Premier League Prediction — Mac & Fromage

Hubert Grillier, *Charles* Kanter, *Arnaud* Maranges, *Thomas* Michel, and *Jack* Rubin

Abstract. We use Premier League Football data from 2003-2022 for exploratory analysis and the creation of machine learning models to predict match outcomes. Through two different data structure approaches, we use in-game match statistics and betting odds from different bookmakers to construct betting strategies to generate significant profit.

1 Data Description

The Premier League is the highest level in the English football league system. The dataset from *football-data.co.uk* consisted of Premier League data from the years 2003-2022, with over 100 features in four categories: Results Data (match results), Match Statistics, Betting Odds Data, Total Goals Betting Odds and Asian Handicap Betting Odds.

All data is structured in sequential order, by season, from the first game played by the first team to the last game played by the last team. It is ordered by home versus away from the start of the season to the end of the season, by date and time in ascending order.

2 Data Preprocessing

Data Cleaning:
Certain seasons' datasets (2003-2004, 2004-2005, 2014-2015) contained missing data, so we deleted the rows and columns which gave us issues.

Feature Selection:
We only used features that appeared consistently in each season. In-game statistics were documented well, but only a certain few bookmakers had consistent data, [1], and we only used features that appeared consistently in each season[2].

Rather than attempting to develop a model to make better predictions than the bookmakers, we aimed to create models that would identify the best betting strategy using any pre-game information publicly available. Thus, we used odds of several betting sites to develop our models as this information would be available before each game.

3 Exploratory Analysis of Data

The English Football league has a relegation/promotion system in which the bottom three teams are replaced each year from the Premier League, the top division. Thus, a total of 40 teams played in the league since 2003 (Figure 1), with only seven teams present in every season (highlighted in red).

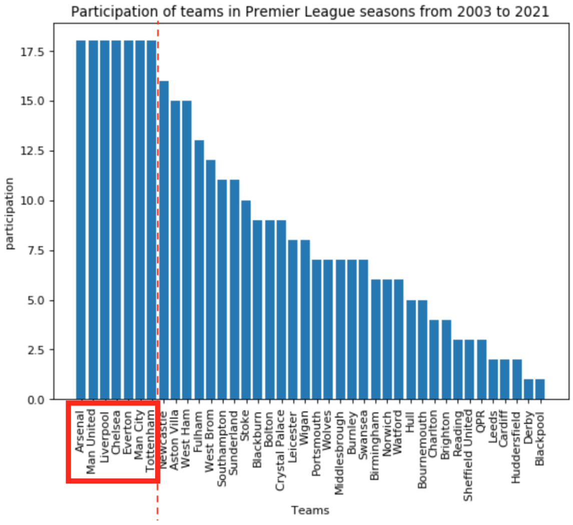


Figure 1. Team participation in Premier League from 2003 to 2021

Game location had impact as well; teams behaved differently when playing at home vs. away (Figure 2). The strongest teams, especially the aforementioned seven, win their games at home and loose or draw about the same number of games (similar for away). Therefore we may conclude that the stronger a team is, the less the location matters. This was consistent across all seasons.

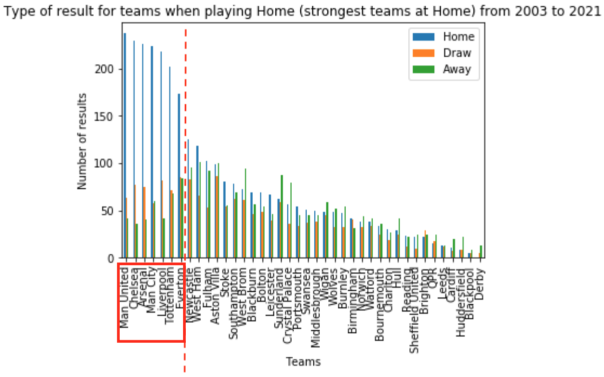


Figure 2. Result for each team when playing at home in the Premier League from 2003 to 2021

Accordingly, the odds follow this pattern (Figure 3). The strongest teams always have the weakest odds, as

[1] Consistent bookmakers were Bet365 ('B365'), William Hill ('WH'), BetWin ('BW'), and Interwetten ('IW').

[2] Features chosen: 'WHD': William Hill Draw Odds, 'B365D': Bet365 Draw Odds, 'IWD': Interwetten Draw Odds, 'HY': Home Yellow Cards, 'HS': Home Team Shots, 'B365H': Bet365 Home Odds, 'HF': Home Team Fouls Committed, 'HST': Home Team Shots On Target, 'HR': Home Red Card, 'WHH': William Hill Home Odds, 'FTHG': Full Time Home Goals, 'HC': Home Team Corners, 'HTHG': Half Time Home Goals, 'FTAG': Full Time Away Goals, 'FTR': Full Time Result

they have a higher probability of winning. Respectively, weaker teams have lower probabilities to win, and thus higher odds. This was consistent across all seasons for both home and away.

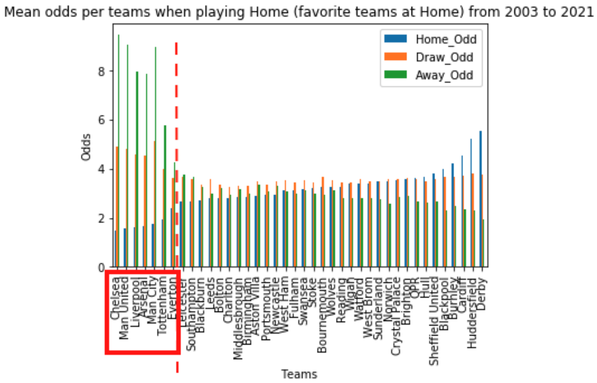


Figure 3. Odds for each team when playing home in the Premier League from 2003 to 2021

We then examined how goals scored and goals conceded affected odds (Figures 4, 5). Indeed, better teams concede less or the same amount of goals and score at least as many goals as their opponents. Thus, the best teams are those with the best attack (goals scored) as well as the best defense (goals conceded). The top seven teams (in red) are consistently among the best in both categories.

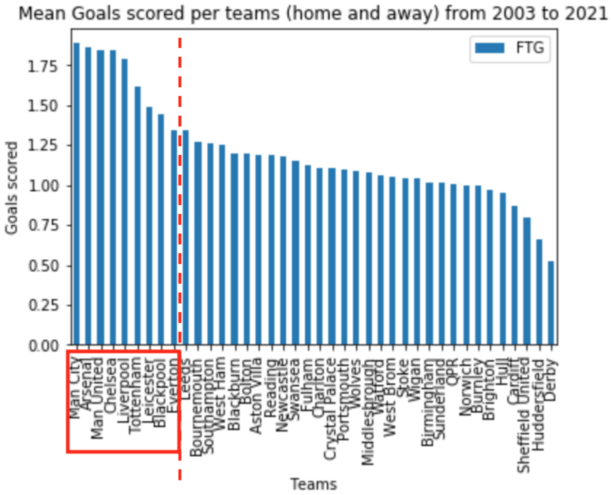


Figure 4. Odds for each team when playing home in the Premier League from 2003 to 2021

Finally, since our goal was to ultimately beat the bookmaker, we looked at how bookmakers determined their odds. Figure 6 compares perfectly fair odds (red) with bookmaker odds (blue). Evidently, bookmakers are precise in predicting outcomes and on average, betting odds have a slight anti-player bias; on average one would get negative returns when betting randomly.

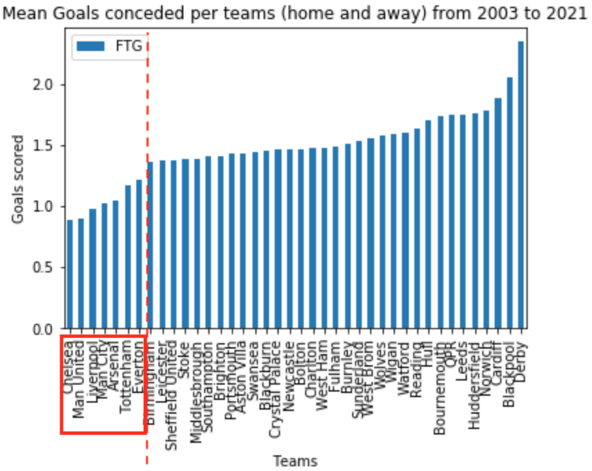


Figure 5. Odds for each team when playing home in the Premier League from 2003 to 2021

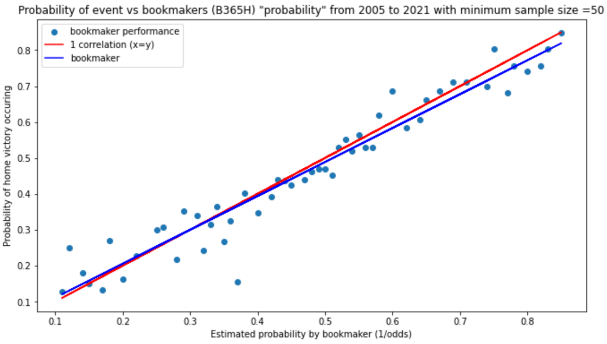


Figure 6. Bookmaker performance, 2005-2022

4 Approach I (Split by Team)

We wanted to predict the outcome of future games without the use of in-game statistics or the results from the game. Thus, we used a rolling average to simulate the in-game statistics of near future games with the averages of the prior n games. The moving average creates a smoothness to the data that can help determine near future results.

