Sports Analysis of FIFA Football World Cup Tournament using Logistic Regression

P. Sudhandradevi
Department of Computer Applications
Bharathiar University
Coimbatore, India
psudhandradevi@gmail.com

V. Bhuvaneswari
Department of Computer Applications
Bharathiar University
Coimbatore, India
bhuvanes_v@yahoo.com

***Abstract-* Sports became a prominent part of the human life. Sports analysis provides the expertise on sports-related events. Sports participants have the higher levels of physical activity, psychological health and social welfare. In current strategy sports analytics became a buzz word. The sentiment by sports journalist Grantland Rice said that "not that you won or lost but how you played the game". Sports science is a widespread academic discipline, applied to areas including athletes performance and Olympic game. The sports data is fine tuned from the fine-tune technique or wearable technology. The objective of the paper is sto find the team who gives their contribution in FIFA Football tournament from the year 1872-2018. The Logistic regression technique is used find the probability of win and loses. This technique has been implemented in R tool based on logistic regression model. The outcome of the prediction gives 76% of accuracy of the model design and also it contributes continent wise football interest among the globe.**

Keywords- AIC, EDA, FIFA Tournament, Football, Logistic Regression, ROC-Curve.

1. INTRODUCTION

Sports analytics has been gaining pervasive trend in the current digital era. It's a growing area of interest from both computer system to manage the technical challenges from the sports performance view to aid the developments of sports and athletes. The sports logs generate the more of the live data using Zebra Technologies. The generated data is about equipment's, balls, player's details, track moment, distance, speed and strike zones. These data slice and dices of specific games play to predict the insights of fan's preference. Rapid development of digital world, sports tags blink 25 times / per second and deliver the data 120 per milliseconds. (www.greatmomentsof sportsmanship.com)

In current scenario football is a family team sports which is most popular in the regional context. The collection of data is relevant to men's International football tournament from the year 1872 to 2017 (http://r-statistics.co / Logistic-Regression-With-R.html). This dataset contains 39662 instances and 9 attributes. The data includes tournaments like FIFA world cup, FIFA wild cup and Regular friendly matches in and around of their home town. The objective of this paper helps for sports analytics to predict best team in the tournament based on year wise and dominating team in the tournament. It also predicts the city wise dominating team which qualifies for FIFA world cup as well as FIFA wide cup. In Machine Learning, more specifically the field of predictive modeling is preliminary concerned with minimizing the error of a model or making most accurate prediction for sports with the ability to learn without being explicitly programmed. Based on recommender tracking (fans or technologies), it predicts the likelihood reviews of the tournaments wins/loss in city or country wise outcomes.

Logistic Regression is performed in R tool. Regression Analysis is a statistical model used for estimating the relationship between the sports attribute like (tournament, city, country, away score, and home score). In this paper we implement a logistics regression model for pattern analysis and also find the probability of event=success and even= failure.

2. METHODOLOGY

A framework is designed with four phases to determine the event success in Sports of FIFA world cup tournament.

- Data Acquisition and Data Pre-Processing
- Exploratory Data Analytics (EDA)
- Logistic Regression
- Validation

