Movie Success Predictor using Regression Analysis

Manav Agarwal
Department of CSE
PES University
Bangalore, India
manav.ag.3052@gmail.com

Shreya Venugopal
Department of CSE
PES University
Bangalore, India
shrey1010.svg@gmail.com

Rishab Kashyap
Department of CSE
PES University
Bangalore, India
rishabkashyap14@gmail.com

Abstract - The world today runs on various forms of entertainment to keep people lively, and a major contribution to the entertainment industry would be movies which dominate the sector and hence gaining information about the success of a movie is a primary concern for enthusiastic viewers. Hence, prior knowledge about certain movies and the factors that influence them would be a great asset to determine how successful a movie might be and how many viewers would it receive. Through this paper, we aim to predict the success of a given movie based on the few influencing features we know about it to create a success rating as accurately as can be predicted. We shall be referencing the IMDb dataset for this purpose, and use it over regression models and to compare the accuracies. From this we shall conclude the best model that can be used.

Keywords - Data cleaning, Pre processing, Regression, SVM, Budget, Votes

I. INTRODUCTION

In today's world we have plenty of movies, and almost a million times the amount of viewers hungry for more and better content. Viewers usually prefer to watch the movies which hold certain characteristics such as a specific type of genre, the actors who have acted in it, certain producers and directors, and sometimes even the budget and the locations of the movie where it was shot. Majority of the people decide this by looking into the movie ratings such as that of Rotten tomato, or IMDb, as the scores given by them are highly critiqued and recognized by the public, representing the quality of content.

In the following project we will be using the IMDb extensive dataset from Kaggle[1] to create the prediction model, which consists of a variety of movies from various countries, along with the names of the actors, producers and directors as well as the number of votes per movie and the variety of age groups that have voted for them, and finally the budget of each movie. From the data collected, we analyze the popularity of earlier movies based on the votes to gain an understanding of the various factors that contribute to its success. On analysis we can find that factors such as the Budget spent on the movie, the Actors who have acted in it, the directors and producers responsible for the creation of the story, and even the genre of the movie are major contributing factors to the success of a movie.

We thus perform exploratory analysis to uncover the strength as well as the contribution of each of these factors to determine the best possible course of action to use on our model so as to get the most accurate prediction of the success rate.

II. LITERATURE SURVEY

Various references show us the use of regression models and SVMs to show the accuracy of the model and how it performs in each case. Some notable works include the gross box office as their attributes([2], [3]), and fewer works include other attributes ([4]) in their models to show the varying results in accuracy of prediction.

A. PAPER - I

We take the example of the paper [2] where they predict five different measures of success, based solely on the attributes known about a movie before its debut, such as the movie title, plot, budget, characters, and the gross rates. Even extra features such as the time period during which a movie came out (Such as on festivals or school summer holidays) are a beneficial factor to determine the success rate. They then cleaned the data and applied the respective models to them. The most prominent model used was the SVM in this case, where the dataset was divided into a total of 176 vectors with the help of the forward search paradigm, which is a process that selects a subset of features which allow us to make the best predictions. The SVM Kernel they used runs on L1 Regularization, which allows them to get a classification rate well above 0.5 for each output % for each of the features. This hence results in an accuracy of 70% success rate for each of the rating outputs and 88% success rate for each of the monetary gross outputs.

B. PAPER - II

The next reference [3] shows us a different attribute being used in the prediction process. In this case they have considered the ROI (Return of Investment) as the major contributing factor in the model, along with various related features to it which included Opening Date, Movie Name ,Budget, Domestic Gross, International Gross, Total Gross, Trailer Views, Studio , Cast and Crew, Genre, Medium (Live action or Film), Trailer Views, Wikipedia Views and Rotten Tomatoes Score. Data pruning is then performed to remove all the unnecessary values from the dataset.

The first method that is used is the technique of linear regression, first SLR for the individual comparison of a few attributes, but the main technique used is the concept of MLR (Multiple linear regression). They applied multiple regression to predict the success of the movie, with the independent variables being Trailer Views, Budget, Critic Ratings and Wikipedia Page Views. From finding the t-statistic and the p-value, certain conclusions are drawn such as: We see that budget has the highest t-Stat value of 8.26, and RT scores with lowest t-stat value (3.33) and highest p-value (0.0011) indicating that it has the least impact on accurate prediction, which is consistent with previous findings. SVM is the next technique used for regression analysis and classification of the attributes. In SVM, the models consist of points in space, where each point represents an example. Mapping is done in a way that tries to maximize the difference between examples from different classes, so that the examples of a particular class are very different from examples from another class. New examples are then as-signed to a class based on whichever side they are closer to. Making the class or category have high intra-class similarity but low inter-class similarity is a goal of SVM. Here, they have used 10-fold cross-validation on their data. This resulted in a total accuracy of 56.25%, higher than any of the previous attempts of SVM over a single variable.

