Predicting Football Matches Outcome with Machine Learning

Conrado Boeira
Dalhousie University
conrado.boeira@dal.ca

***Abstract*—As football is seeing an increase in technology usage for tracking players, we can see a substantial increase in the number of statistics being collected in every match. This measurements can help teams and fans understand the values of certain players and try to predict outcomes of matches and cups. In this project, we have used historical data to try and predict the winner of a certain matches using machine learning methods.**

I. Introduction

Football, also know as Soccer, is the most popular sport throughout the world, with FIFA estimating that 265 million people practice it while more than 3 billion consider themselves as fans [1]. Moreover, major football events such as the quadrennial World Cup amass extremely large crowds to view it, with the 2018 Russia World Cup final achieving a number of viewers of 3.572 billion [2].

Having this in mind, it is not surprising the increased usage of technology in today's game. Leagues such as the British Premier League, have teams tracking their players every move on the pitch, collecting data and statistics in order to maximize the performance of their team and identify weaknesses and strong points in their opponents. Fans can also reap benefits from this modernization of the sport, as they can enhance their experience by using statics to compare and follow their favourite players and teams. Statistics like Expected Goals (xG), the probability of a shot resulting in a goal [3], are becoming more and more popular between fans and analysts, for example.

Therefore, in this project, we have worked on a method to try and predict the outcome of a match using historical data on both teams playing. This is the definition of the problem to be tackled, which correlates to the Business Understanding phase in the CRISP-DM methodology. Many factors need to be taken in consideration for this type of prediction, such as the form of both teams, which team is playing at home and current stand in the competition. This type of method could help coaches better understand which factors influence in their teams performance as well as help them prepare for big games.

This report is organized as follows: Section II outlines some of the papers that tackle the same problem, Section III describes the dataset to be used for this project, and Section IV defines the steps we took in order to solve this problem, from the pre-processing of data to the models chosen to solve it. Finally, in Section V, we present the obtained results.

II. Related Work

Multiple works have been published regarding this type of problem. Rahman [4] proposed a model for predicting the outcome of matches taking into consideration data such as the ranking of both teams, whether the match as a friendly or not and how many days the teams had between the current match and the previous one. The author developed a LSTM based model for predicting which team would win a match, achieving 70% accuracy.

Danisik et al. [5] also proposed a method for predicting match result but focusing on the specific players that compose the starting lineup for both teams. The authors used the players stats defined by EA developers for the football simulation game series FIFA as well as previous team results. This data was feed to an LSTM model, which achieved around 55% accuracy. However, it is hard to say how reliable the data on individual players taken from a game are.

Prasetio and Harlili [6] also proposed a model that take advantage of teams stats defined in the FIFA game. The authors used this information coupled with historical team performance to train a logistic regression model. They achieved accuracy results of 69.5%. However, they still fall in the same category as the previous paper as there is no guarantee of the quality of the data taken from the FIFA game.

III. Dataset

For this project, we decided to use the dataset provided in a Kaggle competition [7] for football match prediction. This dataset contains information regarding a match and previous results for the teams involved, and has the result (home team win, away team win or draw) as the target for prediction. Some of the columns included in it are as follows:

- Home and away team names
- League or cup name
- Date of the last match for both teams
- Number of goals scored in previous matches
- Rating of previous opponents

Moreover, this dataset consists of data on a large number of teams, playing multiple leagues throughout the world, including national teams, men's and woman's teams, and under 20 competitions. In Figure 1, we can see the distribution of data between the 10 most popular leagues in this dataset. Between these leagues, Premier League is the one with the most amount of matches recorded. However, if we look at

Number of matches recorded	110938
Number of different leagues	727
Number of different home teams	9813
Number of different away teams	9892
Oldest match recorded	2019-12-01
Newest match recorded	2021-05-01

TABLE I: Dataset statistics

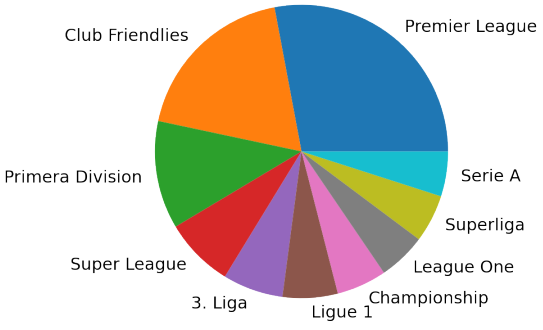

Fig. 1: Distribution of data between the 10 leagues with the most amount of data points

Figure 2, we can see that the distribution between leagues is fair, having the sum of the top 10 leagues correlating to less than one quarter of the whole dataset. Moreover, in Figure 3, we see that the dataset is slightly unbalanced, with the home team winning the match more often than tying or losing. More statistics regarding the dataset can also be seen in Table I. This step follow the Data Understanding phase in CRISP-DM.

