Predicting Premier League Outcomes Using Machine Learning Models

Won Barng
INFO 370
Seattle, Washington
wjbarng@uw.edu

Kangwoo Choi
INFO 370
Seattle, Washington
kangwooc@uw.edu

William Kwok
INFO 370
Seattle, Washington
wkwok16@uw.edu

Vincent Widjaya
INFO 370
Seattle, Washington
vwidjaya@uw.edu

ABSTRACT

Football match outcomes are difficult to predict. Many predictions happen daily in order to increase earnings through betting, but not a lot of these predictions attempted to use feature engineering in combination with machine learning. Therefore, we created many features from a dataset on the Football Data [2] in an attempt to predict game outcomes and betting odds. We ran these new features through machine learning models and produced a model that could predict which team would win games with over 50% accuracy as well as a model that predicted betting odds with decent accuracy. These results provide a good baseline for future predictive models.

1 INTRODUCTION

Sports is an integral part of nearly all cultures. One sport that has lots of fans in the world is soccer, which we will call football throughout this paper. For years individuals have tried to use statistics to figure out what makes teams win, or to try to find out if their favorite teams are the best. [7] suggests that "the research for predicting the results of football matches outcome started as early as 1977 by [Stefani R]".

The English Premier League (EPL) is one of the most popular and largest sports leagues in the world. In 2017/2018 season, around 39.3 millions Americans watched the EPL [1]. Also, The EPL is one of the profitable sports league in the world. In 2017, there were five clubs from EPL in top ten teams by revenue among European clubs [9]. Our goal is to predict EPL outcomes.

In order to produce predictive models for EPL outcomes, we will be exploring a few research questions:

- RQ1: *Is there an association between the betting odds and statistics of a game?* Studying this may also result in finding cases of illegal match fixing. For example, we can see if certain teams are under-performing in some games where bets are higher.
- RQ2: *Is there an association between the betting odds, statistics of a game and the end results of a game?* We may be able to find cases of illegal match fixing such as the scandals that occurred in late 2013 in a different league [6] and produce a model that predicts if match fixing is happening.
- RQ3: *Is there an association between the team stats and the end results of a game?* With this, we will be able to predict what factors are associated with winning the game. Individuals will be able to use this to their betting advantage.

In answering these questions, we will be working with data from all EPL seasons from the 2014/2015 season to the current 2018/2019 season provided by [2]. Each dataset contains data pertaining to the game itself as well as a lot of bets associated with the game.

2 RELATED WORK

A previous study [5] explored the usage of a Poisson Regression in an attempt to model team scores to determine if Manchester United was the best team in the EPL. While their predictions were decent, they state their approach to modeling was simplistic and could definitely be improved. [5] suggests to take into account the "fact that teams differ from game to game due to injuries, trades, and suspensions" and to not "assume that [the] model leads to probabilities for winning/losing/drawing games".

Another study [8] explored the usage of a simple Poisson model as well, suggesting that while "overly simplistic", and "cliché", it is a "good starting point and a nice intuitive way to learn about statistical modelling". Some suggestions to improve the model involve weighting recent matches stronger in the model, using a bi-variate Poisson distribution, or try a Weibull distribution. We believe another improvement could be to use different features. In [8], the formula used only involved goals, what teams were involved, and what the home team was. We can improve the features chosen by using our domain knowledge or different methods of feature selections in order to produce a better outcome.

A third study *also* used a Poisson Regression on one model that used just the home goals and away goals parameters, and another model that used one additional "home advantage" parameter [3]. Again, there are many feature parameters that can be included to improve the accuracy of the predictions, which the aforementioned study also believes. Another suggestion is to use other models such as the "Hierarchical Bayesian Poisson Model" or "step-wise prediction" [3].

We can take the advice of many of these studies and provide our own domain knowledge towards answering our research questions. We will outline how we did this in our methods section below.

3 METHODS

RQ1

In order to predict betting odds from game statistics, we created several models. We chose to use regressors instead of classifiers because betting odds are continuous variables. The training data we used was an aggregate of all games from the previous seasons, grouped into their respective seasons for feature engineering. Our models were then tested against the games that have already been played in the current 2018/2019 season, to prevent over-fitting.

The first model was a quick grid search that utilized k-neighbors regression and 10-fold cross validation with scaled training values. This model did not include feature engineering and instead took into account only the provided number of goals scored by each team as well as the ratio of goals to number of tries. We then realized it did not make sense to predict odds with current game statistics. This is because by the time a game's statistics have been finalized, betters would have already locked in their odds meaning we were making predictions based on values that would not have mattered anymore.

Because of this, we turned towards feature engineering for our second model. We created six variables that represented the aggregate running statistics of a teams' past games in the season to predict the odds in a current game. These six variables are, for each of the home and away teams in a game, the number of wins, the win rate, and the average difference in goals, all based on past games up to the current. The average difference in goals would be negative if a team had more goals scored against them than for them. This approach would allow us to make more relevant predictions.

Our third and final model to answer this research question was a simple multivariate linear regression using forward selection on the six engineered features. This model was made to further assess the performance of our second model and see if the performance of a method without machine learning.

