# Using Machine Learning to predict soccer outcomes

Rahul Kumar, University of Hartford

## Abstract

With over 1.1 billion people watching the 2018 World Cup, there is no doubt that soccer is the most popular sport on the planet—specifically at the FIFA World Cup. This project aims to derive features that are significant in predicting soccer outcomes and whether there is a relationship between betting odds placed by bookkeepers and the result of the match. This is important because betting is extremely prevalent in the biggest competition in the sport. Hence, the results from this project can better advise those placing bets on such tournaments and whether it is worthwhile. Ultimately, this project found that with the best model, the odds that bookkeepers make along with other features are accurate 75% of the time in predicting the outcome of World Cup matches. This implies that while there is a relatively high accuracy rate for this particular model, there is an average correlation between the betting feature and the full-time result.

## Introduction

The World Cup is a national tournament that showcases the best 32 teams on the planet as they battle it out to try and win the most prestigious award in the sport. As a result, a great deal of money can be made through betting as people aim to make some extra cash on the side. This project aims to determine the relationship between betting odds and the final result of a soccer match and whether the odds that bookkeepers put up truly reflect a correct full-time result

prediction. This idea is interesting because oftentimes, gambling is seen as a game of pure chance where it is unlikely to constantly win. For example, playing at a casino or buying a lottery ticket are known for their extremely skewed winning odds, ultimately convincing people to come back to win the big jackpot. But what if there was some hidden way to shift the odds to make it more likely to win rather than vice versa? This project examines this and tries to look for a relationship between the odds and the full-time result. This is a novel work since betting odds are usually avoided due to their lack of accuracy so there are not many scholarly papers written on it.

## Related Work

There has been a lot of work when it comes to predicting the outcomes of sports matches. Kampakis et al.(Kampakis, Adamides, 2014) used Twitter hashtags and tweets to predict the outcomes of a regular Premier League Season. Additionally, a historical dataset of statistics about the Premier League was collected and different models were created with the Twitter dataset, the historical dataset, and both datasets combined. Ultimately, the classifiers that provided the best results were Random Forest and Naïve Bayes.

Another work (van der Zaan, Alfons, 2017) used data from European national competitions to create multiple prediction models for different tournaments. By using an ordered probit prediction model, the probability of a full-time result can be analyzed by observing a normal distribution and how the relationship between the variables varies within a normal distribution. From there, attributes with a p-value less than 0.075 were kept and incorporated into an Akaike Information Criterion (AIC). The AIC model ultimately provided the best results for predictions. Van der Zaan also provides betting data and incorporates different strategies that can

be used to maximize profits. Strategies listed include Kelly's betting strategy, equal payout, and the variance-adjusted betting strategy.

On the direction of using machine learning for sports outcomes, Babak Hamadani experimented with logistic regression and SVM with varying kernels to predict the outcome of NFL games. The dataset consists of web scraped data of 2004, 2005, and 2006 NFL seasons and tested two feature sets: the features concerning the probability of a team winning and the game statistics for each team. Ultimately, logistic regression proved to have the best accuracy with 67%.

## Data

There were two datasets that were used over the course of this project: *International football results from 1872 to 2021* and *2018 World Cup Betting Odds* (oddsportal.com). Created by Mart Jurisoo, *International football results from 1872 to 2021* is a collection of basic tabular data statistics for every official match played since 1872. Features documented include the date of the match, the names of the home and away teams, the final scores of the home and away teams, the name of the tournament, the city and country where the match was played, and whether the location of the match was neutral for both sides. Since this project was examining matches at the World Cup level, 64 matches were extracted from this dataset. The second part of the dataset was web-scraped off of OddsPortal, a prominent sports betting website. Specifically, betting odds of the 2018 World Cup were extracted and added to the subset of matches from the previous dataset.

| Feature: | Description: |
|---|---|
| Neutral | True if home team is from the country where the game is being played at. |
| FTR | Full Time Result (Win, Loss, Draw) |
| HTW Odds | Betting odds that the home team wins |
| HTD Odds | Betting odds that the home team draws |
| HTL Odds | Betting odds that the home team loses |
| Penalties | If match went into a penalty shootout |

