Predict Football Matches Result Using FIFA Players Statistics

Sylvia Le, Emanuel Eagle, Riah Malachi

Connecticut College

**Abstract**

Being able to accurately predict the results of a football match is important. Analysts, coaches and investors need reliable information to make decisions related to changes in strategy and possible future investments. Bettors need to have a better method than random guessing to increase their chances of winning, which can have a positive financial impact on a large scale. This study aims to build a machine learning model that can predict the result of a football match. Specifically, using the statistical data of the players and teams in a particular match, we wish to correctly classify the match result into three different possibilities. Two types of models were built, K-Nearest Neighbors and Decision Tree; each was cross-validated using either a validation set, 10-fold or both. The result shows a decent increase in accuracy over randomly guessing the result. The data and the result obtained from the models also suggest several issues and possible improvements for future research.

Predict Football Matches Result Using FIFA Players Statistics

Accurate and reliable prediction of the outcomes of sports games is of interest to many groups. Whether it be sports analysts or teams, many people could benefit from knowing what it takes to win. By creating models that can be predictive and interpretable we can benefit each of these groups in very important ways.

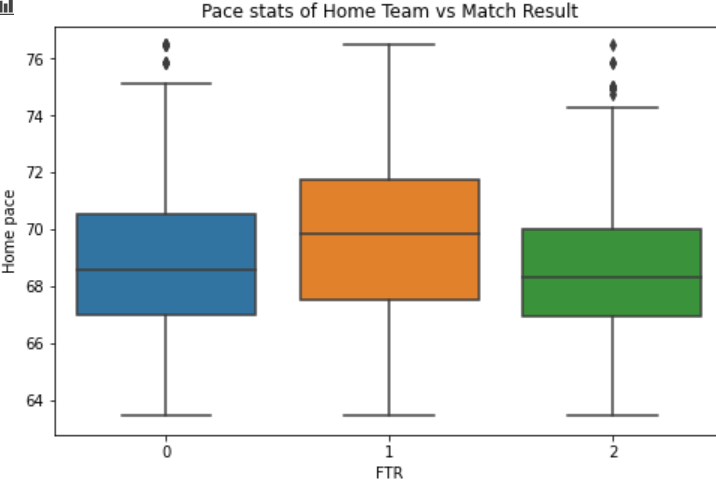
Analysts are commonly asked to predict the outcomes of games and are oftentimes wrong. The analysts who work for companies like ESPN have several data analysts working with them to determine who will win each game and why. The creation of accurate predictive models is very important for these organizations to maintain their status as the go-to place for sports knowledge. While they are asked to predict the outcomes of games, they also have to explain why, this is where interpretability comes into play. Many analysts you see on TV are not statisticians, models need to be explainable to these analysts, not only does it need to be explained to analysts but it needs to be explained to fans. The models need to be interpretable to the extent that analysts can explain which variables are most important for team success.

Professional teams are also very important stakeholders in creating more interpretable models. Teams are increasingly using data analytics to determine what they need to succeed against which team. However, while they use data for each match-up, they also would benefit from knowing in general, which variables are most correlated with winning games. For example, if a model finds that teams win more when they have better shooting skills, this has many benefits for the organization. First, if shooting skill is the most important variable for winning games, coaches can prioritize working on shooting in practice. Also, from an economic standpoint, general managers can focus their salary cap on players that are the best shooters.

