Enhancing Football Match Predictions through AI and Machine Learning in the English Premier League

Langdon Huynh

Received April 12, 2024
Accepted July 31, 2024
Electronic access August 31, 2024

Football is undoubtedly the most popular sport in the entire world. With over 3.5 billion worldwide fans, the sports betting market for football or soccer is enormous. The FIFA World Cup is the most bet-on sporting event in the world. In this quadrennial event each match on average can generate over 2 billion euros in bets wagered. The sport is also very unpredictable as top ranked teams could lose to the lowest ranked teams despite being greatly unfavored. Additionally, results are very unpredictable as unexpected injuries, individual player performances, or even referee ruling may affect the outcome. As there is unpredictability and room for human error in predicting outcomes for football matches, it is vital to have a reliable tool to predict these results with high precision to make smarter and calculated decisions. This paper introduces three Machine Learning models - Neural Networks (NN), Multi-layer Perceptron classifier (MLP), Decision Tree Classifier (DTC), and Random Forest Classifier (RTC) - to enhance the precision of predicting football match outcomes, aiding smarter decisions in sports analytics. AI and machine learning models are particularly suited for this task as they can capture complex dependencies and patterns in the data that traditional statistical methods may miss. Applying these models to the 2021-2022 English Premier achieved an accuracy of 61.54% which surpasses the baseline of random guessing, which stands at 33% due to the three possible outcomes— win, loss, or draw.

Keywords: Football match prediction, machine learning, neural networks, decision tree, random forest, English Premier League, sports analytics.

Introduction

The sports betting industry is one of the largest markets in the world with a size of about $66.8 billion, as of late 2020 and is expected to rise to $106.25 billion by 2025[1]. With football being the most popular sport across the globe, its market is tremendous. Arguably the most competitive football league and most popular for football betting is the English Premier League (EPL). When examining the most money wagered on European football, the bets on the EPL generated over 68.5 billion euros worldwide in the 2021/2022 season[2]. Predicting soccer game outcomes is very important in the realm of sport analytics. These predictions influence various aspects of the game, spanning from fan engagement to betting markets and even team strategies. In football betting there are many parameters that could be factored to increase the odds of making a calculated decision such as historical match data, home and away performance, player injuries, historical performance between the two teams, and current rankings of both teams. With so many factors affecting the outcome of a football match it is difficult to make a consistent accurate decision when predicting. This leads to the question of how can we predict football match outcomes accurately and consistently despite the uncertainty and randomness?

The integration of Artificial Intelligence (AI) and Machine Learning (ML) can provide more accurate predictions for football match outcomes by analyzing a plethora of historical match data and its trends. The objective of this paper is to create a predictive model for football match outcomes, by using AI techniques, and historical data. I primarily focused on the English Premier League as it is one of the most prominent and exciting European football leagues.

Recent advancements in artificial intelligence (AI) and machine learning (ML) have significantly enhanced the accuracy of predictions in football match outcomes. The use of historical data to model team strength and identify trends has proven effective in improving prediction precision. For example, platforms like Kickoff AI[3] have demonstrated that dynamic modeling of team strength can substantially increase the accuracy of football match predictions. This approach leverages a combination of historical data and real-time performance metrics, showcasing the potential of AI and ML in sports analytics.

In academic research, various studies have explored the application of AI and ML in sports predictions. Notably, Fialhoa, Manhães, and Teixeira[4] employed machine learning models such as Random Forest and Gradient Boosting to predict football match results, illustrating how these ensemble methods can uncover complex patterns in data that traditional statistical methods might overlook. Additionally, research by Ulmer and

Fernandez[5] from Stanford University focused on predicting English Premier League soccer matches using multiple classifiers, including Linear Stochastic Gradient Descent, Naive Bayes, Hidden Markov Models, Support Vector Machines, and Random Forest. Their study highlighted the comparative performance of these methods and emphasized the effectiveness of Random Forest and Support Vector Machines in handling the intricacies of match predictions.

These developments underscore the growing importance of AI and ML in sports analytics. The ability to analyze large datasets and identify intricate patterns allows for more accurate predictions, which is crucial for teams, coaches, and analysts aiming to gain a competitive advantage. As research continues to evolve, the integration of advanced AI and ML techniques promises to further enhance the precision and relevance of sports predictions.

This research seeks to address the specific gap of predicting football match outcomes using only recent forms and match data, excluding other external factors. This approach aims to simplify the model while still providing accurate predictions, understanding the limitations and strengths of this method. The exclusion of factors such as player injuries and weather conditions was primarily due to the lack of access to such complex datasets.

By training models on historical matches from the EPL I created 3 different models that predict whether a match would be a win, loss, or draw. The three models I used were the Multi-layer Perceptron classifier (MLP), Decision Tree Classifier (DTC), and Random Forest Classifier (RTC). I trained each model with the match data from the English Premier League 2021/2022 season and created a running standings table to simulate the fluctuating rankings in a normal English Premier League season. Each model predicts a match outcome to be either a win, loss, or draw.

Results

During testing, this study utilized scikit-learn library models to classify matches as either a win, loss, or draw[6]. Each model was trained on a randomized set of matches from the season and was tested on the rest of the matches from that season. To calculate the prediction accuracy of each model I compared the list of predicted match outcomes of a specific model to the actual EPL 2021/2022 match outcomes. If a model predicted the right outcome for a match its number of correct predictions would increase by one and after all the matches were predicted the number of correct predictions would be divided by the total number of matches to achieve a prediction accuracy. In addition to assessing the overall match outcomes, I evaluated each model's performance in predicting wins, draws, and away outcomes. This involved calculating the accuracy of each model's predictions by comparing the number of correctly predicted Wins, draws, and away results against the actual occurrences. The best results for each category are shown in Table 1. To mitigate overfitting, this model doesn't solely rely on the most probable outcome, such as a home win. Instead, it considers all the potential outcomes which ensures a balanced approach.

Model	Overall Accuracy	Wins Accuracy	Draws Accuracy	Away Accuracy
MLP	57.69%	62.50%	42.00%	50.00%
RFC	61.54%	66.67%	36.64%	57.14%
DTC	46.15%	62.50%	33.33%	44.44%

Table 1 Overall results on test set

Table 1 illustrates the overall accuracy of the general performance of each model across all outcomes. The RFC model achieved the highest overall accuracy at 61.54%, followed by the MLP model at 57.69%, and the DTC (model with the lowest accuracy at 46.15%. This suggests that, on average, the RFC model performed relatively better in predicting match outcomes compared to the other models.

Model	**Metric**	**Lower Bound (%)**	**Upper Bound (%)**
MLP	Overall Accuracy	55.22	60.16
	Wins Accuracy	60.78	64.22
	Draws Accuracy	39.84	44.16
	Away Accuracy	48.35	51.65
RFC	Overall Accuracy	58.73	64.35
	Wins Accuracy	64.33	68.01
	Draws Accuracy	34.59	38.69
	Away Accuracy	55.12	59.16
DTC	Overall Accuracy	44.12	48.18
	Wins Accuracy	59.80	65.20
	Draws Accuracy	30.45	36.21
	Away Accuracy	42.13	46.75

Table 2 Confidence intervals for each model

Table 2 displays the confidence intervals for each model's performance metrics, offering insights into their accuracy and reliability. Confidence intervals are crucial as they indicate the range within which the true model performance metrics are likely to fall, offering a measure of the uncertainty associated with each model's predictions. Specifically, these intervals are calculated with a 95% confidence level, meaning that if the process of generating these intervals were repeated many times, approximately 95% of the intervals would contain the true performance metric.

Overall Accuracy: The Random Forest Classifier (RFC) exhibits the highest overall accuracy with a narrow confidence interval of 58.73% to 64.35%, indicating consistent and reliable

predictions. The Multilayer Perceptron (MLP) follows with an accuracy interval of 55.22% to 60.16%, showing strong performance but slightly less certainty. The Decision Tree Classifier (DTC) has a broader interval of 44.12% to 48.18%, reflecting greater variability.

Wins Accuracy: RFC excels in predicting wins, with intervals of 64.33% to 68.01%, showing high precision. MLP also performs well, with intervals of 60.78% to 64.22%. DTC's intervals of 59.80% to 65.20% indicate good performance but with more variability.

Draws Accuracy: The MLP model provides the most reliable draw predictions (39.84% to 44.16%), while RFC's intervals (34.59% to 38.69%) are narrower but less consistent. DTC shows the highest uncertainty with intervals of 30.45% to 36.21%.

