Case Studies Project 2 Part 1

Sian Brady and Katy Reid

February 2024

1 Introduction

This report investigates whether football match results can be accurately forecasted based on the recent form of the two playing teams. The dataset contains the results of matches played in the English Premier League between the years 2000 and 2015. The results of around 6000 matches were used to train and test a prediction model with the following results:

- The model had an overall prediction accuracy of only 51%. Given that only around half of the predictions are correct, this is fairly unreliable, and tells us that this model cannot accurately predict results based only on past games.
- The model does not predict draws, as it is always more likely for a team to either win or lose. This means that the model can never accurately predict a draw.
- The model is better at predicting home wins than away wins.

2 Dataset

The data that we are using to build and test our model contains information on 6840 English Premier League football matches from the years 2000 to 2015. For each match, we are given the names of the teams that played, as well as information on which team won, or if the result was a draw. We also have information regarding the recent performance of each team; namely the results of their last five matches. We aim to predict the outcome of a match based on a team's performance in the past five games, so any observations for which there are missing match results is removed. This leaves us with 5940 observations which we use to build our model. Removing this data may introduce bias into our model and reduce the accuracy of our predictions.

We assigned point values to previous match results - 3 points for a win, 1 point for a draw, and no points for a loss. The sum of these was found for both teams in each match, and the point difference was calculated as home team points minus away team points. A greater point difference value indicated that

the home team had better recent form than the away team, and vice versa for lower negative scores.

3 Method

Figure 1: Match Result against Point Difference.

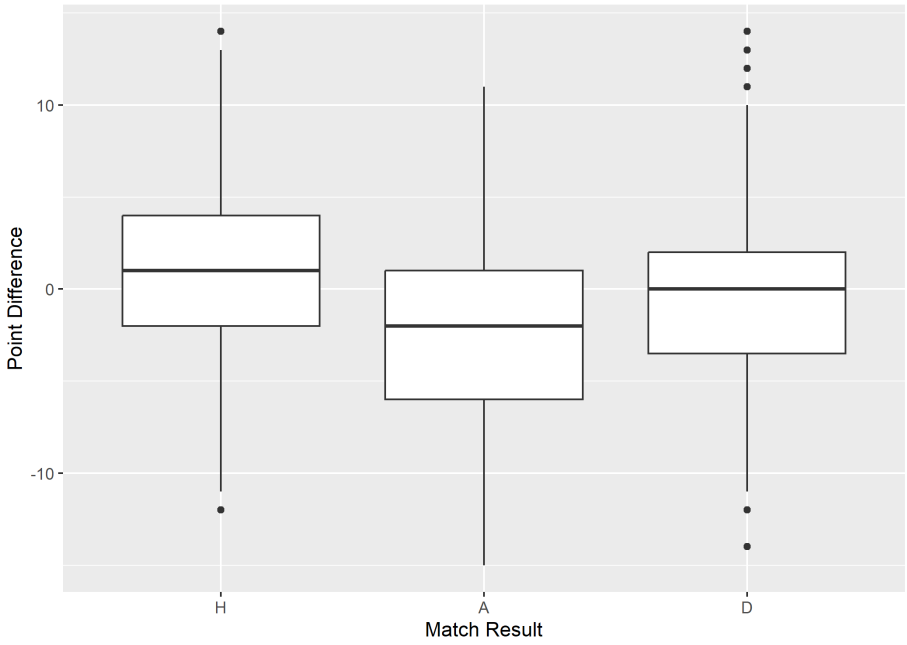

Figure 1 compares the point difference with the match result. As expected, the average point difference in matches where the home team won is greater than in those where the away team won. Point difference for matches with a draw falls in between. All three boxplots show fairly similar quartiles, meaning that variability is consistent across match result groups.

The outcome that we wanted to predict was categorical - either the Home team or the Away team wins or there is a draw. We also considered Linear and Quadratic Discriminant Analysis models, however both predicted less accurately than logistic regression. Therefore, we chose to use logistic regression for our model, using the point difference variable for prediction.

The data was split randomly into ten equally sized subsets of 594 observations, known as 10-fold cross-validation. Training and test sets were then formed by combining subsets; 594 observations (10%) in the test set and 5346 observations (90%) in the training set. The training set was used to estimate the model parameters using the multinomial logistic regression function. The

learned model was then used to predict outcomes, and the error rate of predictions against the test set was calculated. This was repeated ten times, using each subset as the test set. We carried out this process 100 times to find an average accuracy rate.

4 Results

Table 1: Confusion Matrix.

		Reference		
		H	A	D
Predicted	H	246	89	115
	A	44	54	45
	D	0	0	0

Table 1 shows the confusion matrix for the logistic regression model. The model will always predict the match result with the highest probability, so the model always predicts a win with higher probability than a draw. Hence the model does not predict draws, which explains why this is a zero-row. The accuracy of this model is 50.59%, which is quite low.

54.67% of home team wins are correctly predicted, but only 37.76% of away team wins. This could potentially be explained by there being fewer observations of away team wins than home wins in the data set (27.74%), resulting in less data for the model to train with.

5 Model Comparison

We compared our logistic regression model to a baseline model, where we assumed that the home team always wins. The baseline model gave an average accuracy of 46.9%. This suggests that the logistic regression model does add some value to the predictions, as it has a slightly higher accuracy of 50.59%. To compare the models, we also calculated their logarithmic and Brier scores, as shown in Table 2. We can see that the logistic regression model gives lower scores, which further supports the conclusion that it is a more accurate prediction model. However, we see that the difference in scores between the two models is relatively low, as well as the accuracies being fairly similar. This implies that the logistic regression model is only marginally better at predicting the match outcomes than the baseline model. We could improve the accuracy of the logistic regression model by adding more factors than solely the teams' recent form.

Table 2: Model Scores.

	Logistic Regression	Baseline
Logarithmic	410.98	450.48
Brier	0.20	0.28

6 Conclusion

The results verify that a logistic regression model forecasts football match results based on recent form with relatively weak accuracy. Given the significant computational effort, a 50.59% accuracy rate is low, and so further alterations to the model are needed before it could be implemented for prediction. The model is better at predicting home team wins than away team wins. Due to missing observations, 13% of the original data-set was removed, which is likely to have introduced bias which would affect accuracy. Further factors such as team location and finances could influence match results, and may be considered. Adding more variables however does risk over-fitting and increasing inaccuracy, so should be done with caution. Use of a larger sample size, perhaps including other football leagues, could also improve the accuracy of the model.