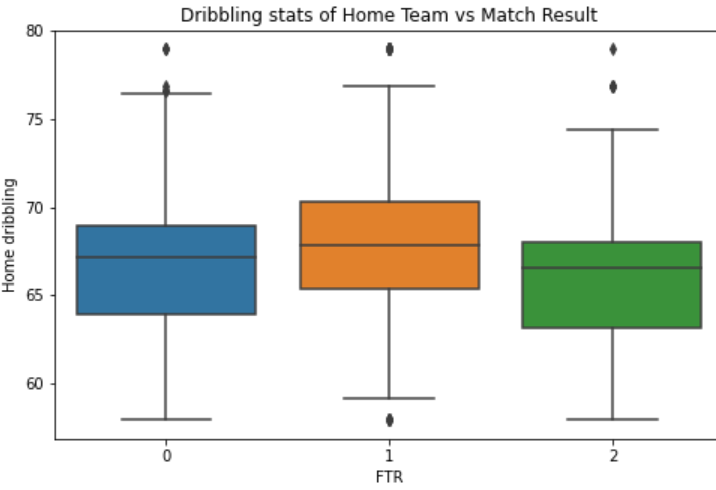
We found three related datasets on Kaggle for this project. There are two players datasets, which include FIFA players data in 2018 and 2019. There are over 70 columns, including the player’s position, attributes (Attacking, Skills, Defense, Mentality, GK Skills, etc), and personal data (Wage, Worth, Date of Birth, etc). Each row is a player, ordered by the overall score, descending from highest to lowest. A third dataset is a group of nine other datasets, which provides a total of nearly 3000 match results from major leagues in Europe in the 2018/2019 season. Fields included are number of goals scored (half-time and full time), match result, number of cards, and betting rates.

There were several key tasks we needed to perform in the data cleaning process. First of all, the players dataset has too many attributes, and we only wanted to focus on a small subset of attributes that are important for building the model. The next task was to merge the player table with the match result table and check for possible N/A values. There are several clubs with Latin names, which led to some font problems and resulted in name conflicts between the dataset. We manually fixed this in Excel. We also encoded the letter values that represent the match result as a numerical value: home team wins as 1, away team wins as 2, and draws as 0. Finally, we scale the X and Y data frame and split them to get the training and the test set.

Since we wanted to predict the match result using the players’ statistics, we chose two attributes, home pace, and home dribbling to construct a box plot graph between the attribute and the result. As expected, in cases where the home team wins, the median values of home team pace and home team dribbling are the highest; and if the home team loses, these values are the lowest. Both of the graphs show that there can be some exceptions: home teams with exceptionally high pace or dribbling stats can still lose a match or have resulted in a draw.



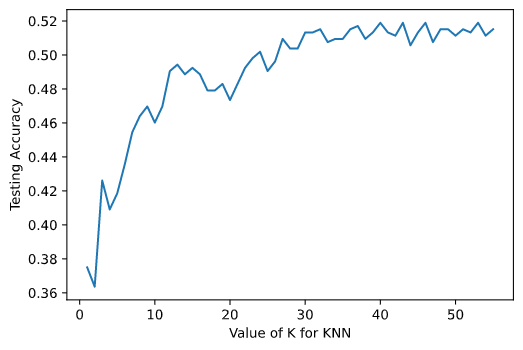
*Figure 1*. Mean pace of Home Team vs. Match result plot



*Figure 2*. Mean dribbling of Home Team vs. Match result plot

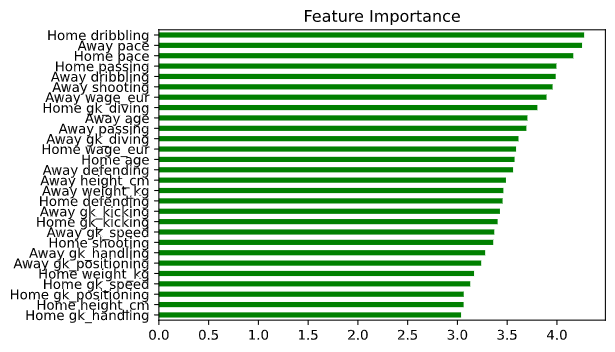
Since our question is a classification problem, we decided to build a K-Nearest Neighbors model and a Decision Tree Classification model. For the decision tree approach, we tried two ensemble methods, random forest and gradient boosting, as one method provides functionality that is more convenient in understanding the model.

For the K-Nearest Neighbor model, we used two cross-validation approaches. In the validation set approach, we looped through a range of K values and picked the K that returns the highest accuracy score. The graph of values of K versus testing accuracy is shown below. The optimal K, in this case, is 40. Using this value, we refitted the model. The accuracy score returned by this approach is 51.89%, performed on the original test set. This means the model can accurately predict the result of the match with the probability of 51.89%. In the second approach, we use grid search to pick the best value for K, in this case, 55, and then 10-fold cross-validation to get the mean accuracy score. It’s worth noting that the values might vary due to the random seed; for instance, using the random seed of 234, the mean accuracy score using this approach is 50.18%. The probability of accurately predict the match result is 50.18% in this case.



