DATA SCIENCE 871

Sahil Bhugwan

https://github.com/SBhugwan/Data-Science-871-ML-project-

Table of Contents

https://github.com/SBhugwan/Data-Science-871-ML-project-

Introduction

Machine learning is being a prominent feature in today's society as it can be used in a variety ways that can improve the decisions making. This project will try to utilize a machine learning on the previous FIFA world cup. During this project my main idea which I will explain later was to try and predict who would have won the FIFA world cup, however due to unforeseen circumstances I had to pivot and decided to predict the survival rate of the teams meaning from the 32 teams which 16 teams will make it to the knock out round. I tried using a 6 different models to see which one was the best.

Idea

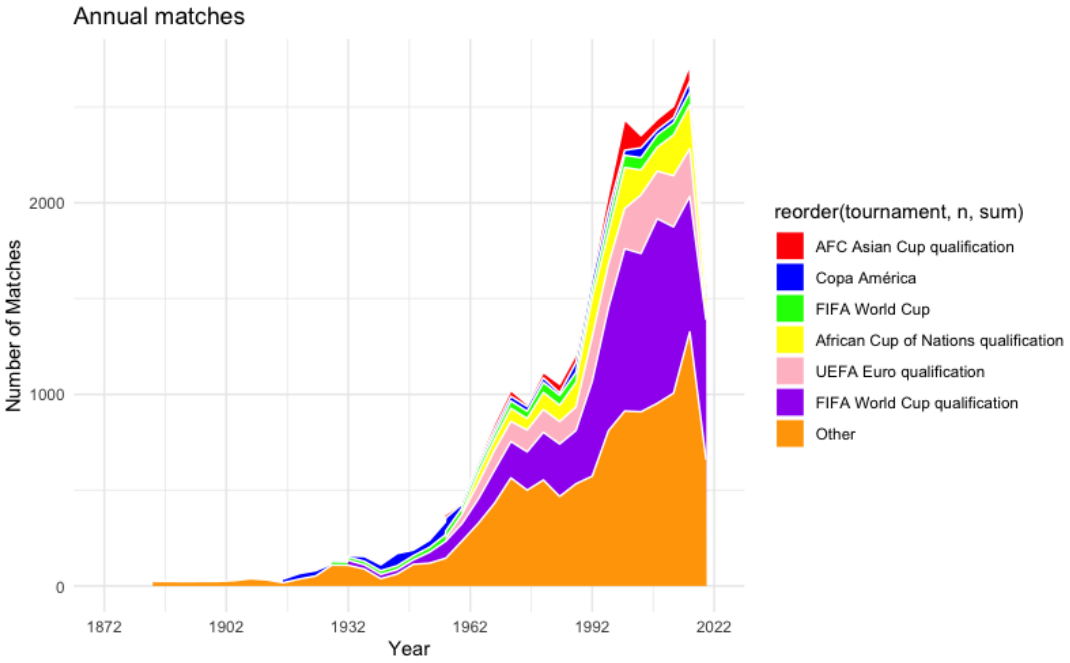
The main idea of this project is to try and predict who should have won the 2022 FIFA world cup in Qatar. My idea was to look at all the past international matches that has been played since 1872. The reason for this is sometimes there tends to be patterns that emerge, as in sports there is an old saying that history repeats itself. As well as when one views sport on TV many analysts tend to compare current matches to past ones. I then will also look at the ranking of the teams in the World cup as that usual as some indication on how well a team is doing leading to a world cup. This is because if a team is ranking high it generally indicates that they are performing well and winning consistently.

