*Marcel Socorro*  
*Big Data Analytics Capstone*  
*St. Thomas University*

*EPL Prediction Model*

Abstract—Analytics in sport has become a strong element for managers and directors to set strategies and planning seasons. In fact, it is a process which uses statistics, mathematics, computer science, data visualization, predictive methods, and some other study’s fields. This project is developed to get the best predictive model for a match in the English Premier League (EPL). Using sports-related data to find meaningful patterns like strong correlations, hidden trends and communicate those patterns (using graphs, charts, essays, etc.) to support a high accuracy for a three-outcome prediction is the goal for this project.

# Introduction

This project is related to sports analytics and more specifically to the outcome prediction for a soccer game in the English Premier League. A soccer game in this league has three possible outcomes:

-Home Team Victory

-Away Team Victory

-Draw

These three outcomes are what makes the prediction, for a soccer game in EPL, more complex than for some other sports like baseball or basketball. The main goal for this project is described as follows:

What is the best classification model, in terms of accuracy, to predict an EPL game using data in real time?

Adding live-data pulls for this project and storing in a database hosted on AWS will allow to use the data in real time.

To get this model, there must be a previous analysis of features selection, statistically proven what variables should be used, and test and validation of the model.

Sport analytics is a current popular trend, and the use of it helps teams make useful decisions to enhance their performance. Decisions involving player acquisition, fielding the best possible rotation on a given game, and which strategy to use against the opposite team evaluating their strength and weakness, are nowadays supported by analytics.

Predicting the outcome for an EPL game is not only useful in the Gambling Industry, it is also useful to know the variables/features that could be impacting these outcomes, so they can be properly managed to get the desired result. This is a critical purpose for people making decisions in the teams playing the League.

This problem is approached using machine learning techniques like Deep Learning Artificial Neural Network, compared to Naïve Bayes, Regression and Classification Trees and Logistic Regression. This analysis is supported by significance test and predicted outcomes accuracy.

1. BACKGROUND

The problem defined is not a new problem; since competitive sports were born, it has been a remarkable interest trying to predict the outcome of the game. However, this interest has evolved from its simplest way to a more complex process involving statistics and machine learning algorithms. Sport analytics is a current popular trend, and the use of it helps teams making useful decisions to enhance their performance. Decisions involving player acquisition, fielding the best possible rotation on a given game, and which strategy to use against the opposite team evaluating their strength and weakness, are nowadays supported by analytics.

This specific project is a third iteration from 2 previous projects covering the same topic. The objective for the first iteration was developed through the statement: How many hidden layers can a soccer-match predicting artificial neural network have before accuracy significantly degrades?

The variables used for this prediction were based on a general performance of the teams the League analyzed, and it does not take in count the history of matches between two teams when the outcome is predicted for a game. The best accuracy it could be obtained was around 62% matching the results with the predicted outcomes. It was only used the technique of Deep Learning with different configurations of node numbers. The approach for this project is to include other classification techniques.

The second project was done based on the EPL team’s general performance together with individual players’ variables. The accuracy registered in this study was around 35% with a configuration of 10 layers.

For this third iteration, the prediction is obtained using the general performance for a team in the current league, but also adding the historic performance of this team against the rival involved in the game predicted. This problem is approached using machine learning techniques like Deep Learning Artificial Neural Network, compared to other classification techniques and supported by significance test, overfitting analysis and predicted outcomes accuracy.

The challenge for sports analytics is to quantify a team performance. Besides the statistics obtained in the game, the people studying this topic have brought interesting ideas and metrics trying to reflect the performance of a Team. The Elo rating system was first introduced to calculate the various skill levels of chess players. Originally these ratings were only used in chess; however, Elo ratings now have many other real-world applications. Most other real-world applications of the Elo rating system involve the rating of sports teams and players in gaming tournaments, with a few other applications. This rating system has been used in such things as the Bowl Championship Series system in college football, the League of Legends videogame tournament, and soft biometrics [1].

Like it was explained before, soccer game in this league has three possible outcomes:

-Home Team Victory

-Away Team Victory

-Draw

These 3 outcomes are what makes the prediction, for a soccer game in EPL, more complex than for some other sports like baseball or basketball.

Due to the dynamic nature of soccer, it is difficult to separate classes especially when a game is not dominated by a team. The game could easily end with any of the three outcomes. However, as it the middle ground between “win” and “loss, “draw” intuitively is the most difficult to predict [2].

In other literature could be find the use of software designed to predict the outcome for a soccer game where the software can provide the implementation state-of-the-art data mining and machine learning algorithms like Bayesian Networks [3].

Other algorithms weigh the parameters of the estimation in 3 steps:

-Calculating the number of goals expected each team to score during the match. These projected match scores represent the number of goals that each team would need to score to keep its offensive rating exactly the same as it was going into the match, and they are adjusted for a league-specific home-field advantage and the importance of the match to each team.

-Using the projected match scores and the assumption that goal-scoring in soccer follows a Poisson process, which is essentially a way to model random events at a known rate, is generated two Poisson distributions around those scores. These results in the likelihood that each team will score no goals, one goal, two goals, etc.

-Taking the two Poisson distributions and turn them into a matrix of all possible match scores, from which it can be calculated the likelihood of a win, loss or draw for each team. To avoid undercounting draws, the corresponding probabilities are increased in the matrix to reflect the actual incidence of draws in a given competition [4]

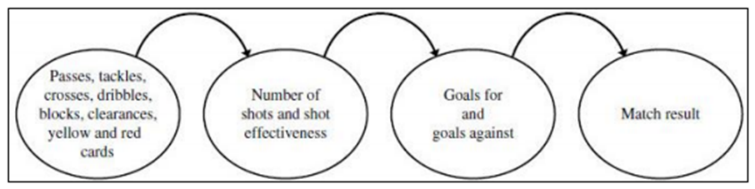
The objective for this project is not to predict score, only the outcome, as a classified variable, therefore the previous approach is not used.

There are external variables linked to the game that could make an impact on the outcome of the game. These variables are if a team is playing home or away, the attendance of the people to the game, the referee designed to guide the game. There are previous articles pointing out the significance of a referee performance for a team playing at home [5]. In this project this variable won’t be analyzed from its impact, but it could be a suggestion for future studies.

As every sports team, the individual actions performed as a Team will impact on the result of the match. A soccer game requires majorly four sets of player skills which are carefully chosen within the playing eleven. They are – goalkeeper, defenders, midfielders and strikers [6]

Sports analytics experts understand that The Game is still human. It is why they got into the field in the first place. It is what all the formulas, numbers, and analyses are about-measuring, managing, and making the most of the people who get to play The Game [7].

The figure below, summarizes in a simple way the essence of the game, however this study also weighs the history of previous games between teams and external variables as playing in home or away, referees and people attendance to the stadium.

*Figure 1. Elements impacting a soccer game*

1. MODEL DESIGN

Six different techniques are developed in this paper: Linear Regression, Logistic Regression, Clustering, Naïve Bayes, Decision Trees and Deep Learning Artificial Neural Network. All of them supporting the predictive analysis of the problems previously defined.

The purpose of using a Linear Regression model is to predict the quantity of points of a team in the last round of a given season. In the scatter plot for both variables could be seen a strong positive linear relationship between them

The input variable: Cumulative goal difference.

The output variable: Cumulative quantity of points per team.

The sum of quantity of points for a team in a season is a parameter measuring the combination of possible outcomes for this team in every match played.

The purpose of using a Clustering model is labeling teams according to the relationship between the average of cumulative quantity of goal difference and points in the last round of each season studied.

Clustering these variables into different labels will allow the models to weigh the previous performance for each team in the previous five seasons

The input variable: cumulative average goal difference and cumulative average quantity of points per team.

The output variable: the label for each team as an integer variable declared as factor

The sum of quantity of points for a team in a season is a parameter measuring the combination of possible outcomes for this team in every match played.