Away Accuracy: RFC again leads with intervals of 55.12% to 59.16%, reflecting stable performance. MLP's intervals (48.35% to 51.65%) and DTC's (42.13% to 46.75%) are broader, indicating less reliability.

In summary, RFC is the most reliable model across all metrics, with narrower confidence intervals suggesting high accuracy and stability. MLP shows strong performance, particularly in predicting draws and wins, while DTC, although effective, has broader intervals, indicating more variability. The 95% confidence intervals provide a reliable estimate of model performance, ensuring that the observed metrics are both statistically significant and indicative of true performance.

MLP Outcome	**Accuracy**
Overall Accuracy	57.69%
Wins Accuracy	62.50%
Draws Accuracy	42.00%
Away Accuracy	50.00%

Table 3 MLP test results

1. **Multi-layer Perceptron classifier (MLP):** The MLP model demonstrated an overall accuracy of 57.69%. Additionally, it achieved the highest Draws accuracy at 42.00% and 50.00% and 62.50% for Away and Wins accuracy respectively.

RFC Outcome	Accuracy
Overall Accuracy	61.54%
Wins Accuracy	66.67%
Draws Accuracy	36.64%
Away Accuracy	57.14%

Table 4 RFC test results

2. **Random Forest Classifier (RFC):** The RFC model showed the highest accuracy overall with 61.54%. It also achieved the highest Wins and Away accuracy with 66.67% and 57.14% respectively. It also came in second place for draws accuracy with 36.64%. This shows the RFC performed significantly better than the other 2 models.

DTC Outcome	Accuracy
Overall Accuracy	46.15%
Wins Accuracy	62.50%
Draws Accuracy	33.33%
Away Accuracy	44.44%

Table 5 DTC test results

3. **Decision Tree Classifier (DTC):** The DTC model seemed to have the lowest performance out of all the models with an overall accuracy of 46.15%. However, it achieved the same Wins accuracy of the MLP model at 62.50% but achieved a mediocre accuracy for Draws and Away Accuracy at 3.33% and 44.44%.

Discussion

The goal of this study was to create a model that accurately predicts football match outcomes to help clarify the randomness of sports betting and to create calculated decisions. I found that the Random Forest Classifier was the most successful at predicting the outcomes of matches with the highest overall performance with an overall prediction accuracy of 61.54%. If a normal person were to guess the outcome of a match, they would have a 33.33% chance of guessing either a win, draw, or loss exemplifying how this model can help sports bettors to make more informed decisions.

The success of the RFC model can be attributed to its inherent characteristics. Unlike simpler models, the RFC utilizes an ensemble of decision trees, each built on different subsets of the data. This approach allows it to capture a wide range of complex, non-linear relationships within the data. The nested "if-else" decision rules of each tree in the forest effectively address the intricate patterns and interactions between various factors influencing match outcomes. This capability enables RFC to provide more nuanced predictions compared to models like the MLP and DTC, which may struggle with such complexity.

Even though this model predicts more accurately than the common bettor, it is essential to understand that the model cannot predict each match completely accurately and it is vital to also use personal judgment when making a bet or decision.

Some limitations to this study could be using only one English Premier League season for both training and testing. By only using the 2021/2022 season I left out 31 seasons which could have helped to achieve a greater prediction accuracy. Another limitation would be the narrow scope of factors implemented.

Because I only used the match outcomes from the 2021/2022 season, I excluded factors such as player injuries, recent performance, and home and away performance. As soccer is an unpredictable sport, there could be many missing players due to yellow cards, red cards, injuries, or personal issues. For further research, new experiments could base off these limitations which may create a more robust accurate model for predicting match outcomes for football or maybe for other sports such as American football, baseball, and basketball.

Methods

In this study I used the historical match data from the English Premier League 2021/2022 season which I collected from Kaggle, an online data science community that provides datasets for machine learning [7]. First, I imported the Pandas and NumPy python libraries to help handle and preprocess the data. From the dataset, I extracted the names of the home and away teams and the outcomes of the 380 matches in an EPL season. Since the match outcomes came in non-numerical values (H, A, D abbreviated for "Home","Away", and "Draw"), I preprocessed the data by changing the non-numerical values to numerical values (0,1,2) corresponding respectively to be able to feed the data into the ML models.