RQ2

In order to predict the result of games, we used a Poisson model modified from David Sheehan's model [8]. Since we predicted the remaining games of the current 2018/2019 season, we thought using the dataset only from 2018/2019 season will give us the most accurate prediction. We used the name of the home team and away team, the number of goals of each team, and the average betting odds of each results from 6 betting websites from the dataset [2]. Using these features, we converted into Team, opponent, Number of goals scored, betting odds for a team, draw, and opponent, and whether the team was home or away.

Using the dataset that we mentioned above, we used the Poisson Model to predict the number of goals the team will score in each match. The team that scored more goals in the previous matches are likely to score more goals in future matches. Also, we used the betting odds because we could figure out audiences' thoughts about which team is likely to win. Since the number of goals cannot be a decimal number, we rounded each number of goals to the nearest integer when calculating the result. Following the rules of the league, the winner of the match gets 3 points, if there is a draw, each team gets 1 point, and the losing team gets 0 points.

RQ3

In order to produce a guess for if games ended up in a draw, Home team win, or Away team win, we ran classification models against each season as well as all seasons combined using features we generated. We used the data points at that time for the average all time score, highest all time score, lowest all time score, number of goals total, number of shots total, and accuracy, for both the home teams and away teams. For the combined dataset, all of the factors are localized to the season the data point came from, meaning there are no lifetime data points.

We ran a support vector classification (SVC), scaled support vector classification (Scaled SVC), and a scaled k neighbors classification (Scaled KNC) model. We chose these models based on a resource by Scikit Learn that suggests what models to use against certain datasets and outcomes [4]. We split each dataset into 80% for training and 20% for validation. We ran a grid search on these models, varying the kernel and degree on the SVC and Scaled SVC models, and the number of neighbors and weights for the Scaled KNC model.

4 RESULTS

RQ1

For all three models, we used the negative mean absolute errors of their predictions to score all three types of betting odds (home, draw, away). Table 1 compares the results of all three models, on all three types of betting odds where closer to 0 is better.

Table 1: Model Accuracy for Each Type of Betting Odds

Model	Home	Draw	Away
1	-1.279	-0.938	-2.747
2	-0.822	-0.645	-1.681
3	-1.007	-0.919	-2.235

Table 2 displays the first few predicted betting odds for the upcoming games for the 2018/2019 season. It also shows the home (H) or away (A) team that has the higher win rate (WR), as well as which has the lower odds (O). Table 3 shows the home and away teams playing at each of the games.

Table 2: Predicted Betting Odds for Upcoming Games

Game	Home	Draw	Away	WR	O
1	2.665556	3.300877	3.101404	A	H
2	2.231759	3.344211	3.743860	H	H
3	3.110556	3.361754	2.630263	A	A
4	1.815000	3.637368	5.066667	H	H
5	1.380926	5.542368	10.513684	H	H

Table 3: Upcoming Games as of Time of Writing

Game	Home Team	Away Team
1	Cardiff	West Ham
2	Crystal Palace	Brighton
3	Huddersfield	Bournemouth
4	Leicester	Fulham
5	Man City	Watford

RQ2

We see in Table 4 the predicted home goals (HG) and away goals (AG) for the first 5 results of the matches before March 9, 2019. Table 5 shows the actual results. For the first game in the table, we predicted that Cardiff will score 1 goal, and West Ham will score 2 goals and win the game. The predicted goals do not entirely match with the actual result. As of March 17, our accuracy of predicted match results is 40%.

We see in Table 6 the comparison of current top 6 teams and predicted top 6 teams. As a result, our model predicted that Manchester City will win the season with the 98 points.

Table 4: Predicted Match Results in 18/19 season

Home	Away	Result	HG	AG
Cardiff	West Ham	A	1.0	2.0
Crystal Palace	Brighton	D	1.0	1.0
Huddersfield	Bournemouth	A	1.0	2.0
Leicester	Fulham	H	2.0	1.0
Man City	Watford	H	4.0	1.0

Table 5: Actual Match Results on March 9, 2019

Home	Away	Result	HG	AG
Cardiff	West Ham	H	2	0
Crystal Palace	Brighton	A	1	2
Huddersfield	Bournemouth	A	0	2
Leicester	Fulham	H	3	1
Man City	Watford	H	3	1

Table 6: Top 6 on March 2nd / Predicted Top 6 in 18/19 season

Ranking	Team	Points	Team	Points
1	Man city	71	Man city	98
2	Liverpool	70	Liverpool	97
3	Tottenham	61	Tottenham	82
4	Man United	58	Chelsea	78
5	Arsenal	57	Man Utd	74
6	Chelsea	56	Arsenal	74

RQ3

We see in Table 7 the accuracy of each model against each dataset, along with a baseline of always betting for the Home team. The SVC models all performed better than their respective Scaled KNC models. However, between SVC and Scaled SVC, the more accurate model varies. We see that the SVC models are able to predict about 50% of game results.