The objective of using a Naive Bayes, Decision Trees, Logistic Regression and Deep Learning is to predict one of the three possible outcomes in an EPL soccer game

1. DATA SCHEMA

The variables used in the analysis come directly from scraping the website *https://fbref.com/en/comps/9/schedule/Premier-League-Fixtures* with the following variables for each match. The variety of the data is basically structured data as per table below.

|  |  |  |
| --- | --- | --- |
| Field | Type | Description |
| Wk | int | Round of the season |
| Day | str | Day of the week |
| Date | date | Date of occurrence |
| Time | time | Time of the match |
| Home | str | Team playing at home |
| Score | str | Score ending the game |
| Away | str | Team playing away |
| Attendance | int | Attendance of people |
| Venue | str | Name of the Stadium |
| Referee | str | Referee designed for the match |

*Table 1. Data set scraped*

The parameter Volume is described with the previous data recorded in 3 different tables. The tables Match, Team and Season are linked among them previously designed in a database. The table Team has 30 rows, the table Season has 6 rows and Table Match has 4376 rows.

From the previous table the data was transformed into a greater quantity of variables in order to get an accurate prediction model. These new variables were determined according to results in the exploratory analysis. The dataset and different variables used for this analysis are defined in the table below for each match.

|  |  |  |
| --- | --- | --- |
| Field | Type (Indepent / Dependent) | Description |
| Season | Int (Ind) | Season played |
| Wk | Int (Ind) | Round of the season |
| Team | Str (Ind) | Team involved in the match |
| GF | Int (Ind) | Goals for |
| GA | Int (Ind) | Goals against |
| HomeAway | Str (Ind) | Team playing at home or away |
| Dif | Int (Ind) | Goal difference |
| CumPoints | Int (Ind) | Cumulative quantity of points in one season |
| CumDifGoal | Int (Ind) | Cumulative quantity of goal difference in one season |
| CumPointsHA | Int (Ind) | Cumulative quantity of points in one season when the team plays at home or away |
| CumDifGoalHA | Int (Ind) | Cumulative quantity of goal difference in one season when the team plays at home or away |
| CumPoints Normalized | Float (Ind) | CumPoints normalized with function MinMax for all teams by each round |
| CumPoints NormalizedHA | Float (Ind) | CumPoints normalized with function MinMax for all teams by each round and by playing at home or away |
| CumDifGoal Normalized | Float (Ind) | CumDifGoal normalized with function MinMax for all teams by each round and by playing at home or away |
| AvgPointsL5G | Float (Ind) | Average of points in last 5 games per Team |
| AvgDifGoalL5G | Float (Ind) | Average of goals difference in last 5 games per Team |
| HistOpTeamPerf | Int (Ind) | Sum of points against opposite team in previous seasons |
| HistOpTeamGoals | Int (Ind) | Sum of Goal Dif against opposite team in previous seasons |
| Cluster | Int (Ind) | Cluster determined from the performance of all teams in previous seasons |
| PrevGamePerf |  | Average of number of points a home team got vs the opposite team in previous seasons |
| Outcome | Str (Dep) | Defeat/Victory/Draw |
| All the previous variables are repeated for the opposite Team in the same row | | |

*Table 2. Schema Dataset EPL Prediction*

One row in the data transformed belong to a team performance in one match; this is the reason why one match is divided into two rows for each team involved in the match. This way the dataset is enlarged to better train the data.

The previous variables were designed weighing patterns according to results obtained in the exploratory analysis like the significative difference between the quantity of points for teams playing at home and away. Another parameter taken in count was the clusters labelling the Teams according to their performance in previous seasons.

All these variables will be involved in a feature selection process in which is intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable, this way the parameter Veracity could be handled and controlled for variables that authentically bring value to the model avoiding overfitting to keep the accuracy results for future data.

V. DATA HOSTING, ETL AND DATA TRANSFORMATION

Data will be hosted remotely on Amazon Web Service (AWS) using a Relational Database Service (RDS). RDS is fully managed, monitored and updated by AWS. It is chosen MySQL 8+ as Database Engine. Every week in an active season the model is prepared to scrap the website in R and then will store in the database the result of the last round performing the Velocity as one of the elements of Big Data 5 V’s.

The Value of this data is set to predict the outcome of future games with an acceptable (greater than 70%) accuracy. To accomplish this goal the data is split in a proportion of 80% for training and 20% for testing. The testing set is split in a random way to diminish the impact of a team performance’s trend.

R uses two libraries to connect the instance on RDS and extract the data from it. Connection to MySQL 8+ is encrypted in transit and protected by user and password. For the purpose of this project it is opened to the security group (firewall) to accept all incoming connections.

The load of the data is accomplished with DBeaver. A schema is created with the fields or variables that were selected and imported the csv values from the local computer to the database hosted in MySQL 8+.

Extraction of data is done through R, by using the following libraries: RMySQL and DBI.

The following main libraries are used in R

RMySQL--- Connection with the database in DBeaver

RODBC--- Uploading data into the database in DBeaver

Rvest--- Scrapping website to get the data

tidyverse--- Transformation and visualization of the data

magrittr--- Forward piping operator

caTools--- Splitting and sampling the data

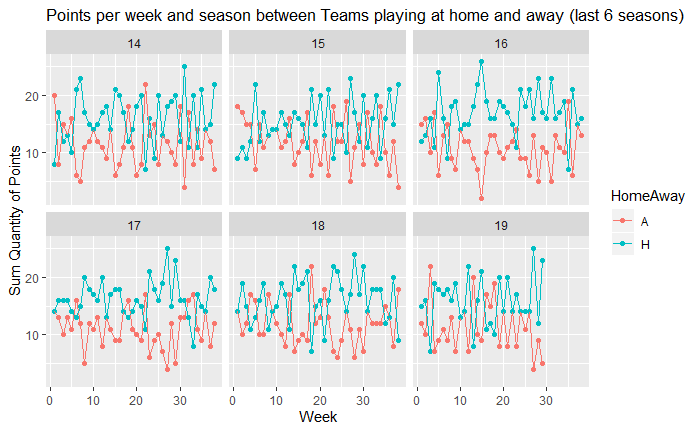
e1071--- Naïve Bayes Model

rpart & rpart.plot--- Regression and Classification Tree

nnet--- Artificial Neuralnetwork for classes

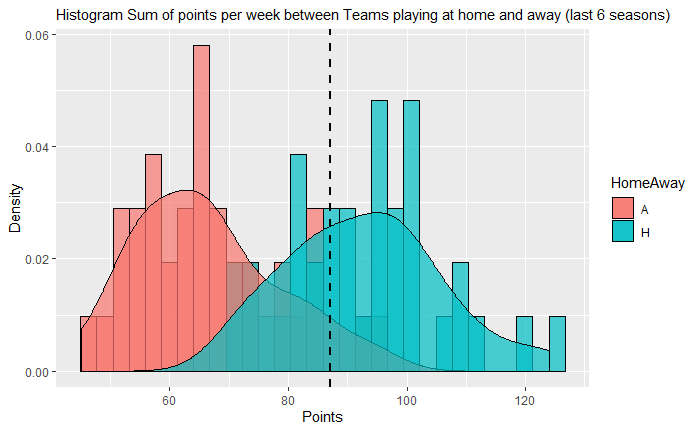
VI. RESULTS AND EXPLORATORY ADATA ANALYSIS

Teams performing different at home or away is one of the variables to be suspected to impact the game. Plotting the data for the last six seasons, allows to confirm that indeed a Team playing at home tends to perform better than playing away.



*Figure 2. Points per round between teams playing at home and away.*

Comparing this data through a density histogram allows to observe the difference between the means of points obtained per week for Teams playing at home and away in a given season.



*Figure 3. Histogram Sum of points per week between Teams playing at home and away (last 6 seasons)*

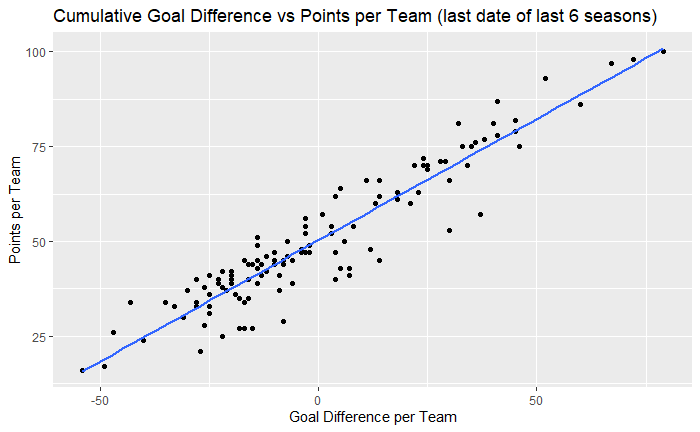
Is this difference between the means significant?.