Table 1. Table of selected features from the dataset

## Methodology

There were four main stages of this project that led to the results found. These stages consist of data cleaning, data preprocessing, implementing models as well as analyzing results, and visualizing the data. The data cleaning stage first started off with accessing the results.csv file and examining a statistical summary of the data. From there, an extra column was generated that determined if the home team won, lost, or drew with the away team. Additionally, the web scraped data that was added had to be manually checked for each match to make sure that each betting odd matched the correct tea. Finally, the features that were selected for the models consisted of the home and away teams, whether the match was played at a neutral location, the betting odds, and whether the match went into penalties. The data preprocessing consisted of implementing the LabelEncoder from SciKit-Learn and encoding the descriptive variables for the models. The features encoded were the home and away teams, the neutrality of the location, the full-time result, and whether the match went into penalties. The default training and testing

set sizes were set to 80% and 20% of the data, respectively. The models used for this project were a multi-layer perceptron, a decision-tree, a SVM, and a Random Forest Classifier.

## Experiments & Results

The tools and models used were derived from the Scikit-Learn library provided by Python and the code was run on the Jupyter Notebook IDE for Python. Four models were implemented in total for this data set. The Multi-Layer Perceptron and Decision Tree models had an optimal accuracy with 70% of the data as training data and 30% as testing data. The SVM was optimal with 65% of the data as training data and 35% as testing data. Finally, the Random Forest Classifier had 60% of the data as training data and 40% as the testing data.

Ultimately, the Multi-Layer Perceptron had the best results for the accuracy, precision, recall, and f-score. Seen in the figures below, the MLP model had the best results, followed by the Decision-Tree, SVM, and Random Forest Classifier.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 1 |
| 1 | 1.00 | 0.60 | 0.75 | 10 |
| 2 | 0.64 | 1.00 | 0.78 | 9 |
| accuracy | | | 0.75 | 20 |
| macro avg | 0.55 | 0.53 | 0.51 | 20 |
| weighted avg | 0.79 | 0.75 | 0.73 | 20 |

Figure 1. Classification report for the Multi-Layer Perceptron model.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 1 |
| 1 | 0.62 | 0.80 | 0.70 | 10 |
| 2 | 0.57 | 0.44 | 0.50 | 9 |
| accuracy | | | 0.60 | 20 |
| macro avg | 0.40 | 0.41 | 0.40 | 20 |
| weighted avg | 0.56 | 0.60 | 0.57 | 20 |

Figure 2. Classification report for the Decision-Tree model

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 2 |
| 1 | 0.62 | 0.67 | 0.64 | 12 |
| 2 | 0.56 | 0.56 | 0.56 | 9 |
| accuracy | | | 0.57 | 23 |
| macro avg | 0.39 | 0.41 | 0.40 | 23 |
| weighted avg | 0.54 | 0.57 | 0.55 | 23 |

Figure 3. Classification report for the SVM model.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 2 |
| 1 | 0.80 | 0.57 | 0.67 | 14 |
| 2 | 0.43 | 0.60 | 0.50 | 10 |
| accuracy | | | 0.54 | 26 |
| macro avg | 0.41 | 0.39 | 0.39 | 26 |
| weighted avg | 0.60 | 0.54 | 0.55 | 26 |

Figure 4. Classification report for the Random Forest Classifier model.

As a result, it can be determined that at its best, the MLP model can predict the FTR of a match given data for each feature approximately 75% of the time. This better than the other model accuracies which were 60%, 57%, and 54% respectively. Additionally, the visuals shown in Figures 5 and 6 demonstrate that there was no strong correlation between each of the variables (except for the HTW and HTL odds, which were expected). Hence, even though there may be no direct relationship between betting and the full time result, the inclusion of multiple features for a given match can assist in determining the final match result along with the betting odds.