*Figure 3*. Value of K in K-Nearest Neighbors mode vs. Testing Accuracy plot

In the decision tree approach, we used two ensemble methods. In the first method, we built nine different Gradient Boosting Classification models with different learning rates ranging from 0.01 to 1 and chose the one with the highest accuracy score. We chose the learning rate of 0.01, which corresponds to an accuracy of 51.70%. This model accurately predicts the match result 51.70% of the time. In the other ensemble method, we built a Random Forest model with 500 trees, max number of features in considering the best split is 6, and use 10-fold cross-validation to validate. The accuracy of this model is 46.97%. We also obtained the feature importance plot of this model, which is shown in figure 4.



*Figure 4.* Feature importance graph

Since there are three possible results for a match, the probability of correctly guessing the result is 33.33%. We found that using our KNN model, we can correctly predict the match result 51.89% of the time. In our worst model, Random Forest Classifier, the accuracy is 46.97%, which is still better than random guessing. This result is close to the level of accuracy achieved by a previous study, which produced 51.1% accuracy (Herbinet, 2018). In addition, due to the feature importance graph, we can see that the difference between the attributes’ importance is not noticeably significant, but the overall trend is that attributes related to goalkeeping are not as important.

One of the big drawbacks of this study is that it only took into account the players’ and team’s statistics. In reality, the result of a match is affected by more factors than just the people on the field. For instance, the performance of a team on a particular match may be greatly affected by its previous performance in another match, due to changes in players or strategy. Moreover, many studies have suggested that the presence of audiences on the field has a reciprocal impact on the performance of athletes. This suggests that the model might not work as well given the data from the year 2020/2021, because due to the COVID pandemic, many sports events are carried out without audience participation. Even without the global pandemic, we believe that the accuracy of the model can increase if there can be data on audience participation at football matches.

Another problem we observed from the data analysis step is that the players datasets are greatly skewed, especially the wage and player worth attributes. There is only a very small set of players with extremely high wages, while most other players’ wages cluster in the very lower end of the scale, regardless of their skills. There are players with exceptional skills set but a rather low salary. This phenomenon not only suggests that we need to come up with better methods to handle this skewness but also points at the reality that there exist great inequalities in giving the footballers what they deserve. This is also an important factor driving the result of a football match. Further research in the future can look at how this extreme wage distribution might affect the chance of winning, and possible methods to preserve that probability while reducing the pay gap.

Finally, the feature importance graph also suggests that it might be beneficial to look at the preprocessing step again. The gk-attributes are specified only if the player is a goalkeeper, hence the gk-attribute columns include quite a lot of N/A values, which could have been better taken into account. This could alter their importance in the feature importance ranking and further increase the accuracy of the models.

**References**

Gambling Sites. “Sports Betting Research & Analysis - How Important Is It?” *GamblingSites.ORG*, 3 Dec. 2019, www.gamblingsites.org/sports-betting/research-and-analysis/.

Palmer, Catherine. *Sports Betting Research: Literature Review*. www.communities.tas.gov.au/\_\_data/assets/pdf\_file/0025/43945/Sports\_Betting\_Literature\_Review.pdf.

Sichel, Adam Robert. *The Economic and Social Effects of Legalized Sports Gambling*. 2020, kb.osu.edu/bitstream/handle/1811/91813/1/Adam\_Sichel\_HonorsThesis.pdf.

Wann, D. L., & Hackathorn, J. (2019). Audience effects in sport: The reciprocal flow of influence between athletes and spectators. In M. H. Anshel, T. A. Petrie, & J. A. Steinfeldt (Eds.), APA handbooks in psychology series. APA handbook of sport and exercise psychology, Vol. 1. Sport psychology (p. 469–488). American Psychological Association.

Herbinet, Corentin. *Predicting Football Results Using Machine Learning Techniques*.

Department of Computing, Imperial College of Science, Technology and Medicine, June 2018, www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf.