Prediction of Rank of a Football Team in a Season Based on Machine Learning and Statistics

(Sumit Sunil Khopkar, Student Number – 21261643)

(Pradnyesh Bhalange, Student Number – 21262382)

***Abstract* - Machine Learning and Statistics have been used in various aspects of sports one of which the most popular one is the prediction of the outcome of a match. Based on relevant historical data, a prediction can be devised using this data and Machine Learning or Statistical methods. In previous studies, the focus was largely upon predicting the outcome of a match. However, in our study, we are predicting what the rank of the team would be in its corresponding league for that particular year based on certain attributes which we obtain from our dataset. We have performed Random Forest Regression to predict the ranks of the football teams.**

***Keywords –* Seaborn, Python, R, sklearn, Random Forest Regressor, Accuracy, Ranking, Correlation, Mean, ANOVA, Prediction**

1. INTRODUCTION

The game of football has three basic high-level outcomes after it is played out. These are either a win, a draw, or a loss for a team. Sports analytics is a growing field in the field of analytics and analytics football is one such example. Finding the result of a match is a subject of interest in the research community. The motivation behind this is not just because of curiosity but also two important factors which are –

1. For football experts to analyze a team
2. Betting purposes [1]

The betting market in football quantifies to around $700 billion to $1 trillion by a BBC article [2]. Our study takes a different approach to the prediction side since we do not intend to predict the outcome of one match, rather we intend to predict at what rank will the team finish its season after the conclusion of the season. This we are calculating by first predicting the number of wins all the teams will have in that particular season and then ranking them accordingly. This we are performing using Random Forest Regressor model upon the data and its features selected based on the Exploratory Data Analysis that we have performed. We have also performed Hypothesis Testing to confirm the dependency of the number of wins on our selected attributes.

1. RELATED WORK

We reviewed previous works to identify which methods were used to perform similar predictions. All these related works predicted the outcome of a match that is only slightly different from our study. LSTM has been used to find out the result of a match on the assumption that old matches results, player attributes, and the impact of those players’ attributes would prove to be useful for the final prediction [1], [3]. Although, the results weren’t significant. Or at least the use of Deep Learning wasn’t too useful. This we can confirm by comparing the results of these studies with other two studies which made use of Machine Learning methods such as Gaussian Naïve Bayes, Support Vector Machine, Random Forest, and Gradient Boosting [2], [4]. The kernels of SVM namely Linear SVM, as well as RBF SVM, were tried [2]. After looking at all these there were three main conclusions drawn out that data is too old is insignificant, a model trained on only one type of league generates better results but overfits the model, and Deep Learning methods have no evident edge over Statistical and Machine Learning methods. We used these related works to decide the methods we should use to perform our prediction.

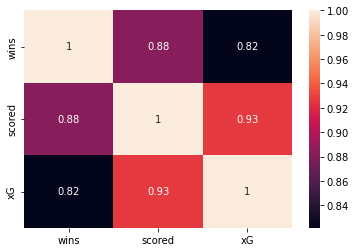
1. DATASET AND EXPLORATORY ANALYSIS

***Dataset* –**

Based on our analysis of the related works we proceeded with our problem statement to determine the best dataset that we can use. We downloaded our dataset from Kaggle. The dataset consists of two CSV files. One file has the aggregated data for an entire season for every league and the other file has data for every match. We decided to use the former one because we need to predict the ranks predicted using the data of the final ranks of teams. This CSV file consists of 684 rows and 24 columns. The dataset has features such as season, league, team, the rank of the team, total matches played, total matches won, drawn, lost, and other features which are crucial statistics of a game such as total goals scored, total expected goals (xG) and others. This data can be grouped based on leagues and seasons to obtain the rankings of the teams of any particular league of one season. The seasons range from 2014 to 2019. We haven’t included any old data based on our analysis of related works. But we also decided to use a dataset of different leagues since we didn’t wish to overfit the model on just one league and that our model should work as a generalized predictor. Our dataset also includes the number of matches because the number of matches in one season is not the same for every league. Thus, it can be observed that we have tried to make a very generalized model.