First, Shapiro test is run for the variable Quantity of Points per week, the pvalue obtained is 0.225. The results don’t show enough evidences to reject null hypothesis; the variable follows a normal distribution.

Once it is proved that the variable is normally distributed, it is confirmed that there is a significant difference between the Sum of quantity of points obtained by teams playing at home and away each week. Pvalue equal to 7.892e-09 is less than 0.05, which means that there is enough evidence to reject null hypothesis (No significant differences between both variables). It could be confirmed that playing at home or away, it is a factor impacting the outcome of the game.

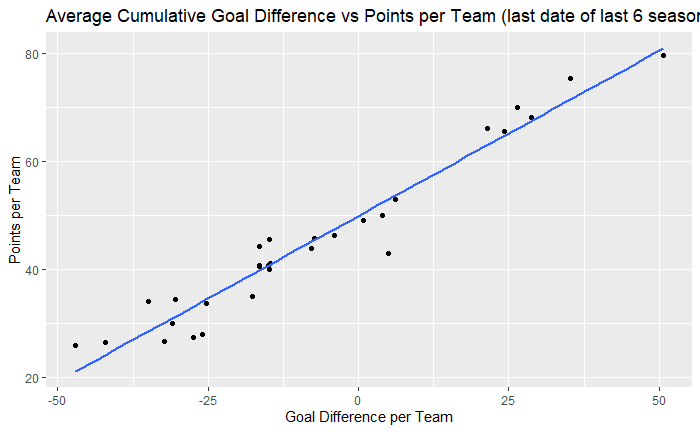
This is the reason why, any variable used for predictions in the model is split into 2 different variables for teams playing locally or as visitor.

Analyzing the relationship between the variables Goal difference per Team and Points obtained per Team at the last round of each of the last six seasons it could be obtained the following graph.



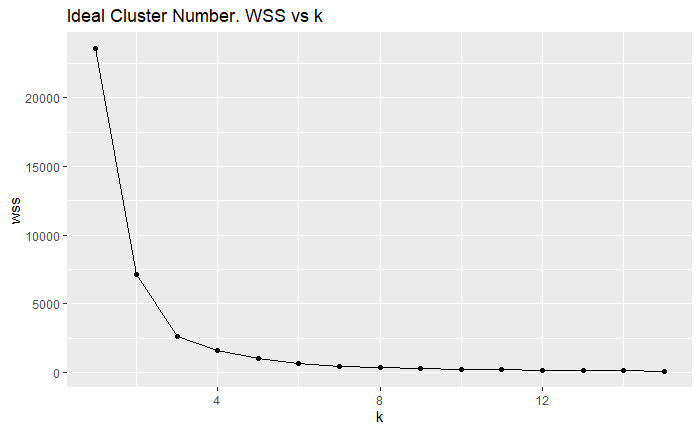
*Figure 4. Cumulative Goal Difference vs Points per Team (last date of last 6 seasons)*

Based on this analysis the performance of any team could be estimated by the average of the previous variables in the last six seasons and getting the same plot by now based on the averages.



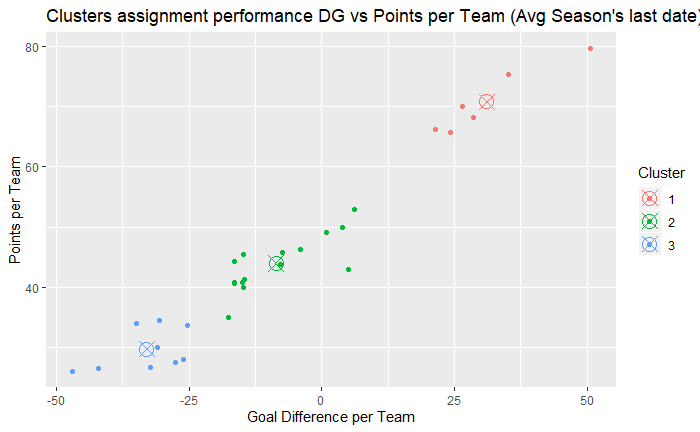
*Figure 5. Average Cumulative Goal Difference vs Points per Team (last date of last 6 seasons)*

Once we have this data, a clustering analysis is done to assign clusters for the teams and grouping by their performance in the last five seasons.