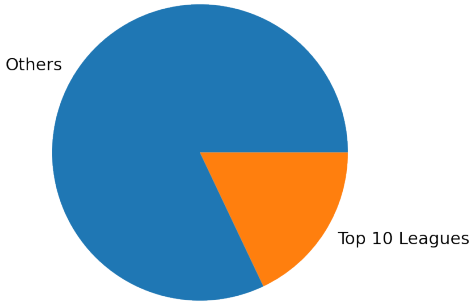

Fig. 2: Distribution of data between the top 10 leagues and the rest of the leagues in the dataset

IV. Methodology

The problem we intend to solve in this project is a supervised multi-class classification one. For this, we will perform a pre-processing step on the features, will explore the correlation between multiple variables and propose two models for prediction.

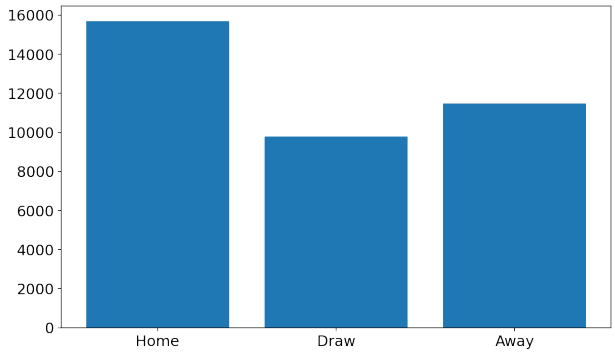

Fig. 3: Distribution between the target values

A github repository [1] has been allocated for all source files of this project, including the code for all the plots presented in this report.

A. *Feature pre-processing*

Before any attempt to create a prediction model, it is important to perform operations of Data Preparation, according to the CRISP-DM method. Having this in mind, an important first step to be performed while pre-processing the data is to drop some of the columns. The names of the teams is an example of features to be dropped as they would not add much information to the model. Moreover, features that act as identifiers or that specify if previous games are a cup or league match also can be safely dropped.

Another detail to take into consideration is the league for which the games are valid for. As we can see in Figure 1, many of the matches recorded are club friendlies. In this type of matches, the results can oscillate more, as clubs normally take these games more lightly, since there is no stakes in them. Therefore, it can be beneficial to remove matches that can fit in this type of category. Having this in mind, we have removed all friendlies, any league with age limit, such as U23, U20, and U19.

Moreover, we have also counted the amount of recorded matches per league. With this, we removed any league with under 100 recordings. This was done in order to avoid smaller leagues with less predictable results due to a lack of quality.

Also, many of the features presented on the dataset needed to be processed in order to get more significant metrics. Using some of the available solution for the Kaggle competition as inspiration [8], [9], we created the following extra variables:

Coach Changes: We have data regarding who was the coach for each team in their previous 10 games. This information by itself does not hold much value. However, from these points, we can derive if there was any changes in the coaching staff for either of the teams recently. This can be impactful in the match outcome as it is common for players to need some time to adapt to a new system.

Rest Days and Trips: Another set of features that initially don't hold much information are the dates the last 10 games

[1] https://github.com/conradoboeira/Football-Match-Results-Prediction

occurred as well as if these games were played at home or not. From this data, we will derive two new features: the average rest amount of rest days a team got in their latest fixtures and the amount of trips (either from their original city to the opposing team city or coming back from it) made in the most recent games. This features can help us understand how tired players were before a match.

Goal Differential and Attack and Defence Rates: Using the amount of goals scored and conceded in the most recent matches, we added goal differential which give us the difference between these two amounts. Also, by comparing to the average league values, we also added two features, attack and defence rate, to measure the relative strength of the team in both sides of the field. Attack ratio was defined as the ration between the average number of goals a team score by the goals per match league, and defence ratio follows the same formula but using the average conceded goals by a team.

Average Rating: Instead of individually evaluating the rating of all previous opponents, we also have merged these features into a single one representing the average rating of the recent opponents.