C. PAPER - III

The third paper we look into [4] takes a slightly different approach into explaining how the querying via Google Trends about the movie titles, actors and various other attributes allows in the prediction of a successful movie. Their basic idea is that Query data from search engines can provide many insights about human behavior. Therefore, massive data resulting from human interactions may offer a new perspective on the behavior of the market. In this paper, they have made use of Linear and ridge regression models to predict movie incomes. Then they make use of the SVM/SVR technique to do the classification and summary of regression models. Multivariate regression model is used to predict multiple dependent variables from the independent variables with the help of the sklearn library in python. The next thing is Ridge regression , which is made use of to improve the performance of the already existing model by adding in the correction factor (This is also known as regularized regression, or RLR). Unfortunately, this resulted in a worse model compared to linear regression with a higher MSE and presence of more outliers than in the previous case. SVM in this case is the most popular kernel regression method and is once again made use of in this case for classification purposes. It provides the best accuracy compared to the previous regression methods.

III. MOTIVATION

As data scientists, we wish to exploit the various techniques and tools used to uncover useful information from a variety of data to infer various useful traits in it. Our aim is to try and ease the workload taken by people by even a small yet significant amount by using the resources at our disposal to our fullest to produce a well-functioning model that serves its purpose.

IV. DATA DESCRIPTION

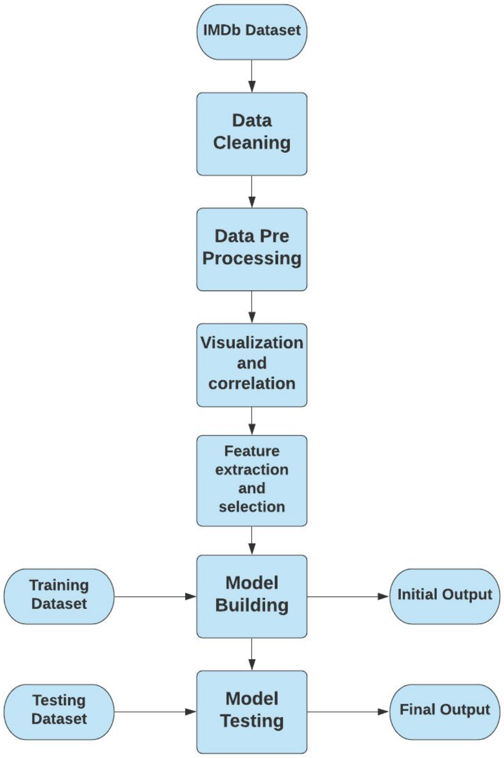
The dataset that we shall be using as mentioned earlier will be the IMDb dataset from Kaggle[1], which consists of 4 separate tables representing the movie names and details, names of all the people contributing to the movie, the number of votes each movie got over a variety of population and finally the titles with respect to teach movie. The dataset as a whole consists of 97 columns, and each dataset has a varying number of rows. The first dataset lists out the movies we are using, which consists of a list of more than 81,000 movies, each with their titles, movie length and various other attributes. The next dataset consists of all the names of the contributors to the movie and their roles such as actors, directors and so on. The third dataset gives a list of all the votes given to each movie, differentiated based on the vote given on 10, the age group, the gender as well as the region. The final dataset gives an overview on the first two datasets by combining the movies and the actors' roles in the movies they played. Using the information given to us from the above dataset, we shall build a predictive model by keeping the target variable as the metascore, and using multiple variables to compare as the independent variables.

V. PROPOSED MODEL

The first step would be to clean the data in order to obtain the right values to work with. Analysis on the data is a very crucial step as the wrong data might lead to various issues and inaccuracies in our model. The next step on extraction of the data would be to pre process them into a usable format. Then we move on to visualizing the features and performing a correlation analysis on the data for feature selection. On selecting the features we need we finally move on to building our model.

The model building is the most important part in the entire process, as the right model gives us the most accurate results, and the wrong model might not be as accurate, and furthermore might result in wrong information being concluded from it. Through this paper, we propose a regression model or a classification model to predict our data, and to further perform a comparative study on many possible models hence choosing the more accurate one. Within each model, a comparison is made based on the accuracy with respect to the number of classes it categorizes a movie's success into. The accuracy is based on a confusion matrix obtained from the predictions. The accuracy with approximately the best number of bins or categories is taken as the accuracy of the specific model.

The flow of the various steps to be followed is displayed as shown in the flowchart. The same procedure is followed to create our final model.