Data

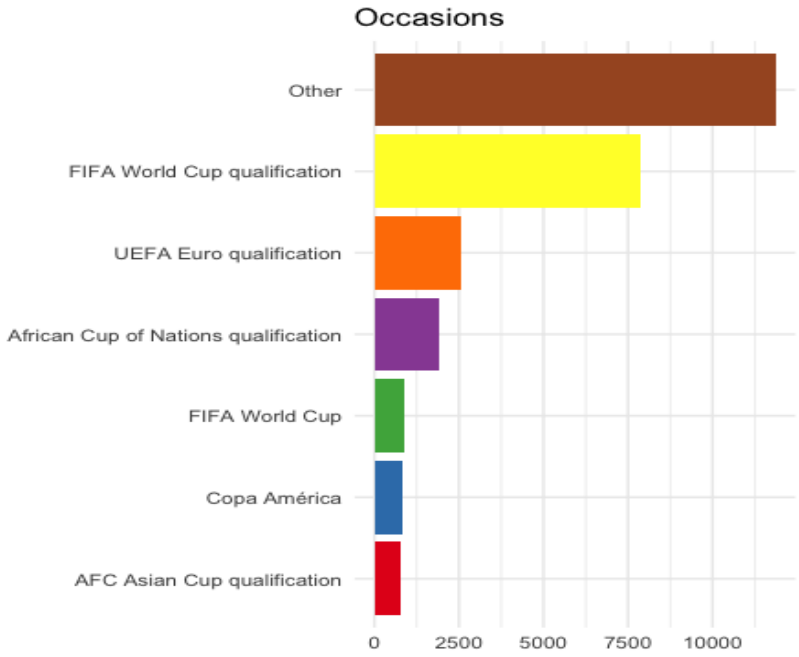
The data that I will be using for this project will be all the international soccer matches which I found on Kaggle. This data set includes all the matches that was played it has both home and away team, there scores as well as the tournament that the match was being played. Given that this data set included data on all soccer matches up an till after the world cup I then filtered the data to just before the world cup as I want to predict the matches in the world cup based on past performance. The next data set I loaded was the FIFA world cup ranking. This data set contains the ranking of teams since its inception in 1992. It is updated based on when international matches therefore there is no consistent pattern to when it was updated. After that I then converted the data to date time as well as filtered the data from August 1, 2018, onward. Given that certain teams in the data set have different names I then filtered it so that there is consistency throughout. I then merged the data set with the games played. Below is an example using Brazil as an example (view code).

Data Exploration

In this section I will do some analysis on the past 150 years of soccer matches. During the international season international teams play a wide range of different games that include friendly matches, world cup/ continent qualifiers as well specific tournaments. However one must take note that the FIFA world cup only takes places every 4 years and if a team plays all the group stage matches and all the knock out matches then the maximum games they will play will be 7 matches therefore in total a FIFA world cup only consists of 64 games every 4 years. Therefore in a 4 year cycle majority of international matches will be friendly games.

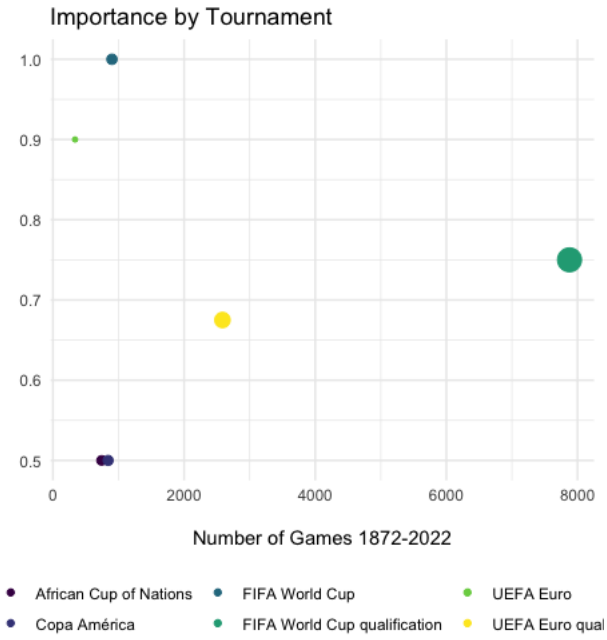


as seen in Figure above it clear that majority of the games that a team plays is friendlies. FIFA world cup qualifiers as well as the teams continent qualifiers are the second most matches that international teams will play. This can be examined more clearly on the figure below.



