SOCCER BETTING SYSTEM USING MACHINE LEARNING

Guillermo López-Areal
glopezareallopezabad@hawk.iit.edu

Eva Solís Giménez
esolisgimenez@hawk.iit.edu

Abstract

It is well known that the big bookmakers, such as William Hill or Bet365 use Artificial Intelligence in their software. They use machine learning algorithms to make calculations and adjust the betting fees for sports events accordingly. Moreover, everyone has heard of the expression 'The house always wins'. This is the main reason that motivated this project. We have developed a system that allows the end-user to know, beforehand, the probability of winning his bet.

1. Introduction

Sport betting systems using machine learning have been gaining popularity in recent years. These systems use advanced algorithms to analyze historical data and make predictions about the outcomes of sporting events. Previous work in this area has focused on developing algorithms that can accurately predict the outcomes of games and provide valuable insights to bettors.

Some common methods used in sport betting systems with machine learning include decision trees, random forests, and neural networks. These methods can be trained on large amounts of data to identify patterns and trends that can be used to make accurate predictions.

Overall, the results of using machine learning in sport betting systems have been promising. These systems have been shown to be effective at providing accurate predictions and helping bettors make informed decisions. However, further research is needed to improve the accuracy and reliability of these systems.

In this project, we have approached this problem by creating both a RNN model and a LSTM model for soccer results predictions..

2. Problem statement

Everyone has heard of the expression *'The house always wins'*, which is undoubtedly true at any given level of gambling. This is due, mainly, to the fact that not-professional gamblers are driven more by their emotions than by their rational thinking. Especially in soccer, most bets are placed on each individual's favorite teams and players rather than on the likelihood.

It is well known that the big bookmakers, such as William Hill, Bet365 and others, use Artificial Intelligence in their software. These algorithms are used to make calculations and adjust the betting fees for sport event. This way, the odds are fixed so that the house always wins.

Figure 1. Well-known bookmakers.

Most AI sport predictive models are designed and built for these great world bookmakers. Nevertheless, the purpose of this project involves developing a system that allows the end-user to know beforehand the probability of winning his

bet. The main reason behind this idea and the key feature that makes it innovative is that is it designed for the end-user and not for the betting houses. Creating successful bets is all about building margins into the odds and balancing the book so that no matter who wins, the bookmaker always makes some profit. The problem approached in this project is to determine a realistic fee to any given sports event, compare it to the big bookmakers fee's and draw conclusions from it.

3. Proposed solution

The main goal is to design a predictive model for sports results. The plan is to build a scalable project, firstly focused on the most popular sport in the world: soccer. This sport is not only where most bets are placed, but also the sport with the highest volatility and failure rates.

Our approach is based on a machine learning model that, given some statistics about a soccer game, outputs a vector with three probabilities: probability of the house winning, probability of tie, probability of the visiting. With this results, the user could know in advanced the estimated probability of every result and use this to make his bet. Some previous works related to this topic have been used as reference [3] [1] [2].

3.1. Theoretical basis

Our first task was to analyze how the bookmakers operates in such a globalized society and how they make a profit. A bookmaker association is an organization that accepts sports bets. Bookmakers hire analysts who evaluate the chances of an outcome in a sporting event and set odds accordingly.

The main idea is pretty straightforward: the players (customers) place bets, and if the bet wins, the bookmaker pays the winnings calculated according to the odds, if it loses, the player loses the amount he bet. Each bookmaker operates according to a mathematical model that calculates the odds necessary to make a profit. The bookmaker seeks to earn income on any event outcome. The bookmaker's task is to compile a variety of events for the conclusion of the bet by bettors.

Betting is largely based on popular opinions, which this means that it is not always possible to get the correct odds just by moving your own probability lines. So the first way in which bookmakers make money is by pricing the market in a way that does not represent the true value, i.e., the true probability of the outcome.

Moreover, although it may seem the odds are calculated exclusively on the probability a certain event might occur, this is definitely not the case. For example, in Figures **??** and **??** two betting strategies are compared, 'Win' and 'Win or Draw'.

Figure 2. Bookmakers' accuracy.

Figure 3. Bookmakers' profit.