For both home and away teams, we created a dictionary of dataframes with the key being team name and the value being a dataframe of the games that team played over the 2003-2021 time period, in ascending datetime order. In addition, we tested this process for different rolling windows (n = 3, 5, 10. . .) and used an exponential moving average (EMA) to account for time decay. Believing 5 games to be a good gauge of a team's "streakiness", we then created our models with a 5-game-span EMA.

Modeling and Betting Strategies:

Following our data preparation, we created a variety of models to help us create a backtested betting strategy to predict win/loss and earn a profit.

We used "One Hot Encoding" to transform the Full Time Result (FTR) feature into 0s for losses (treating

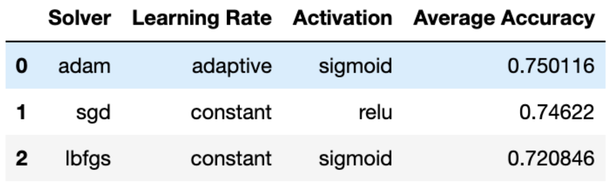
"Draws" as losses for simplicity) and 1s for wins, and used a threshold of 0.5 for classifying predictions. Since EMA created a time series data set, we used 2003-2021 data to train our models and used the current season, 2021-2022, for testing, using "accuracy" as the main determinant of model performance.

We built a simple linear regression model, a neural network, and two ensemble models: a Random Forest classifier and a Bagging classifier. The accuracy results from all four models are depicted in the table below:

Model	Accuracy
Neural Network	0.7601
Regression	0.6707
Random Forest	0.5509
Bagging Classifier	0.5397

Using a neural network required that the testing data and training data be structured identically, so we trained the model using historical team data only with teams that appeared in the testing set.

To determine the optimal neural network parameters, we tested combinations of number of epochs, hidden layers size, solver, learning rate, and activation function[3]. Using 1000 epochs and a hidden layer size of (80,) was most effective, and yielded top results for each solver (Figure 7).



	Solver	Learning Rate	Activation	Average Accuracy
0	adam	adaptive	sigmoid	0.750116
1	sgd	constant	relu	0.74622
2	lbfgs	constant	sigmoid	0.720846

Figure 7. Neural Network accuracy, by parameter selection

Since the Neural Network model was the most accurate, we used it to create the following three betting strategies:

1. **Standard:** Correct prediction wins $*N* and an incorrect one loses $*N*.
2. **Halve Bet on Loss:** If the prior bet was incorrect, then the wager $*N* is decreased to $*N/2* .
3. **Double Bet on Win, Halve Bet on Loss:** If the prior bet was incorrect, then the wager $*N* is decreased to $*N/2* and if the prior bet was correct then the wager is increased to $2*N* (min. bet $1, max. bet $100).

These gave us the following results:

Strategy	Profit
Standard	$1,470
Halve Bet on Loss	$532.97
Double on Win, Halve on Loss	$3,532.50

5 Approach II (One Set)

Our second approach was to create one single dataframe with data from every game in our 2003-2021 dataset (no longer separated by team). This would allow use of more teams and features.

Feature Creation:

For each game, we computed the rolling average, on the ongoing season, for the following features:

- The average number of goals scored by the home / away team when playing home /away
- The average number of goals scored by each team
- The average number of goals conceded by the home / away team when playing home /away
- The average number of goals conceded by each team this season
- The win rate of the home / away team when playing home /away
- The win rate of both teams
- The average odds for home team wins, draw and away team wins

The analysis in Section 3 revealed the existing correlation between the number of goals scored/conceded by each team and their performance in the league. It was therefore important to include the above goal-related features for our prediction model.

Betting odds were again relevant (see Section 2).

Binary predictions

Many betting sites have binary prediction bets (e.g., "will Arsenal win this game?"), so we built binary prediction models using One Hot Encoding to predict win/loss.

Our neural network was unable to process all the data, so using 2015-2021 data for training and 2021-2022 data for testing, we were able to achieve the following:

Solver	L.Rate	Hid.Layer	Activation	Profit
adam	constant	logistic	(30,)	$1250
TPR	**TNR**	**FPR**	**Precision**	**Accur.**
0.507	0.775	0.224	0.565	0.678

For a more comprehensive model, we trained with 2003-2021 data and testing with the 2020-2021 season, the K-Nearest Neighbor (KNN) model turned out to be the most accurate model on the testing set, achieving an accuracy of 0.68 and an f1 score of 0.6 (better than NN).

Figure 7 shows the ROC curve and Figure 8 shows the confusion matrix associated with the KNN model.