Since the objective of this study is to predict the match outcome, I labeled the numerical match outcomes as "Y". Since the names of the EPL teams are not numerical, I initially planned on replacing each team name with their numerical ranking at the end of the 2021/2022 season. This static approach would mean the model would not capture the dynamic of football and the fluctuating performances throughout the season. To counter this problem, I simulated a running EPL standings table for the 2021/2022 season.

How an English Premier League table works is that teams receive 3 points for a win, 1 point for a draw, and 0 points for a loss. To simulate this table, I created a dictionary which assigned each team to their current number of points. Figure 1 illustrates this simulation. By traversing through the season, each team will receive the points depending on their current match performances and after every 10 matches the dictionary would be sorted from most points to least points. For all 380 matches in an EPL season, I created a 2d array with 380 rows and 2 columns where each element would be labeled with their current point standing and either home or away. I then labeled this 2d array as "X" as the teams playing in each match would be fed into the ML models to analyze trends and ultimately predict the match outcome. Additionally, I removed the first 20% of the EPL match fixtures to create a more accurate dataset. At the start of the season, the league standings are based on alphabetical order and the start of the season tends to have unpredictable results in comparison to the middle and end of the season where teams' performance and ranking are more predictable. Additionally, the first league table standings have no correlation with team performance as the season starts with the teams ordered in alphabetical order as every team starts with 0 points.

After preprocessing the data, I needed to split the dataset to train and test. After importing the sklearn train and test split model I applied X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.33) to split the dataset with 33% of the data as test and 67% of the data as train. After preprocessing and cleaning the data, I trained 3 different ML models using the X_train and Y_train. The three ML classifiers I used were MLP, DTC and RFC. The MLP, DTC, and RFC models were chosen due to several criteria. Initially, we started with simple models, such as linear models, but found they lacked the ability to capture the complex, non-linear relationships between inputs and outputs inherent in football match data. The MLP, DTC, and RFC models were chosen because of their ability to develop these non-linear relationships, making them more suitable for addressing the unique challenges of football match outcome prediction. Since the train performance was similar to the test performance, I did not have to develop methods such as dropout or regularization to prevent overfitting.

Given that the distribution of match outcomes (wins, losses, and draws) in the training and testing datasets was consistent, there was no need to apply advanced techniques like undersampling or oversampling. The natural balance within both datasets ensured that the proportions of each outcome type were preserved, allowing the models to learn effectively without the introduction of class weighting strategies or sampling adjustments. This approach maintained the integrity of the dataset and allowed for an unbiased evaluation of the models' performance across all outcome types.

To create the confidence intervals for each model's performance metrics, we used bootstrapping, facilitated by the sklearn library. Initially, we obtained predicted probabilities for each outcome using the predict_proba method from the MLP and RFC models, and decision rules from the Decision Tree Classifier (DTC).

Using the resample function from sklearn.utils, we generated multiple bootstrap samples from the test dataset. For each of these samples, we recalculated the performance metrics, such as accuracy for wins, draws, and away games. We then computed the 2.5th and 97.5th percentiles of these metrics to determine the lower and upper bounds of the 95% confidence intervals. This method provides a range within which the true performance metrics are likely to fall with 95% confidence, reflecting the variability and reliability of the models' predictions.

1. **Multi-layer Perceptron classifier (MLP)** is a type of artificial neural network that is widely used for classification and regression tasks. MLP consists of multiple layers of interconnected nodes or neurons that consist of an input

Date	Time	HomeTeam	AwayTeam	FTR
22/05/2022 3% 11/09/2021 2% Other (362) 95%	 2022-09-20 2022-09-20	20 unique values	20 unique values	H 43% A 34% Other (88) 23%
13/08/2021	20:00	Brentford	Arsenal	H
14/08/2021	12:30	Man United	Leeds	H
14/08/2021	15:00	Burnley	Brighton	A
14/08/2021	15:00	Chelsea	Crystal Palace	H
14/08/2021	15:00	Everton	Southampton	H
14/08/2021	15:00	Leicester	Wolves	H
14/08/2021	15:00	Watford	Aston Villa	H
14/08/2021	17:30	Norwich	Liverpool	A
15/08/2021	14:00	Newcastle	West Ham	A