Table 7: Model accuracy per dataset

Season	SVC	Scaled SVC	Scaled KNC	Home
2014/2015	0.5658	0.5526	0.5395	0.4526
2015/2016	0.4605	0.4474	0.4342	0.4132
2016/2017	0.6447	0.6579	0.5526	0.4921
2017/2018	0.5395	0.5000	0.4605	0.4553
2018/2019	0.6034	0.6552	0.5345	0.4740
All	0.5055	0.5138	0.4779	0.4566

For all the SVC and Scaled SVC models, the best kernel parameter ended up being "linear", and the most accurate degree was 2. The Scaled KNC parameters for neighbors were all different, except the most optimal parameter for weight was always "distance".

5 DISCUSSION

RQ1

The first model, a grid search using variables already provided in the dataset, performed worst. We found this reasonable, as it should not make sense to use the end statistics of a game to determine its odds, which would have been finalized before the game ended. The second model performed best, as we applied feature engineering with more relevant statistics while utilizing 10-fold cross-validation and k-neighbors regression. The third model, a multivariate linear regression with forward selection on the engineered features, came in the middle of the first two. All these results came in line with our expectations, and it arguably shows to a certain extent the validity of our assumptions and approach.

To draw further observations, we calculated the percentage of games predicted to have the same team having both the higher win rate and the lower betting odds. In the real world, the team with the higher win rate normally has lower betting odds. Our predictions ended up with 81% of games having the same team for both higher win rate and lower odds. This indicates that for the remaining 19% of games, the team with the higher win rate is predicted to have higher betting odds. While this does not necessarily indicate match fixing, where under-performing teams will have lower odds, there might be trends here that were unaccounted for.

Our results show that our model is capable of making better predictions as we get deeper into the season, and that it performs very well with minimal overfitting. As for the 19% of games predicted to have different teams possessing the higher win rate and lower odds, this might be due to insufficient knowledge for the final games, but our model will be even more informed and updated of teams' performance by then.

RQ2

Our results addressing our second research question show that our model can predict the results of the game based on betting odds and the number of goals that each team scored in prior matches. We were not able to calculate the entire accuracy of the model because we do not know all the results yet. However, if we compare the predicted results of the games from March 9 to March 17, the accuracy was only 40%.

We cannot conclude any match fixing is taking place using this model.

RQ3

Our results addressing our third research question show that our model can make logical predictions on what the outcome of the game will be based on previous team performance. We show that our models are better estimators than guessing the Home team will win for each game. The dip in model performance in 2015/2016 could be due to a lot of factors. One of our guesses is that home team performance dropped and as such affected certain aspects of the model.

However, roughly 50% accuracy is not good enough to start betting large amounts of money on. We also learn from this that our factors do have a connection to the game outcomes. Our list is not comprehensive, and because football is such a complex game, we can adjust our models with an infinite amount of different parameters for predictions.

6 FUTURE WORK

Overall, our models can be adjusted to use different engineered factors to greatly improve predictions. We currently only use factors like previous game scores and previous shots, but those factors are disconnected from other information we may know about the team itself and not their performance. One example of a factor we could look at are team members and how teams perform with that player on the team or not over years of that player being traded. Also, we could look at the coach's preference and winning odds by tactics that he used in a match. As for models made for RQ1, we would look into why some teams are predicted to have both the higher win rate and higher betting odds, which isn't regular.

Future work could also look into using different methods of feature selection before running each model.

REFERENCES

[1] BWW News Desk. 2018. Record 39.3 Million Americans Watched NBC Sports' Presentation of the Premier League in 2017-18. https://bit.ly/2ObyQVE

[2] Football-Data. 2019. Data Files: England. http://www.football-data.co.uk/englandm.php

[3] Junyuan Gao and David Aldous. 2016. Predicting Premier League Final Points and Rank Using Linear Modeling Techniques. https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Junyuan_Gao.pdf

[4] Scikit Learn. 2018. Choosing the right estimator. https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

[5] Alan J. Lee. 1997. Modeling Scores in the Premier League: Is Manchester United Really the Best? *CHANCE* 10, 1 (1997), 15–19. https://doi.org/10.1080/09332480.1997.10554791 arXiv:https://doi.org/10.1080/09332480.1997.10554791

[6] Clair Newell, Holly Watt, and Ben Bryant. 2013. Football match-fixing: six arrested by police investigating betting syndicate as rigging hits British game. http://bit.ly/2TtvGlT

[7] Nazim Razali, Aida Mustapha, Faiz Ahmad Yatim, and Ruhaya Ab Aziz. 2017. Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL). *IOP Conference Series: Materials Science and Engineering* 226 (aug 2017), 012099. https://doi.org/10.1088/1757-899x/226/1/012099

[8] David Sheehan. 2017. Predicting Football Results with Statistical Modelling. https://dashee87.github.io/football/python/predicting-football-results-with-statistical-modelling/

[9] UEFA. 2017. The European Club Footballing Landscape. , 53 pages. https://www.uefa.com/MultimediaFiles/Download/OfficialDocument/uefaorg/Clublicensing/02/58/98/12/2589812_DOWNLOAD.pdf