In the previous charts it can seen, that the chances of winning a bet placed on a 'Win or Draw' is much larger than the chances of winning a simple 'Win' (not tie). However, in **??** it can be observed that the bookmaker's is not proportional, i.e., it may not always be profitable to place a bet which is largely probable.

3.2. AI in the bookmaker's industry

When picking which team to bet on, there is a lot of facts to consider. Because of this, applying one of the most well-liked machine learning techniques, neural networks, to the betting industry is a great idea.

Bookmakers use machine learning algorithms to analyze large amounts of data related to sports events and betting markets, including historical data on past events, current performance metrics for teams and players, and information on betting activity and odds.

These algorithms can learn complex patterns in the data and use these patterns to make more accurate predictions about the outcomes of future events. For example, a machine learning model trained on data about past soccer matches might be able to predict the likelihood of a particular team winning its next match, or the probability of a certain number of goals being scored in a game.

By using machine learning to improve their predictive modeling, bookmakers can set more accurate odds and make more informed decisions about which events to offer bets on. This can help them to maximize their profits and provide a better experience for their customers.

We may create a neural network for sports betting that has three straightforward categories. The architecture of such a network is shown below.

4. Data set

As mentioned before, this project aims to be broaden to a big collection of different sports betting systems. However, although this model could be projected to other sports, such tennis or basketball, we have started with the most popular sport in terms of sports betting: soccer. Sports data analyst reveals that soccer betting makes up 70% of the global betting market, hence, it was the most adequate choice for our project.

4.1. Selection of the data set

The data set selected for this project has been the *Football Results and Betting Odds Data of EPL* data set, available at *Kaggle*[1]. It contains 6,000 soccer match results, statistics and betting odds data from 2002 to 2018.

This data set includes an extensive set of features, such as soccer team names, matches results, goal scored, matches statistics and betting odds data. Nevertheless, for this project we selected only the features that were considered more important for the results prediction task. The features considered as input were the following:

- HTGS: Home Team Goals Scored, which will tell us how many goals the home team has scored throughout the season. This is a very important feature since, it will predict accurately whether the Home Team is more likely to score more or fewer goals than the away team.
- ATGS: Away Team Goals Scored which is how many goals has the away team scored throughout the season. This is a key feature again, with the same reasoning as the
- HTGC: Home Team Goals Conceded. This indicated the amount of goals conseeded by the home team.
- ATGC: Away Team Goals Conceded. This indicates the amount of goals conseeded by the away team.
- HomeTeamLP: Home Team Leadership Points. This feature will tell us the home teams ranking in the premier league. For example if a team which is first in the EPL is

[1] https://www.kaggle.com/datasets/louischen7/football-results-and-betting-odds-data-of-epl

going against a team which is eleventh in the EPL's ranking. The odds will state the team which is higher in the ranking is more likely to win. This will consequently affect the fee for that certain team.

- AwayTeamLP: Away Team Leadership Points. This feature indicates the team's ranking in the premier league.

On the other hand, given that the goal of the proposed model is to predict the final results of a soccer match, the *FTR* feature of the data set is considered as label. This feature corresponds to the results for the corresponding game and has three possible values: 'H', 'D', 'A. The first value indicates the home team wins, the second refers to a tie and the last one refers to the visitor team winning the match.

4.2. Data analysis

Data analysis is a crucial step in the process of implementing machine learning algorithms. By carefully analyzing the data, we can identify trends, patterns, and anomalies that can help us understand the underlying relationships and phenomena that are being modeled. In this section the results the results of the data exploration are presented.

Firstly, the distribution of the labels has been plotted as a histogram, as shown in 4. This demonstrates that, in the selected data, more than 2,500 matches results in the house winning, almost 1,500 end in a draw and the rest conclude with the away team winning.

Furthermore, 5 shows the correlation between the selected features, which will serve as input for the model.

5. Implementation

In this section, we will be implementing a machine learning model to solve the sports results prediction problem. We will go through the steps of preparing the data, choosing an appropriate

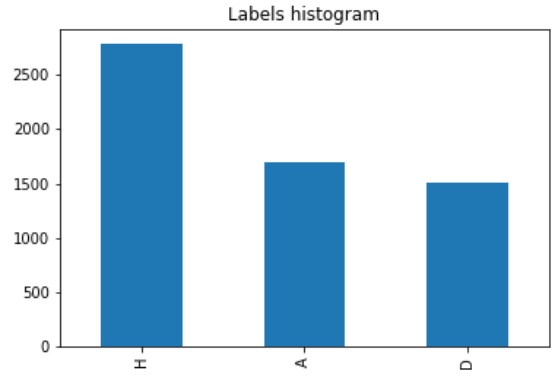

Figure 4. Match results' histogram.

