Sports Analytics for Football League Table and Player Performance Prediction

Victor Chazan - Pantzalis
The Data Mining and Analytics research group
School of Science and Technology
International Hellenic University
Thermi, Greece
v.chazan-pantzalis@ihu.edu.gr

Christos Tjortjis
The Data Mining and Analytics research group
School of Science and Technology
International Hellenic University
Thermi, Greece
c.tjortjis@ihu.edu.gr

Abstract–**Common Machine Learning applications in sports analytics relate to player injury prediction and prevention, potential skill or market value evaluation, as well as team or player performance prediction. This paper focuses on football. Its scope is long–term team and player performance prediction. A reliable prediction of the final league table for certain leagues is presented, using past data and advanced statistics. Other predictions for team performance included refer to whether a team is going to have a better season than the last one. Furthermore, we approach detection and recording of personal skills and statistical categories that separate an excellent from an average central defender. Experimental results range between encouraging to remarkable, especially given that predictions were based on data available at the beginning of the season.**

Keywords–Sports Analytics, Performance Prediction, Machine Learning (ML), Data Mining, Classification, Regression.

I. INTRODUCTION

Sports analytics is the use of historical data and advanced statistics to measure performance, make decisions and predictions regarding performance and outcomes, in order to gain an advantage over competitors [1]. Performance prediction is the commonest task in sports analytics. Sport analysts process data regarding players and teams with an intended goal: the prediction of match results, tournament winners or team and individual player efficiency. Forecasts may be related to short–term or long–term events. For that reason, diverse methods and algorithms have been deployed.

Clubs use sophisticated devices and software (i.e. GPS tracking systems) to gather and analyze data generated by players during training sessions and matches. They process these data to use for short–term decision making and long–term organization development. Also, extensive analysis of all data available is a prerequisite for betting companies. Finally, fans are also very interested in advanced statistics and how they affect football.

For all the above reasons, the use of sports analytics has increased during the last few years. Football was selected for our research because of the abundance of statistical categories and historical data, its fame, as well as the simplicity of its rules and of national championships formats. On the other hand, there are special difficulties, which make football long–term prediction challenging.

The abundance of online data regarding football is an asset, but requires filtering and proper data for team and player performance prediction. Unfortunately, this is not always easy. Additionally, team and player performance can be affected by incidents not depicted in the data collected; a team is rated higher than it should be when their opponents underperform. A player might have a low rating performance when coming into action after a serious injury.

Finally, the nature of football makes statistical recording of match events as well as player and team rating, an ambiguous process. Following the same pattern, performance prediction is not easy and long–term performance prediction is even tougher, but also not sufficiently studied until now.

Nevertheless, as it is shown in this paper, it is possible, up to a certain level, to make some long–term predictions, especially for team performance. Our prediction is relatively good, mainly with regards to the champion and the teams that win European qualification. What makes this research interesting is that the prediction is performed *before* the beginning of the season, with no official matches played, only with historical data and the information gathered during the summer break. Another novelty of this paper is that *advanced statistics* from previous seasons are used for prediction.

The remaining of this paper is structured as follows: Section II reviews the literature providing background information, Section III defines the problem and details our approach, Section IV provides experimental results, evaluated in Section V, and Section VI concludes with directions for further work.

II. BACKGROUND

C. Reep is believed to be the first British notational analyst. He published a statistical analysis of patterns of play in football, along with B. Benjamin in 1968, using 578 matches between 1953 and 1967. During the last 20 years, sophisticated techniques, algorithms, and tools for sports analysis were developed, while articles and papers related to sports analytics are constantly being published. Match outcome prediction is an interesting topic in Sports Analytics. Researchers approach the problem from different angles. A simple, solid but also obsolete prediction strategy is to predict the number of goals scored by the two teams.

The first model sufficient for predicting the result of a match was created in 1997 by Dixon and Coles. The model is considered a classic and was able to extract probabilities for the goals scored in a match, following Poisson distribution [2].

During the last years, researchers, focused on directly predicting wins, draws and losses, instead of trying to predict goals scored or points won. Various Machine Learning (ML) algorithms were implemented in order to discover the most discriminating factors that separate the winning from the losing side; Lago-Penas et al. concluded in shots on goal, crosses, match location, ball possession and opponent team ability, based on a ranking system [3]. Harrop and Nevill supported that the best predictor is pass accuracy, followed by the number of shots, the number of passes and dribbles (the fewer the better) and the venue of the match [4]. Mao et al. claimed that the features that provide the most positive effects are shots on goal, shot accuracy, tackles and aerials won [5].

Tax and Joustra employed a set of factors from public data and used dimensionality reduction techniques, such as Principal Component Analysis (PCA), along with ML algorithms (Naive Bayes and Multilayer Perceptron) to predict the Dutch football championship.

They achieved an accuracy of almost 55% in their predictions and proved that a hybrid model, combining public data and betting odds could improve accuracy [6].

Neural Networks (NNs) have also been used for prediction in football. McCabe and Trevathan dealt with four different sports. Using data from 02–08 and a Multilayer Perceptron, trained with Back Propagation, equipped with conjugative–gradient algorithms, they tried to predict match results. The NNs had 20 input layer nodes, 10 hidden layer nodes and 1 output layer node. The same features were used for every sport. Football had the worst average prediction performance of 54.6% [7]. Then, Hucaljuk and Rakipovic concluded that NNs performed better than any other ML technique they used [8].

Goddard, in 2005, compared the two methods, i.e. modeling the goals scored vs modeling win–draw–lose match result and concluded that a hybrid model achieves the best prediction performance. He also was one of the first to use variables other than previous match results. He leveraged features like the importance of individual matches, geographical distance between the two opponents and more. For the win–draw–lose method he used an ordered probit regression model and exploited a database of English match results for the past 25 years. He also included in his work a comparison of his predictions with the betting odds of the matches and concluded that achieving a positive betting return over time is possible [9].