*Figure 6. Ideal Cluster number. WSS vs k*

According to wss vs k analysis, the optimal cluster is chosen for 3 units, the big drop happens in the third cluster, and from the third cluster, the drop in the wss is less than the half of the previous clusters. This selection is done assuming a big drop for a difference greater than 50%, and seeing in the graph that, for k=3, occurs the greatest inflexion, this visual observation is also known as elbow method.

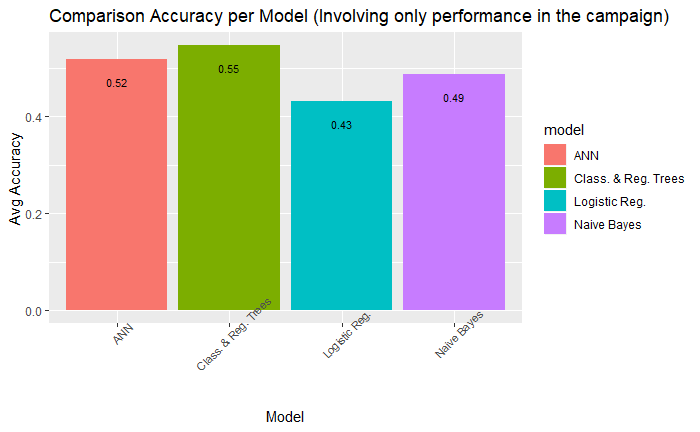


*Figure 7. Clusters assignment performance DG vs Points per Team (Avg Season's last date)*

In the graph above, each point (Team) is assigned to a cluster, this is an element to group teams with similar performance, to use it as a predictor variable in the model.

Once all the variables of team’s general performance are determined, the data was randomly split into 2 proportions, 80% of the data for the training model (1750 observations) and 20% of the data for the testing model (438 observations).

The models described in the graph below were obtained and tested based on the split made.

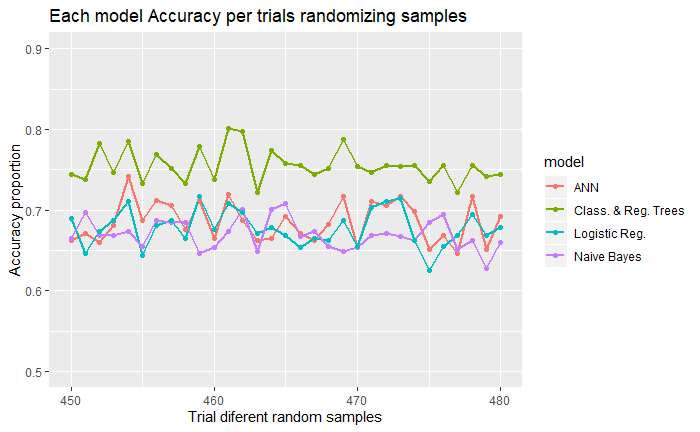


*Figure 8. Comparison Accuracy per Model (Involving only performance in the campaign)*

The values of accuracy obtained from each model are different and they are not good to predict the outcome for an EPL match.

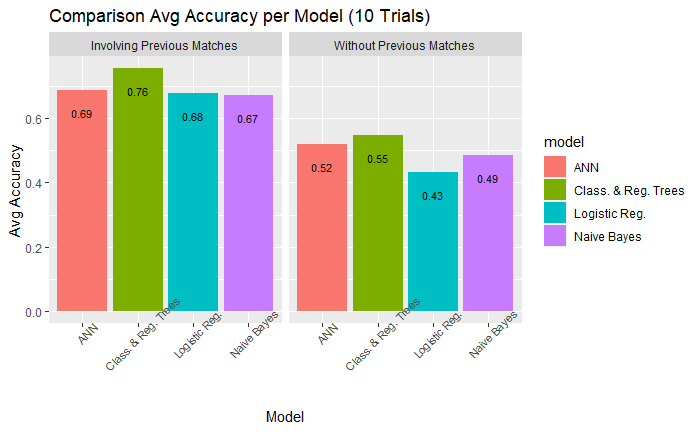
For the same dataset, it is included the variable PrevGamePerf; which is the average of number of points that a home team obtained vs the opposite team in previous seasons in which the data was collected.

