Elo-Enhanced LSTM Model for Accurate Prediction of Football Match Results and Betting Success

Tong Jin

14 March, 2023

Abstract

This paper presents a football match prediction framework that combines LSTM (Long-Short Term Memory) models with the Elo Rating system to predict the number of goals scored by both sides in a match using past match data. The aim is to outperform bookmakers in the betting market. The model is trained and tested on English Premier League data, achieving a mean squared error of 1.2293 and an R-squared of 0.14. In a simulated betting procedure, the model yields a 5.33% total return, demonstrating its ability to outperform bookmakers. However, the model has shortcomings, such as low correct prediction rate for promoted teams and limited input data. Future work may focus on incorporating more data and considering bookmakers' behavior.

1 Introduction

Football match result prediction is an area of research that aims to use data and machine learning techniques to predict the outcome of football matches. Due to the complexity and uncertainty of football matches, traditional prediction methods often need help to capture the complex relationships of football match data. In recent years, with the development of big data technology and advances in artificial intelligence, machine learning techniques to predict the outcome of football matches have become a research hotspot. By analysing match data and historical football match results, researchers can build various machine-learning models to predict the outcome of football matches, such as support vector machines, decision trees, logistic regression and neural networks. These models can use football match-related features to predict match outcomes, such as team performance, player statistics, historical records, etc.

This article will focus on a football match prediction framework that uses LSTM (Long -Short Term Memory) models combined with the Elo Rating method to predict the number of goals (scores) scored by both sides of a match using past match data as input to the model, in the hope of beating the bookies in the betting market for a significant return.

2 Literature Review

Football match result prediction is an area of study that attempts to predict the results of football matches using statistical and machine learning methods. Researchers have recently used a variety of statistical models and machine-learning techniques to forecast football game outcomes. Some new research on football match prediction and betting markets is covered in this literature review.
A statistical model for association football results and betting market inefficiencies was put forth by Dixon and Coles [1997], whose model is still regarded as a proven leader. They found inefficiencies in the betting market and estimated the number of goals made by each team using a Poisson model. At the turn of the century, researchers began experimenting with more single variables to predict the outcome of matches. Goddard [2005] conducted a study using regression models to predict goals and match results. The study compared the performance of different regression models, highlighting the potential of regression models in football match prediction and could help bet market players.
Elo scores were first used by Hvattum and Arntzen [2010] to forecast the results of association football matches. They demonstrated that Elo ratings worked better than other approaches, including logistic regression. HERBINET [2018] evaluated the performance of various machine learning algorithms, showing that the neural network model achieved the highest accuracy, demonstrating the potential of machine learning in football match prediction. Last year, Peters and Pacheco [2022] analysed the role of line-ups in the final scores of football matches using machine learning prediction models they developed and found that the Support Vector Regression model outperformed other techniques in predicting final scores. Additionally, their model was profitable when emulating a betting system using real-world odds data.
In conclusion, over the years, various statistical models and machine learning techniques have been applied to predict the outcomes of football matches. The success of those research highlights the potential of data-driven approaches in football match prediction and provides valuable insights for practitioners in the field of football analytics and betting markets.
This paper will therefore present a football match prediction framework based on the LSTM (Long-Short Term Memory) model, one of the most widely used models in deep learning techniques, combined with Elo Rating, and test the framework's ability to be applied in the real betting market.

3 Data

3.1 Data Resources and Features

The dataset used comes from a football database website: http://www.football-data.co.uk/. It includes:

- From 19 August 2000 to 26 February 2023
- Detailed match data and results from 8,599 English Premier League games

- Betting odds from Bet365 (one of the most famous bookmakers)

The data features we use in our model are present:

- Matches Information
  - Date
  - Home Team Name
  - Away Team Name
  - Home Team Final Goals
  - Away Team Final Goals
- Matches Events
  - Home Team Shots
  - Home Team Shots on Targets
  - Away Team Shots
  - Away Team Shots on Targets
- Other Information
  - Elo Rating
  - Betting Odds

3.2 Data Pre-processing

Data pre-processing is an essential step in the application of machine learning models, including but not limited to removing columns that are not needed in the original data source to reduce the time it takes to read the data and format the original string-type data.
In addition, we need to perform further calculations on the raw data to generate new input variables that can be used for direct input into the model for training and testing. Using the algorithm for calculating the number of shots on target for the away team illustrated in Figure 1 as an example, the same can be done for the average shots-related data for the last three matches between the home team and the away team.