In addition, many remarkable advanced statistics have emerged during the last decade, such as Expected Goals (xG), Packing, Defensive Coverage, Sequences and more. xG is a statistical measure of the quality of chances created and conceded (Expected Goals Against). xG probabilistically assign a score from 0 to 1 to each chance based on several variables. Shot quality evaluation is usually achieved by training NNs over large datasets of shots. xG are calculated for individual players, but also cumulatively for the whole team. The model eliminates some of the randomness of the actual goals scored and gives better insights into team performance. xG has not avoided criticism, but there have been certain cases that the method was implemented with great success.

In 2016, Eggels et al. used xG trying to build a model to classify each scoring opportunity into a scoring probability. They leveraged geospatial data and implemented various classification techniques. They also indicated that xG could be further used for evaluating players and seasons, but they warned that probability estimates of goal scoring opportunities may suffer from high standard deviation [10].

Apart from works on predictive analysis, there are various interesting researches referring to comparative analysis. The main components compared are wins and losses. It appears that a noteworthy attribute that most researchers point out is *efficiency*. Efficiency is defined as the number of goals divided by the number of shots. Shots on goal, pass accuracy, quality of the opponent team, venue of the match and ball possession also seem to be significant variables [11].

Bekris et al. used a different approach; they compared matches with at most one–goal difference (i.e. short range) to matches with at least three–goals difference (i.e. wide range). They found out that wide range winners outplayed their opponents in ball possession percentage, number of passes, “one vs. one” duels won, number of shots, number of shots on target and shooting accuracy. Contrariwise, those findings do not stand for short range matches, which are more sensitive to luck [12].

Researchers have also used the concept of rating. Rating is a single number which is used to describe the strength of a team in comparison to other teams at the time. A famous rating system is the ELO Ratings, which was used by Hvattum and Arntzen [13]. They used ELO Rating differences between teams as covariates in ordered logit regression models. Constantinou and Fenton used the pi-ratings that they had earlier invented for model validation, trying to make long–term prediction over team performance [14].

Predicting the outcome of a match is important, but maybe not as important as the prediction of team performance for the season. It is obvious that it is very hard to predict the long–term performance of a team and it is much harder to predict its performance by comparison with the performance of other teams. Limited work has been done on this challenging task so far. One of the most intriguing but also almost unexplored scopes is the prediction of a championship’s final table.

Van Haaren and Davis emphasized on the difficulty to predict the exact position of a team in the final table, because it depends on the final position of every other team [15]. Another obstacle for their method was the number of matches that ended in a draw. Ranking systems used for simulating match results have difficulties in predicting draws. This resulted in high variance on the predicted number of points for each team. However, they indicated two substantial metrics needed for evaluating the quality of the predicted final tables: the percentage of correctly predicted relative positions and the Mean Squared Error (MSE) regarding positions.

Oberstone developed a multiple regression model, ending up with 6 independent variables which he assessed to be sufficient for predicting the final league table of EPL in terms of points, instead of accurate positions [16]. He also used F distribution to compare means of multiple samples (i.e. one–way analysis of variance) to investigate which pitch actions differentiate the four best teams from all the others in the league. He managed to achieve outstanding results.

There have been some interesting works focused on the financial strand of football clubs, rather than pitch performances. Kringstad and Olsen used data from the Norwegian league and focused on the relationship between financial strength and sporting outcome [17]. They presented some mixed results: evidence suggested that budgeted revenue was a success indicator, but only for bottom–half teams, while static and dynamic regression models they implemented supported the notion of budgeted revenues being a driver of sporting outcome. They concluded that focus on athletics is still vital as money is a significant factor of success, but only to a certain extent.

Coates et al. used data from every team that participated in Major League Soccer (MLS) in USA during 2005–13. They examined the relationship between salary level and dispersion with football success. They revealed that while the wage bill of team has a positive effect on success, salary inequality has a negative effect on success. In that way they proved that cohesion is essential in football [18].

Cintia et al. used pass–based performance indicators and other efficient metrics, like the Pezzali score. The signification of this metric lies on the fact that it rewards teams effective on both sides of the pitch, i.e. in offensive skills and in defensive duties. It is formulated as follows:

$$Pezzali\ score(team) = \frac{|goals(team)|}{|attempts(team)|} \times \frac{|attempts(opponent)|}{|goals(opponent)|} \quad (1)$$

They simulated matches from four major leagues and claimed that they achieved superb results, as they predicted match outcome with an accuracy of almost 60%. They also found that the final rankings in the simulated championships were very close to the true rankings. Nevertheless, some teams had a considerable ranking error, which was explained by very high or very low Pezzali score. Finally, they marked the simplicity of their models and encouraged researchers to work with more complex models as they reckoned that there is room for improvement in accuracy [19].

Constantinou and Fenton, studying predictive accuracy in long–term team performance, proposed a method which they called smart–data [14]. They exploited external factors which might influence the strength of a team (i.e. managerial changes, European qualification, newly promoted teams etc.). With those factors they built new ones, such as “true team strength”, “expected performance” and more. Their goal was to predict the final table in terms of points won by each team. They achieved great results, managed to single out certain external factors that boost or worsen a team performance, focusing on the quality of their data, not on the quantity.

Football passes are important actions. Cakmak et al. introduced a metric, named “Pass Effectiveness” [20].. They based pass evaluation upon mathematic grounds. Pass effectiveness is being extracted from

the combination of five other measurable pass metrics: gain of a pass, pass advantage, goal chance, decision time and pass effectiveness of the next pass.

Passing networks is also a very intriguing subject; players are represented as nodes of a network, while passes between two players are represented as edges between the players. The edges are weighted based on the number of passes being exchanged between players. Cintia et al. leverage passing networks in several papers, in order to predict football matches outcome, but also final league tables. They concluded that networks are more efficient for long–term predictions for whole competitions [21]. Grund analyzed a dataset of 283,259 passes and applied mixed–effects modeling to 76 repeated observations of the interaction networks and performance of 23 soccer teams. He proved that best performing teams were characterized by networks with high intensity and low centralization [22].