Table 6: Snapshot of Kaggle Dataset of 2021-2022 English Premier League Match Outcomes

```
{'Liverpool': 3, 'Chelsea': 3, 'Tottenham': 3, 'Man United': 3, 'West Ham': 3, 'Leicester': 3, 'Brighton': 3, 'Bre
{'Liverpool': 6, 'Chelsea': 6, 'Tottenham': 6, 'West Ham': 6, 'Brighton': 6, 'Man United': 4, 'Brentford': 4, 'Eve
{'Tottenham': 9, 'Liverpool': 7, 'Chelsea': 7, 'West Ham': 7, 'Man United': 7, 'Everton': 7, 'Brighton': 6, 'Leice
{'Liverpool': 10, 'Chelsea': 10, 'Man United': 10, 'Everton': 10, 'Tottenham': 9, 'Brighton': 9, 'Man City': 9, 'W
{'Liverpool': 13, 'Chelsea': 13, 'Man United': 13, 'Brighton': 12, 'Everton': 10, 'Man City': 10, 'Tottenham': 9,
{'Liverpool': 14, 'Chelsea': 13, 'Man United': 13, 'Brighton': 13, 'Everton': 13, 'Man City': 13, 'West Ham': 11,
{'Chelsea': 16, 'Liverpool': 15, 'Man United': 14, 'Brighton': 14, 'Everton': 14, 'Man City': 14, 'Tottenham': 12,
{'Chelsea': 19, 'Liverpool': 18, 'Man City': 17, 'Brighton': 15, 'Tottenham': 15, 'Man United': 14, 'Everton': 14,
{'Chelsea': 22, 'Liverpool': 21, 'Man City': 20, 'West Ham': 17, 'Brighton': 15, 'Tottenham': 15, 'Man United': 14
{'Chelsea': 25, 'Liverpool': 22, 'Man City': 20, 'West Ham': 20, 'Man United': 17, 'Arsenal': 17, 'Brighton': 16,
{'Chelsea': 26, 'Man City': 23, 'West Ham': 23, 'Liverpool': 22, 'Arsenal': 20, 'Man United': 17, 'Brighton': 17,
{'Chelsea': 29, 'Man City': 26, 'Liverpool': 25, 'West Ham': 23, 'Arsenal': 20, 'Wolves': 19, 'Tottenham': 19, 'Ma
{'Chelsea': 30, 'Man City': 29, 'Liverpool': 28, 'West Ham': 23, 'Arsenal': 23, 'Wolves': 20, 'Tottenham': 19, 'Ma
{'Chelsea': 33, 'Man City': 32, 'Liverpool': 31, 'West Ham': 27, 'Arsenal': 23, 'Tottenham': 22, 'Wolves': 21, 'Ma
{'Man City': 35, 'Liverpool': 34, 'Chelsea': 33, 'West Ham': 27, 'Tottenham': 25, 'Man United': 24, 'Arsenal': 23,
{'Man City': 41, 'Liverpool': 37, 'Chelsea': 36, 'West Ham': 28, 'Man United': 27, 'Arsenal': 26, 'Tottenham': 25,
{'Man City': 47, 'Liverpool': 41, 'Chelsea': 38, 'Arsenal': 32, 'West Ham': 28, 'Man United': 27, 'Tottenham': 26,
{'Man City': 47, 'Liverpool': 41, 'Chelsea': 41, 'Arsenal': 35, 'West Ham': 31, 'Tottenham': 30, 'Man United': 28,
{'Man City': 53, 'Chelsea': 43, 'Liverpool': 42, 'Arsenal': 35, 'West Ham': 34, 'Tottenham': 33, 'Man United': 31,
```

Fig. 1 Simulated EPL standings table

layer, one or more hidden layers, and an output layer. Each neuron is connected with neurons with the next layer with wires or weights whose values are continuously updated through backwards propagation. Backwards propagation is the process where the network starts at the output layer through the hidden layers and to the input layer to update and refine the weights to achieve a minimized amount of error[8].

2. **Decision Tree Classifier (DTC)** is a machine learning algorithm used for classification and regression tasks. In comparison to RFC, DTC is just a singular tree which is composed of decision nodes which classify based on a specific feature of the data. A decision tree is composed of decision nodes, leaf nodes, and a root node. The root node encompasses the entire dataset and is where a specific feature is delegated. The decision nodes represent conditions