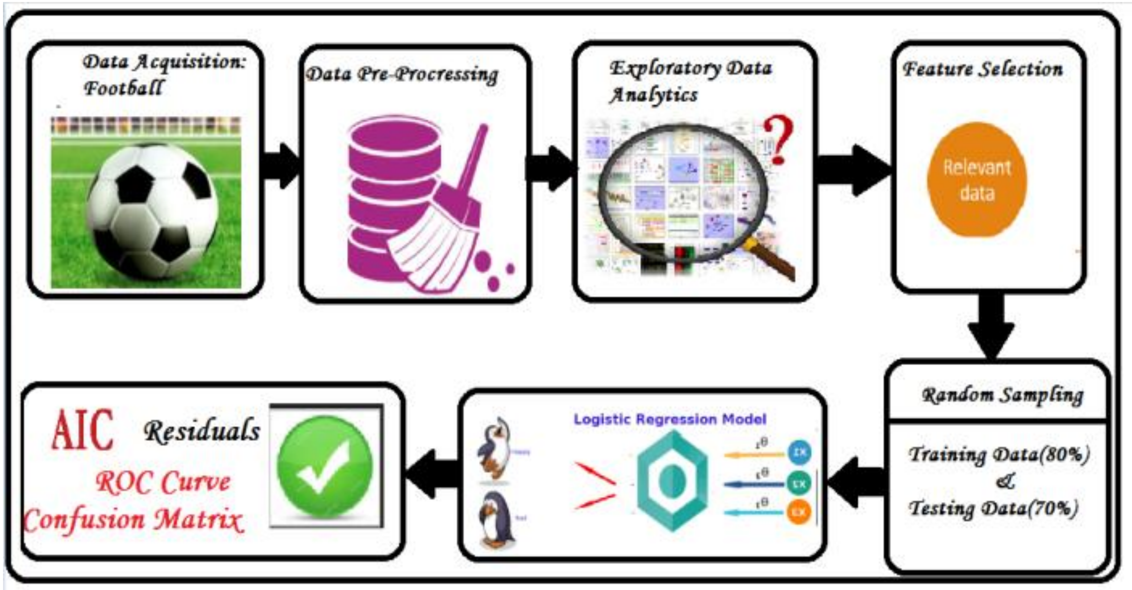

Figure 1. Sports Analysis Success/Lose Prediction on Logistic Regression Model

2.1 Data Acquisition

The international men's football tournament data is collected from Kaggle dataset. The dataset consist of 39,669 instances and 9 attributes. The data is collected and saved as CSV file with the following attributes. (www.kaggle.com).

Table 1: Dataset Description

Attributes	Description
Date	Date of the match
Home_team	Name of the home town
Away_team	Name of the away town
Home_score	Full time home team score including extra time with no penalty shootouts.
Away_score	Full time away team score including extra time with no penalty shootouts.
Tournament	Tournaments like Regular friendly matches, FIFA world cup, and FIFA wide cup, UEFA Euro qualification.
City	Name of the city/town/admin where the matches played.
Country	Name of the country where the matches played.
Neutral	Indicate whether the match played at a neutral venue.

2.2 Data Pre-Processing

This football data is in raw format which is all the given attributes are in categorical format. To perform the further process the noisy data and null values are removed and the data is converted into numeric logistic values to design the model. In football dataset the attribute neutral is created by dummy values. Two categorical values "True and False" will has the dummy values 0 and 1 respectively.

Before Data Pre-Processing

date	home_tea	away_tea	home_scc	away_sco	tourname	city	country	neutral
1872-11-3(	Scotland	England	0	0	Friendly	Glasgow	Scotland	FALSE
1873-03-0	England	Scotland	4	2	Friendly	London	England	FALSE

After Data Pre-Processing

year	home_tea	away_tea	home_scc	away_sco	tourname	city	country	neutral
1872	223	76	0	0	52	625	201	1
1873	77	218	4	2	52	964	70	1

Figure 2. Data Pre-Processing

2.3. Exploratory Data Analytics

Exploratory Data Analytics is an approach to gain deep insight on structures, variables and outliers. Here EDA focus on viewing and interpreting sports data in different dimensional for further analysis. EDA is performed on dataset in a step by step approach by analysing Univariate, Bivariate and Multivariate variables. The summary of dataset is used to drive with the data insights on attributes, types of the variable, levels of the data to formulate insights for analysis. (www.towardsdatascience .com)

2.3.1. Descriptive: Univariate Analysis

Univariate analysis is country wise, city wise and tournament wise analysis is to understand the score of play. (www.jessesadler. com)

2.3.2. Diagnostics: Multivariate Analysis

Multivariate analysis determines empirical relationship between the more numbers of variables. In this dataset the relationship between the various attribute and the correlation between the attribute while determines what is the highly score of the team played in the tournament.