Spatiotemporal data are significant in sports analytics. The advances in image processing made the analysis of positional data a lot easier. Borrie et al. suggest that temporal pattern analysis will lead to a deeper understanding of sport performance. They detected temporal patterns to find similar pass sequences within matches [23].

Player performance prediction is also interesting. Nsolo et al. investigated the attributes which best predict the success of individual players, based on their position, and evaluated different ML algorithms regarding prediction performance. They focused on top players of the top five European leagues and evaluated players based on different attributes for each position. They concluded that forwards tend to have higher performance ratings than other players, so maybe more advanced metrics should be applied on defensive players [24].

Previously, we used past data for long–term performance prediction; we estimated how many goals a certain player will score in a season and the number of a player's shots during each individual match. We also predicted the playing positions of a set of players according to their attributes [25]. We also predicted the best NBA defender as well as the MVP for 2 years [26].

Sîrb et al. presented a set of 54 performance criteria, over different playing positions in order to evaluate the performance of players, consider each player's natural position and the tactical formation that the team deployed in a match [27].

Finally, Pappalardo et al. analyzed player performance from 18 different competitions for several years and presented PlayeRank, a data–driven framework. The dataset contained 31 million matches and 21 thousand players. PlayeRank was found to outperform competitive predictive algorithms. They also discussed what distinguishes top players from others and discovered patterns for excellent performances. One of the limitations was that PlayeRank does not consider off–ball actions, like pressing. The authors also emphasized on the fact that an improved version of the framework should be able to leverage data from other sources, like wearables, GPS and video tracking data [28].

III. Problem Definition, Approach Followed

A. Problem Definition

Long–term performance prediction for teams or individual players are fields requiring exploration. Not only coaches, but also sports agents and bookmakers are interested in how teams or players perform during a season compared to previous ones. What is discussed in this section is the context of this problem. Also, the objectives of the research are set.

The unique components of football matches make long–term predictions very difficult; only few goals are scored per match. Also, there is no clear changeover between the instantaneous change of possession and transition between offense and defense. Moreover, player positions and tactics are not fixed and finally, the game has a continuous flow, which complicates recording of game events [16].

Our research focuses on statistics from the previous season and historical data. Also, some financial data (i.e. transfer spending, team salaries) are exploited to contribute to the team evaluation process.

The novelty of this research is that advanced metrics were used, such as xG and Pezzali score to predict next season performance *before* the season begins, not after matches have already been played and recorded. Additionally, attackers are usually graded higher than defenders, even if they are not always more influential in team strategy. So, regarding to player evaluation, this research attempts to identify skills and features suitable, that make good defenders.

B. Approach Followed

This section showcases the flow of events taking place before we can get any meaningful experimental results, as well as the way the data were acquired and their preprocessing. The block diagram which summarizes the process is depicted in Fig. 1:

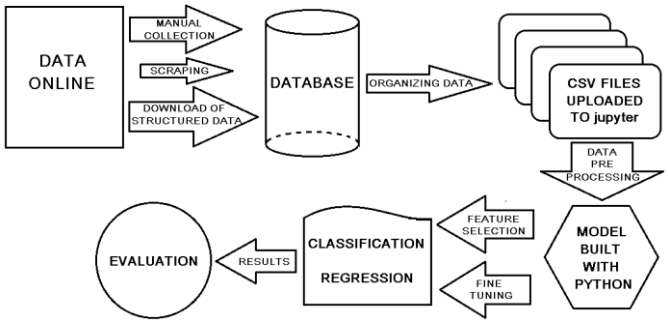

Figure 1: *Block diagram of the process followed for the experiments.*

At first, the appropriate data had to be found. There are a lot of web pages that contain information and statistics regarding football matches and events. The data refer to both teams and players. Some of the data were accessed and collected manually, especially when that was easy. However, some of them were scraped from the internet using various scraping tools. Finally, a free database from an expired online competition had been downloaded and used for the experiments. The database contains data from thousands of players and is extracted from a famous manager simulation game. It demonstrates player ratings for several football skills. Players are rated by domain experts.

After the process of data acquisition, there was a large database which needed to be organized. The database was split into different csv files, according to what data were essential for each experiment. Then, the csv files were uploaded to jupyter, the software that was mainly used for data processing.

Naturally, the data firstly needed to be preprocessed. They were checked for null values, duplicates, noise etc. Python was used to clean the data and build the models. Then, data transformation and data reduction took place to keep only the appropriate features for each classification or regression.

Finally, results were evaluated in terms of accuracy, error rates and bias involved. They were also being compared to results of other similar researches to estimate the value produced by them.

IV. Experiments and Results

A. 1st experiment: Team Performance Prediction

The first experiment is divided in two parts: The first part can be described as follows: Having a dataset with every team from four important European football national leagues, with more than 40 features for every team for each of the last four years (2015-18), predict whether a certain team is going to have a better or worse season than the previous year in terms of points. Every previous season is used as training set and the final season (i.e. 2017–18) is

used as test set. It is handled as a binary classification problem and the evaluation is conducted by measuring *AccuracyP* as follows:

$$AccuracyP = \frac{Number\ of\ teams\ with\ correct\ perfromance\ prediction}{Number\ of\ total\ teams} \quad (2)$$

Then, for the second part of the experiment, another method is presented; using almost the same features as in the first part, a model was built, to simulate every match of the 2018–19 season for the same championships (i.e. 380 matches per championship). Then, the virtual points collected by teams are accumulated in order to predict the final league standings. The predicted league table is compared to the actual league table and the evaluation is conducted by calculating the Root Mean Squared Error (*RMSE*) for the championship:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y_i} - y_i\right)^2} \quad (3)$$

where:

- n is the number of teams participating in the championship.
- $\hat{y_i}$ is the predicted points for the *i*-th team.
- y_i is the actual points for the *i*-th team.