In Figure 4 we present the distribution graphs for these generated features as well as some other already present in the dataset. We can observe that many of the values follow a normal distribution. Using the attack and defence rates as an example, we can see that the values follow the expected pattern, a Gaussian distribution having the value 1 as it's mean, as it represents the average league offense/defence.

B. Correlation

Still inside the Data Understanding and Data Preparation steps of the CRISP-DM methodology, we looked at the correlation between different variables in the dataset. Figure 5 exhibits the obtained results. We can see how some features have strong association, such as the number of rest days for the home and away team, the attack/defence rate of a team to it's goal differential as well as to the average rating of it's opponents.

We also looked exclusively at the target variable, the result of the game, creating the graph seen in Figure 6. We can see that there is not a single defining feature that would help us easily define the winner of a game.

C. Models

For creating the predictor, we have decided to use 2 different types of models. First, a Random Forest (RF) model as it is a fast and accurate model, recurrently used in Machine Learning challenges in Kaggle, such as the one where the used dataset was collected from.

The second model created was a Feedforward Neural Network (FNN) . With this model, we have access to a wider range of configurations and can create a more customizable model to deal with the complicated relationships between features.

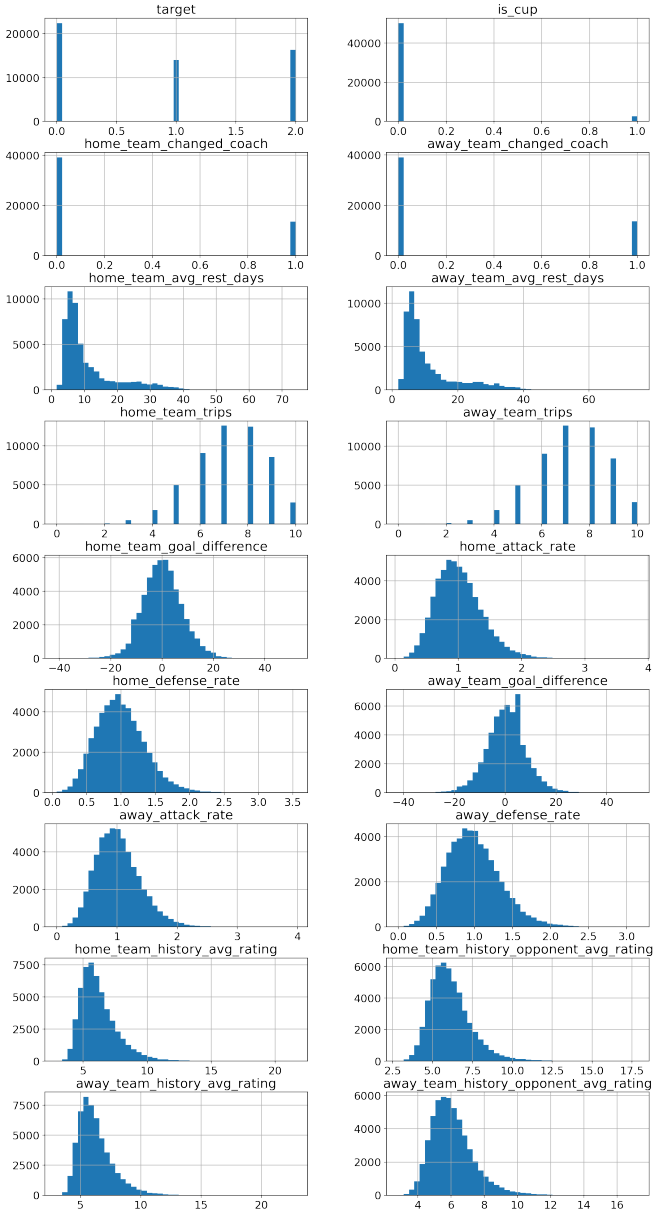

Fig. 4: Distribution for some of the features used in the model training.

V. Evaluation

A. Testbed

In order to run the experiments and models proposed, we leveraged machine learning libraries Scikit-learn [10] for creating the Random Forest model and Pytorch [11] for the FNN model. All experiments were run on an Ubuntu 20.04 machine with an 8 core Intel i7-11390H processor and 16 GB of RAM.

B. Models Parameters

For the RF model, we have performed a Grid Search parameter tuning and have used a model configured with 150 as the number of estimator each with a max depth of 5.