2.4 Logistic Regression Model

In Logistic regression model find the probability of the event=success or event=failure. The formula for using generalized linear model as given in equation (1).

$$g(E(y)) = \alpha + \beta x_1 + \gamma x_2 \qquad (1)$$

Here, g () is link function, E(y) is target variable, $\alpha + \beta x1 + \gamma x2$ is linear predictor, α, β, γ is known as predictor link() function "link" the expectation of the y to the linear predictor.

Logistic regression handles relationship between the independent variables the following models are created to predict the score and country for the players represents in model1 and model2 based on the single variable and multi-variable insights. Then the variables are fit into the model to check whether they are significant variable or insignificant variable. Insignificant values are removed based on the ANOVA and p-value depends on residual error. To remove the variable we use the backward elimination approach, it starts with all the predictors in model and removes the predictors if it is insignificant. The model can add or delete the predictors based on the significant value for better accuracy. Then the model accuracy value and predicted value are compared to validate the model, whether it is predicted I right way or not. The model is validated using ROC Curve. They are mentioned below. (www.hackerearth.com), (www.analyticsvidhya .com)

2.5 Validation Metrics

This validation metrics helps us to find whether the model is fit or not. (www.geeksforgeeks.or gcross-validation-machine-learning)

2.5.1. AIC (Akaike Information Criteria) Value

The analogous metric of adjusted R^2 in logistic regression is known as AIC. AIC measures the fit which penalize the model for the number of model coefficients. The model always prefers the minimum AIC value.

2.5.2. Confusion Matrix

The table represents of Actual vs. Predicted values. It finds the correctness of the model and avoid overfitting.

		Predicted	
		Good	Bad
Actual	Good	True Positive (d)	False Negative (c)
	Bad	False Positive (b)	True Negative (a)

Figure 3. Confusion Matrix (Source: plug-n-score)

The accuracy of the model is evaluated using the equation 2,

$$\text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Here, TP stands for True Positive, TN stands for True Negative, FP stands for False Positive, FN stands for False Negative.

2.5.3. ROC (Receiver Operating Characteristic) Curve

Estimate the deals between the true positive rate (Sensitivity) and false positive rate (1-Specifity). If $p > 0.5$ we concerned about success rate. The area under the curve (AUC), referred as index of accuracy is a perfect performance metric for ROC curve. (www.theanalysisfactor.com)

3. RESULTS AND DISCUSSION

3.1 Data Acquisition

The source of the data has been collected from "Kaggle Dataset". The data consist of world football governing body FIFA, it has been ranking international teams since 1992. The data contains all available FIFA men's international score rankings from the year 1872 to 2017 as in the figure 4. The ranking and score points has been scraped from the official FIFA website. The data doesn't include Olympic Games or nation's B-team, U-23 or league select team only it concentrate on men's full internationals.

date	home_team	away_team	home_scor	away_score	tournament	city	country	neutral
1872-11-3	Scotland	England	0	0	Friendly	Glasgow	Scotland	FALSE
1873-03-0	England	Scotland	4	2	Friendly	London	England	FALSE
1874-03-0	Scotland	England	2	1	Friendly	Glasgow	Scotland	FALSE
1875-03-0	England	Scotland	2	2	Friendly	London	England	FALSE
1876-03-0	Scotland	England	3	0	Friendly	Glasgow	Scotland	FALSE
1876-03-2	Scotland	Wales	4	0	Friendly	Glasgow	Scotland	FALSE
1877-03-0	England	Scotland	1	3	Friendly	London	England	FALSE
1877-03-0	Wales	Scotland	0	2	Friendly	Wrexham	Wales	FALSE

Figure 4. FIFA World Cup: Men's International

3.2 Data Pre-Processing

The data consist of null values and noisy data. For better accuracy we avoid the unnecessary or not applicable values. In Fig 3.2 the missing values for the football dataset is given below. Here 100% of value is observed and there is no missing value (0%), now the data is ready to perform. After removing the null values the data is converts in numeric value because it fails when the response variable is categorical. So the response variable is converted into numeric values by giving dummy values. Here the response variable is “neutral” and dummy values are False=1, True=0 is shown in Figure 5 and 6.