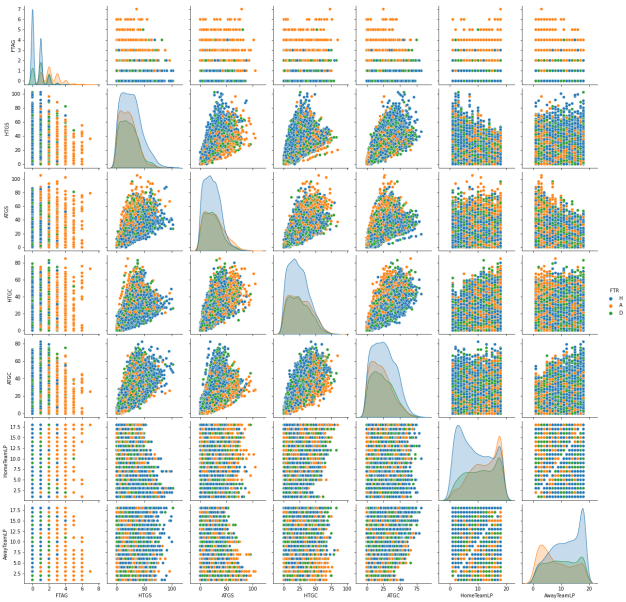

Figure 5. Features correlation.

model, training the model, and evaluating its performance. By the end of this section, we will have a working model that can make predictions on new data. This implementation will serve as a basis for further experimentation and optimization.

5.1. Data preprocessing

Firstly, in order to prepare the input features as suitable tensor for the model, the following steps were performed:

1. Import the data set, remove unnecessary

columns and shuffle the rows.

2. Separate labels and features.

3. Encode labels using the *LabelEncoder* class from *sklearn*, which encode the class labels as: {'A': 0, 'D': 1, 'H': 2}.

4. Normalize the features.

5. Split the data set into training, validation and test sub sets, which contains 3840 , 960 and 1200 samples respectively.

Finally, it is important to mention that, in order to prepare the data sets as inputs for the RNN model, the features needed to be resized.

5.2. RNN Model

In this section, we present the proposed solution for approaching the sports betting problem and predict soccer matches results. We decided to approach this problem by implementing a recurrent neural network (RNN) model to predict soccer matches' results. The reason behind this approach is that RNNs could be used to analyze historical data and identify patterns that can be used to make accurate predictions about the outcome of future games.

The implemented model has been specifically designed for use in soccer betting systems and trained in the mentioned European Premier League data set. In the following sections, we will describe the specific design of our RNN and evaluate its performance on a set of test data. Figure 6 shows the architecture of the RNN model used in this project.

5.2.1 LSTM

Furthermore, for a deeper analysis of the problem, a LSTM model has also been implemented for this project. LSTM (Long Short-Term Memory) is a type of recurrent neural network that is well-suited for time series data and predicting sequential events. In a sport betting system, LSTM

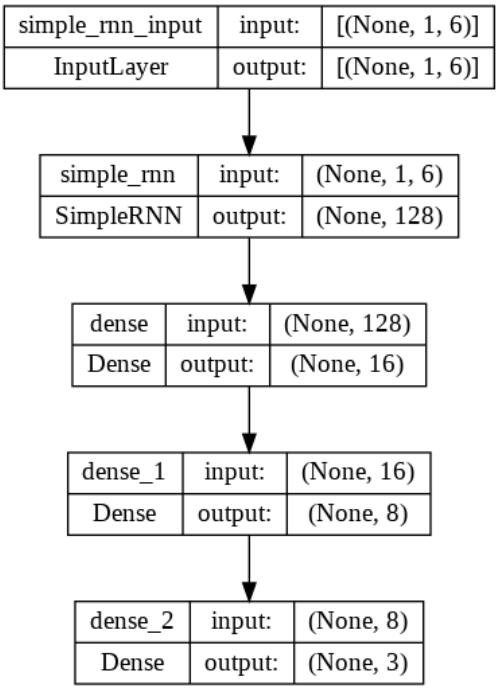

Figure 6. RNN model architecture.

