PREDICTIONS OF TOP 5 EUROPEAN FOOTBALL LEAGUE MATCHES

Vladimir MKRTCHYAN
Ph.D. in Economics
Tigran GHLIJYAN
Master of ANAU, Armenia

Key words: football, odds, monte carlo simulations, predictions, machine learning.

Introduction. Sports gambling industry has always been an interesting research area for both scientists and hobbyists who are trying to beat the bookmakers. While several strategies are claimed to beat the bookmakers using arbitrage or expert predictions, long term returns of those strategies are not consistent. Very interesting approach is used in [Kaunitz, Zhong and Kreiner, 2017] which does not aim to outperform the forecasting models of the bookmakers but uses the odds throughout the betting market to find mispriced odds. It is shown that the inefficiency of the betting market can be exploited to consistently beat the bookmakers. Some relatively accurate models to predict outcomes of football matches are proposed in [Melnykov, 2013] and [Dyte and Clarke, 2000, pp. 993–998].

This particular project aims to develop a web application which can give betting predictions on the upcoming matches of the major European leagues. We do not aim to outperform bookmakers in their predictions, rather we aim to create a simple predictor, test and deploy it. Firstly, we collect the data of matches from the past seasons. Than we propose a learning model based on the indicators which we find relevant in predicting the outcome of a certain match. After estimating the performance of the proposed model we implement Monte Carlo simulations to understand the profitability of the predictions. Finally, we deploy the program in the form of simple web application which provides predictions on the upcoming matches of top 5 European leagues.

Methodology and literature review. Since we aim to not only develop a predictive model but also simulate bets, we also need the closing odds on the outcomes that is 1X2 odds. For that purpose, we crawled the sports statistics web page [Odds Portal, 2019], which contains both historical data and closing odds on the outcomes of the matches. The sample of the crawled data is shown in Table 1 below. We have crawled past matches starting from season 2010 to 2018 for the top 5 European leagues, namely, La Liga (Spain), Bundesliga (Germany), Legue-1 (France), Serie-A (Italy) and Premier League (England).

In Table 1 “coef1”, “coefX” and “coef2” show the closing odds corresponding to win of first team, draw and win of the second team. The variable “matchDate” will be used to figure out the evolution of the club rating during the season.

country	league	season	matchDate	team1	team2	goals1	goals2	ceof1	coefX	coef2
france	ligue-1	2015-2016	2015-09-13	Marseille	Bastia	4	1	1.50	4.19	7.25
france	ligue-1	2015-2016	2015-09-13	Nantes	Rennes	0	2	2.64	2.90	3.09
france	ligue-1	2015-2016	2015-09-13	GFC Ajaccio	Monaco	0	1	4.28	3.13	2.04
france	ligue-1	2015-2016	2015-09-12	Lorient	Angers	3	1	1.99	3.20	4.29
france	ligue-1	2015-2016	2015-09-12	Montpellier	St Etienne	1	2	2.82	3.10	2.72
france	ligue-1	2015-2016	2015-09-12	Nice	Guingamp	0	1	2.26	3.14	3.50
france	ligue-1	2015-2016	2015-09-12	Toulouse	Reims	2	2	1.97	3.38	4.10
france	ligue-1	2015-2016	2015-09-12	Troyes	Caen	1	3	2.37	3.21	3.16
france	ligue-1	2015-2016	2015-09-12	Lyon	Lille	0	0	1.88	3.32	4.61
france	ligue-1	2015-2016	2015-09-11	Paris SG	Bordeaux	2	2	1.26	5.85	12.73

Table 1. Sample of 10 scrapped entries

Now we present the transformation of the raw data to be able to create a predictive model. As mentioned above we aim to predict the outcome of a given match. First of all, we need to figure out what are our predictor variables. For our model we chose