A. CORRELATION ANALYSIS

The first step into analyzing the data is to perform a correlation analysis on the data. The correlation shows us the relationship between the various attributes present in the data, and returns the values of the Pearson correlation coefficients for comparison.

On retrieving the values, the higher the correlation values, the more reliable the attribute is to be taken for the model building process. This is because when we have strongly related attributes, they tend to complement each other in the process of analyzing values better than most other features can, and hence will guarantee a certain degree of accuracy to the model we propose.

The drawback of correlation would be the fact that it can work only with quantitative data types and not qualitative, as qualitative data might result in spurious correlations leading to various inaccuracies in our model output. The correlation matrix can be used to find the relationship of all the attributes present in our dataset with each other in the form of a heatmap.

VI. MODEL TYPES

A. SIMPLE LINEAR REGRESSION

The first type of model we look into would be the Linear Regression model using a single variable. This is the simplest form of regression which shows us the plot between one independent variable and the dependent variable we have chosen as our target variable. Higher the correlation, better the regression model, and higher the accuracy. We can calculate the residual plots as well.

B. MULTIPLE LINEAR REGRESSION

This is a regression model which uses multiple independent variables to predict a single target or dependent variable. Each one has a certain degree of correlation with the target, higher the better, and all of them contribute to improving the accuracy. MLR shows us how well the variables relate to each other. A major drawback would be the effect of multicollinearity between the independent variables which might result in inaccurate values.

C. SUPPORT VECTOR MACHINE

SVM is a popular supervised learning model where labeled data is given to it, each data point that we feed to it is known as a “Support Vector”. These points help to determine the plane of division for the data points thereby classifying them into two or more distinct groups based on the dimension=ions of the data fed to the model.

D. K-MEANS CLASSIFIER

KMeans is a model used for classification purposes. It is a partitioning clustering algorithm, this method is to classify the given date objects into k different clusters through the iterative process, converging to a local minimum. So the results of generated clusters are compact and independent [5].

It has a time complexity of $O(n^{dk+1})$, where n is the number of clusters, k is the number of clusters, and d is the number of dimensions.

E. LOGISTIC REGRESSION

The logistic model (or logit model) is used to model the probability of a certain class or event existing. It was developed in the 1970s as an alternative to the OLS methods to provide a solution to dichotomous outcomes. It is well suited for studying the relations between a categorical or qualitative outcome variable and one or more predictor variables [6].

F. RIDGE REGRESSION

Ridge Regression is a regression model that is implemented in cases where the values of the features considered from a dataset suffers from the condition of multicollinearity, which is a condition where there exists a near-linear relationship among independent variables we consider. Ridge regression standardizes the variables first and then adds a bias on the OLS method used to predict improved output values.

G. LASSO REGRESSION

LASSO (Least Absolute Shrinkage Selector Operator) is a regression model similar to Ridge regression, and uses the concept of “shrinkage”, that is we shrink the data to a central tendency such as mean and then operate on the variables. It performs L1 Regularization, which adds a penalty to the absolute value to the magnitude of the coefficients and outputs the predicted values.

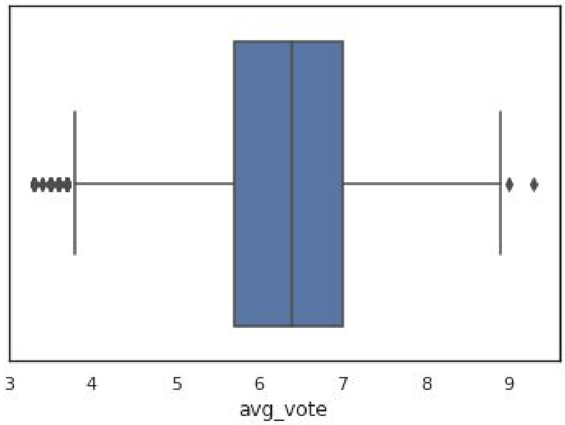
H. TIME SERIES ANALYSIS

Time Series analysis[6] is a statistical methodology performed on a response variable to obtain it’s value over different points of time. It is commonly used for Forecasting our data, which is one of the most frequently used forms of predictive analytics. A time series analysis can help us to understand the underlying naturalistic process, the pattern of change over time, or evaluate the effects of either a planned or unplanned intervention. In our model, we shall be using a Seasonal Auto Regressive Integrated Moving Average Exogenous (SARIMAX) model, which allows us to deploy a Seasonal model to our dataset for forecasting analysis.