can be used to analyze past betting patterns and trends in order to make more accurate predictions about future outcomes. LSTM is able to effectively capture long-term dependencies in data and make predictions based on these dependencies. This is important in a sport betting system where historical data and trends can be used to inform predictions about future outcomes. Additionally, LSTM is able to handle large amounts of data and can be trained on multiple layers, allowing for improved prediction accuracy. Overall, LSTM is a strong candidate for use in a sport betting system due to its ability to handle time series data and make accurate predictions based on historical trends. In Figure 6, the architecture of the LSTM model used for this project is presented.

6. Results

This results section presents the findings of the study. In this section, the performance of the model is evaluated using metrics such as accuracy to determine how well it can predict the outcome of future games. The results are discussed in the

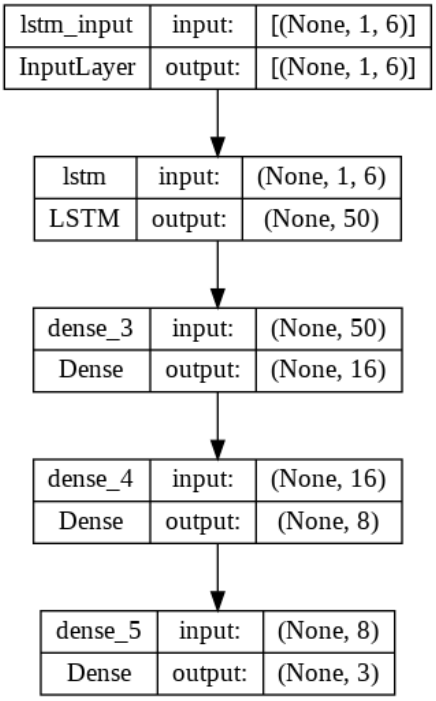

Figure 7. LSTM model architecture.

context of the sport results prediction problem and compared to any baseline or previous approaches.

After training both RNN and LSTM models with the selected soccer data set, the training and validation accuracy is analyzed in the following plots (Figures 8 and 9).

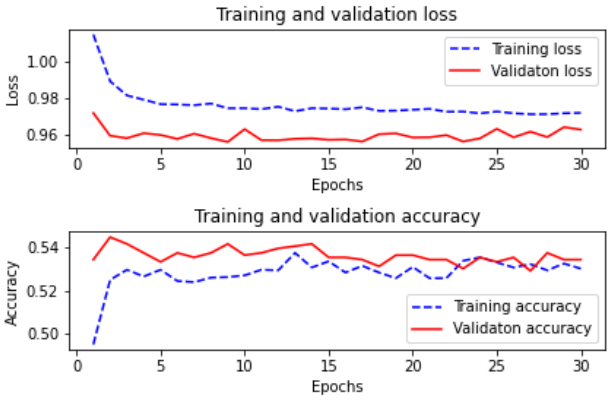

Figure 8. RNN model results.

Furthermore, when evaluating both models on the test data, we get an accuracy of 56.250% with the RNN model and an accuracy of 55.833% for

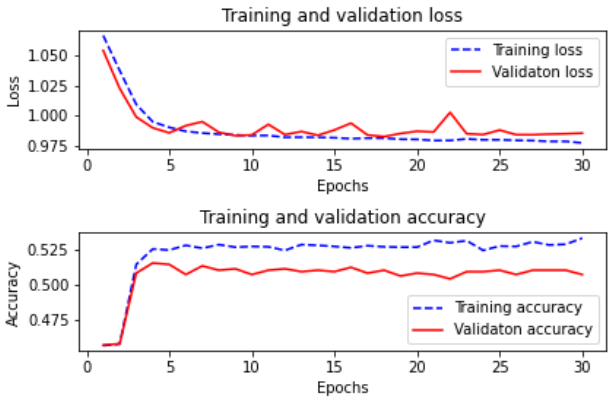

Figure 9. LSTM model results.

the LSTM model. The comparison of the RNN and the LSTM showed that the RNN model performed slightly better. This is likely due to the RNN's ability to effectively model time-series data, such as the sequence of results in a sports league, and its ability to "remember" past events and use this information to inform its predictions.

