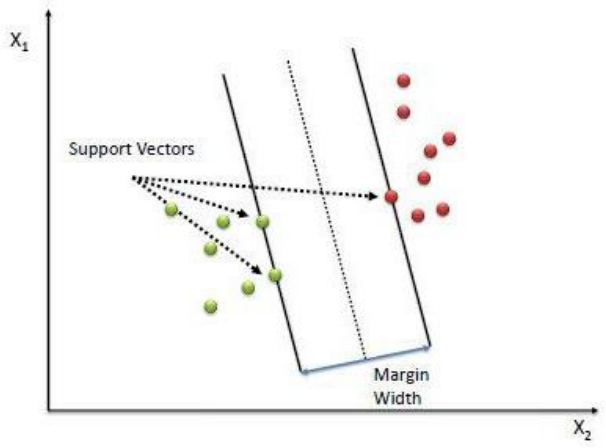

Fig.2: Support Vector machines

iii) Ensemble Models

a) GBM(gradient boosting)

It is a highly used model by data scientists. Apart from fitting the model, and finding the model with the highest predicting power, parameter tuning is also important. This method is one of the most powerful techniques in predicting models.

Gradient boosting involves three methods.

- A loss function which needs to be optimized.
- A weak learner which will be used to make predictions.
- An additive model to append weak learners to reduce the loss function.

b) Random Forest

- It is an ensemble algorithm that amalgamates multiple decision trees and navigates compounded problems to give us the final answer. There are various parameters that we have taken care of in this such as max_depth,min_sample_split,max_leaf_nodes,min_samples_leaf, etc.
- Random forest is a bagging technique. There is parallel processing of trees in the random forest and while building the trees there is no interaction between these trees.
- There is parallel processing of trees in the random forest and while building the trees there is no interaction between these trees.

c) XG boost

It is a really popular tool and widely used among Kaggle competitors. It has already been tested in production for large scale problems. It can work through most f the regression problems due to its versatility and flexibility. It can also be used with different platforms and interfaces due to its portability and compatibility. Large-scale cloud dataflow systems such as Flink and Spark can also be connected with Xgboost. It can be used with various languages such as python, java, c++, R, etc and here in our case, we have used Python as our programming language. It intends to exploit every bit of hardware resources and memory for tree boosting algorithms. It was developed to decrease the computing time by a great extent. Handling missing value is also an important feature of this algorithm.

FIG. 3: XG BOOST

After looking into all the metrics almost we have seen that XGBRegressor has given the best results with a mean squared error of 0.404 and with "{'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 500}". The Feature Importance given by this model is shown in the image. Some of the most important features are -num_voted_users, duration, budget, main_genre. Language.

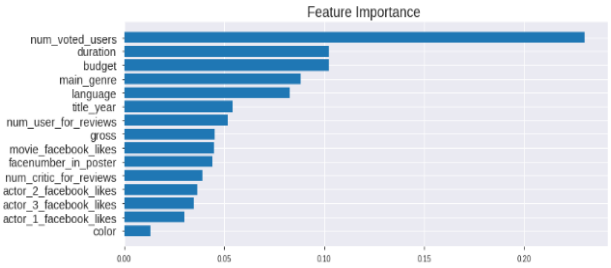

Fig. 4: Selected Features

B) CLASSIFICATION MODEL BUILDING

For building a classification model we have used the preprocessed data that we have used to build the regression model and have to change the target variable. As we have used three techniques to build a classification model and then find out the best among them, here also while building a classification model we have done the same.

The methods that we have used here to build the model are:

i. Logistic Regression
ii. Support Vector Classifier
iii. Ensemble Models
a) Random forest classifier with Hyper Parameter Tuning.
b) Gradient boost classifier with parameter tuning.
c) XG Boost classifier with Hyperparameter tuning.

Logistic Regression:

This method is used when the desired target variable is categorical. It is known as logistic regression because of the function it uses, 'Logistic function'. The logistic function was developed by the statisticians to find out the properties of population growth. It is an S-shaped curve map between value o and 1 and can take any real-valued number but not at the limits.