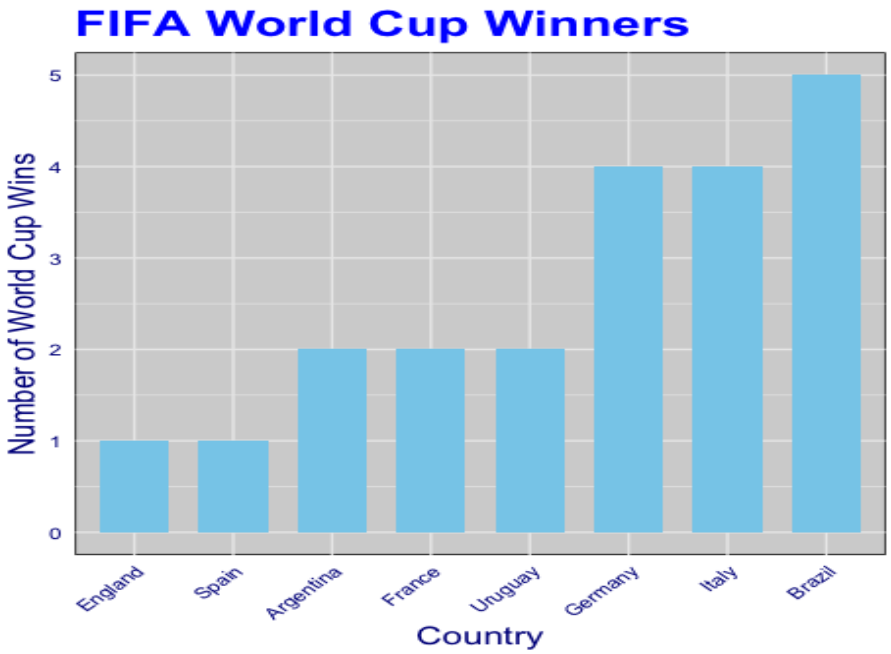
Therefore one can see that it is vital that teams perform well in qualifications and friendly matches as it will allow the team to build momentum to going into a major tournament. As when a team makes the tournament they play a handful of games and once they make the knock out matches there is no room for error.

Given that teams play a lot of friendly games one most ask the question how important these games are in the context of winning a world cup. Intuitively it makes sense that the qualification games are important, as a team has to qualify for a world cup in order to win a world cup. looking at the continental tournaments such as Euro or African cup of nations they are important matches in terms of finding out who the best team in in that specific continent. However those matches don't have an impact on the world cup, as a team can win that trophy and still not qualify for the world cup. an example of that most recently was Italy that won the Euro in 2021 but failed to qualify for the 2022 world cup. The figure below is a breakdown of importance of each tournament.



Therefore if a countries sole purpose is to win the FIFA world cup they should be looking to perform their best during the FIFA world cup qualifications therefore increasing their chance of making it to the world cup. One important consideration that should be noted is that friendlies and any qualification have an impact on the teams rankings in the world, this is a psychological factor that can impact on how the teams perform in crunch matches.

Therefore I will next look at all the previous FIFA world cup winners as illustrated in the graph below.



It should be noted that the first FIFA world cup was played in 1932 and is played every 4 years. There was however only two world cups that where played this was due to World War 2. Looking at this graph it is clear that Brazil is by far the best performing team at the world cup. Just for interest Brazil is also the only team in the world to have played at every single world cup since its inception in 1932, which clearly shows that by playing in world cups it does increase the likelihood of winning the tournament.

Initial Plan for Feature Engineering

my initial idea was to look at a wide range of features to try and predict who should have one the FIFA world cup. To do this I was going to look ranking , points, home/away fixtures as well as a variety of other features. However I ran into a few issues with the code not being able to producing spuriousness results. The code for this can be viewed on GitHub on previous commits

Feature Engineering

Given this I decided to simplify it even further I decided to predict the teams survival of making out the group stages of the World cup. Given that 32 teams start the world cup, I wanted to see after the group stage matches with 16 teams will remain. To do this I will only be looking at only a few features that being matches from 1992 the reason why I decided to look from matches from there is due to the fact that teams rankings were first introduced back in 1992. Therefore from this I can see the changes in the rank of the team and work out the average rank

of the team these are all done by looking at the total points that a team gets from playing international matches. The other additional features that I will be looking at is where the match that they playing has some level of stake (importance of the game) and it was played at a world cup.