This time a function was developed to get 31 different values of accuracies for each model. Also, the seed was changed for every run making sure the data is split in the same proportion but with different observations. Applying the same procedure for splitting the data into equal proportions of training and testing sets in the function, and running every model previously run the following graph shows the different values of accuracies for each model.



*Figure 9. Each model Accuracy per trials randomizing samples*

The values of accuracy were impactedby using the previous outcomes between matches for the same teams playing the match to be predicted as it can be seen in the graph below.



*Figure 10. Comparison Accuracy per Model with/without variable previous matches performance*

1. DISCUSSION

The accuracy for the outcome prediction models has been impacted by the variables included in the analysis, and all models have increased in almost 20% their accuracies, but the objective for this project it’s to find the best proven statistically model.

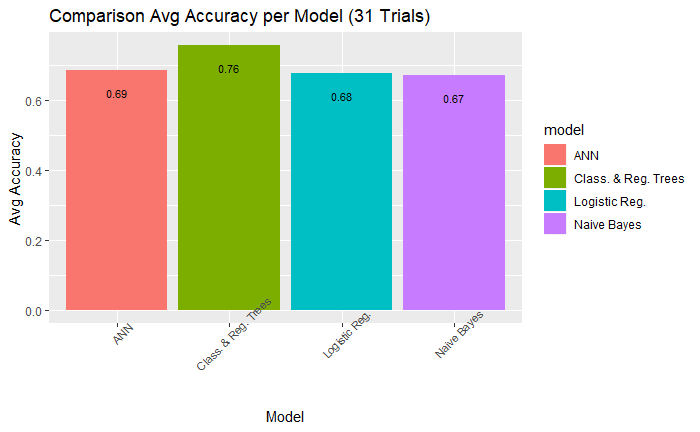
Is the size and the seed for configuration of the Artificial Neural Network model a variable to consider for optimizing the model?

To answer this question, a function was developed to get different accuracy values for different values of sizes (hidden layers) and seeds.

Anova was run for these values and the pvalue obtained is equal to 0.0263, less than a significance level of 0.05. The conclusion is Pvalue lower than alpha, there is significant difference for the Size, let's find out if this difference is enough to define the best size to work with or the worst size the work with.

According to Tukey results the only pair with a pvalue less than significance level is the pair size (7-3), but to define the number of hidden layers equal to 7 as worse than the model with 3 hidden layers. There is no significant difference between size equal 3 and the rest of the models with size different than 3 except 7. It can't be confirmed statistically that size impacts on the model. However, let's set the size equal 3 for the comparison with the rest of the models.

Going back to the 31 different values of accuracy obtained for each model changing the seed, the objective is to confirm if there is a significant difference among the four model as a result.



*Figure 11. Comparison Accuracy per Model (31 Trials)*

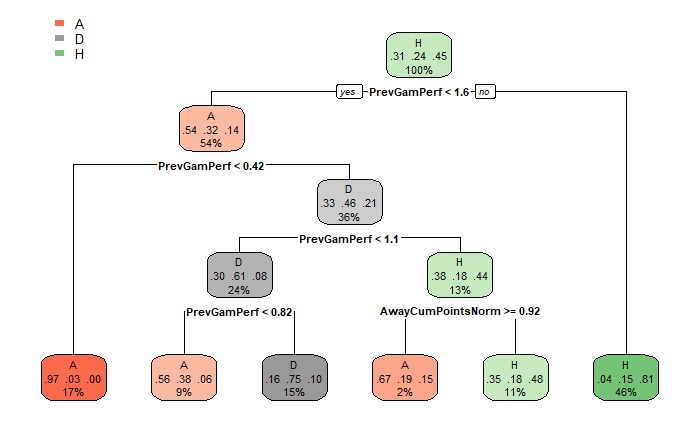
This is the main question of the project. To have it answered it, the approach is to run Anova and comparing the pvalue with a significance level of 0.05. The pvalue resulting for this analysis is <2e-16, less than 0.05, which means that there is significant difference between at least one model and the rest.