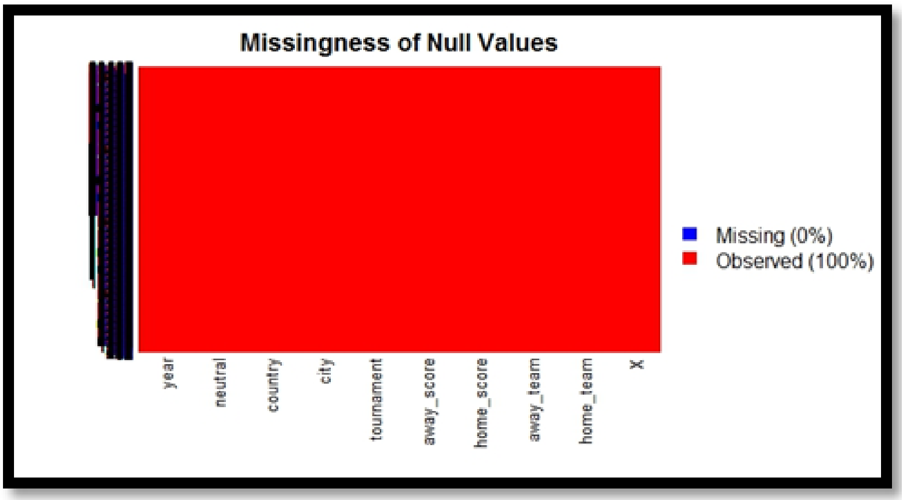

Figure 5. Missingness of NULL Values

home_team	away_team	home_score	away_score	tournament	city	country	neutral	year
223	76	0	0	52	625	201	1	1872
77	218	4	2	52	964	70	1	1873
223	76	2	1	52	625	201	1	1874
77	218	2	2	52	964	70	1	1875
223	76	3	0	52	625	201	1	1876
223	277	4	0	52	625	201	1	1876

Figure 6. One hot encoding: Binarization

3.3 Exploratory Data Analysis: EDA

EDA is an approach to analyze the data, which gives the in-depth analysis of the data in the datasets and also summarizes statistical value.

3.3.1. Univariate Analysis

```
    country
USA      : 1087
France   :  775
England  :  659
Malaysia:   633
Sweden   :  632
Germany  :  575
(Other)  :35301
```

Figure 7. Country wise Analysis

In FIFA world cup the more number of countries are participated in the international Football tournament. Here USA participated frequently. Germany has played for 575 times in the tournament as in the figure 7.

```
      world_cup.city world_cup.country df_continent
39657         Moscow            Russia       Europe
39658 Nizhny Novgorod           Russia       Europe
39659         Samara            Russia       Europe
39660   Rostov-on-Don           Russia       Europe
39661  St. Petersburg           Russia       Europe
39662         Moscow            Russia       Europe
```

Figure 8. Continent wise Analysis

In this continent wise analysis is performed as in figure 8. Here the country is integrated with the continent by using geocode. It predicts that the “Europe” continent gives more contribution to play in football and Oceania continent gives the less contribution in football. Oceania is the continent which combines east pacific, malaises and some island near to east pacific.