VII. EXPERIMENTAL ANALYSIS

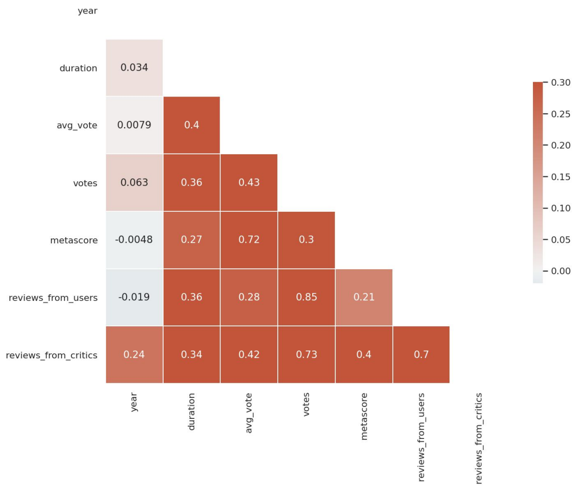
A. DATA CLEANING AND FEATURE SELECTION

Our dataset as explained before consists of various movie names, ratings and the various attributes that influence them. We carefully distinguish and select the features that we shall be using in our models after cleaning and filling in missing values. We check for the possibilities of any outliers in the dataset and clean them. We then create our training(movies from 2005 to 2015) and testing(movies from 2015 to 2019) dataset for each individual model. For classification purposes we split the result between 1-100% based on the success rate of the movie in both the testing and training datasets (named as 'hitflop' in both cases). On successful creation of the same, we move on to the correlation analysis.



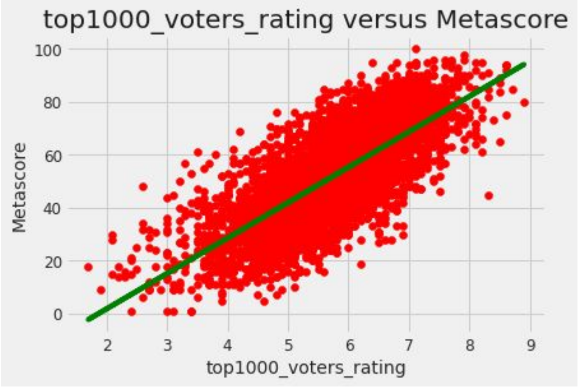
B. CORRELATION ANALYSIS

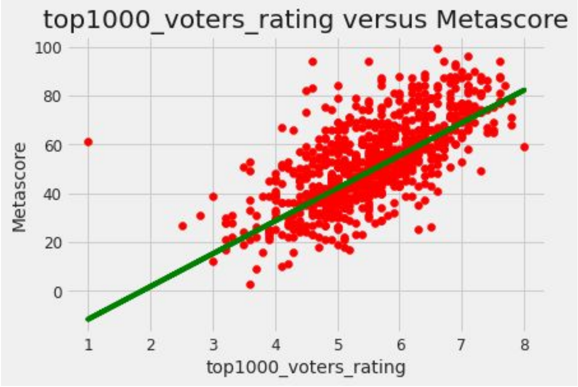
In the diagram, we have plotted the correlation of the selected features which we thought would bring about a certain degree of change in the model. We select our target variable to be the metascore, the reason being that the metascore is the value which tells us whether a movie is going to be a success or not. In this case, we plot the correlation heatmap using seaborne library in python and we conclude that the variable that is highly correlated with the metascore in this case would be the average votes. We use this to plot our linear regression model.



C. SIMPLE LINEAR REGRESSION

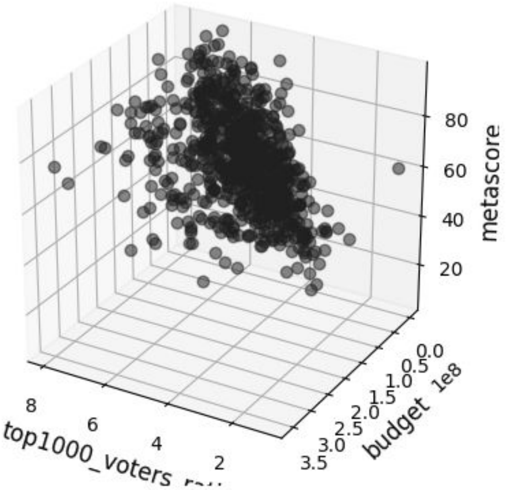
Linear regression is performed here on the target variable being "metascore" and the independent variable chosen with respect to the correlation analysis, created as shown in the diagrams. This model returns a residue of 0.55 with an accuracy of around 55%. The graphs obtained from the training and testing datasets respectively are as shown.