Algorithm 1: Calculate average of last three away shots on target

Input : Dataframe containing football match statistics
Output: Dataframe with average of last three away shots on target

```
Function main():
    // Read data from csv file into pandas DataFrame
    teamshome ← read_csv('football_stats.csv');
    // Sort the DataFrame by match date
    teamshome ← teamshome.sort_values(by=['Date'], ascending=True);
    // Initialize dictionary to store last five away shots on
       target for each team
    teamawayshotsontarget ← {};
    for row in teamshome do
        teamaway ← row['AwayTeam'];
        // If team not in dictionary, initialize empty list
        if teamaway not in teamawayshotsontarget then
            teamawayshotsontarget [teamaway ] ← [];
        end
        awayshotsontarget ← row['AST'];
        // Append away shots on target to list
        teamawayshotsontarget [teamaway ].append(awayshotsontarget);
        // If list is longer than 3, remove the oldest item
        if len(teamawayshotsontarget [teamaway ]) > 3 then
            teamawayshotsontarget [teamaway ].pop(0);
        end
        // If list is exactly 3 items, calculate average and
           add to DataFrame
        if len(teamawayshotsontarget [teamaway ]) == 3 then
            row['AST3'] ← sum(teamawayshotsontarget [teamaway ]) / 3;
        end
        // Update row in DataFrame
        teamshome.iloc[row.name] ← row;
    end
    // Write modified DataFrame to csv file
    teamshome.to_csv('football_stats_modified.csv', index=False);
end
```

Figure 1: Calculation of away team shots on target

4 Methodology

4.1 LSTM model architecture

LSTM (Long Short-Term Memory) is a recurrent neural network model commonly used for sequential data processing that can excel at long-term dependency problems. Its core idea is to introduce a memory unit that can add or remove information when needed, thus avoiding the usual RNN problem of gradient disappearance or explosion.

The LSTM consists of three gating units: the input gate, the forget gate and the output gate. They control the amount of information entering and leaving the memory unit. Suppose x_t is the input to the current time step, h_{t-1} is the hidden state of the previous time step and c_{t-1} is the cell state (memory cell) of the previous time step.

Input gates

The input gate controls the extent to which the input signal x_t enters the cell state c_t. It calculates the output of a Sigmoid function and a tanh function, representing the importance of the input and the current candidate value, respectively. It is formulated as follows:

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \tag{1}$$

$$\widetilde{c}_t = \tanh\left(Wxcx_t + W_{hc}h_{t-1} + b_c\right) \tag{2}$$

where W and b are the parameters to be learned and σ denotes the Sigmoid function. $\widetilde{c}_t$ is the candidate cell state at the current time step.

Forgetting gates

The forgetting gate controls the extent to which the cell state c_{t-1} from the previous time step is preserved at the current time step. It calculates a Sigmoid function and an elemental product of c_{t-1}, indicating the information to be forgotten. It is formulated as follows.

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t \tag{4}$$

where $\odot$ denotes the product of elements.

Output gates

The output gate controls the hidden state h_t of the current time step, which is calculated based on the cell state c_t. The output gate computes the output of a Sigmoid function and a tanh function, indicating the importance of the hidden state and the current candidate value, respectively. It is formulated as follows.

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \tag{5}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{6}$$

where o_t is the output of the output gate.

4.2 The Elo Rating System

The Elo Rating System is a commonly used rating algorithm that is widely used in football leagues. The algorithm updates each team's Elo rating to reflect its relative strength based on the results of the tournament.

Specifically, the Elo algorithm calculates an updated rating based on each team's initial rating and the results of the match. The initial rank can be determined by historical performance or other metrics. Match results are then expressed as wins (1), draws (0.5) or defeats (0).

The basic formula for the Elo algorithm is as follows.

$$R_n = R_o + K\left(S_n - E_n\right) \tag{7}$$

where R_n is the updated team rating, R_o is the initial rating, K is a constant (used to adjust the degree of each update), S_n is the match result (1 for a win, 0.5 for a draw and 0 for a defeat) and E_n is the predicted result, which is calculated as

$$E_n = \frac{1}{1 + 10^{(R_m - R_n)/400}} \tag{8}$$

where R_m is the rank of the opponent.

According to the Elo algorithm, the rank of the winning team increases while the rank of the losing team decreases. Also, the degree of victory or defeat will affect the updated rank change. In this paper's calculation, we choose $K = 10$ and the initial rating $R_o = 1500$.

4.3 Model Settings and evaluation methods

The data are split into training and test dataset (the last 1,759 matches, four total seasons and all matches of the 22/23 season until 26 February). We used a total of six variables - home and away team shots, shots on target and real-time Elo Rating - as input variables to the model, and the number of home team goals and away team goals as target variables (output variables) to the model. After several tests with adjusted parameters, details of the parameters and hyperparameters used in the model are shown in Table 1.