It is important to mention that machine learning algorithms are designed to learn from data and make predictions based on that data, however, the outcome of a soccer match can be influenced by many factors that are difficult to capture and model using data alone. For example, the skill level of the players, the strategies used by the teams, and the physical condition of the players are all important factors that can impact the outcome of a soccer match, but they are difficult to accurately measure and represent using data.

Moreover, given the there are three possible outcomes, achieving an accuracy higher than the 50% means that the model is at least performing quite better than a random guess. Additionally, sports events are often highly unpredictable, making it challenging for any model to achieve a high accuracy. Therefore, a 50% accuracy in soccer results prediction may still be considered to be a good result.

Overall, this results section provides a clear and concise summary of the model's ability to make accurate predictions and inform successful bets.

7. Future work

This study has shown that the proposed method is effective in predicting the results of soccer matches. Consequently, there are several directions for future work that could improve the performance of the method and extend its applicability to other scenarios. For instance, predicting the odds of a certain match to be won, tied or lost, we can counteract the fees with the giant bookmakers such as William Hill or Bet365. Theoretically such calculation is done simply by inverting the odds of each particular outcome, however the results won't show that.

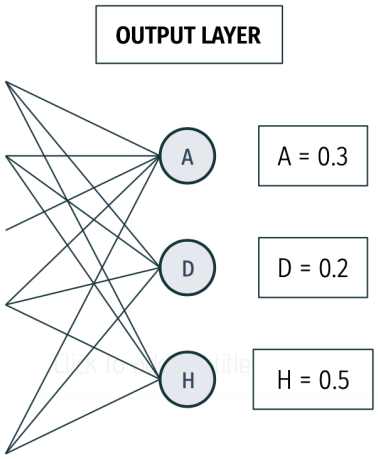

Figure 10. Probabilities predictions.

Hypothetically, and with a well-performing network, the relationship between the odds of winning tying or losing a certain match and the fees from the bookmakers could be extensively studied. Bookmakers take into account betting activity and market movements for a particular event. For example, if there are a large number of bets on a particular team to win, the bookmaker may adjust rates to reflect the increased likelihood of that outcome. Furthermore, when a match who is going to be followed by thousands of people, such as, a Champions League match or a World Cup match, bookmakers do know that most of the bets done, are done, not by people with empirical knowledge of statistics, but from people who bet with their emotions, because they want to see their team win, and that is a chance for bookmakers to adjust the odds for those teams, and make huge profit.

8. Conclusion

In conclusion, the comparison of RNN and LSTM models in a sports betting system using machine learning showed that the RNN model performed slightly better. This is likely due to the RNN's ability to effectively model time-series data and its ability to "remember" past events and use this information to inform its predictions. These characteristics make the RNN well-suited to the task of predicting the outcome of future games in a sports betting system. While LSTM models also have the ability to model time-series data, they did not perform as well in this specific application, suggesting that RNNs may be the preferred choice for building sports betting systems using machine learning. However, further research and experimentation may be necessary to fully understand the potential benefits and drawbacks of using these models in a real-world setting.

Moreover, the results of this studied reflects that Machine learning could be used for predicting sports results and beating bookmakers by training a model on a large dataset of past sports results and related information, such as the teams involved, the players' performance, and the conditions of the game. The model could then use this information to make predictions about the likelihood of different outcomes for future sports events. These predictions could be used to inform betting decisions and help bettors choose the most likely winners and losers. By making more accurate predictions than bookmakers, bettors could potentially increase their chances of winning and earn more money. Additionally, machine learning models could be used to automatically place bets on behalf of the bettors, further increasing the efficiency and accuracy of the betting process.

References

[1] Gabriel Fialho, Aline Manhães, and João Paulo Teixeira. Predicting sports results with artificial intelligence – a proposal framework for soccer games. *Procedia Computer Science*, 164:131–136, 2019. 2

[2] Ondřej Hubáček, Gustav Šourek, and Filip Železný. Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, 35(2):783–796, 2019. 2

[3] Miltos Petridis Max Bramer. *Artificial Intelligence XXXV*. Springer, 2018. 2