Also, every model is evaluated for its ability to predict the outcome of matches played. The evaluation metric is *AccuracyM*, defined as follows:

$$AccuracyM = \frac{Number\ of\ games\ with\ correctly\ predicted\ outcome}{Number\ of\ total\ games} \quad (4)$$

The features used for the experiment are the ones that were considered more relative to team performance. Those features can be divided in three categories:

1. Past data generated during the last five years. This mainly refers to performance indicators from previous seasons (e.g. team average points).
2. Team statistical features from the season that has just ended (e.g. wins, xG, shots, possession percentage, Pezzali score and more).
3. Data not measurable by team performance (e.g. financial). These attributes are generated during the summer break, so most of them are independent from the previous season, but very likely to have an impact on the new season's performance.

Finally, in the dataset, there is the target attribute. It is binary and corresponds to whether the team is going to have a better or worse season than the previous one in terms of points won.

After data preprocessing, some attributes were removed from the original datasets, being irrelevant with the research or noisy, adding limited value to the outcome. Those were team statistics, like cards, interceptions, offsides, fouls etc.

The first problem to handle was that not every team of the previous championship takes part in the new one; there are teams that are relegated and teams that are promoted. It is meaningless to have historical data about newly promoted teams, because the data would refer to a different league than the one studied. So for the newly promoted teams, some adjustments had to be made. Indicatively, calculating the average team points of the last five seasons, if a team was playing in a lower division during that time, the points of the bottom league team were assigned to it.

Those adjustments caused certain problems; newly promoted teams do not necessarily have the same strength as teams that have just got relegated. Thus, the way they are described by features and attributes assigned to them might not be representative of their actual status. Furthermore, the three newly promoted teams are all assigned with the same values for the corresponding variables, which is not efficient. Therefore, the validity of this method is questionable.

The data were split into train and test set. Then, multiple classifiers were used to classify the test set teams into two classes (i.e. better season / worse season). Grid search was used for model tuning and 10–fold Cross Validation was used for testing the effectiveness of the model. A feature importance graph was also deployed to track the most valuable features; difference between goals and team xG in the previous championship turned out to be the most important features.

On the other hand, managerial change was not deemed an important performance indicator. However, it must be noted that the attribute used in this research does not factor what circumstances caused the managerial change in the club. Similar researches in the future may deal with this issue.

On the model build, Random Forest was the classifier that achieved the highest accuracy, with more than 70% AccuracyP and with standard deviation less than 10%.

For the second part of the experiment, more databases were used. That data pertained to the results of every match of the four championships presented in the first part of the experiment. Every unnecessary attribute was again removed and datasets were merged with the datasets of the first part of the experiment. This process resulted into a new dataset, which contained every match from a football season with its full time result and with statistical, financial, and historical data about the two teams involved in each match.

Naturally, a problem came up: some of the teams participating in championship matches lack any data, because they are newly promoted. Thus, there were some missing values in the dataset to be handled properly, by the following method. Newly promoted teams were considered to be the weakest ones in the league and were assigned the maximum, the minimum or the mean value of the corresponding attribute, according to the nature of each attribute.

The next step was to combine the attributes of home and away team by subtracting the corresponding pairs. Some attributes from the first part of the problem were excluded from the second one as the subtraction was not meaningful. Finally, every team was encoded using dummy variables and the first three seasons of each national championship were concatenated.

The dataset was split into a training set and a validation set. Last season was kept separately from others, the target (i.e. the full-time result) was hidden and was used as a test set. Training/validation set consisted of 1140 rows (380 rows for the test set) and almost 40 attributes (team dummy variables were not included).

Standardization of the data, parameter tuning, and Cross Validation techniques were used again, exactly as previously. Multiple classifiers were deployed to predict the match outcome and therefore the leagues' final standings. Team market value percentage, expected points and non–penalties xG turned out to be the most important features, but not by a big margin from other attributes.

During this process two problems came up; the first problem was that almost every classifier used had the tendency to favor big teams over smaller ones. The other problem was that most of the models built faced difficulties in predicting draws.

Despite the drawbacks, the results achieved could be considered promising. They are comparable to results from similar researches, while the advantage of this research is that the experiments can be concluded at the beginning of the season, with no official matches played and recorded. The best AccuracyM for the outcome of the matches was 57% for the English Premier Division and the smallest RMSE for team points was 9, achieved for Spanish La Liga. French Ligue 1 produced the worst results, both in terms of AccuracyM and RMSE. The results from each league are presented in the following Tables 1 to 4. The best result for each league is noted with green color and the worst one with red.

Table 1: *Results from the English Premier League.*

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	55	17
Decision Tree	45	12.9
Random Forest	56	14.3
KNN	48	15.3
SVM (rbf kernel)	54	18.2
SVM (poly kernel)	**57**	**11**
XGBoost	52	17.3

Table 2: *Results from the Spanish La Liga.*

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	47	23.7
Decision Tree	39	14.9
Random Forest	48	17.7
KNN	46	13.3
SVM (rbf kernel)	**51**	13.8
SVM (poly kernel)	47	**9**
XGBoost	45	17.4

Table 3: *Results from the Italian Serie A.*

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	**53**	19.7
Decision Tree	41	**11.3**
Random Forest	40	14.4
KNN	47	14.7
SVM (rbf kernel)	52	19
SVM (poly kernel)	50	12.2
XGBoost	42	14.5

Table 4: *Results from the French Legue 1.*

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	42	28.2
Decision Tree	39	17.6
Random Forest	**45**	24.8
KNN	39	20.9
SVM (rbf kernel)	43	22.2
SVM (poly kernel)	43	**17.3**
XGBoost	44	21.8

It is obvious that SVM with polynomial kernel is observed to steadily achieve good results in every league studied, so it is regarded as a benchmark for this research from now on. Overall, the best result in terms of RMSE was achieved from the Spanish La Liga, where the classifier predicted the final league table with surprisingly high accuracy, given the few attributes used. In Fig. 2, the real vs predicted league tables are shown and compared.