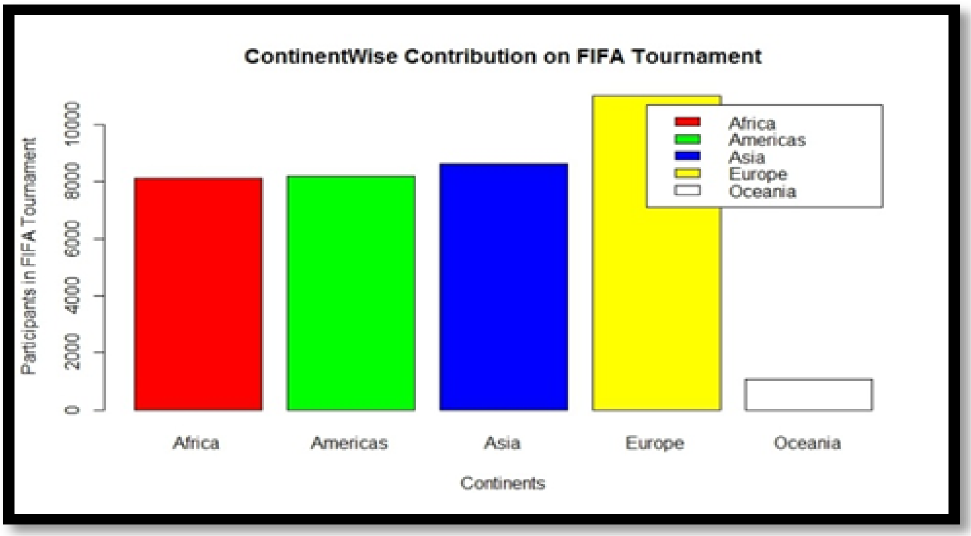

Figure 9: Continent wise Contribution in FIFA World Cup Tournament

3.3.2. Multivariate Analysis

In this dataset the relationship between the various attribute has been performed. The correlation between the country and home team is highly correlated because home team always depends upon the country. Then the attribute away team and city is next highly correlated (0.64). The red negative value is represents in red colour.

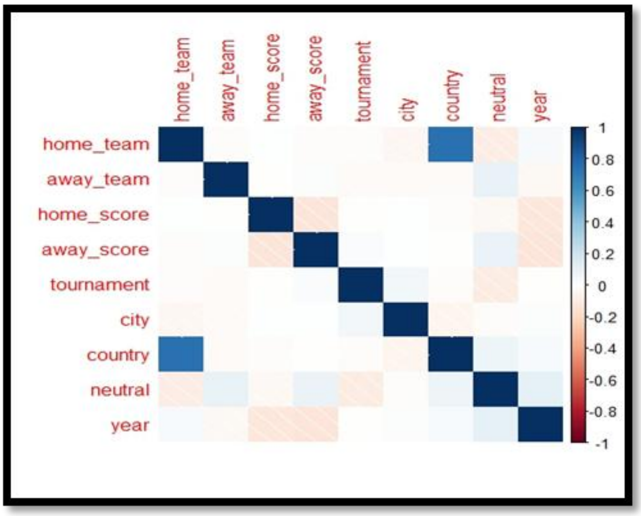

Figure 10: Correlation Analysis among attributes

3.4 Data Sampling: Feature Selection

The data has been split based on the random sampling method. In random sampling method 80% of data has been trained and 20% of data has been used for tested in the first model and 70% of data has training data and 30% of testing data has been taken for second model The training and testing data for two models is compared in the table 2.(www.coursera.org)

Table 2: Random Sampling Method

Sampling Data	Train Ration (%)	Test Ratio (%)
Model1	80%	20%
Model2	70%	30%

Sampling Data / Ratio	Training Data	Testing Data
Model1 (80%, 20%)	> dim(train) [1] 31729 9	> dim(test) [1] 7933 9
Model2 (70%, 30%)	> dim(train) [1] 27763 9	> dim(test) [1] 11899 9

	home_team	away_team	home_score	away_score	tournament	city	country	neutral	year
34471	93	73	0	0	52	1668	84	0	2012
11345	264	145	1	2	52	744	239	0	1980
5190	288	25	3	2	91	210	261	0	1962
23244	120	253	2	0	2	1715	129	1	2000
36376	18	34	1	2	52	1779	13	0	2014
20701	176	235	1	3	82	807	161	0	1997

	away_team	home_score	away_score	tournament	city	country	neutral	year
34471	73	0	0	52	1668	84	0	2012
11345	145	1	2	52	744	239	0	1980
5190	25	3	2	91	210	261	0	1962
23244	253	2	0	2	1715	129	1	2000
36376	34	1	2	52	1779	13	0	2014
20701	235	1	3	82	807	161	0	1997
25788	224	4	0	3	882	125	0	2003