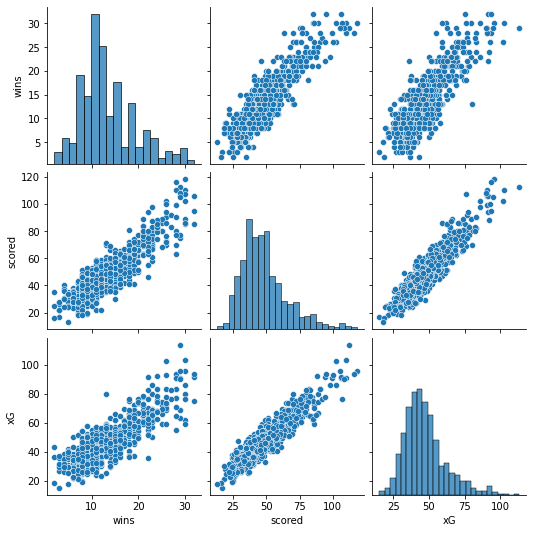
***Exploratory Data Analysis* –**

Our first step in Exploratory Data Analysis was to check for any null values. We found out that there were none. We dropped the features 'pts', 'xG\_diff', 'xGA\_diff', 'xpts\_diff', 'npxG', 'npxGA', 'npxGD' because points and rank of the team are redundant and similarly the other attributes are differences between actual and expected values of goals, goals against and so on. This difference is again redundant since we do have the values of actual goals and other respective attributes. We then drew a correlation matrix using the seaborn library in Python of all the remaining attributes. As a result of poor correlation, we dropped the features 'draws', 'loses', 'missed', 'xGA', 'ppda\_coef', 'deep\_allowed', 'oppda\_coef'. Thus, there was a good correlation only between the features that focused on scoring goals and winning matches. We ended up with the following features and visualized its correlation matrix. All these features have a good correlation.



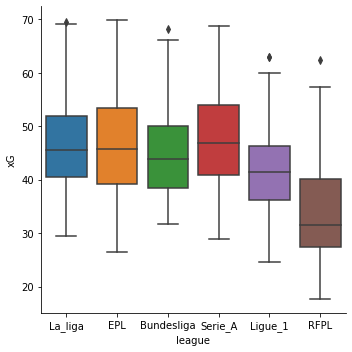
**Fig. 1 – Correlation Matrix after Removal of Features**

After this, we plotted a Pair Plot to further analyze the distribution of these features. In that, we realized that the feature ‘deep’ which represents how many passes were played in the final 20 yards of its opponent’s goal was scrapped out because of its wide distribution in the scatter plot.



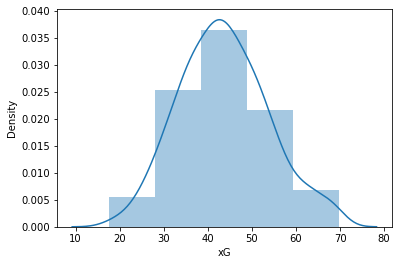
**Fig. 2 – Pairplot of features after removal of ‘Deep’ feature**

We then drew a boxplot for every league for its xG (Expected Goals) and noticed outliers. So, we removed those outliers and drew the boxplot again.



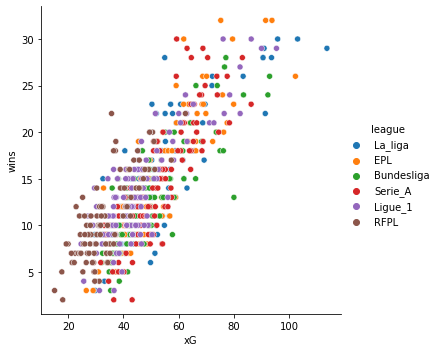
**Fig. 3 – Boxplot of leagues against xG after removal of outliers**

We observed that xG was higher in the ‘La Liga’ league but after removing outliers ‘Serie A’ had the highest mean. Thus, xG could also be related to the league.



**Fig. 4 – Histogram of xG**

We also made a histogram of xG to check whether it is normally distributed. And our final visualization was to draw our hypothesis. This is a scatter plot for xG against the number of wins for every league.



**Fig. 4 – Scatter Plot for xG Wins for Every League.**

1. HYPOTHESES AND RESEARCH QUESTION

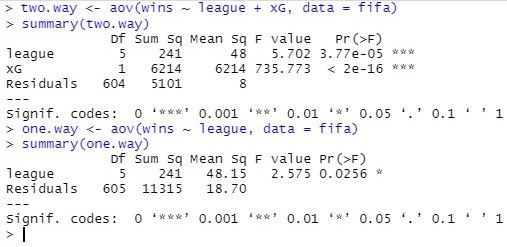
On exploring the dataset and visualizing it, we could observe a relation between the xG, the league in which the team played in and the number of wins. There can be a case that a team could have scored a different number of goals (or have a different xG) and finished at a different rank if the team was playing in a different league. This gave rise to our hypothesis testing statement. We assume that there is no relation between the leagues, xG, and the number of wins. Thus, we form our hypothesis statements –

**Null hypothesis () – There is no relation between xG, league, and the number of wins because of these.**

**Alternative hypothesis () – There is a relation between xG, league and, the number of wins because of these.**

Since in our case we consider a numerical (xG) and a categorical variable (league) as our independent variables and a numerical variable (wins) as a dependent variable we performed our hypothesis testing using ANOVA.

We calculated the value using the R language. After calculation, we obtained the following results –



Thus, in both cases of Two-Way and One-Way ANOVA, the P-Value is below 0.05. Thus, our Null Hypothesis can be rejected. We then use this assertion to include both the ‘league’ and ‘xG’ features for the prediction since there is a dependency between the number of wins and these features.

***Research Question –***

Can the rank of a team in a season be predicted based on historical data containing goal-scoring features and the league to which a team belongs?

1. METHODS USED AND WHY

After knowing the results of our hypothesis we went forward with our chosen features to perform a prediction using the Machine Learning model ‘Random Forest Regressor’. Our data consisted of the independent variables – league, team, matches, scored, and xG. Our prediction feature was the number of wins. In these features, league and team are categorical features. Hence, we used ‘One Hot Encoder’ to encode the data into N number of columns since N 2. We did the same for the teams as well. Doing this increased the number of features to 180 columns. Random Forest Algorithm is a good algorithm for performing predictions on high-dimensional data. This is only in the case of the number of predictors () being less than the number of observations () [5].

This is true in this case. The choice of using Random Forest was also because of the desirable results obtained in previous studies which used Random forest [2], [4].

We used the library sklearn to build the Random Forest Regressor model and to predict. We also used this library to preprocess our data. The data was split into training data and testing data with a test ratio of 80:20 respectively. The Random Forest is trained with n\_estimators = 10.

1. RESULTS AND FINDINGS

Our test data accuracy was low on the accuracy metric of Jaccard Score which calculates the exact number of matches. But our approach to the prediction was different anyway. Our purpose was to find out the rank of the team.

For this purpose, we created a new test dataset that comprised all the teams from the ‘La Liga’ league for the season 2014. We then ran this data through our prediction model and then based on the number of wins we sorted the rank of the teams. This gave us a 60% accuracy for the same metric of Jaccard score. Thus, we could successfully predict the ranks of 12 teams out of a total of 20 teams for the 2014 season of the La Liga league. Following are the results that we obtained –

|  |  |  |
| --- | --- | --- |
| **Actual vs Predicted League Table for 2014 Season of La Liga** | | |
| **Rank** | **Actual Team** | **Predicted team** |
| 1 | Barcelona | Real Madrid |
| 2 | Real Madrid | Barcelona |
| 3 | Atletico Madrid | Atletico Madrid |
| 4 | Valencia | Valencia |
| 5 | Sevilla | Sevilla |
| 6 | Villarreal | Villarreal |
| 7 | Athletic Club | Athletic Club |
| 8 | Celta Vigo | Celta Vigo |
| 9 | Malaga | Rayo Vallecano |
| 10 | Rayo Vallecano | Espanyol |
| 11 | Espanyol | Malaga |
| 12 | Real Sociedad | Real Sociedad |
| 13 | Elche | Elche |
| 14 | Getafe | Getafe |
| 15 | Levante | Levante |
| 16 | Deportivo La Coruna | Almeria |
| 17 | Granada | Deportivo La Coruna |
| 18 | Eibar | Eibar |
| 19 | Almeria | Granada |
| 20 | Cordoba | Cordoba |

Thus, it can be observed that there aren’t any drastic misplacements for any team in the table.

1. CONCLUSION

We can conclude from our above observations that the Random Forest Regressor Algorithm with the right selection of features can produce above-average results. These predictions can be used by the teams to determine if they need to improve in their area of attacking. We can infer from the above results that ‘Real Madrid’, even after being predicted as the winner did not win the league. This means that they would have to concentrate on areas other than attack.

Although our results are above average there need to be improvements made to the model selection. If we would have included the other features, would it have improved our accuracy even if they had a low correlation? This remains an area to explore for future work. We can also make a model to predict every match. The data for our current problem statement wasn’t much which could have affected the accuracy as well.

1. REFERENCES

[1] N. Danisik, P. Lacko, and M. Farkas, “Football match prediction using players attributes,” *DISA 2018 - IEEE World Symp. Digit. Intell. Syst. Mach. Proc.*, pp. 201–206, 2018, doi: 10.1109/DISA.2018.8490613.

[2] R. Baboota and H. Kaur, “Predictive analysis and modelling football results using machine learning approach for English Premier League,” *Int. J. Forecast.*, vol. 35, no. 2, pp. 741–755, 2019, doi: 10.1016/j.ijforecast.2018.01.003.

[3] M. A. Rahman, “A deep learning framework for football match prediction,” *SN Appl. Sci.*, vol. 2, no. 2, pp. 1–12, 2020, doi: 10.1007/s42452-019-1821-5.

[4] Q. Zhang, H. Z. Xu, L. Q. Zhou, and L. Wei, “Prediction of football match results based on model fusion,” *PervasiveHealth Pervasive Comput. Technol. Healthc.*, vol. Part F1481, pp. 198–202, 2019, doi: 10.1145/3319921.3319969.

[5] L. Capitaine, R. Genuer, and R. Thiébaut, “Random forests for high-dimensional longitudinal data arXiv : 1901 . 11279v1 [ stat . ME ] 31 Jan 2019,” pp. 1–22, 2019.