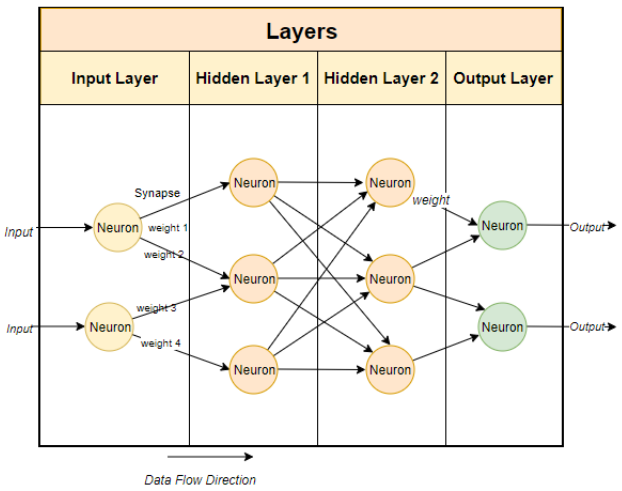

Fig. 2 Neural Network

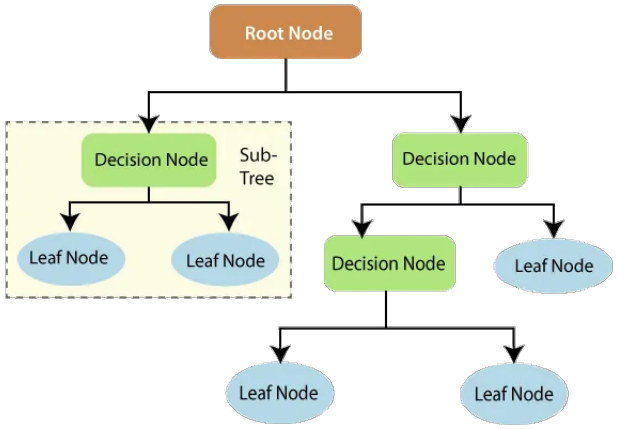

Fig. 3 Decision Tree

of a feature and pose a question on that feature which leads to branches or a split. The branches represent a possible outcome of the feature, and the leaf node is the final node where the classification or regression prediction is made[9].

3. **Random Forest Classifier (RFC)** is also a machine learning technique widely used for classification tasks. It is different from a MLP or DTC as it is an assembler learning algorithm that consists of many individual decision trees. Each decision tree is constructed based on the training data. Each tree is restricted based on the results from previous trees, so the trees produce different results12. For classification, the Random Forest combines the predictions of the individual decision trees through voting where the class that receives the most votes from the trees is the final prediction. In our study, the classification is either a win, loss,

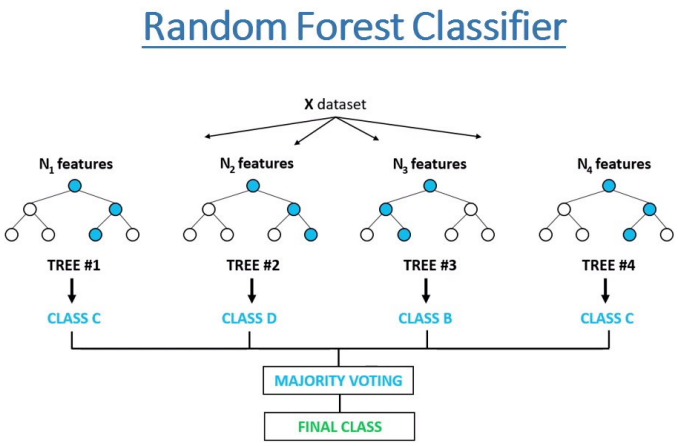

Fig. 4 Random Forest

or draw.

After training each model with X_train and Y_train I gathered the predictions for each respective model and compared the predictions to the actual results of the EPL 2021/2022 season.

References

1 *Breaking the Lines. Last modified May*, **9**, year.

2 S. R. Department, *Statista. Last modified September 7*.

3 L. Maystre and V. K. K. A. AI, `https://kickoff.ai/`., Accessed April 12, 2024.

4 G. Fialho, A. Manhães and J. P. Teixeira, *Procedia Computer Science*, **164**, 131–136.

5 B. Ulmer and M. Fernandez, *Predicting Soccer Match Results in the English Premier League*, School of Computer Science, Stanford University.

6 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg and J. Vanderplas, *the Journal of machine Learning research*, **12**, 2825–2830.

7 S. Uddin, *Kaggle*.

8 F. Malik, *Medium. Last modified May*, **18**, year.

9 F. Khushaktov, *Medium. Last modified August*, **26**, year.