

Numerical Algorithms for Predicting Sports Results

by Jack David Blundell,¹ School of Computing, Faculty of Engineering

ABSTRACT

Numerical models can help predict the outcome of sporting events. The features within these models rely on data associated with the competitors. Here, a logistic regression model was used to predict the outcome of American Football matches. As well as using data such as the two teams' previous results, novel features such as stadia size and the distance the away team had to travel were incorporated. This model was seen to perform better than two baselines: (i) simply predicting the home team or (ii) using the previous result between the teams. By adding more informative features to the model, a prediction accuracy of 65% was produced. These predictions matched that of bookmakers showing that the information enclosed within the model could be used to predict games to a highly successful rate.

Keywords: American Football; bookmakers; logistic model; machine learning; regression; sporting predictions

¹ email address for correspondence: jackblundell@hotmail.co.uk

Numerical Algorithms for Predicting Sports Results

INTRODUCTION

The aim of this work was to see if numerical data could be used to predict the outcome of American Football matches. This was done through a logistic regression model. Informative features were required to model each football match so that a successful prediction could be made. Using these features, the model would produce a binary output representative of whether the home or away team was more likely to win.

During the background research, it was noticed that few authors had attempted to predict the outcome of sporting events using the logistic regression approach. Therefore, this project aimed to see if this method was as successful as the alternatives used in other papers. As well as this evaluation, I compared the predictions obtained from my model with that of the bookmakers'. It was shown that my model fared no worse than these betting forecasts showing that the model was a good predictor of American Football games.

DATA

American football

The sport that was to be used in this project needed to have a caveat that match outcomes are rarely tied. This would make predicting matches easier through simply selecting the home or away team to win. American Football was therefore chosen as ties are infrequent within the sport.

Games

I used the 8063 American Football matches played within the NFL between 1970 and 2006. The results were downloaded from <http://www.pro-football-reference.com>. As 45 of these were tied matches, they were ignored in this project, leaving a final 8018 matches.

STADIUM INFORMATION

I obtained the stadium that each team played at within the relevant 37 years from http://en.wikipedia.org/wiki/Chronology_of_home_stadiums_for_current_National_Football_League_teams. This enabled me to store the size of each stadium and using an online city-to-city distance tool (<http://www.geobytes.com/CityDistanceTool.htm>), I was able to get the distances between each of these stadia.

BOOKMAKER SPREADS

To assess the accuracy of the final model, as well as comparing it to the final outcome it was evaluated against the forecasts put forward by the bookmakers for each game. This data was retrieved from Prof. Philip Gray who had carried out similar work in the field of American Football predictions (Gray & Gray 1997).

METHODOLOGY

Baselines

To get an idea how successful the numerical model was, I used simple prediction methods to act as system baselines. The first relied on the theory that the home team within a sporting event has a certain advantage over the away team (Vergin & Sosik 1999, Stefani 1980, Harville 1980). Therefore, this basic predictor chose the home team to win within every match (HOME).

The other baseline algorithm used previous results between the two competing teams to achieve a prediction. After analysing the accuracy of differing amounts of data, it was found that the forecasts became less accurate when older results were used. Thus, the most accurate predictions came from just using the previous year's results between the teams (PREV_RES).

When tested on all of the matches between 1970 and 2006, the unsupervised HOME baseline achieved an accuracy of 57.8% and the supervised PREV_RES obtained 58% accuracy. Using the McNemar test (Dwyer 1991) to find the statistical significance at the 0.05 level, they both were found to be significantly better than a random prediction approach and therefore good baselines for the complex model.

Logistic model

A logistic regression model was used to improve on the 58% accuracy found by the baselines. The output of a logistic model is a binary result (1/0) so as tied games were ignored, this made the approach suitable to forecast either a home or away win.

The logistic model encompasses features (attributes) that represent data relevant to the result that is trying to be predicted. Then during the training of the model, the relationship between each feature and the result is assessed to see if it has a strong or weak correlation with the end result. This relationship is then used when trying to predict the outcomes within the testing phase.

The next step was to choose the model's features. These needed to represent information such as the overall strength of each team, the recent form of each team, etc. In order to obtain a set of features that achieved the most accurate representation of an American Football game, an iterative prototype approach was used. This meant potential match attributes could be analysed and either kept or ignored depending on their performance.

