Team Prediction in the English Premier League

SENG 474

Alastair Beaumont
University of Victoria
a.beaumont11@gmail.com

Kolby Chapman
University of Victoria
kol.j63@gmail.com

Graeme Nathan
University of Victoria
graemenathan@gmail.com

Cole Peterson
University of Victoria
colpeterson@gmail.com

I. Introduction

This project will mine English Premier League (EPL) data to predict the winners of soccer games. One factor we are particularly excited to consider is the threat of relegation, as it is a rather unique and important element of the league. We intend to join information from multiple data sources and test many different algorithms on the project, as many freely available data set are lacking, and many different algorithms have shown success on similar problems.

II. Related work

Predicting sports performance is a well studied field. Recently, fantasy sport websites like FanDuel and Draft Kings have opened up new gambling opportunities where machine learning techniques can be applied to predict positive returns [1], and there has been increased interest in the area [2] as the industry grows and becomes more profitable. Sports have valued data mining techniques as it helps eliminate the high emotional stakes the field carries for many of its experts which tend to bias predictions [3]. Sport results can have tremendous world-wide impact, the results of soccer matches have even been shown to affect stock investor behaviour [4], and so accurate prediction can be incredibly valuable outside of the stadium. Most research has investigated predicting the outcome of matches, but now there has been emphasis on selecting a fantasy lineup. Other areas of research include using biomechanical measures to gauge player health and fitness to negotiate contract deals [5], and gauge injury risk [6], [7], or examine the impact of mid-season changes to the coaching staff [8], all of which can help make better predictions of game outcome, or select better players for a lineup.

Sports data exhibits interesting paradoxes where classic gambling fallacies such as the "hot-hand" fallacy actually turn out be real, due to the psychological element of sport. Although the authors of [9] found that in general making one basketball shot did not increase your chances of making another, some baseball players were found to be more "streaky" than could be statistically accounted for in [10]. Making more accurate predictions may rely upon teasing apart some of these more complicated and unintuitive elements of sport data, and verifying conventional wisdom like "home-field advantage" is legitimate [11]. Identifying the right features to use in training a machine learning model for sport can be tricky. Some approaches have relied upon expert opinion to select features [12], while others have added single features iteratively, only keeping them if they increased prediction accuracy [13]. Within the game of basketball, a team's performance in the area under the net (short shots and rebounds) was found to have the higher impact on final result than performance anywhere else on the court [14]. As statistics of soccer tend to be relatively sparse compared to other sports, much effort to apply data mining techniques to the game of soccer have relied upon automated processing of video to detect events in matches [15], [16], player speed [17], or more complex data like player position on the pitch to discover offensive patterns [18].

Many classification techniques have been experimented with including Support Vector Machines [19], Neural Networks [20], Bayesian Methods [13], Logistic Regression [13], [19], Fuzzy Models [21], Decision Trees [12], and Markov chain Monte Carlo Methods [22]. It can be difficult to compare the success of these models in different studies because most of them use different data. Additionally, some studies aim to predict binary win/loss, whereas others [23] classify with margin of victory granularity. Many different methods have shown to have statistically significant predictive powers, with some studies reporting prediction accuracy $> 80\%$ [14]. There remains no consensus or conclusion over which methods provide the highest accuracy for sports in general, or the game of soccer. One study [19] found logistic classifiers, naive Bayes, support vector machines, and artificial neural networks to all produce accuracy of $66.82\% \pm 1.00\%$, demonstrating that no method is clearly better over another on the same data.

III. Data description

The data used for this project will be the team stats of soccer teams. Theses stats will include the win/loss record for the team, which team was the home team, the goals scored by each team in the game, and the team's place in the standings. Using these stats we can potentially predict the outcome of the match. The data source for this project is a website called 'www.football-data.co.uk' [24]. This website has a csv file containing the scores of every game in a season. The website contains csv files for many different European leagues. There is a csv file for every season dating back to 1993. There are also other websites that have APIs that can be used if more information is needed

IV. Proposed project

As outlined in the introduction, our project will be predicting the outcomes of two teams meeting in the EPL.

We have two goals. The first goal is to build a classifier that is able to predict the outcome of a game between to member teams of the Premier League with acceptable accuracy (this level of accuracy will be defined further into the project).

The second goal is the investigation of the distance to relegation datum. Some questions we are interested in answer regarding relegation are: Does a team's impending relegation improve their performance? Does the inclusion of data about closeness to relegation improve the accuracy of a classifier?