Figure 11. Train data and Test data

3.5 Logistic Regression Model

Once the random sampling is done. Then the trained data is fit into the model. Here the independent variable is neutral which has two categorical values 0 and 1. The dependent variable and always depends the independent variable. (www.guru99.com)

```
> model1<-glm(neutral~., data=train, family="binomial")
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.636e+01  1.362e+00 -19.345   <2e-16 ***
home_team   -9.052e-03  2.692e-04 -33.625   <2e-16 ***
away_team    3.174e-03  1.800e-04  17.634   <2e-16 ***
home_score   1.713e-03  8.738e-03   0.196    0.845
away_score   1.585e-01  1.033e-02  15.345   <2e-16 ***
tournament  -1.132e-02  6.769e-04 -16.730   <2e-16 ***
city        -1.986e-05  2.744e-05  -0.724    0.469
country      9.486e-03  2.890e-04  32.829   <2e-16 ***
year         1.259e-02  6.809e-04  18.492   <2e-16 ***
```

Figure 12. Model 1

Fit all the predictors into the model to check the coefficient of the model. Variable "neutral" is the target variable.

```
> model2<-glm(neutral~home_team+away_team+away_score+tournament+country+year, data=train
, family="binomial")
```

Model	Model1	Model2
AIC Value	28660	28657
Residual Deviance	28642	28641

Figure 13. Model 2

In # based on asterisk values we can find significant or not. Here home_score and city variables are insignificant. So they are eliminating from the model by using backward elimination approach. It goes iteratively still insignificant value arrives.

3.6 ANOVA: Comparative Analysis

In ANOVA tests two models to find the best fit. The residual difference between two models is -2 and deviance is -0.52239.

```
Analysis of Deviance Table

Model 1: neutral ~ home_team + away_team + home_score + away_score + tournament +
    city + country + year
Model 2: neutral ~ home_team + away_team + away_score + tournament + country +
    year
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     31720      32471
2     31722      32471 -2 -0.52239   0.7701
```

Figure 14. ANOVA: Comparative Analysis

3.7 Validation Metrics

To optimize our model, we use this validation metrics. The model consists of insignificant and significant predictors by eliminating the insignificant values directly (city, home score) may change the model. So, we concentrate on Null Deviance, Residuals and AIC parameters.

Model1 & Model2 in [70% , 30%] Sampling

Model	Model1	Model2
AIC Value	28660	28657
Residual Deviance	28642	28641

Model1 & Model2: [80%, 20%] Sampling

Model	Model1	Model2
AIC Value	32528	32526
Residual Deviance	32510	32510

Figure 15. Comparison of Models with Validation Metrics

3.7.1. Residual Deviance

It performed based on the independent variable in the model. The outcome of the model always gives the minimum value for better accuracy. The Residual deviance of model and model1 is (32471).

3.7.2. AIC

In this AIC value for model is calculated as “32489” and for model1 is “32485”. While iterating the value of AIC is decreased in the model1. This is fit because AIC always prefers the minimum value. Residual deviance is same for both models.

3.7.3. Model Optimization

In model it has two insignificant predictors so need to eliminate the predictors based on the significance. By eliminating home score and city the AIC value is minimized from “32489” to “32485”. So, we conclude that the elimination of these predictors is not related to the results of the match.