ACTUAL TABLE		PREDICTED TABLE	
1. Barcelona	87	1. Barcelona	83
2. Atletico Madrid	76	2. Atletico Madrid	75
3. Real Madrid	68	3. Real Madrid	64
4. Valencia	61	4. Valencia	58
5. Sevilla	59	5. Sevilla	57
6. Getafe	59	6. Getafe	57
7. Espanyol	53	7. Real Betis	57
8. Athletic Bilbao	53	8. Eibar	57
9. Real Sociedad	50	9. Celta Vigo	57
10. Real Betis	50	10. Villarreal	55
11. Alaves	50	11. Athletic Bilbao	54
12. Eibar	47	12. Real Sociedad	54
13. Leganes	45	13. Leganes	54
14. Villarreal	44	14. Espanyol	51
15. Levante	44	15. Alaves	51
16. Celta Vigo	41	16. Levante	51
17. Valladolid	41	17. Valladolid	51
18. Girona	37	18. Girona	51
19. Huesca	33	19. Vallecano	50
20. Vallecano	32	20. Huesca	49

Figure 2: *Spanish La Liga 2018–19 actual vs predicted table.*

Green fonts are used for teams that won European qualification through Champions League, blue for teams that won European qualification through Europa League and red for teams relegated after the end of the season. The classifier has done an outstanding job in predicting these teams' performance. It correctly predicted the champion, but also the ranking of the first six teams in the league.

In this example, SVM with polynomial kernel succeeded not to overestimate the top teams (i.e. a problem which was often observed throughout most of the classifiers), but on the other hand overestimated the bottom teams instead. One other problem was its inefficiency in predicting draws, as very few match outcomes were predicted as "draw".

Despite their divergence and how small or big AccuracyM and RMSE were in every case, most of the classifiers correctly predicted the league champion. The equivalent accuracy was very good, regarding teams that won European qualification and mainly those amongst them that qualified for Champions' League, as shown in Table 5. Results for the relegated teams were also acceptable. Europa League teams were the exception, as the prediction accuracy was poor.

Table 5: *Accuracy in predicting champion, teams that won European qualification and teams relegated.*

	Premier League	La Liga	Serie A	Ligue 1	Overall
Championship Winner	71%	71%	57%	57%	64%
European Qualification	86%	76%	82%	46%	75%
Champions League	79%	86%	71%	57%	74%
Europa League	38%	29%	29%	0%	29%
League Relegation	52%	48%	57%	10%	42%

Finally, as far as AccuracyM is concerned, another aspect of the experiment is the following: Instead of using the previous three seasons as training set and the last season as validation set, use the first 10 match days of 2018–19 season as training set and the remaining 28 match days as test set. In that case, AccuracyM of the Spanish La Liga rose from 51% to 70%, as seen in Fig. 3. Therefore, it is shown that present season's data can boost the accuracy of the model in a very beneficial manner.

```
              precision    recall  f1-score   support

          -1       0.72      0.38      0.50        76
           0       0.59      0.78      0.67        79
           1       0.83      0.90      0.86       125

   micro avg       0.72      0.72      0.73       280
   macro avg       0.72      0.69      0.68       280
weighted avg       0.73      0.72      0.71       280

[[ 29  35  12]
 [  6  62  11]
 [  5   8 112]]
Accuracy: 0.725
Mean is 0.7043939393939395
Standard deviation is 0.11155148041316686
```

Figure 3: *Accuracy in predicting match results after 10 match days from the Spanish La Liga have been analyzed.*

B. 2nd experiment: Player Performance Prediction

This experiment focuses on individual players, specifically central defenders. In rating systems, there is a bias toward forwards and attacking midfielders. Goals are considered the most important element of football, so defenders' contribution to a team is usually underestimated. Consequently, there is very limited research on central defenders. Additionally, while it is easy to rate attacking players, according to the goals, key passes and assists, it is not straightforward what makes a good central defender.

The purpose of this research is to examine the characteristics and the statistics for central defenders in comparison with their season rating and decide which of them contribute more to distinguish a central defender as a top class player.

The data collected refer to player attributes, playing positions and some demographic features. The database was narrowed down to 59 players, as only central defenders, playing in English Premier League and having participated in at least 10 league matches for the 2016–17 season were selected.

The next step was to collect season statistics for those players. The main focus was on statistics regarding defensive player actions, but also, some team statistics were collected; despite demanding to build a model based on player performance, it must be acknowledged that a footballer's team has an impact both on his statistics but also on his overall rating.

The initial approach to the problem was to normalize every numeric value of the dataset, so every attribute's range was transformed to range 0 to 1. Then a multiple regression model was built with every possible feature. Despite the simplicity of this approach, some useful early conclusions were drawn regarding to which features contribute more to central defenders' competency. It seems that for the examined dataset, interceptions are the most important characteristic, followed by team overall rating, as expected. Players' best attributes turned out to be their jumping reach, versatility, acceleration and first touch on the ball.

Another approach that was followed was to split the dataset's features into three categories:

1. Player characteristics and attributes.
2. Player statistics.
3. Team statistics.

Again, the target was to build three multiple linear regression models (i.e. player attributes based, and statistics based), but with fewer independent variables than in the first approach. The method used for the implementation of this part of the experiment was backward elimination. For the first category of features, the final model was built with seven features, which seem to be the most influential for a central defender.

The five assumptions of linear regression were also verified for this model; there was an indication of linearity in the model. Also, the expectation (mean) of residuals was almost zero and it appeared that there was no (perfect) multicollinearity between features. Additionally, by performing a Breusch–Pagan test, it was proven that there is no heteroscedasticity in the model. Nevertheless, the final assumption was not verified; The Durbin–Watson test gives a value much lower than 2, which implies that there was positive autocorrelation between features. Also, the R–squared and the adjusted R–squared were relatively low (under 0.5). However, considering that dependent and independent variables emerged from two different sources, the results could be described as encouraging.