- ratings of the competing teams
- head to head statistics
- home advantage

As for the ratings of the clubs we chose the Elo rating system which is quite popular choice in creating the ratings of competing sides in different kinds of sports. Ratings of the competing clubs and home advantage are more or less obvious indicators which affect the outcome of a football match. Let us introduce the head to head statistics indicator which aims to estimate the advantage between two particular clubs based on their past games. Why this indicator is important? Take for example Barcelona and Real Sociedad. Starting from season 2013 until 2016, Sociedad which has low rating as compared to Barcelona, played incredibly well against the latter (4 wins, 4 losses and 2 draws). We see that there is a total equality during this particular period of time. However, Barcelona does not demonstrate this kind of head to head statistics against all the teams with ratings similar to that of Real Sociedad. It means that this statistic describes the advantage between two particular competing sides. We calculate head to head statistics indicator between team 1 and team 2 matches with the following formula:

$$H2H_{t1t2} = \frac{1}{N}\sum_{i}^{N}\left(O^{i}{}_{t1t2} - O^{i}{}_{t2t1}\right)$$

where $O^{i}{}_{t1t2}$ is equal to 1 if on the i^{th} match - $t1$ vs $t2$ the winner was $t1$, -1 if the winner was $t2$ and 0 if the match was ended with draw. By this definition of head to head statistics 1 and -1 mean absolute advantage of either home or away team and 0 would mean an absolute equality in past matches. It is evident that the head to head statistics lies between -1 and 1. Now that we have cleaned the data and figured out our predictor variables we create the predictive model. First what we have tried is Artificial

Neural Networks. So as we can see from Table 2, we have four predictor variables - $Elo1$, $Elo2$, $H2H$ and $Home\ Team$. To build the NN we used Machine Learning library Keras. Because, we have only 4 predictor variables our network is rather simple - it is a NN with 3 Dense layers. As the outcome of the match is a categorical response variable, in the last layer we should use the "softmax" activation function which return probabilities corresponding to every outcome category.

team1	team2	elo1	elo2	h2h	home
Celta Vigo	Valencia	1677.64294434	1806.46362305	-0.625	team1
Mallorca	Osasuna	1671.00817871	1706.91992188	-0.1111111111111111	team1
Atl. Madrid	Betis	1850.27404785	1735.38085938	0.6363636363636364	team1
Real Sociedad	Mallorca	1737.5546875	1678.74633789	0.16666666666666663	team1
Valencia	Barcelona	1801.70471191	2078.6640625	-0.7857142857142857	team1

Table 2. Sample data after processing and adding the predictor variables

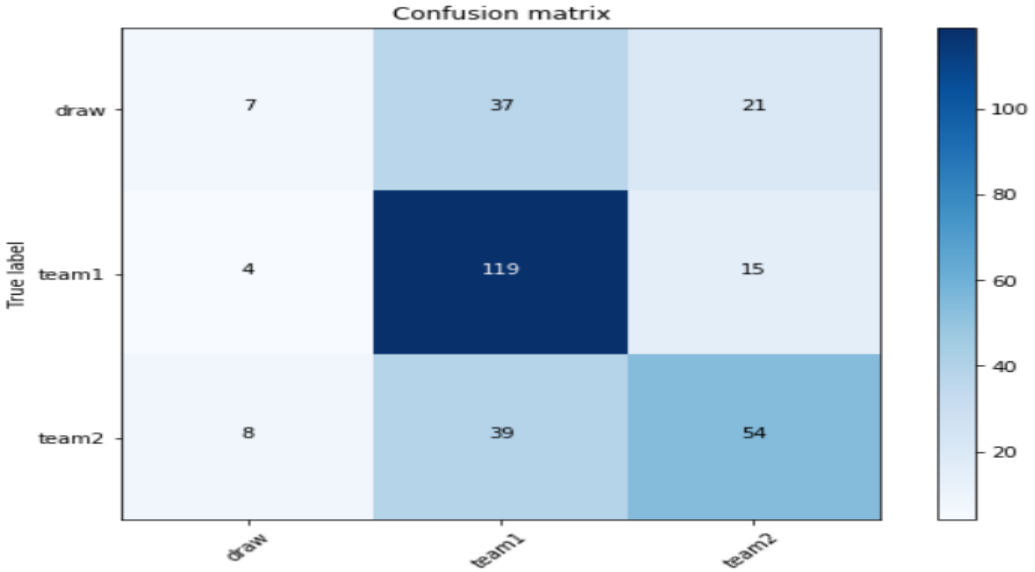

Figure 1. Confusion matrix of Premier League predictions

For building the model we take the matches of past 9 seasons of top 5 European leagues starting from 2010 until 2018. We split the data into train, test and validation sets. 20 percent of the data we took as a testing data. The rest of the data was split to training and validation sets by the proportion 80: 20. Let us look at the model performance for Premier league. As we can see from Fig. 1, the model predicts very few draws. This is rather expected because the probability of the draw before the match is the least, that is bookmakers always favor the winning odd of one or the other side and not the draw. From the other hand, it is good to have a model which predicts draws since the odds for draws are always higher as those are considered the least probable outcomes