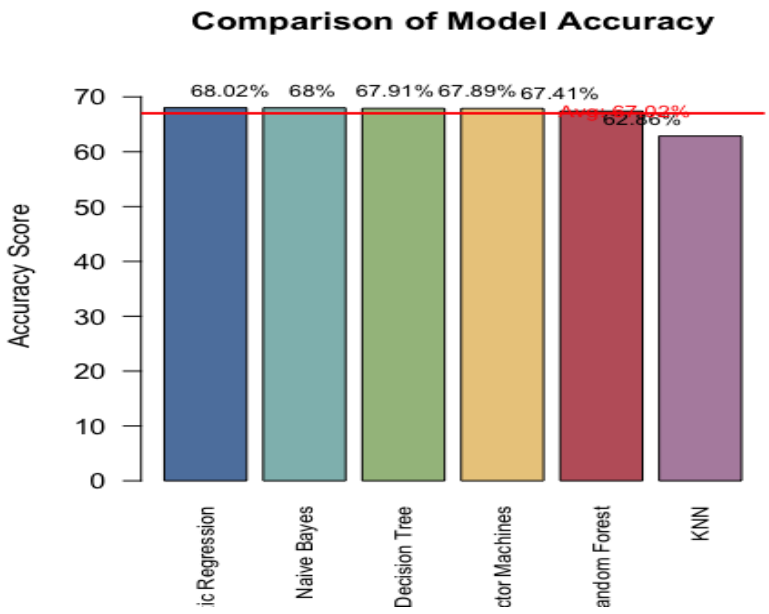
Training

In this section I will discuss what I did with regards to predicting the matches, these features that I mentioned earlier (that being average rank, rank difference , point difference , stake of match and lastly if it was played at a world cup). These features I will be using for training the machine learning models. looking at the code (that is in the feature engineering and prediction) I first split the model into two different sections that being training and testing sets for machine learning. By splitting the data into different sets I can use the X_train and y_train sets to evaluate the performance on the unseen data using the X_test and y_test sets. Therefore by separating these sets will allow for the model to asses it ability with regards to being able to generate new, unseen data that and thereby avoid overfitting with regards to the training data. I did also create a function that ensure that it takes the target variable "y" which ensure that there is a 80%-20% split in favour of the training data. Another important note is that the variable y is the dependent variable in which the machine learning model aims to predict. I set the seed to 42 this to ensure reproducibility by doing this it ensure that I am able to be able to obtain the same train/ test split every time I run the code, this will ensure that my results are consistent and reproducible.

Models

The next thing that I did was look at different models. The models that I looked at are the following logistic regression, support vector machines, K-nearest neighbours, Gaussian naive Bayes, decision trees and random forests. Therefore each of these models have their different strengths and weakness, by trying different models I am able to compare certain metrics such as the area under the ROC curve which will allow me to identify which models are best suit for the task at hand. Another reason for looking at different models is that each of them has different hyperparameters that has an impact on its performance.

The results of each of the models can be illustrated below.



The results from my different tests clearly show that the logistic regression performs the best out of all the different models. Given this and the fact that accuracy is a common metric used which measures the proportion of correct predictions from the total number of predictions. In all the different models the logistic regression is able to correctly predict the outcome of the match 68.02% of the time on my test set.

Cross Validation

A cross validation technique is a re-sampling technique that is used in machine learning that is used to assess the performance of the models. The main idea of behind this to once again split the data into multiple sets these are referred to as folds. The model is once again training on some of the data (for example say K folds) and then evaluates on the remaining sets (K-1). This process is then repeated multiple times however it uses different sets of the data this allows for it to serve as the validation each time. The benefits of this is that it is able to provide a more robust estimate of the model when compared to a single train split, thus allowing for a better understanding of how the model is able to generalize the unseen data. This method is also used with regards to tuning hyperparamters, as it is able to allow to find the optimal combinations of hyperparemters.

I performed two different cross validations the first being of the random forest. I got a bias value of -66.76 what this results means is that the model is in fact performing worse without

cross validation. which is unexpected given the fact that cross validation is generally expected to improve the models performance. Therefore what can be concluded from this is that the random forest model without cross validation actually performs better. There are a few possibilities to consider why this may be the case such as data issues meaning that the data used in the cross validation might differ in some ways from the data used to train the random forest model. what I could find is that there was missing values in the prepossessing. Another possibility is the number of folds used (K) which will affect the results. I then ran a test to get the variance which was 0.0003822415. This results indicate that there is a very low variability with regards to the accuracy obtained in the cross validation process. The positive outcome in my results means that the model is not heavily influenced by the training and validation subsets. I then did a cross validation with regards to the logistic regression. I however also obtained a negative bias value of -67.34309 this suggests that once again this model performs worse that the normal logistics regression. This once again suggests that there may be issues with the cross performance validation.