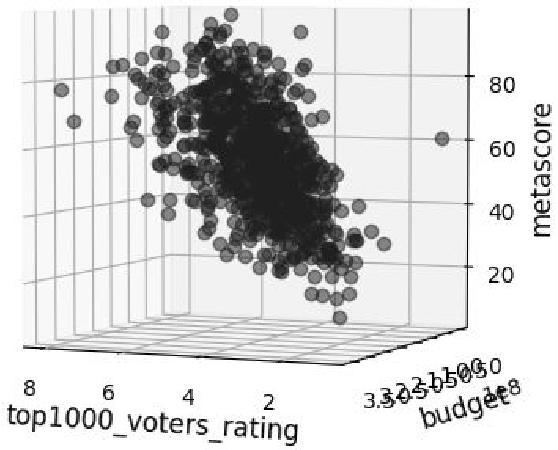


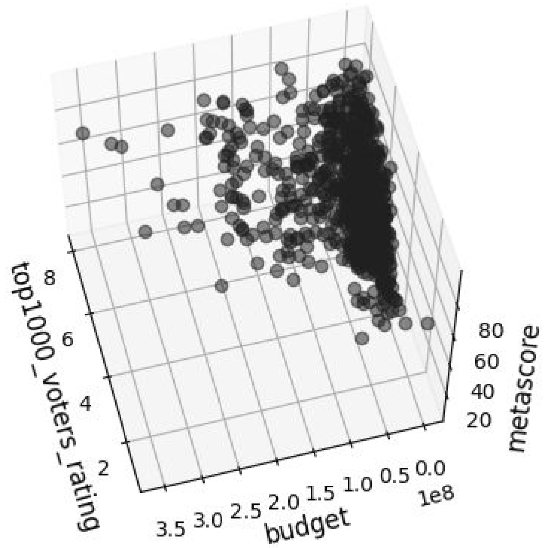


D. MULTIPLE LINEAR REGRESSION

This model makes use of all the variables that are present in the movies dataset and plots the model wrt all the attributes as shown in the correlation graph. Each one holds its own weight in the model, and we end up with a residual of 0.6203, and an improved accuracy of 70.77% on classifying our movies based on the 4 domains created. The models for the training and testing datasets are as shown below.

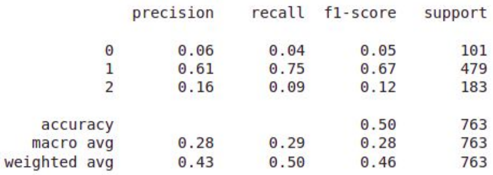




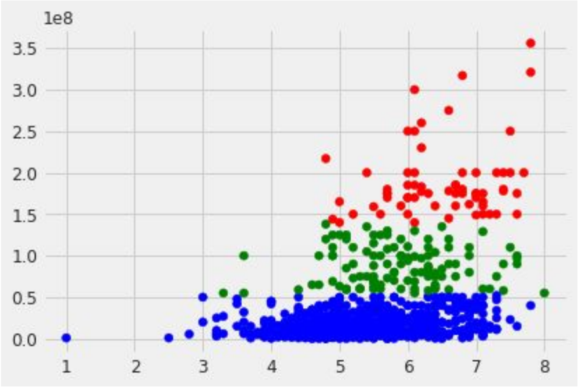


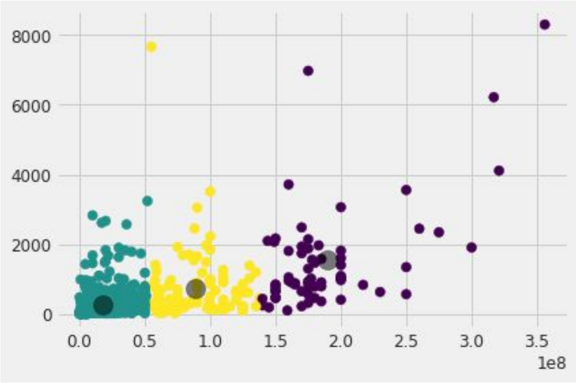
E. K-MEANS CLASSIFIER

The KMeans classifier classifies our datasets fed to it by dividing it into a number of clusters, in this case being three clusters. These clusters divide our result into 3 clusters, one cluster below 0.33 rating, one between 0.33 and 0.66 and the final one having anything above 0.66. These values are bounded between 0 and 1 after being scaled. The proposed model here thus results in an accuracy of 0.5 with the statistics as shown below.