Features

The features used within my model centred on work carried out within soccer predictions where these features were encompassed in an ordered-probit model (Goddard & Asimakopoulos 2004). Here, the authors implemented novel features within their model to predict the outcome of soccer games. For instance, they looked at the two competing teams' recent form. The decision was made that American Football and soccer have similar traits and therefore these features could be transferred across the two sports. See Table 1 for a list of features used.

Feature	Description
awaycapacity	The stadium capacity of the away team
awayrecentawayresultn	The result in the n th recent away game for the away team
awayrecenthomerresultn	The result in the n th recent home game for the away team
awaywinratio_1	The away team's win ratio for last year
awaywinratio_2	The away team's win ratio for 2 years previous
distance	The distance the away team had to travel to play the match
homecapacity	The stadium capacity of the home team
homerecentawayresultn	The result in the n th recent away game for the home team
homerecenthomerresultn	The result in the n th recent home game for the home team
homewinratio 1	The home team's win ratio for last year
homewinratio 2	The home team's win ratio for 2 years previous
lastyearresult	The result of last year's corresponding game between the two teams

Table 1. Original Model Feature Set

Here we can see that the recent form of the two teams is captured through the 'recentawayresultn' and 'recenthomeresultn' features. The win ratios of the previous two years represent how successful the teams have been overall in their matches during recent seasons. The capacities of the two stadia try to capture the number of fans a team has, and how this affects the chances of them beating the other team. The 'distance' feature shows how the number of miles an away team has to travel can influence the probability of them becoming victorious. Lastly, the 'lastyearresult' feature borrows from the PREV_RES algorithm in that it looks to see if predictions can be aided through previous encounters between the teams.

The features associated with each match were extracted from the list of results and the stadium data. Then the features were placed into vectors with the result of the match appended at the end. Then using the machine learning software WEKA (Witten & Frank 2005), the model was trained using 20 years of NFL matches and tested on the remaining eight years.

The logistic model reached an accuracy of 62%, which proved to be significantly better than the two baselines described previously. This showed that the features incorporated in the model, when used together held more predictive qualities than simply picking the home team or relying on last year's results between the teams.

Attempts were then made to try to build on this model and improve its forecasting ability. The first included adding 'power rankings' to the feature space. Power rankings are a way in the media to represent a team's current strength based on current form, players injured, etc (Boulier & Stekler 2003). The power rankings at the end of each season were obtained and used within match vectors for the following season. However, this extended feature set decreased the accuracy of the model. Upon further inspection, it was deduced that the rankings should be used and updated on a week-by-week basis. Thus when last year's power rankings were used to represent a team's strength in the following season, the predictive qualities of the rankings were lost.

A further attempt was made to improve the system by altering the feature set from the original model. One feature represented the recent home and away games for the two teams. Here, the result of the recent game is stored, i.e. home/away/tie. This is more suited to soccer as the differences in score are rarely large. So to make the model more suited to American Football, the differences in scores were stored rather than simply just the result. Although this slightly improved on the 62% achieved in the original model, the difference was not found to be statistically significant. See Table 2 for the list of features used in the final model.

Feature	Description
awaycapacity	The stadium capacity of the away team
awayrecentawayscdifn	The score difference in the n th recent away game for the away team
awayrecenthomescdifn	The score difference in the n th recent home game for the away team
awaywinratio_1	The away team's win ratio for last year
awaywinratio_2	The away team's win ratio for 2 years previous
distance	The distance the away team had to travel to play the match
homecapacity	The stadium capacity of the home team
homerecentawayscdifn	The score difference in the n th recent away game for the home team
homerecenthomescdifn	The score difference in the n th recent home game for the home team
homewinratio 1	The home team's win ratio for last year
homewinratio 2	The home team's win ratio for 2 years previous
lastyearresult	The result of last year's corresponding game between the two teams

Table 2. Final Model Feature Set

EVALUATION

The model shown here was found to have more predictive qualities than simply choosing the home team. Furthermore, the model is more accurate than just allowing the results of last year's encounters to decide the forecast.