The features of the second category (all derived from the same source) were the independent variables, while player rating (also derived from the same source) was, again, the dependent variable. This time, the final model consisted of 12 features, after Backward Elimination, with very low P–values, while, as seen in Figure 4, R–squared was 0.867 and adjusted R–squared was 0.833, a vast improvement from the first model.

Out[191]:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.867
Model:	OLS	**Adj. R-squared:**	0.833
Method:	Least Squares	**F-statistic:**	25.03
Date:	Mon, 11 Nov 2019	**Prob (F-statistic):**	3.39e-16
Time:	02:24:20	**Log-Likelihood:**	64.550
No. Observations:	59	**AIC:**	-103.1
Df Residuals:	46	**BIC:**	-76.09
Df Model:	12		
Covariance Type:	nonrobust		

Omnibus:	1.400	**Durbin-Watson:**	1.912
Prob(Omnibus):	0.496	**Jarque-Bera (JB):**	1.161
Skew:	-0.136	**Prob(JB):**	0.560
Kurtosis:	2.369	**Cond. No.**	1.41e+05

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.9602	0.142	49.122	0.000	6.675	7.245
x1	-0.0327	0.004	-7.776	0.000	-0.041	-0.024
x2	0.0032	0.001	5.650	0.000	0.002	0.004
x3	-0.0113	0.002	-4.635	0.000	-0.016	-0.006
x4	6.782e-05	2.04e-05	3.317	0.002	2.67e-05	0.000
x5	0.2326	0.039	6.017	0.000	0.155	0.310
x6	0.0972	0.028	3.525	0.001	0.042	0.153
x7	0.1261	0.021	6.127	0.000	0.085	0.168
x8	-0.1559	0.042	-3.749	0.000	-0.240	-0.072
x9	-0.0481	0.019	-2.588	0.013	-0.086	-0.011
x10	0.7259	0.124	5.862	0.000	0.477	0.975
x11	0.9718	0.226	4.292	0.000	0.516	1.428
x12	2.1514	0.743	2.896	0.006	0.656	3.647

Figure 4: *Model built with player statistics as independent variables.*

Additionally, all five assumptions of linear regression were met; Linearity of the model was obvious, as seen in Fig. 5. The expectation (mean) of residuals was found almost zero and there was no (perfect) multicollinearity between the features. The Breusch–Pagan test gave a p–value of 0.44, so there was no heteroscedasticity and the Durbin–Watson test gave a value of 1.91, so there was almost no autocorrelation between the features.

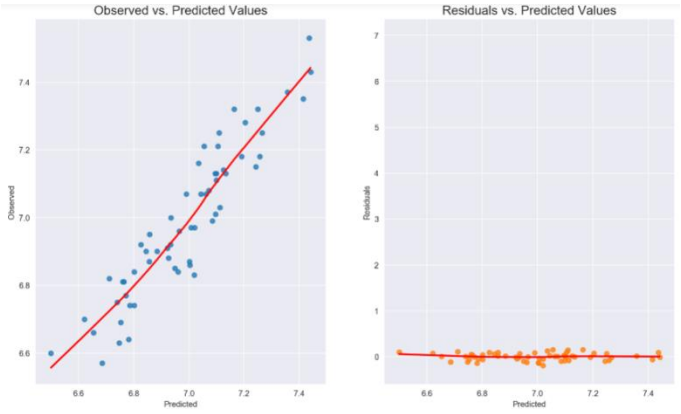

Figure 5**:** *Linearity of the second model.*

The third set of features (i.e. team statistics) did not help to build a satisfactory model. There was an indication that the only of those variables worth noting is "TeamRating". It was decided to incorporate "TeamRating" in the second model, in order to exploit that feature. Indeed, by updating the model, adding "TeamRating" as its 13th feature, there has been a slight improvement to the model; R–squared rose to 0.907 and adjusted R–squared rose to 0.88.

In conclusion, summarizing the results of all models deployed, the most critical attributes and game actions for predicting the performance of a central defender can be described in the following list. It must be highlighted that attacking skills are not absent from the list, following the way modern defenders are expected to play:

Interceptions,
Clearances,
Aerials Won,
Tackles,
Jumping reach,
Versatility,
Acceleration,
First touch on ball,
Age,
Passing,
Vision,
Determination,
Strength,
Professionalism and ability to perform well in important matches,
International Caps,
Minutes Played,
Fouls,
Inaccurate short passes,
Key passes,
Goals,
Team's rating.

V. Discussion

This section reviews and discusses our approach and results. Problems that came up during the process and the solutions given are debated. Results of the experiments are evaluated and threats to validity are mentioned, too.

The first problem encountered was the abundance of data. It was practically impossible to use every free online data acquired, so data

selection was challenging. Fortunately, the models produced from the datasets were not computationally intensive, so the approach followed was to include as many attributes relative to the research as possible, in order not to miss out important information. Later, during the feature selection phase, some less important attributes were removed.

Conversely, the acquisition of data substantial for research on football analytics was very difficult. Data regarding player injuries and data from wearable devices are mostly defined as private personal details. Thus, there are no such free online data to be used for the experiments.

Another problem was the handling of newly promoted teams, as statistics from previous seasons were generated for a lower division, so they could not be used. Concerning those teams, values were automatically assigned to some variables. That was a necessity, but in certain circumstances the predicted team performance did not meet some teams' real potential.

Additionally, most models were biased in favor of big clubs. An attribute that could be used as a penalization factor in cases of overestimation could balance out the aforementioned bias.

A similar problem was encountered because of the models' difficulty in predicting draws. The solution behind that problem usually lies on the proper usage of cost sensitive classifiers or by tuning the weights of classes.

In both cases, the proposed solutions were tested. Even though they resolved the issues that were deployed for, they both failed to extract better results than the ones already achieved. Hence, they were not included in the models.

Every championship has its own particularities, so rules extracted from one league do not necessarily apply to others. Therefore, despite the feasibility of building a universal model, proven by experimental results, the exploration of the differences between leagues would probably provide better research opportunities. This issue also reflects the second experiment. The database used consisted of defenders based only in England, because it would be inefficient to include players from different leagues.