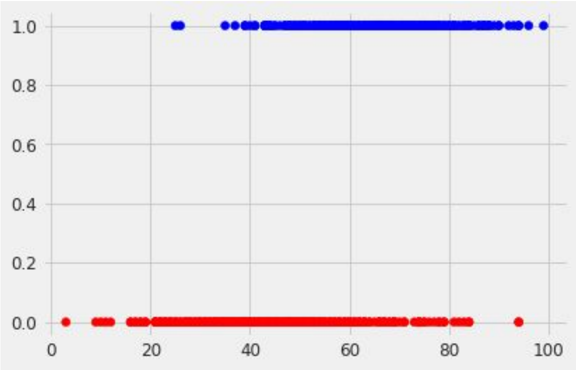
```
              precision    recall  f1-score   support

           0       0.06      0.04      0.05       101
           1       0.61      0.75      0.67       479
           2       0.16      0.09      0.12       183

    accuracy                           0.50       763
   macro avg       0.28      0.29      0.28       763
weighted avg       0.43      0.50      0.46       763
```





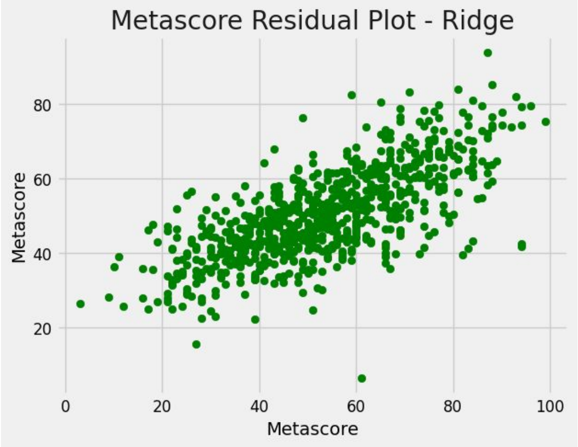
F. LOGISTIC REGRESSION

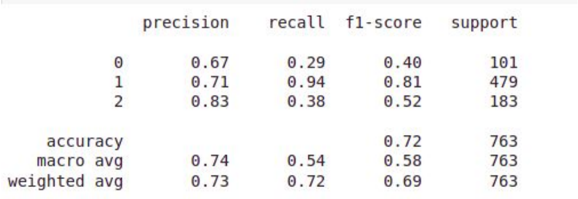
Here we have made use of the sklearn library to implement the LogisticRegression model. We train the model against our training datasets and test it with the test split to conclude a final accuracy of 75.23%.

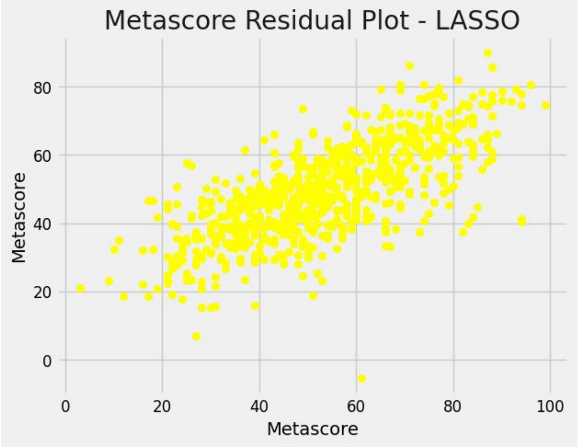


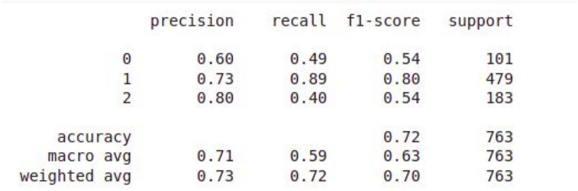
G. RIDGE AND LASSO REGRESSION

As defined before, these regression models are used to add a certain correction to the OLS regression methods (Simple Linear and Multiple Linear regression models). The bias is varied in the former while the data is varied with respect to a central tendency in the latter. For the Ridge regression model, we get a residual score of 0.48, and a residual of 0.46 with LASSO. Both models give us a confusion matrix accuracy of 71.

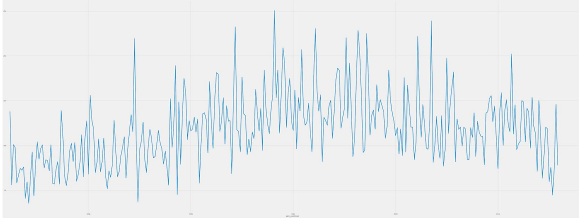

Metascore Residual Plot - Ridge


```
              precision    recall  f1-score   support

           0       0.67      0.29      0.40       101
           1       0.71      0.94      0.81       479
           2       0.83      0.38      0.52       183

    accuracy                           0.72       763
   macro avg       0.74      0.54      0.58       763
weighted avg       0.73      0.72      0.69       763
```


Metascore Residual Plot - LASSO

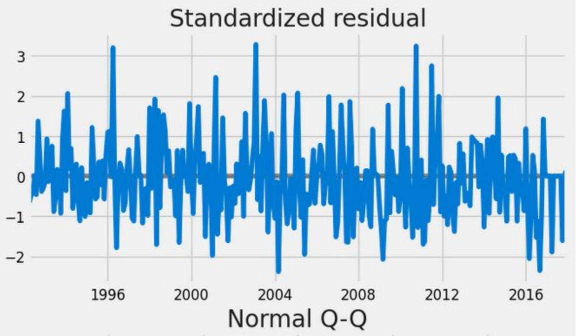

```
              precision    recall  f1-score   support

           0       0.60      0.49      0.54       101
           1       0.73      0.89      0.80       479
           2       0.80      0.40      0.54       183

    accuracy                           0.72       763
   macro avg       0.71      0.59      0.63       763
weighted avg       0.73      0.72      0.70       763
```

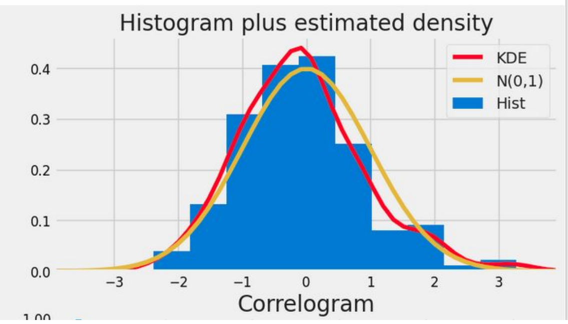
H. TIME SERIES ANALYSIS

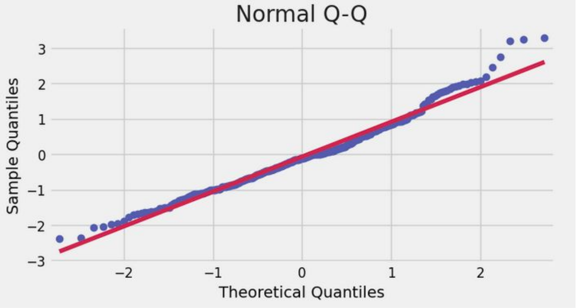


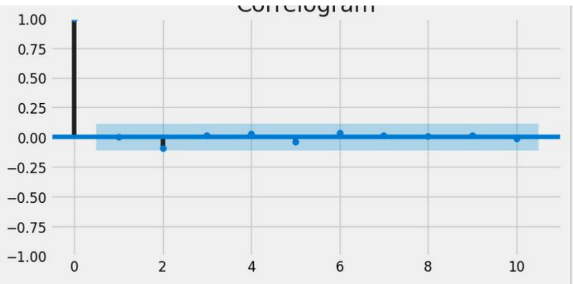
From the above we can see that the data shows minimal seasonality, but show peaks during certain seasons for each year. Also, we can see that most good movies were released between 1997 - 2011 as they have the highest scores

Now we make use of the SARIMAX model to try and forecast our predictions.

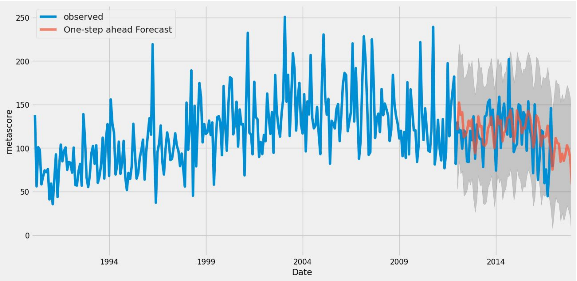


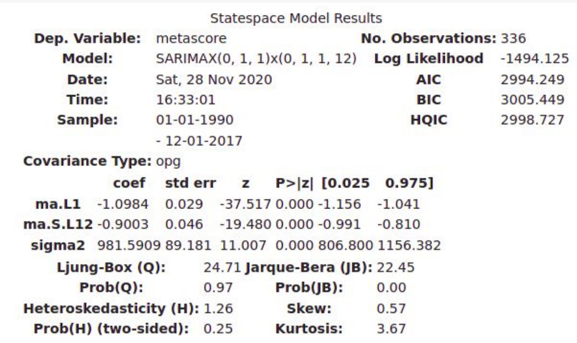




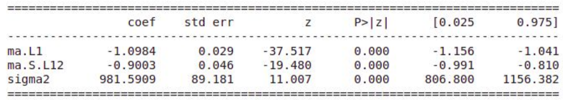


Using the above values we can forecast our predictions for the next few years. The plot is as follows:



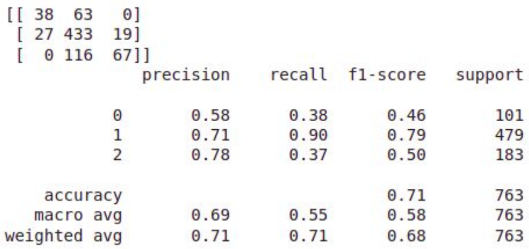


```
                        Statespace Model Results
 Dep. Variable:  metascore                           No. Observations: 336
         Model:  SARIMAX(0, 1, 1)x(0, 1, 1, 12)  Log Likelihood   -1494.125
          Date:  Sat, 28 Nov 2020                           AIC   2994.249
          Time:  16:33:01                                   BIC   3005.449
        Sample:  01-01-1990                                HQIC   2998.727
                 - 12-01-2017
Covariance Type: opg
               coef   std err     z     P>|z|  [0.025   0.975]
   ma.L1    -1.0984   0.029   -37.517  0.000  -1.156   -1.041
 ma.S.L12   -0.9003   0.046   -19.480  0.000  -0.991   -0.810
  sigma2   981.5909  89.181    11.007  0.000  806.800  1156.382
       Ljung-Box (Q):          24.71 Jarque-Bera (JB): 22.45
           Prob(Q):            0.97      Prob(JB):      0.00
 Heteroskedasticity (H): 1.26          Skew:          0.57
   Prob(H) (two-sided):  0.25        Kurtosis:        3.67
```



```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -1.0984      0.029    -37.517      0.000      -1.156      -1.041
ma.S.L12      -0.9003      0.046    -19.480      0.000      -0.991      -0.810
sigma2       981.5909     89.181     11.007      0.000     806.800    1156.382
==============================================================================
```

I. SVM

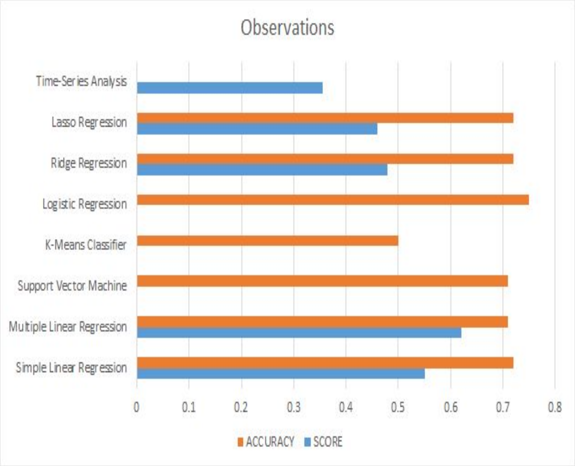


```
[[ 38  63   0]
 [ 27 433  19]
 [  0 116  67]]
              precision    recall  f1-score   support

           0       0.58      0.38      0.46       101
           1       0.71      0.90      0.79       479
           2       0.78      0.37      0.50       183

    accuracy                           0.71       763
   macro avg       0.69      0.55      0.58       763
weighted avg       0.71      0.71      0.68       763
```

SVM returns the below attributes in comparison with the various attributes used. The result is not as strong as that of the regression models seen before.

RESULT

MODEL	SCORE	ACCURACY
Simple Linear Regression	0.55	0.72
Multiple Linear Regression	0.62	0.71
Support Vector Machine	-	0.71
K-Means Classifier	-	0.5
Logistic Regression	-	0.75
Ridge Regression	0.48	0.72
Lasso Regression	0.46	0.72
Time-Series Analysis	0.3552	-



From the observations obtained above, it is seen that most of the models(Simple Linear Regression, Multiple Linear Regression, Logistic Regression, Ridge and Lasso Regression) seem to give almost the same accuracies. Hence other factors must be considered to determine the best model like: Number of attributes required for testing and training, Availability of attributes, and Complexity of the model. Taking these into consideration the best model is- Logistic Regression Model taking the attributes:

1.'top1000_voters_rating' 2.'Action'
3. 'Adventure' 4.'Animation'
5.'Biography' 6.'Comedy'
7.'Crime' 8. 'Drama'
9. 'Family' 10.'Fantasy'
11.'History' 12. 'Horror'
13.'Music' 14.'Musical'
15.'Mystery' 16.'Romance'
17.'Sci-Fi' 18.'Sport'
19. 'Thriller' 20. 'War'
21. 'Western'

CONCLUSION

From the analysis we have done so far we can conclude that Logistic Regression gives the most accurate model as seen from the results above. The main application of the success predictor would be to not just help the audience, but to rather allow theatre owners to decide, based on the values of our outputs above, the number of screens they can keep for a movie in multiplexes or the amount that is to be paid to the distributors in return. Conversely, it helps the distributors too. This would also allow them to thus maximize their profits based on this.

REFERENCES

[1]IMDb_Dataset

[2] Predictor for Movie Success

[3Predicting movie box office success

[4]PREDICTING MOVIE SUCCESS FROM SEARCH QUERY

[5] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74. (citation)

[6] Velicer, Wayne & Fava, Joseph. (2003). Time Series Analysis. 10.1002/0471264385.wei0223. (citation)