logistic Regression ROC Curve

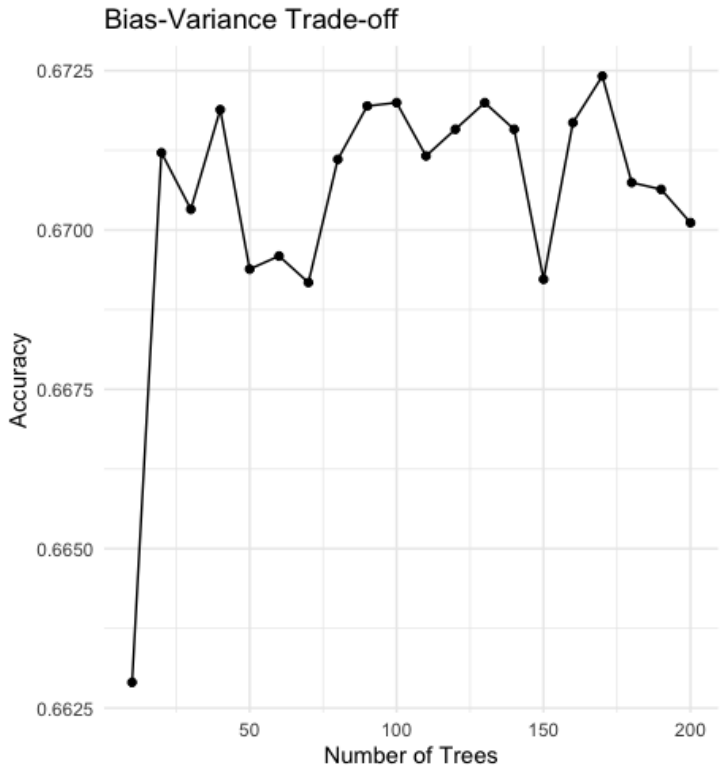
The ROC curve show the trade-off between the true positive rate and the false positive rate at different thresholds. For some reason I wasn't able to print the curve. I was able to calculate the area under the curve in my code it is print(roc_obj) which is 0.7541. What this mean is that my model has some level of discriminatory power. It performs better than a random classifier however it may have room for improvement to achieve a higher accuracy rate.

Bias Variance trade off.

An two important concept for machine learning are bias and variance. Bias is used to describe the error that occurs when a simplified model is used to approximate a real-world problem. It expresses how well the model's forecasts match the actual values. A model with a high bias makes significant assumptions about the data by oversimplifying it and possibly underfitting the training set. This means that the model is unable to capture underlying patterns and consistently under- or overestimates true values. Low accuracy and poor generalisation to previously unseen data can result from high bias.

Variance refers to the variability or the sensitivity of the models prediction with regards to the fluctuations in the training data. Therefore it measures how well the models predictions will change with regards to the trained data on different subsets. If the model has a high variance it indicates that the model is very sensitive with regards to random fluctuations in the training

data which may result in overfitting. This will generally occurs when the model fails to generalize the new unseen data.



The figure shows the bias variance trade off what one can see is that of the number of fold increase there is fluctuations with regards to the accuracy of the model. However it should be noted that the variance tends to fluctuate between 0.67 and 0.6725 which shows that even though there is fluctuations it tends to be small therefore we can see that that of average the accuracy of the model is about 0.67125

Prediction

The next thing that did was create a function to predict the probability of each of the group stages matches the results are as followers :