pre-match. For Premier League predictor we have got an accuracy of nearly 60%. From confusion matrix we can also calculate the number of matches on which the predictor totally failed. For 39 matches the predicted value was 'team1' but the actual value was "team2" and for 21 matches the predicted value was 'team2' but the actual outcome was "team1". The predictor totally fails on 19% of the time, and the rest of the predictions fail only by one step. Completing similar analysis on the other 4 leagues we can give approximate estimation on leagues which are more or less predictable than the others.

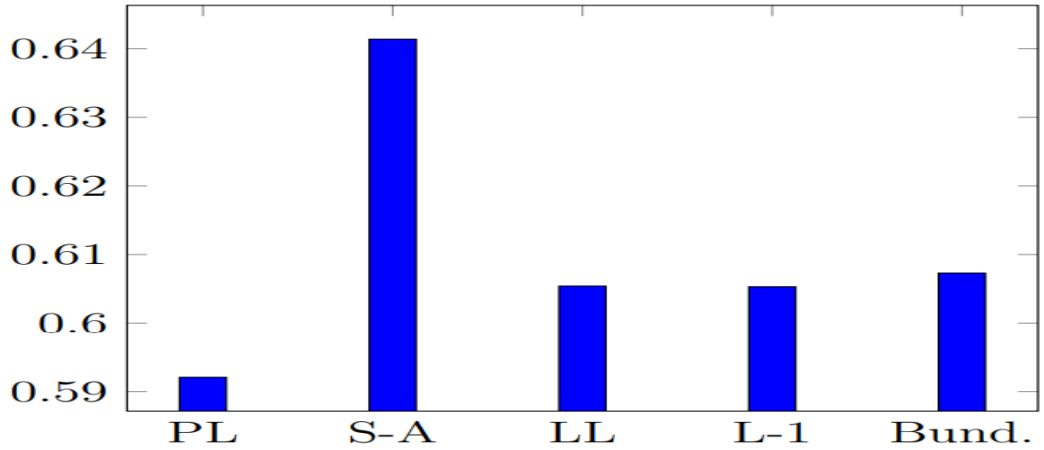

Figure 2. Predictability by top 5 national leagues

From Fig. 2 we can see that the least predictable national championship is the Premier League. The matches of La Liga, Bundesliga and French league are almost equally predictable. Finally, the most predictable of all with prediction accuracy of nearly 65% is Italian championship.

Scientific novelty. The scientific novelty of this article is that a model is proposed here, which predicts the probability of winning the matches of the main European league. Using the predicted odds, we present betting strategies, analyze their profitability. Article also assesses the predictability of the top 5 European football tournaments. Finally, using the results of the predictions, a web application was created, which gives predictions about the upcoming football matches of the top 5 European leagues.

Analysis. From Fig. 2 we have the prediction accuracies for top 5 European league matches. The problem is how to figure out, if it is bad or good. Considering the fact that our model does not tend to predict draws, it is quite reasonable result. However, from the perspective of profitability analysis accuracies do not illustrate anything. This is the point, where we need to use betting simulations. For example, we can take 10 matches randomly from the test set, predict their outcomes, simulate 1 dollar bet on the coefficient that corresponds to the predicted outcome and if it is correct get a profit equal to that particular coefficient minus our bet of 1 dollar. However, the profit, estimated in

this way, is completely ra//////ndom and if let's say in one set of games we achieve a 6 dollars of profit from 10 matches, for next set of 10 matches we can lose 10. The reason behind this is the ultimate randomness of the outcome of a match. Moreover, different profit values are the results of different realizations of outcomes of 10 observed matches.