Which is the model or models significantly different? The answer is explained through the results of Tukey test. According to this test. Classification and Regression Trees Model provides an accuracy in the prediction greater than the rest of the models

|  |  |
| --- | --- |
| Model | p adj |
| Class. & Reg. Trees-ANN | 0 |
| Logistic Reg.-ANN | 0.510625 |
| Naive Bayes-ANN | 0.032345 |
| Logistic Reg.-Class. & Reg. Trees | 0 |
| Naive Bayes-Class. & Reg. Trees | 0 |
| Naive Bayes-Logistic Reg | 0.510625 |

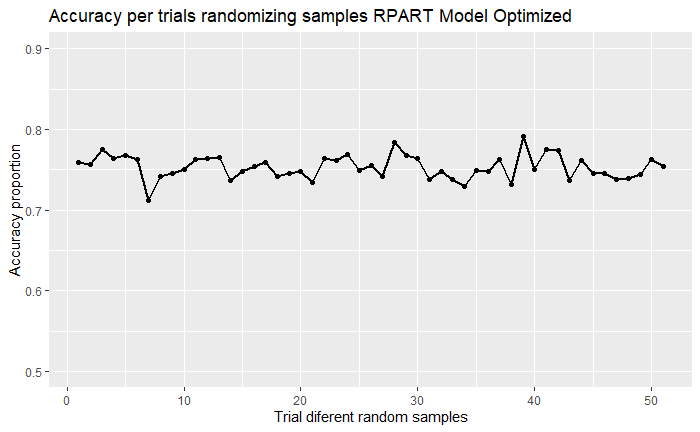
*Table 3. Tukey Results for prediction Models*

It could be concluded that the best model to predict the outcome for an EPL match is the Classification and Regression Trees with an accuracy of 76%. In the graph below the model chosen is plotted



*Figure 11. RPART Tree Model*

For all the models were used more than 10 variables. According to the previous graph the variables impacting the model are only 2 (PrevGamPerf and AwayCumPointsNorm). The optimization of the model is done by erasing all the variables that are not used, this allows to avoid overfitting, and this way the model is simplified. Once the function is run again based on the two variables left the accuracy obtained is similar to the previous model involving all the variables.



*Figure 12. Accuracy per trials randomizing samples RPART Model Optimized*

VIII. CONCLUSIONS

The values of accuracy were impacted, in almost 20%, b*y* using the previous outcomes between matches for the same teams playing the match to be predicted.

The Tukey test made to the several accuracy values, obtained from different parameters of sizes (hidden layers) and seeds, confirmed that there is not significance difference in the accuracy according to the structure of the ANN model.

Running Anova for different values of accuracy obtained for each model and changing the seed, allowed to confirm that there is significant difference between at least one model and the rest.

According to Tukey test, it could be confirmed that the best model, in terms of accuracy, is the Classification and Regression Trees model with a significant difference statistically proven with the rest of the models.

Plotting the Classfication and Regression Trees, was confirmed that the model is using only two variables to make the prediction. The rest of the variables were removed with the objective of simplify the model, and same accuracy was obtained.

For future extensions of this study is recommended to use the economical movements that teams do in summer and winter, this variable is suspected to impact in the first five round of each season.

##### References

1. Zachary Andrews. Comparing Predictive Models For English Premier League Games. Submitted to the Graduate School Appalachian State University in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE. May 2019. Department of Computer Science
2. Enes Eryarsoy and Dursun Delen. Predicting the Outcome of a Football Game: A Comparative Analysis of Single and Ensemble Analytics Methods. Published online by Proceedings | 2019
3. Jay Boice. How Our Club Soccer Predictions Work., a computational journalist for FiveThirtyEight. LAST EDIT SEP. 13, 2018
4. Nazim Razali, Aida Mustapha, Faiz Ahmad Yatim, Ruhaya Ab Aziz. Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL). IOP Conf. Series: Materials Science and Engineering 226 (2017)
5. Ryan H. Boyko, Adam R. Boyko & Mark G. Boyko. Referee bias contributes to home advantage in English Premiership football. Published online: 19 Jul 2007
6. Syed Tabrez Hassan An analysis of strikers ranking in Barclays premier league football using multi-criterion decision making approach. International Journal of Yogic, Human Movement and Sports Sciences 2018.
7. Benjamin Z. Alamar. Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers. Columbia University Press. 2014 Edition.