___Starting group F:___

Morocco vs. Croatia: Croatia wins with 0.60

Morocco vs. Belgium: Belgium wins with 0.66

Morocco vs. Canada: Morocco wins with 0.55

Croatia vs. Belgium: Belgium wins with 0.61

Croatia vs. Canada: Croatia wins with 0.60

Belgium vs. Canada: Belgium wins with 0.65

___Starting group C:___

Argentina vs. Saudi Arabia: Argentina wins with 0.69

Argentina vs. Mexico: Draw

Argentina vs. Poland: Argentina wins with 0.57

Saudi Arabia vs. Mexico: Mexico wins with 0.73

Saudi Arabia vs. Poland: Poland wins with 0.67

Mexico vs. Poland: Draw

___Starting group A:___

Senegal vs. Qatar: Senegal wins with 0.62

Senegal vs. Netherlands: Netherlands wins with 0.61

Senegal vs. Ecuador: Senegal wins with 0.59

Qatar vs. Netherlands: Netherlands wins with 0.74

Qatar vs. Ecuador: Ecuador wins with 0.57

Netherlands vs. Ecuador: Netherlands wins with 0.64

___Starting group E:___

Germany vs. Japan: Draw

Germany vs. Spain: Spain wins with 0.58

Germany vs. Costa Rica: Germany wins with 0.55

Japan vs. Spain: Spain wins with 0.64

Japan vs. Costa Rica: Draw

Spain vs. Costa Rica: Spain wins with 0.58

___Starting group H:___

Uruguay vs. South Korea: Draw

Uruguay vs. Portugal: Portugal wins with 0.58

Uruguay vs. Ghana: Uruguay wins with 0.69

South Korea vs. Portugal: Portugal wins with 0.65

South Korea vs. Ghana: South Korea wins with 0.62

Portugal vs. Ghana: Portugal wins with 0.71

___Starting group B:___

Iran vs. England: England wins with 0.63

Iran vs. USA: USA wins with 0.57

Iran vs. Wales: Wales wins with 0.56

England vs. USA: Draw

England vs. Wales: Draw

USA vs. Wales: Draw

___Starting group G:___

Switzerland vs. Cameroon: Switzerland wins with 0.60

Switzerland vs. Brazil: Brazil wins with 0.63

Switzerland vs. Serbia: Draw

Cameroon vs. Brazil: Brazil wins with 0.75

Cameroon vs. Serbia: Serbia wins with 0.65

Brazil vs. Serbia: Brazil wins with 0.56

___Starting group D:___

Denmark vs. Tunisia: Denmark wins with 0.55

Denmark vs. France: France wins with 0.59

Denmark vs. Australia: Denmark wins with 0.60

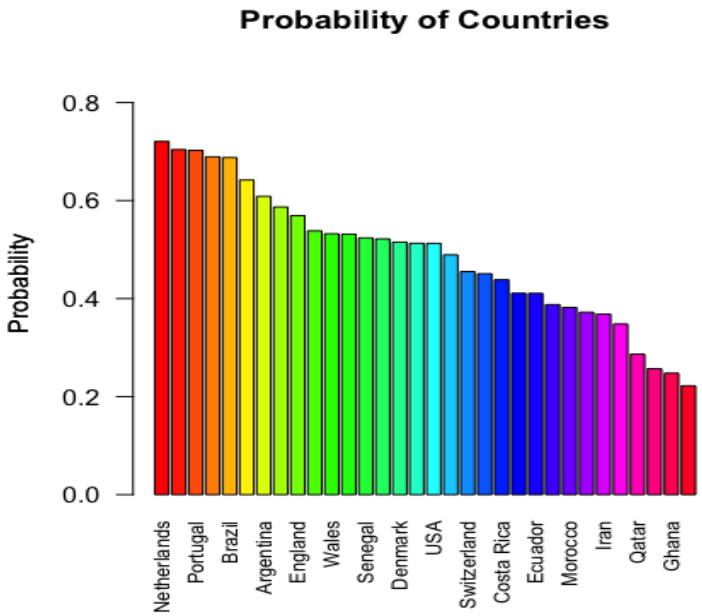
Tunisia vs. France: France wins with 0.68

Tunisia vs. Australia: Draw

France vs. Australia: France wins with 0.63

From these results we can see that some of these results actually happened, however a clear noticeable difference is that during the Saudi Arabia vs Argentina game my results predict that Argentina will win with a probability of 0.70 however we know that this in fact that didn't happened. However it should be noted that predict sports games is rather difficult due to the fact that there are always major upsets in competitions however given that my model did give an accuracy of 68.02% one can conclude that there may be some discrepancies in the predictions.

Given these results I then created a function that will calculated the expected points given the probability of each of the teams results. This was done so that I was able to see the survival of each of the countries that will continue out of the world cup. The graph below show the probability of each of the teams



The survival of the top 16 teams are very similar to that of the teams that actually made it out the group. However one thing that my model did predict is that a top team like Germany will not make it out of the group. This in factual fact did happen during the world cup, however a surprise of the world cup is that Morocco made it all the way to the semi-finally, however in my model it predicted that they won't even make it out of the group stages.

Conclusion

Therefore given this we can see that some of my results given from my model where able to predict some of the survival rate of the teams from the world cup. However given the unprecedent nature of soccer it can be very hard to predict matches. Though it should be noted that from all of my models the logistic regression had the highest accuracy of 68.02% implying that it is not perfect as there can be some improvements. A possibility can be to try and use different features that in my model which may improve the accuracy of my model.