3.7.4. Accuracy: Confusion Matrix

	Actual	
Predicted	0	1
0	20805	5171
1	161	1626

Figure 16. Confusion Matrix

3.7.5. Accuracy of Actual Vs. Predicted

Accuracy is to find whether the predictors are fit in the model. Here accuracy for model and model1 is compared. The models have 76% of accuracy. (www.sthda.com)

```
> accuracy
  conf_mat1 conf_mat2
1  0.763245 0.7632765
```

Figure 17. Confusion Matrix

```
 Accuracy Cuttoff.37891
0.8164490     0.3942644
```

Figure 18. Identifying Best Fit

3.7.6. ROC Curve

To increase the accuracy of the model threshold value should be accurate. ROC curve helps us to find the threshold value. It summarizes the model's performance by validating between true positive rate (sensitivity) and false positive rate (1-specificity). The area under curve is a perfect performance metric. This neutral variable has two values TRUE and FALSE, so it lies between 0 and 1 respectively.

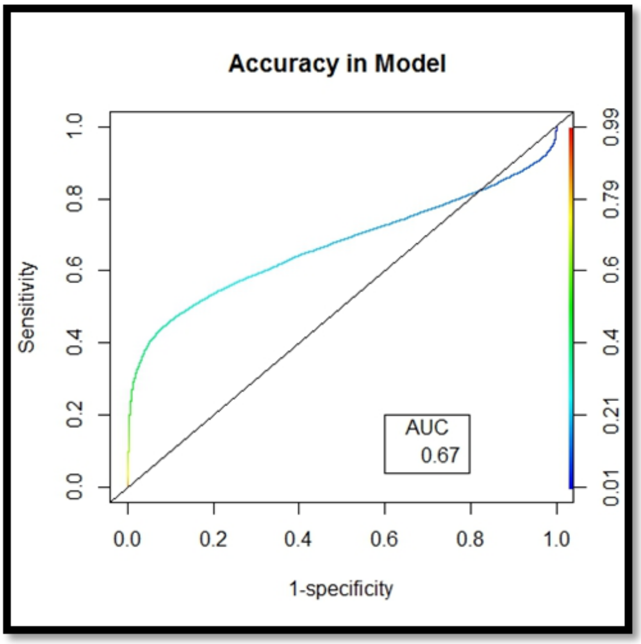

Figure 19. ROC Curve: Accuracy metrics

3.8 Discussion

Brazil	England	France	Germany	Malaysia	Sweden	USA
501	659	775	575	633	632	1087

Brazil	Germany	Malaysia	Sweden	England	France	USA
511	575	600	658	666	775	1087

Figure 20. Before Predicting Model and After Predicting the Model

From the results,

- USA has the maximum number of contributions in Football tournaments.
- Brazil has the minimum number when compare to other countries.
- England, Brazil these countries predictions are increased when compared to original data.
- Most of the Countries are falls in Europe Continent. They give the contribution in playing and participating in football tournament.

4. CONCLUSION

In this paper, we predict that the outcome of the match is depends upon the predictors like city, country, home score, away score, etc. In this analysis we contribute on Football tournament based on the FIFA World cup. The model we have designed gives prediction of football around the Globe. The prediction says that Europeans are contributed a lot when compare to other continents. From this analysis the people who wants to make a career in sports management and analytics, it is important to note that there is no sign of this field being less important and it is far from being a fad. The time is right to step into and contribute to this dynamically changing field.

REFERENCES

Cited at " www.greatmomentsof sportsmanship.com/"

Cited at " http://r-statistics.co / Logistic-Regression-With-R.html"

Cited at " https://www.kaggle.com/ tadhgfitzgerald/fifa-international-soccer -mens-ranking-1993now"

Cited at " https://www.towardsdatascience .com/exploratory-data-analysis"

Cited at "https://www.jessesadler. com/post /geocoding-with-r/"

Cited at " https://www.hackerearth.com/ practice/machine-learning/machinelearning-algorithms/logistic regressionanalys is-r/tutorial/"

Cited at "https://www.geeksforgeeks.or gcross-validation-machine-learning/"

Cited at "https://www.analyticsvidhya .com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/ "

Cited at " https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/"

Cited at "https://www.coursera.org /lecture/wharton-quantitative-modeling/4-7-logistic-regression-PYkDQ"

Cited at "http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/"

Cited at "www.guru99.com/r-generalized-linear-model"