Like linear regression, logistic regression uses an equation for the representation.

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

Where b1 is the coefficient for the single input value y is the output predicted and b0 is the bias(intercept term).

The testing data is used to train the coefficients (b values) in the logistic regression algorithm and is calculated by using maximum-likelihood estimation.

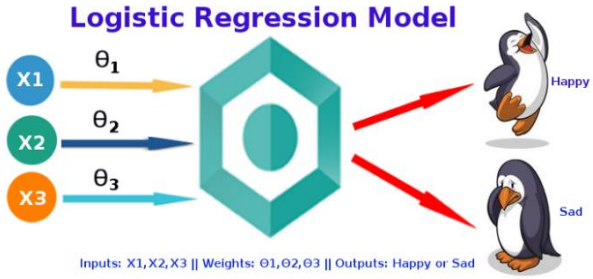

Fig. 5: Logistic Regression Model

In our modeling process, the Gradient Boosting classifier is the best one with the final model with 83 % accuracy.

Now after development and analyzing the results of both the classification and regression models we can see that both of them have given the same amount of importance to the respective features. The results of both models are mentioned in fig 6 and fig 7 respectively.

Regression Model	Mean_squared_error
Simple Linear Regression	0.70
SVRegressor Linear	0.72
SVRegressor Polynomial	0.93
SVRegressor RBF	0.68
Gradient Boost	0.43
Random Forest	0.45
XGBoost	0.40

Fig. 6: Regression Model Result

Classification Model	MisClassifications	Accuracy	Precision	Recall	F1-Score
Logistic Regression	190	0.75	0.47	0.40	0.41
SVC Linear	181	0.76	0.47	0.45	0.46
SVC Polynomial	143	0.81	0.52	0.50	0.51
SVC RBF	146	0.81	0.51	0.50	0.50
Random Forest	130	0.83	0.54	0.50	0.51
Gradient Boosting	127	0.83	0.54	0.51	0.52
XGBoost	139	0.82	0.52	0.51	0.51

Fig. 7: Classification Model Result

As we can see from the results in fig 6 and fig 7 the best model among the regression model is XG Boost with the lowest mean squared error and the best one among the classification models is Gradient Boosting with an accuracy of 83%.

IV. CONCLUSION

In this work, we have made two different types of models and identify the best of both of these categories for this particular task. Our classification model predicts the result with 83% accuracy which is quite good and effective at the same time. The datasets that we have used can also be amalgamate with newly released movies datasets to increase the efficiency of the model further.

V. REFERENCES

[1]. M. H. Latif and H. Afzal, “Prediction of movies popularity using machine learning techniques,” International Journal of Computer Science and Network Security (IJCSNS), vol. 16, no. 8, p. 127, 2016.

[2]. Javaria Ahmad, Prakash Duraisamy, Amr Yousef, Bill Buckles, "movie success prediction using data mining" July 2017

[3]. Ajay Siva, Santosh Reddy, Pratik Kasat, Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining in International Journal of Computer Applications, Foundation of Computer Science, New York, USA, vol.56, no. 1, October 2012.

[4]. Nithin VR, Pranav, Sarath Babu PB, Lijiya A" Predicting Movie Success Based on IMDb Data", October 2017

[5]. Anantharaman V, Ebin G.Job, Neha Sam, Asst. Prof. Sheryl Maria Sebastian, “Movie Success Prediction Using Data Mining” Global Research and Development Journal for Engineering.

[6]. https://towardsdatascience.com/

[7]. https://machinelearningmastery.com/

Cite this article as :

Prayas Dixit, Shahzeb Hussain, Gaurav Singh, "Predicting the IMDB rating by using EDA and machine learning Algorithms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 441-446, July-August 2020. Available at doi : https://doi.org/10.32628/CSEIT206481
Journal URL : http://ijsrcseit.com/CSEIT206481