Table 1: Model Parameters and Hyperparameters

Parameter/Hyperparameter	**Value**
Input Size	6
Hidden Size	512
Output Size	2
Number of Layers	2
Loss Function	Mean Squared Error
Optimizer	FusedAdam
Learning Rate	0.0000001
Betas	(0.9, 0.999)
Number of Epochs	100
Batch Size	1
Data Type	Half-Precision Floating Point (Float16)

For model performance evaluation, we have looked at two different metrics and stimulating betting's return rate.

- **Mean Squared Error**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{9}$$

  MSE (mean squared error) is a measure of the magnitude of the error between the predicted and actual results of a model and is the average of the squares of the differences between the predicted and actual values.

- **R-squared (goodness of fit)**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

The R-squared (goodness of fit) measures the degree of model fit, which indicates the proportion of the total variance in the model prediction results that the independent variables can explain.

- **Simulating betting return rate**

$$ROI = \frac{total\ revenue - total\ cost}{total\ cost} \tag{11}$$

Simulating betting return rate is a measure of the profitability of a particular betting strategy and represents the average return per unit bet amount. In addition, it is essential to explain that during the simulated betting process, we chose to take the result of any match with a predicted goal of less than 0.5 as a draw, as the predicted value is not a whole number and cannot be judged as a draw.

5 Results and Analysis

For our final model, we obtain a MSE value of 1.2293 and a R-squared of 0.14, which means that our model predicts the scorelines for a match is 1.10 goals away from the truth. This is quite large, considering the limited input variables and the unpredictability of some unusual scores (such as the recent Liverpool 7-0 Man Utd game). This could be acceptable.
We also ran a simulated bet based on the predicted results (win, draw, and loss) for the 21/22 season to evaluate our model performance in the betting market. We bet an equal amount on 380 matches in the 21/22 season and ended up with a total return of 5.33%. This may not be a significant return in the investment market, but it proves that our model beats the bookies. In addition, during the calculation process of conducting simulated bets, we found very poor predictions for the three upgrade teams, possibly because our independent calculation of the Elo Rating did not consider the possible effects of the promotion and relegation system, leading to a decrease in the accuracy of the model. After removing the data from the matches containing the three upgrade teams, our simulated betting return came in at 14.88%, which is a respectable return.

6 Conclusion and Future Work

In conclusion, this paper shows an Elo-enhanced LSTM model for successful football match prediction and betting market performance. The complex relationships between football match statistics are often tricky for traditional prediction methods to capture. Still, the combination of deep learning techniques and the Elo Rating system offers a good answer. To outperform the bookmakers in the betting market, our model predicts the number of goals scored by each team using past match data, team performance, player statistics, and Elo ratings.
After reviewing the literature, we discovered that different statistical models

and machine learning techniques had been successfully used to predict football match outcomes. The success of these studies demonstrates the promise of data-driven methods for predicting football games and offers valuable information for those working in football analytics and betting markets.

We use data from the English Premier League to train and assess our model, and the results are a mean squared error value of 1.2293 and an R-squared of 0.14. The model's performance is acceptable despite the comparatively large prediction error, given the few input factors and the unpredictable nature of some unusual scores. Additionally, using the model's predictions for the 21–22 season as the basis for a simulated betting procedure, we obtained a total return of 5.33%, demonstrating that our model outperforms the bookmakers.

In summary, our research shows the potential of deep learning techniques combined with the Elo Rating system to predict football match outcomes and achieve success in the betting market. Our model's ability to outperform the bookmakers demonstrates its usefulness in the real world and offers insightful data for football analytics and betting markets.

However, our model has several shortcomings, such as a meagre correct prediction rate for promoted teams, which could perhaps be solved by obtaining real-time third-party club Elo ratings; only shot-related data is used to measure the team's offensive strength, and no data on shots taken or other defensive data is collected. The yield of a model that seeks to beat the bookies might be increased by taking into account the bookies' behaviour rather than pre-match data from the known betting market, such as Kelly's index and odds.

References

Mark J. Dixon and Stuart G. Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 46(2):265–280, 1997. ISSN 0035-9254 1467-9876. doi: 10.1111/1467-9876.00065.

John Goddard. Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21(2):331–340, 2005. ISSN 01692070. doi: 10.1016/j.ijforecast.2004.08.002.

Corentin HERBINET. Predicting football results using machine learning techniques. Report, Imperial College London, 2018.

Lars Magnus Hvattum and Halvard Arntzen. Using elo ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3): 460–470, 2010. ISSN 01692070. doi: 10.1016/j.ijforecast.2009.10.002.

George Peters and Diogo Pacheco. Betting the system: Using lineups to predict football scores. *arXiv preprint arXiv:2210.06327*, 2022.