Domain expert opinion was used in rating player attributes. Scout reports, despite generally describing well enough player ability and potential, have often been misleading, while intentional tampering with ratings and attributes should not be ruled out. Additionally, financial–based data usually suffer from inaccuracies and cannot be fully trusted.

Unfortunately, football matches are not affected only by team ability and player skills. There are some external factors that cannot be predicted. Luck is an imponderable factor. Long term injuries of important players are also part of the game. "Strange" results in matches where one or both teams are not in real need of victory are often observed. Finally, betting odds inevitably have an influence on match outcome. All those drawbacks, which can be viewed as threats to validity, prove that long–term sports prediction is very demanding and may not always provide meaningful results. Nevertheless, the results of the experiments conducted in our research can be described as good or even impressive in certain occasions.

AccuracyP level for the first part of the first experiment can be described as satisfactory, given the fact that it is a long–term prediction with no official match data and statistics available. A professional expert could exploit the experimental results, along with his own intuition and make certain decisions.

The main achievement of the research is the second part of the experiment, where the models used predicted some famous champions' final table with great accuracy. Also, classifiers are able to predict almost 2 out of 3 match outcomes when the model is applied in the midst of the season. Consequently, this implementation can be vastly used for betting purposes under certain circumstances. Provably, planning a profitable betting strategy based on experimental results and –apparently– in some human expertise, is possible.

Finally, the second experiment succeeds into locating a set of attributes and skills that a central defender must improve in order to be considered a top class player. Of course, every player is different and has his/her own playing style, but it would be very useful for coaches to have a specific targeting when training a player. Long–term player prediction performance could also be a huge contribution to fantasy sports games. The experiment resulted in a variety of features. Unsurprisingly, some of them were the main defending actions and attributes, but in an interesting manner, some were also found to be attacking actions or attacking attributes.

VI. Conclusions and Future Work

A. Conclusions

In this research, two fundamental cases of sports analytics were studied: team performance prediction and player performance prediction.

For the first experiment, the goal was to predict how each team of four important European leagues would perform during the 2018–19 season. The data available were only historical data (from 2015 onwards) and information about team actions (transfers, managerial changes etc.) during the summer of 2018, just before the beginning of the season. Two approaches were followed to address this issue.

For the first approach, the target was to classify teams in those that would perform better than last season and those that would perform worse than last season in terms of points collected. Results could be described as satisfactory, but not impressive, as AccuracyP of the classifiers deployed reached the level of 70%. In this approach, no distinction between the examined championships was made, so the model used could be described as universal.

Another approach for team performance prediction achieved remarkable results; the idea was to simulate every match of the season and classify their results as home win, draw or away win. At the end, each team's points were accumulated, and a predicted league final table was extracted. The effectiveness of the model was measured with two metrics: AccuracyM of the predicted match outcomes and RMSE of predicted vs actual team points in the league table. The highest AccuracyM achieved was 57% for the English Premier League and the lowest RMSE was 9 for the Spanish La Liga. Additionally, the champion was correctly predicted in 64% of the times and the teams that won European qualification were correctly predicted in 75% of the times. Also, this time, the four championships were separately studied and differences between them were evident.

Our results are very satisfactory and comparable to results of similar researches. Regarding prediction of match outcome, Tax and Joustra achieved 56% accuracy [6], while McCabe and Trevathan achieved 54.6% accuracy [7]. Joseph et al. achieved their best result using Baysian Networks, with 59.2% accuracy [29] and Eggels et al. achieved 54% accuracy [10]. Cintia et al. predicted match outcome with 60% accuracy and team points with 9.1 RMSE [19]. Our results have the advantage of being obtained without any current official match data available.

Additionally, applying prediction after using the first ten match days of the season as a training set was suggested as an alternative. In that case, AccuracyM of predicted match outcomes was impressively raised to 70%.

The second experiment was about defining which attributes and match actions are mainly influencing a central defender's match rating. The dataset consisted of 59 central defenders having played at least 10 matches for the English Premier League 2016–17 season. The method used was Multiple Linear Regression with Backward Elimination and the evaluation metrics were R–squared and adjusted R–squared.

Findings were noteworthy, as for a quite satisfying 0.907 R–squared and 0.88 adjusted R–squared, thirteen features were proved to be statistically significant. Classic defensive actions like interceptions and clearances were amongst them, along with player attributes more suitable for defenders, such as jumping reach and strength. The interesting part was that some attacking skills, such as passing, and

some attacking match actions (i.e. key passes made, goals scored) were also found to have an impact on rating central defenders. This fact stresses the change of playing approach from central defenders nowadays.

B. Future Work

The experiments have shown that it is possible to make long–term predictions about team and player performance, so it is reasonable that researchers will work in the same direction in the future, trying to resolve some issues or trying to improve the experimental results.

Data from cameras and wearables would be an invaluable asset to any sports analytics research. Future works on sports analytics should focus their attention on gathering and leveraging data from those devices.

Another idea would be the evaluation of newly promoted teams' ability and the study of their performance to comprehend what are the factors that lead them to be successful or not.

Problem with models' bias in favor of bigger clubs and difficulties in predicting draws were not fully resolved. Cost sensitive classifiers and tuning of the classes' weights did not improve the experimental results. Hence, it is suggested to scientists to delve deeper into those methods or implement a different approach to solve the aforementioned problems.

What was generally observed in this research and must preoccupy researchers in the future is the major divergence displayed on results extracted from different leagues. Therefore, fundamental differences between leagues should be specified, otherwise models could only be applicable on individual leagues and not become universal.

Additionally, it would be very useful if future researchers took into consideration some aspects that were not examined in this research; player fatigue or starting lineup rotation due to consecutive matches and important player long–term injuries are factors that can affect players or teams, but at the same time can make a model very complex. However, if the complexity is confronted, those data could be great assets for the research.