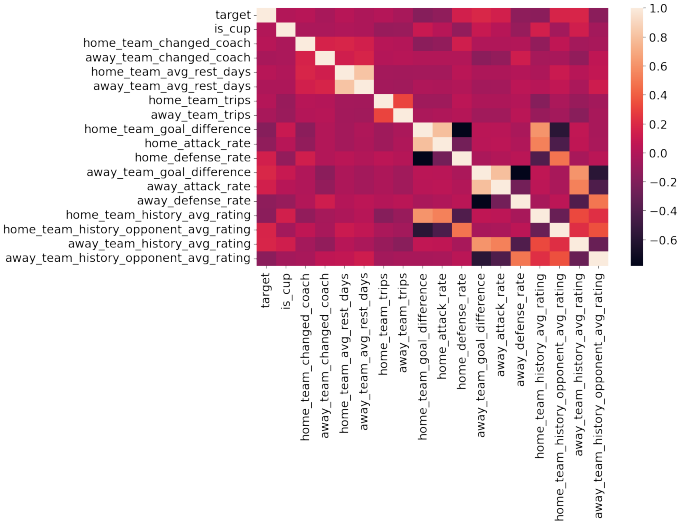

Fig. 5: Correlation heatmap for some of the features present in the dataset

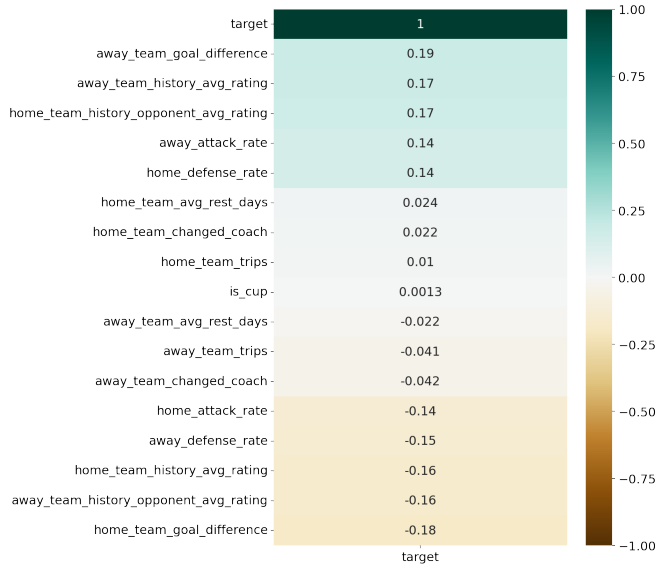

Fig. 6: Correlation between some of the dataset features and the target variable.

For the FNN model, after manual tests, we have decided to use a model with two linear layers, one with 50 and another with 25 neurons. The data is fed to the model in batches of 32 and we use a Stochastic Gradient Descent Optimizer.

C. Results

In Table II we can see the results for precision, recall, F1 Score and Accuracy for both trained models. As we can see, RF outperforms the FNN model in terms of accuracy and recall, while FNN exhbits better performance in terms of recall and F1 Score.

If we look further and investigate the confusion matrices for both models, as we can see in Figure 7 and 8, we discover that both models favor the dominant class, home team win.

Metric	Random Forest	FNN
Precision	0.32	0.38
Recall	0.41	0.38
F1 Score	0.34	0.38
Accuracy	0.48	0.41

TABLE II: Results for the two models used

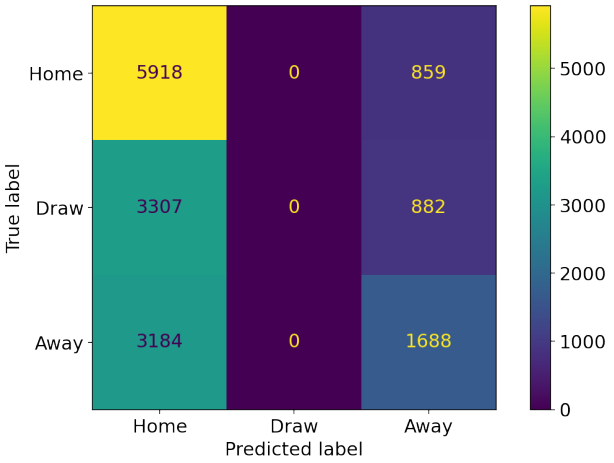

Fig. 7: Confusion matrix for the RF model

However, we can see that, although RF has higher accuracy, it never actually predicts any draw. The FNN model is still heavily biased in favor of a home team win, but it does not suffer from a complete lack of draw predictions.