More Granular Predictions

For a more complex betting strategy, we considered the three possible outcomes of a soccer game (win, loss, draw).

[3] solvers: ("sgd","lbfgs","adam"), learning rates: ("constant", "invscaling", "adaptive"), and activation functions: ("tanh","relu","sigmoid").

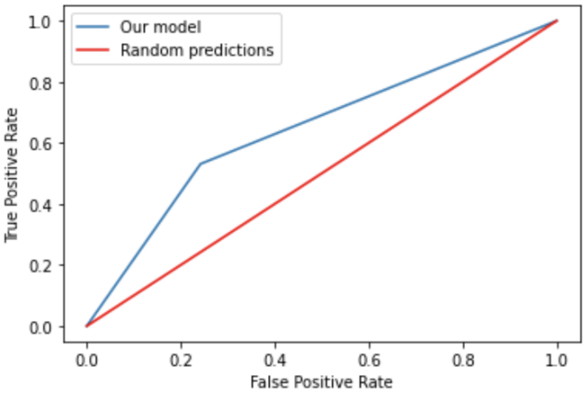
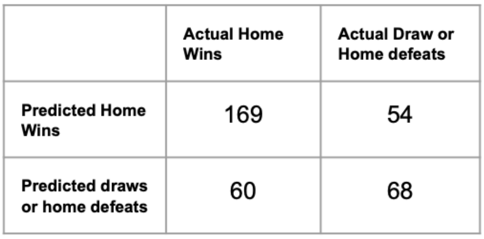
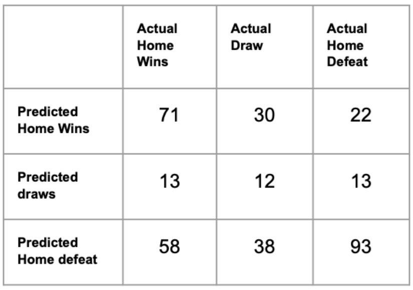


Figure 8. ROC curve



	Actual Home Wins	Actual Draw or Home defeats
Predicted Home Wins	169	54
Predicted draws or home defeats	60	68

Figure 9. Confusion matrix for a KNN binary prediction



	Actual Home Wins	Actual Draw	Actual Home Defeat
Predicted Home Wins	71	30	22
Predicted draws	13	12	13
Predicted Home defeat	58	38	93

Figure 10. Confusion matrix of the model based on Random Forrest classifier

The Random Forrest classifier turned out to be the most accurate model, and after running a cross validation we reached the conclusion that 100 estimators was most effective.

With this we achieved an average accuracy of 0.49. The model was quite effective in predicting win/loss, but inaccurate in predicting draws (see confusion matrix in Figure 9). Draws in soccer are hard to anticipate; games can go from a 0-0 draw to a 1-0 win with a single goal. This is why odds for draws are complex and a key point in our betting strategy.

Betting Strategy:

Finally, we tried to set up a betting strategy based on the prediction made by this model and the odds of the betting sites. The idea was to find some arbitrage in the odds, and therefore a strategy that would make us make some profit during the 2021-2022 season. Betting $10 USD on every game of the 2021 season, according to the predictions of our model, we would have make a profit of **$ 700 USD**. These are satisfying results, as bookmaker odds are designed so the gambler loses money. Indeed, if our model used odds only and no other parameters (betting every time on the most likely result, i.e. the one with the smaller odd), we would end up with no profit.

6 Conclusion

Our study of the Premier League's online betting market revealed that there indeed exists an opportunity to create a winning betting strategy using in-game statistics and the odds provided by bookmakers. The betting strategies we implemented through our predictions models generated significant profit and in the end, our most effective prediction model was the NN from Approach I, with the following results:

Model	**Accuracy**
Neural Network	0.7601
Bet Strategy	**Profit**
Double Win, Halve Loss	$3,532.50

Future Improvements:

To further develop these performance models, we could construct new features, either with current data or sourcing new data, such as the following:

1. **Ball Possession:** Track ball possession per game to analyze a team's general dominance in a game.
2. **Individual Player Stats:** Track player performance to evaluate the general strength of a team as well as its per-game lineup.
3. **Motivation:** Estimate each team's motivation for each game depending on current record, relegation possibility, etc.
4. **Other Competitions:** Track each team's performance in other competitions such as league cups and European Champions League for more details on a team's strength.

Furthermore, we might consider other models such as a model to predict the score spread of a game, rather than just win/loss/draw. From these, we would have more insight into the game outcomes and could make safer bets by betting larger spreads.

Another option would be to create a Contribution Weighted Model based on the *Wisdom of Crowds* theory to identify experts within the crowd in order to outweigh them and make a better forecast.

References

football-data.co.uk