REFERENCES

[1] **Holman, V..** What is Sports Analytics? *Agile Sports Analytics.* [Online] November 15, 2018. https://www.agilesportsanalytics.com/what-is-sports-analytics/.

[2] **Dixon, M.J. and Coles, S.G.** *Modelling Association Football Scores and Inefficiencies in the Football Betting Market.* 1997.

[3] **Lago-Peñas, C. - Lago-Ballesteros, J. - Rey, E..** *Differences in performance indicators between winning and losing teams in the UEFA Champions League*, 2011, Journal of Human Kinetics, Vol. 27, pp. 135-146.

[4] **Harrop, K. and Nevill, A.** *Performance indicators that predict success in an English professional League One soccer team*, 2014, Int'l Journal of Performance Analysis in Sport, Vol. 14, pp. 907-920.

[5] **Mao, L. - Peng, Z. - Liu, H. - Gómez, M.-A.** *Identifying keys to win in the Chinese professional soccer league.* 2016, Vol. 16, pp. 935-947.

[6] **Tax, N. and Joustra, Y.** *Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach.* 10, 2015, Transactions of knowledge and data engineering, Vol. 10.

[7] **McCabe, A. and Trevathan, J.** *Artificial Intelligence in Sports Prediction* Information Technology: New Generations, 2008. pp. 1194-1197.

[8] **Hucaljuk, J. and Rakipovic, A.** *Predicting football scores using machine learning techniques.* 2011. MIPRO, 2011 Proc. 34th Int'l Convention.

[9] **Goddard, J.** *Regression models for forecasting goals and match results in association football*, Elsevier B.V., 2005, Int'l Journal of Forecasting, Vol. 21, pp. 331-340.

[10] **Eggels, H. - van Elk, R. - Pechenizkiy, M.** *Explaining soccer match outcomes with goal scoring opportunities predictive analytics*, 2016.

[11] **Lepschy, H. - Wasche, H. - Woll, A.** *How to be Successful in Football: A Systematic Review*.1, 2018, The Open Sports Sciences Journal, Vol. 11, pp. 3-23.

[12] **Bekris, E. - Gioldasis, A. - Gissis, I. - Komsis, S. - Alipasali, F.** *Winners and losers in top level soccer. How do they differ?* 2014, Journal of Physical Education and Sport, Vol. 14, pp. 398-405.

[13] **Hvattum, L.M. and Arntzen, H.** *Using ELO ratings for match result prediction in association football.* 2010, Int'l Journal of Forecasting, Vol. 26, pp. 460-470.

[14] **Constantinou, A. and Fenton, N.** *Towards smart-data: Improving predictive accuracy in long-term football team performance.* 2017, Knowledge-Based Systems, Vol. 124, pp. 93-104.

[15] **Van Haaren, J. and Davis, J.** *Predicting the Final League Tables of Domestic Football Leagues.* 2015. 5th int'l conf. mathematics in sport. pp. 202-207.

[16] **Oberstone, J.** *Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success.* 2009, Journal of Quantitative Analysis in Sports, Vol. 5.

[17] **Kringstad, M. and Olsen, T.-E.** *Can sporting success in Norwegian football be predicted from budgeted revenues?*

[18] **Coates, D. - Frick, Bernd - Jewell, T.** *Superstar Salaries and Soccer Success: The Impact of Designated Players in Major League Soccer.* 2014, Journal of Sports Economics, Vol. 17, pp. 716-735.

[19] **Cintia, P. - Pappalardo, L. - Pedreschi, D. - Giannotti, F. - Malvaldi, M.** *The harsh rule of the goals: Data-driven performance indicators for football teams.* 2015. IEEE Int'l Conf. Data Science and Advanced Analytics,

[20] **Cakmak, A. - Uzun, A. - Delibas, E.** "Computational Modeling of Pass Effectiveness in Soccer," *Advances in Complex Systems*, vol. 21, no. 3-4, 2018.

[21] **Cintia, P. - Rinzivillo, S. - Pappalardo, L.** *A network-based approach to evaluate the performance of football teams.* 2015. Machine Learning and Data Mining for Sports Analytics workshop (MLSA'15), ECML/PKDD conf. 2015.

[22] **Grund, T.U.** *Network structure and team performance: The case of English Premier League soccer teams.* 2012, Vol. 34, pp. 682-690.

[23] **Borrie, A. - Jonsson, G.K. - Magnusson, M.** *Temporal pattern analysis and its applicability in sport: an explanation and exemplar data.* 2002, Journal of Sports Sciences, Vol. 20, pp. 845-852.

[24] **Nsolo, E. - Lambrix, Pa. - Niklas, C.** *Player Valuation in European Football.* 2018. 5th Workshop on Machine Learning and Data Mining for Sports Analytics co-located with ECML PKDD 2018.

[25] **Apostolou, K. and Tjortjis, C.** *Sports Analytics algorithms for performance prediction.* IEEE 10th Int'l Conf. on Information, Intelligence, Systems and Applications (IISA 2019), pp. 469-472, 2019.

[26] **Sarlis V. and Tjortjis C.,** *Sports Analytics – Evaluation of Basketball Players and Team Performance*, Information Systems, Vol. 93, November 2020, doi: 10.1016/j.is.2020.101562..

[27] **Sîrb, L. - Molcuț A. - Nastor, F.** *The Exercise of Prediction Process o fPerformance within Football Sports Management by Using Fuzzy Logic from the Perspective of Value Analysis on Tactical Compartments of Game of the Football Players.* 2015, Journal of Knowledge Management, Economics and Information Technology, Vol. 5

[28] **Pappalardo, L. - Cintia, P. - Ferragina, P. -Massucco, E. - Pedreschi, D. - Giannotti, F.** *PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach*, ACM Transactions on Intelligent Systems and Technology September 2019 Article No.: 59.

[29] **Joseph, A. - Fenton, N. - Neil, M.** *Predicting football results using Bayesian nets and other machine learning techniques*. 2016, Knowledge-Based Systems, vol. 19, pp. 544-553.