In order to guarantee that the models are fitted, we can look at the learning curve for the RF model and to the training loss curve for the FNN model.

For RF, we can see in Figure 9 that the model training score seems to stagnate after a certain point, and more data

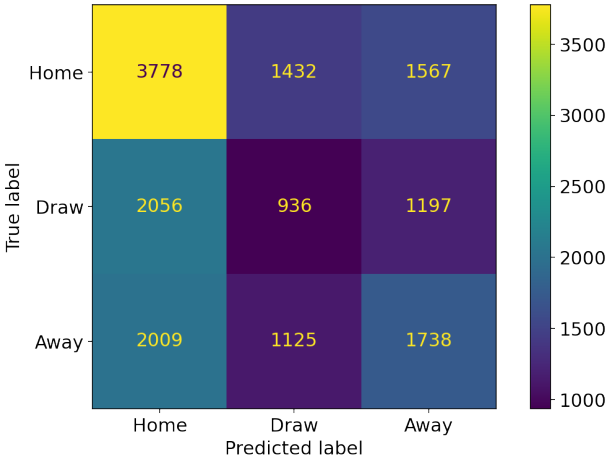

Fig. 8: Confusion matrix for the FNN model

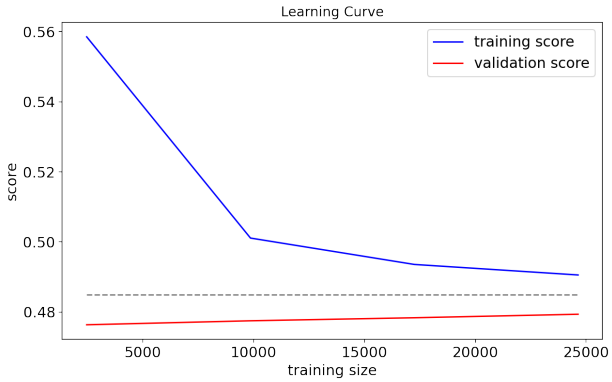

Fig. 9: RF learning curve

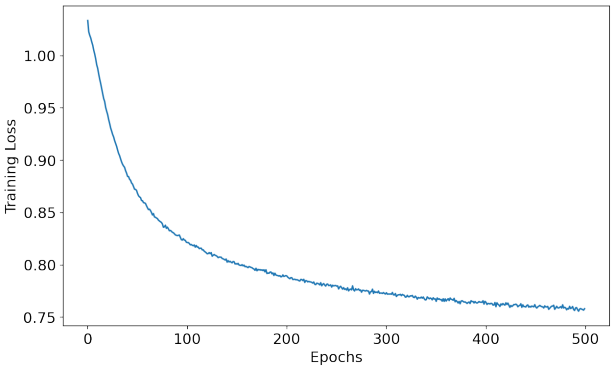

Fig. 10: FNN training loss

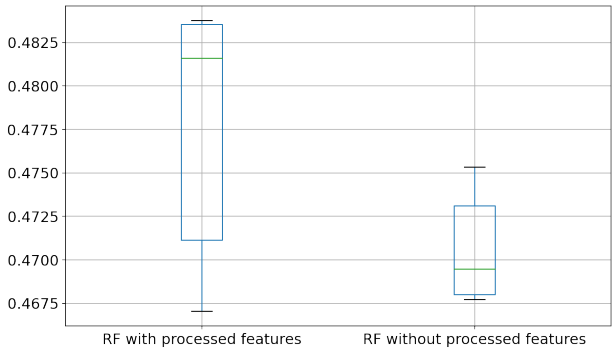

Fig. 11: Comparison between two RF versions

does not help with it. We can see that the model is not clearly not overfitting as well.

For the FNN model, we are evaluating the effect of the number of epochs on the loss observed during training. We can see that the curve seem to have reached a plateau by the end of the training.

D. Processing Significance

In order to access the effect of the pre-processing stage conducted, we also prepared a test comparing the RF model with and without the new columns. The boxplot presented in Figure 11 show a comparison across 5 folds of the dataset for the two model versions. We can see that the model with the new features seem to slightly outperform the one without it. However, when comparing both through a statistical test, we found a p value of 0.116, which implies that we cannot safely assume that the first model is better.

E. Discussion

The results obtained might seem underwhelming initially. An accuracy of less than 50% is not a very appealing outcome. However, it is important to put some context into these measurements. The second ranked result in the Kaggle competition [9] achieved an accuracy of only 0.5015.