Date from: 18.05.2021 League: Italy Get predictions

Team 1	Team 2	Date	Prediction
Fiorentina	Genoa	26 May	**Fiorentina**
AS Roma	Parma	26 May	**AS Roma**
Atalanta	Sassuolo	26 May	**Atalanta**
Bologna	Napoli	26 May	**Napoli**
Cagliari	Udinese	26 May	**Draw**
Torino	Lazio	26 May	**Lazio**
Frosinone	Chievo	26 May	**Chievo**
Inter	Empoli	26 May	**Inter**
Spal	AC Milan	26 May	**AC Milan**
Sampdoria	Juventus	26 May	**Juventus**

Table 3. Predictions of the upcoming matches of Serie-A

So, to avoid this kind of delusions and to be more general we conduct random simulations of 10 matches N times. We chose 10 matches in similarity to 10 matches that are played during every match week. The key points of the algorithm of Monte-Carlo simulations are introduced below. For every round of simulation generate a uniformly distributed random number r. If r is less than the away win probability of the home team we consider the outcome of the match as away team win. If r is less than the away win probability plus the draw probability the outcome is considered to be a draw. Otherwise, it is considered as home team win. This technique comes from probability theory and is quite natural approach in simulating random events with some probabilities (see [Wang, 2012]). After having the outcomes of the simulated matches, we simply compare them to the outcomes that were predicted by our model. In case of correct predictions, we add a profit equal to value of corresponding coefficient minus our 1-dollar bet. Using this algorithm we conducted 1000 simulations of 10 matches chosen from the test set. For every round of simulation, we obtained a profit that corresponds to one of 1000 realizations of the simulation. So, to obtain the long term profit estimate we need to average the profits over all of the realizations. For arbitrarily chosen 10 matches we obtained profit estimate equal to $0.51. However, this technique is limited to the specific domain of the training data. Here, we must note that even if we had a model which is 100% accurate based on past data that does not in any way mean that it will succeed on unseen data. So, beating the bookmakers is something that is nearly impossible task to achieve. We chose to create simple web page which gives users the opportunity to get predictions on future matches of their chosen league. When user

selects a league and a date, the application returns the list of matches that will take place in an interval of 7 days after the specified date with the predictions on the outcomes of those matches. We used Java framework - Spring MVC to serve as a back-end of our application. The predictions on future matches are done beforehand after match week.

Conclusion. We proposed a simple predictive model for the matches of top 5 European leagues. The predictor variables were Elo ratings, home advantage and head to head statistics. It turns out that by our model the less predictable league is English Premier League and the most predictable is Italian Serie-A. We also deployed our predictor by creating a simple web application which gives users the opportunity to get predictions on the upcoming matches.

Disclaimer*: Authors of this paper are not in any way responsible for the monetary losses caused by following any ideas and procedures included in this project. Bet responsibly.*

References

1. Lisandro Kaunitz, Shenjun Zhong, and Javier Kreiner. "Beating the bookies with their own numbers-and how the online sports betting market is rigged". In: *arXiv preprint arXiv:1710.02824 (2017).*
2. Yana Melnykov. "Examining influential factors and predicting outcomes in European soccer games". PhD thesis. North Dakota State University, 2013.
3. David Dyte and Stephen R Clarke. "A ratings based Poisson model for World Cup soccer simulation". Journal of the Operational Research society 51.8, 2000, pp. 993.
4. Odds Portal - Betting Odds Monitoring Service. https://www.oddsportal. com/ Accessed: 2019-04-25.
5. Hui Wang. Monte Carlo Simulation with Applications to Finance. CRC Press, 2012.

Vladimir MKRTCHYAN. Tigran GHLIJYAN

Predictions of top 5 European football league matches

Key words: football, odds, monte carlo simulations, predictions, machine learning

Beating the bookmakers is something that is nearly impossible task to achieve. However, in this paper we propose a model to predict winning odds of major European league matches. Particularly, the predictions were done on top 5 (England, Spain, Italy, France and Germany) football league matches. To create an accurate predictor, we used Machine Learning to build an Artificial Neural Networks. Using the predicted probabilities, we introduce betting strategies and analyze their profitability using Monte Carlo simulations. To create more precise model, we set predictor variables such as ratings of the competing teams, head to head statistics and home advantage. We also estimate the predictability of top 5 European football competitions. The analysis is done based on historical data of matches and the closing odds on outcomes of matches. Finally, we deploy the results of the predictions by creating a web application which gives the predictions on the upcoming football matches of the top 5 European leagues.