We will use multiple data sources. Data sets for soccer are limited. To mitigate this, we will combine data from multiple sources and sets to build our own training and testing sets.

We will also use multiple algorithms. Each algorithm will be run on the same sets, and then have their accuracy compared. The limiting factors here will be time and creativity; we will implement as many algorithms as we can think of that we have time for.

We will gather the data of previous games for each team over multiple years. We might even branch out to other leagues to acquire as much information as possible to increase our training data. Once we have acquired as much training data as possible we will make predictions using our algorithms for past games.

Finally, we will explore the effects of the distance to relegation datum on both the accuracy of our predictions and the performance of the teams.

V. Our Approach

VI. Results

A programming framework has been developed which parses the data so that it is easy to work with. The stats from a given season are loaded into a `Season` object, which can be queried for stats for any given team or the league at any date in the season. For example, you can request the number of points that a team has on a given date, or the standings of the league, and the object will ensure that it does not use any future data. The `Season` object can be expanded to include more features, and serves as the interface between the raw data and our developed code. Additionally, a `Predictor` object has been created which uses features from the `Season` object to predict the outcome of matches. At this point, relatively naive predictors have been programmed to determine a baseline. One predictor always chooses the home team, one always chooses the away team. One always predicts the team higher in the standings, and one predicts based only on whether teams had won or lost their last game. The baseline results are shown below over five seasons of EPL soccer.

Predictor	Accuracy
High Points	0.456599190283
Last Game	0.381983805668
Home Team	0.43979757085
Away Team	0.308259109312

It is note worthy how well a home-team predictor performs, rivaling a high-points predictor, and confirming that home field advantage in EPL soccer.

A simple decision tree machine learning approach was also tested, using LOOCV where one season was left out as testing data. Limiting the depth of the tree improved results and avoided over fitting, but the accuracy was not much better than the high-points model. This indicates that more advanced features need to be used, most obviously score.

Decision Tree	Accuracy
Depth = 2	0.479028340081
Depth = 3	0.48012145749
Depth = 4	0.462874493927
Unlimited Depth	0.39044534413

VII. Conclusion and Future Work

Future work will look to create a new points measure which is more accurate at predicting the a team's abilities, which reflects the following intuition: losing to a good team is not as bad as losing to a bad team, and losing by a small margin is not as bad as losing by a large margin, as well as coming up for a model for the influence of the threat of relegation. To take into account relegation, we will create a relegation score for each team. An example relegation score would use the following formula which is a ratio between (0,1):

1 - (`points/team_max_points`)

Where `points` is the current number of points that the team has so far in the season and `team_max_points` is the maximum number of points a team can get based on the number of games they have left in the season. The higher the relegation score, the closer the team is to relegation. This relegation score formula may still need to be improved. We will also want to create an algorithm that determines the outcome of the game by looking at the team's last X number of games. We will want to find an X that will provide the most accurate results. For example, if team A has won 3 of its last 5 games and team B has won 1 of its last 5 games then we will predict that team A will win the game. If it is predicted that both teams will win or both teams will lose, then we will predict a draw for that game. We will run this algorithm for different number of X games to find the X that gives the most accurate results. We will compare these results with the results of the naive algorithms to see if there are any improvements. We will run this algorithm with the relegation score and without the relegation score to see if there are any differences in the outcomes. This will determine if a team being in relegation will affect the outcome of the match.

VIII. Task Breakdown

Based on the feedback from the initial proposal, we have addressed the following concerns. For each game we will try to predict the outcome of the game, not each teams potential outcome.

We will be using features from previous games to predict the outcome of the games. We will look at the win/loss/draw

record of the previous games for each team. Based on the teams previous number of wins, losses, or draws, we will predict the outcome of the game. For example, if team A has won 3 of its last 5 games and team B has won 1 of its last 5 games then we will predict that team A will win the game. If we see in the history of two teams meeting that there is a tie between wins and losses, this will result in a draw prediction.

We will compare against the naive algorithms that we implemented. We created four naive algorithms. One will always predict the home team will win, another will predict that the away team will win, another will predict that the team higher in the standings will win, and the last one will pick the team that won its last game to win.

REFERENCES

[1] G. Sugar and T. Swenson, "Predicting optimal game day fantasy football teams."