The final model was then evaluated against the predications of the bookmakers. This was based on the view that the bookmakers are generally the most accurate forecasters when compared with statistical systems or expert predictions (Harville 1980, Bouiler & Stekler 2003). The model achieved an accuracy of 65.2% whereas the bookmakers obtained a rate of 67.4%.

Having said this, the difference in accuracies was not found to be statistically significant. This means that the forecasts made by the model developed in this work were no worse than that of the bookmakers'.

When compared to the accuracies of other models used within similar research, the system described here is not as accurate as some of the more complex models. However, during the analysis of these more complex models the respective authors found that their models were less accurate when compared to bookmaker's predictions. For example, Stefani's least square model obtained an accuracy of 68.4% but was subsequently overshadowed by the bookmaker's accuracy of 71% on the same matches (Stefani 1980).

The system can also be seen to be more accurate than subjective predictions of an expert found within one piece of research. Here, it was discovered the New York Times Editor had a forecasting accuracy of 59.7% (Bouiler and Stekler 2003).

Further analysis was carried out to assess the informative qualities of each feature. Known as feature ablation, this involved taking out each of the features one at a time to create a new 'temporary' model. This temporary model was then used to predict the results of each game and if it was more accurate than the original then it can be concluded that the feature removed is a damaging one. Whereas a feature that decreases the accuracy is proposed to be an important feature. Although the feature ablation studies were not wholly conclusive, it could be seen that a team's win ratios from the previous years are very informative and thus could be extended within the model. Furthermore, the model uses each team's previous nine home games to represent their current form. The studies suggested that cutting this down to around

four or five would be more suitable. These would be interesting avenues to pursue if the work of this project were extended.

RELATED WORK

As mentioned within this work, the features within the model described here were adapted from that of the model illustrated by Goddard and Asimakopoulos (2004). Used to predict the outcome of English soccer matches, they found that their ordered-probit model could compete with the forecasts made by the bookmakers.

With regard to American Football modelling, one piece of research used an Ordinary Linear Squares approach to forecast NFL results (Stefani 1980). However, this approach fell short when compared to the forecasts made by the bookmakers.

Moreover, Stefani used this same model to predict the outcome of soccer matches and basketball games. This allowed him to achieve similar prediction accuracies compared to the American Football version. This suggests that most sports can be interchangeable when representing them within a model.

Harville's approach relied on a much more complex system of linear models to predict the winner of an American Football match (Harville 1980). He obtained a prediction accuracy of 70.3% which in turn was overshadowed by the corresponding bookmakers forecast rate of 72.1%.

CONCLUSION

I have shown that American Football matches can be accurately modelled through the use of features within a regression model. Furthermore, I have found that a simple logistic model can achieve just as accurate forecasts compared to some more complex alternatives. This is shown by the model's ability to provide forecasts that match those of the bookmakers.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Katja Markert for all her time, help and support throughout the project. Also, I want to acknowledge Dr. Andy Bulpitt for his important comments within the development of the project.

Lastly, I would like to thank Prof. Philip Gray for his time and effort providing bookmaker's data which enabled me to make huge strides in the evaluation of my work.

REFERENCES

- Boulier, B.L. and Stekler, H.O. (2003), 'Predicting the outcomes of national football league games', *International Journal of Forecasting*, 19(2), 257–270.
- Dwyer, A.J. (1991), 'Matchmaking and mcnemar in the comparison of diagnostic modalities' *Radiology*, 178(2), 328.
- Goddard, J. and Asimakopoulos, I. (2004), 'Forecasting football results and the efficiency of fixed-odds betting', *Journal of Forecasting*, 23 (1), 51–66.
- Gray, P.K. and Gray, S.F. (1997), 'Testing market efficiency: Evidence from the NFL sports betting market', *Journal of Finance*, 52(4), 1725–1737.
- Harville, D. (1980), 'Predictions for national football league games via linear-model methodology', *Journal of the American Statistical Association*, 75(371), 516–524.
- Stefani, R.T. (1980), 'Improved least squares football, basketball, and soccer predictions', *Systems, Man and Cybernetics, IEEE Transactions on*, 10(2), 116–123.
- Vergin, R.C. and Sosik, J.J. (1999), 'No place like home: an examination of the home field advantage in gambling strategies in NFL football', *Journal of Economics and Business*, 51(1), 21–31.
- Witten, I.H. and Frank, E. (2005), *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.