But most importantly, it is important to notice how the game of football itself is uncertain in it's nature. The stronger team in a tie can see itself losing a match on pure luck or chance. To illustrate this point, we have selected a couple of matches wrongly predicted by the RF model. Both matches are from the English Premier League (EPL) and feature Manchester City, the current EPL champion and a 4 time champion in the last 5 seasons.

The first match happened at December 27, 2019, against Wolverhampton Wanderers. Man City was coming to this match with a 8-game unbeaten streak, in multiple competitions. Their team featured many international level players and was expected to win. However, in the 12th minute of the game, Ederson, Man City stating goalkeeper, got sent off. With this unusual and unpredictable event, the favourites found themselves a man down through almost the entirety of the game. They were able to score 2 goals but ended up conceding 3 and lost the game.

Another clear example of the unpredictability of the game was seen in April 10, 2021. Manchester City occupied the 1st spot in the table with 71 points, 8 points clear of the second place team. Their opponent was Leeds United, which were the 11th in the EPL standings with only 45 points. In this game, not only was Man City the favourite to win, but they also showed it in the game. They held the ball for 71% percentage of the time, compared to Leeds 29%, and shot the ball an impressive 29 times in the direction of the goal, compared to Leeds merely 2 times [12]. Not only that, but, as we can see in the shot chart in Figure 12, Man City had the advanced stats in it's favor, with almost 1.99 expected goals to Leeds 0.18 [13]. However, Leeds was able to score on both of it's attempts, including a last minute goal, and went away with the win.

With this examples, we intend to illustrate the difficulty of the problem in hand. There are so many factors that influence the results of a football match, and many of them can be boiled to simple luck. It is near impossible to predict if a player is going to have an off night, or if someone might receive a red card due to a careless tackle, or even if the referee won't commit any errors that might influence in the result of the match. Having this in mind, the results achieved can be seen as an interesting case study to how much we can predict only using historical team data.

Fig. 12: Shot chart and expected goals for the match between Manchester City and Leeds United [13]

VI. Conclusion

In this project, we have tackled the problem of predicting football matches results using historical data. We have created two models, a Random Forest and a Feedforward Neural Network, to predict the result of a given match, home team win, away team win or a draw. We have achieved results slightly bellow the 50% mark for accuracy with both models. This can be attributed to the unpredictable nature of the game as well as the limitations of the dataset used.

We would like to expand this project in the feature by incorporating individual players statistics. Having measurements of players characteristics and their availability for matches might help the model better understand matchups and avoid missclassifcations in cases where a team has their star player unavailable due to injury, for example.

References

[1] "Allianz and football," Accessed in 2022. [Online]. Available: https://www.allianz.com/en/about-us/sports-culture/football/allianz-football.html

[2] "More than half the world watched record-breaking 2018 world cup," Accessed in 2022. [Online]. Available: https://www.fifa.com/tournaments/mens/worldcup/2018russia/media-releases/more-than-half-the-world-watched-record-breaking-2018-world-cup

[3] W. Spearman, "Beyond expected goals," in *Proceedings of the 12th MIT sloan sports analytics conference*, 2018, pp. 1–17.

[4] M. Rahman *et al.*, "A deep learning framework for football match prediction," *SN Applied Sciences*, vol. 2, no. 2, pp. 1–12, 2020.

[5] N. Danisik, P. Lacko, and M. Farkas, "Football match prediction using players attributes," in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*. IEEE, 2018, pp. 201–206.

[6] D. Prasetio *et al.*, "Predicting football match results with logistic regression," in *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*. IEEE, 2016, pp. 1–5.

[7] "Football match probability prediction," Accessed in 2022. [Online]. Available: https://www.kaggle.com/competitions/football-match-probability-prediction/overview

[8] S. Yeung, "Football prediction by xgboost," Accessed in 2022. [Online]. Available: https://www.kaggle.com/code/szeyeung/football-prediction-by-xgboost/notebook?scriptVersionId=94822183

[9] M. Trivedi, "Lstm and bi-dir base," Accessed in 2022. [Online]. Available: https://www.kaggle.com/code/manavtrivedi/lstm-and-bi-dir-base/notebook?scriptVersionId=92011165

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[12] "Transfer markt," Accessed in 2022. [Online]. Available: https://www.transfermarkt.us/manchester-city_leeds-united/statistik/spielbericht/3429798

[13] "Understat - manchester city 1 - 2 leeds," Accessed in 2022. [Online]. Available: https://understat.com/match/14740