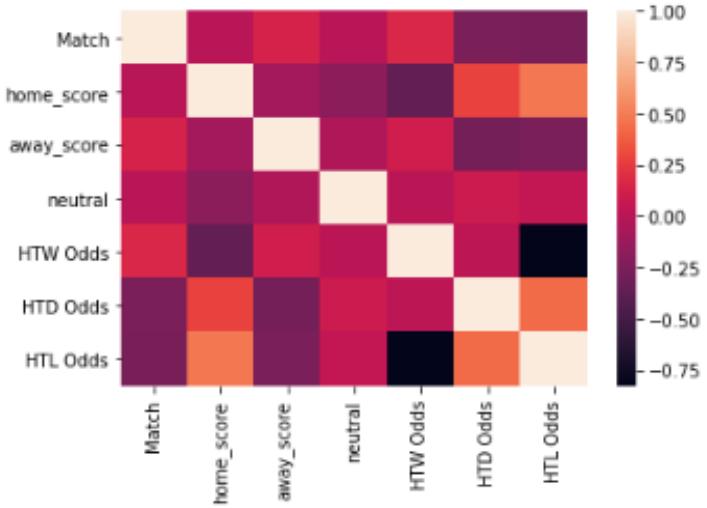


Figure 5: Correlation matrix heatmap between features

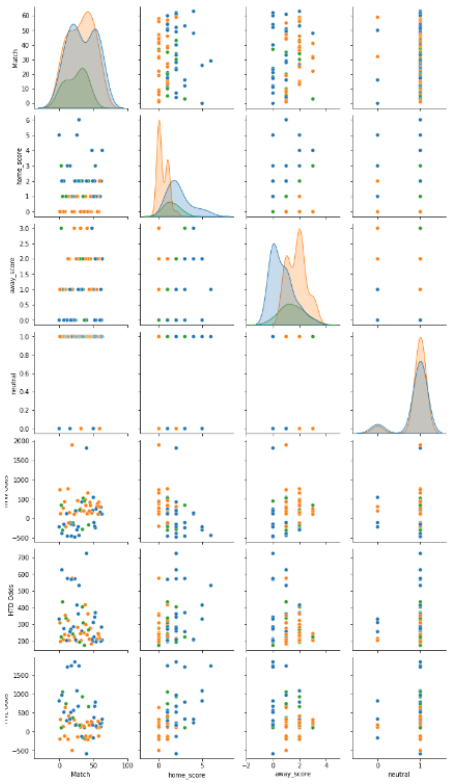


Figure 6: Scatterplot between various features with an emphasis on FTR

## Conclusion

Something that becomes clear from the results is that the dataset formulated after data cleaning and preprocessing has enough information to determine the final result of a World Cup match 75% of the time using a MLP. While this may seem low, the fact that the accuracy rate for a model containing features that have no relationship with each other was this high means that there may be an implicit relationship that was not shown directly in the visualizations.

There were multiple challenges that I faced during the duration of this project. The most challenging part of the project had to be the web scraping for the betting odds and cleaning the data. Even though it looked straightforward, there were many operations that I did not expect

when going through the CSV file. This included matching the odds with the correct match-up, cleaning the data of null values, as well as encoding descriptive data columns. Additionally, determining the best machine-learning models to implement in this project was difficult as there were many viable options.

Given more time and resources at hand, it would be beneficial to implement more statistics about the team rankings and the caliber of players playing on each team. This project treated each team as being equal in ranking and team structure, but in reality, this is far from true. As a result, this project was biased to teams with lower rankings than teams that are higher ranked because everyone was treated with equality. Instead, more emphasis for winning should be placed on higher ranked teams and vice versa for low ranked teams. Future research should incorporate these ideas to create the best model for accurately predicting soccer match outcomes.

## Bibliography

Hamadani, Babak. "Predicting the outcome of NFL games using machine learning." URL http://cs229. Stanford. edu/proj2006/BabakHamadani-PredictingNFLGames. pdf (2006).

Kampakis, Stylianos, and Andreas Adamides. "Using Twitter to predict football outcomes." arXiv preprint arXiv:1411.1243 (2014).

Van der Zaan, Tim, and Andreas Alfons. "Predicting the outcome of soccer matches in order to make money with betting." Rotterdam: Erasmus University Rotterdam (2017).

“World Cup 2018 Results; Historical Odds.” Oddsportal.com.