[2] S. Bishop, "Method and system for conducting fantasy sports games on a wide area computer network," 12 2004. [Online]. Available: https://www.google.com/patents/US20040266530

[3] M. Haghighat, H. Rastegari, and N. Nourafza, "A review of data mining techniques for result prediction in sports," *Advances in Computer Science: an International Journal*, vol. 2, no. 5, pp. 7–12, 2013.

[4] A. Edmans, D. Garcia, and Ø. Norli, "Sports sentiment and stock returns," *The Journal of Finance*, vol. 62, no. 4, pp. 1967–1998, 2007.

[5] T. H. Davenport, J. Harris, and J. Shapiro, "Competing on talent analytics," *Harvard Business Review*, vol. 88, no. 10, pp. 52–58, 2010.

[6] C. Carling, F. Le Gall, and G. Dupont, "Are physical performance and injury risk in a professional soccer team in match-play affected ower a prolonged period of fixture congestion?" *International journal of sports medicine*, vol. 33, no. 1, pp. 36–42, 2012.

[7] A. Owen, G. Dunlop, M. Rouissi, M. Chtara, D. Paul, H. Zouhal, and D. P. Wong, "The relationship between lower-limb strength and match-related muscle damage in elite level professional european soccer players," *Journal of sports sciences*, vol. 33, no. 20, pp. 2100–2105, 2015.

[8] A.-L. Balduck, M. Buelens, and R. Philippaerts, "Short-term effects of midseason coach turnover on team performance in soccer," *Research quarterly for exercise and sport*, vol. 81, no. 3, pp. 379–383, 2010.

[9] T. Gilovich, R. Vallone, and A. Tversky, "The hot hand in basketball: On the misperception of random sequences," *Cognitive psychology*, vol. 17, no. 3, pp. 295–314, 1985.

[10] J. Albert, "Streaky hitting in baseball," *Journal of Quantitative Analysis in Sports*, vol. 4, no. 1, 2008.

[11] R. P. Schumaker, O. K. Solieman, and H. Chen, *Sports Data Mining*. Boston, MA: Springer US, 2010, ch. Predictive Modeling for Sports and Gaming, pp. 55–63.

[12] E. Zdravevski and A. Kulakov, *System for Prediction of the Winner in a Sports Game*. Springer, 2010, pp. 55–63.

[13] D. Buursma, "Predicting sports events from past results towards effective betting on football matches," vol. 21, 2011.

[14] Z. Ivanković, M. Racković, B. Markoski, D. Radosav, and M. Ivković, "Analysis of basketball games using neural networks." IEEE, 2010, pp. 251–256.

[15] M. Sykora, P. W. H. Chung, J. P. Folland, B. J. Halkon, and E. A. Edirisinghe, *Computational Intelligence in Information Systems: Proceedings of the Fourth INNS Symposia Series on Computational Intelligence in Information Systems (INNS-CIIS 2014)*. Cham: Springer International Publishing, 2015, ch. Advances in Sports Informatics Research, pp. 265–274.

[16] S.-C. Chen, M.-L. Shyu, M. Chen, and C. Zhang, "A decision tree-based multimodal data mining framework for soccer goal detection," vol. 1. IEEE, 2004, pp. 265–268.

[17] A. Redwood-Brown, W. Cranton, and C. Sunderland, "Validation of a real-time video analysis system for soccer," *International journal of sports medicine*, vol. 33, no. 8, p. 635, 2012.

[18] J. Van Haaren, V. Dzyuba, S. Hannosset, and J. Davis, *Automatically Discovering Offensive Patterns in Soccer Match Data*. Springer, 2015, pp. 286–297.

[19] C. Cao, "Sports data mining technology used in basketball outcome prediction," 2012.

[20] A. McCabe and J. Trevathan, "Artificial intelligence in sports prediction." IEEE, 2008, pp. 1194–1197.

[21] K. Trawiński, "A fuzzy classification system for prediction of the results of the basketball games." IEEE, 2010, pp. 1–7.

[22] H. Rue and O. Salvesen, "Prediction and retrospective analysis of soccer matches in a league," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 49, no. 3, pp. 399–418, 2000.

[23] A. P. Rotshtein, M. Posner, and A. Rakityanskaya, "Football predictions based on a fuzzy model with genetic and neural tuning," *Cybernetics and Systems Analysis*, vol. 41, no. 4, pp. 619–630, 2005.

[24] Football-Data, "Football-data.co.uk." [Online]. Available: http://football-data.co.uk/