3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11619

Investigating inefficiencies of bookmaker odds in football using machine learning

Benedikt Mangold[1], Johannes Stübinger[2]
[1]GfK, Germany, [2]Department of Statistics and Econometrics, University of Erlangen-Nuremberg, Germany.

Abstract

The efficient-market hypothesis states that it is impossible to beat the market, as the price reflects all available information. Applied to bookmaker odds for football games, there should not be a systematic way of winning money on the long run. However, we show that by using simple machine learning models we can systematically outperform the markets belief manifested through the bookmakers odds. The effect of this inefficiency is diminishing over time, which indicates that the knowledge that has been derived from and the pure amount of the data is also reflected in the odds in recent times.

We give some insights how this effect differs across major football leagues in Europe, which algorithms are performing best and statistics on the ROI using machine learning in football betting. Additionally, we share how the simulation study has been designed in more detail.

Keywords: *Machine Learning; Football.*

1. Introduction

What if you could beat the market? Many tried, many failed. This paper analyses the efficiency of the market manifested in bookmaker odds in the domain of football. For that, we use publicly available information about football clubs and their active players to train a machine learning model that predicts the outcome of a match. The simulation covers the five major European football leagues with corresponding second leagues and data from twelve seasons. We use the predictions to (virtually) place a bet on historic odds to analyze if we could systematically outperform the believes of the market, measured by the return of the placed bet. In the following, we revisit the results of the paper by Stübinger et al. (2020) with respect to the success of the betting strategy, the implications for the efficiency of the offered odds and some detailed analyses from the perspective of the bookmakers during the covered period.

The first section revisits the simulation design and important results of Stübinger et al. (2020), the second section discusses the findings with respect to the well-known market efficiency hypothesis. The third section concludes this manuscript.

2. Modelling a football match with machine learning

Our simulation study is mainly based on two data sources we crawl from the internet. First, we consider all football matches from Primera Division, Segunda Division (Spain), Premier League, Football League Championship (England), Bundesliga, Bundesliga 2 (Germany), Serie A, Serie B (Italy), Ligue 1, Ligue 2 (France) from season 2006/2007 to 2017/2018. This record of 47,856 football matches provides a true hardness test for any back-testing study, as investor interest and analyst scrutiny are particularly high for these football nations. Second, we take into account 40 features for each player who was active in the respective matches. To be more specific, we consider skills from the areas “General”, “Ball Skills”, “Passing”, “Shooting”, “Defense”, “Physical”, “Mental”, and “Goalkeeper”.

In the spirit of Stübinger and Knoll (2018), the data set is sliced into 12 overlapping study periods, each shifted by one season. Each study period consists of a formation period and an out-of-sample trading period. The formation period identifies complex relations between the features of the players and the corresponding match result by fitting different machine learning models Random Forest (RAF), Boosting (BOO), Support Vector Machine (SVM), Linear Regression (LIR). Furthermore, we introduce a weighted ensemble method (ALL) by integrating the information of the four baseline approaches. The trading period (1) predicts football matches with the help of the mentioned data-driven methods and (2) exploits the obtained information in order to construct a statistical arbitrage strategy. If our assumptions hold, the trading algorithm would be able to find market anomalies and to generate positive profits.

Table 1 provides an overview of the results achieved by the different machine learning models. The ensemble strategy ALL results in the highest accuracy of 81.77% and an average payoff with a value of 1.0158. Note that with an average payoff that is greater than 1 one can beat the bookmaker over the long term as for each monetary unit spend, on average the return is strictly positive. Overall, the higher the complexity of a strategy, the higher the quality of our predictions which is reflected in higher average payoffs. It should be mentioned that we benchmark these strategies against baseline betting algorithms, e.g., randomly betting or always betting on the event with the lowest odd (most probable outcome) or placing bets on the home team. All these benchmarks perform significantly worse than strategies based on ML. We carefully conclude that (a) player characteristics contain information about the outcome of football matches and (b) our ML methods can capture and exploit these signals from the data.

Table 1. Statistical performance indicators for the betting strategies for the football seasons 2006/2007 to 2017/2018.

Key figure	RAF	BOO	SVM	LIR	ALL
Accuracy	0.8126	0.7912	0.6971	0.7292	0.8177
Average Payoff	1.0043	1.0072	0.9757	0.9933	1.0158

Source: Stübinger et al. (2020).

Figure 2 analyzes the performance of the strategies RAF, BOO, SVM, LIR, and ALL over time. We sort the time series top down in order of their cumulative return at the latest point in time series. As expected, the strategy ALL is best in class with a cumulated return value of 9.47. We observe that time series of ALL, BOO and RAF are flattening out since 2013. This may be the influence of better odds, as the bookmakers themselves started using machine learning algorithms which results in a more precise estimation of odds capturing the increasing amount of available information. These findings, especially the decrease in profit over time, confirm the weak efficiency hypothesis as pointed out in the following section.

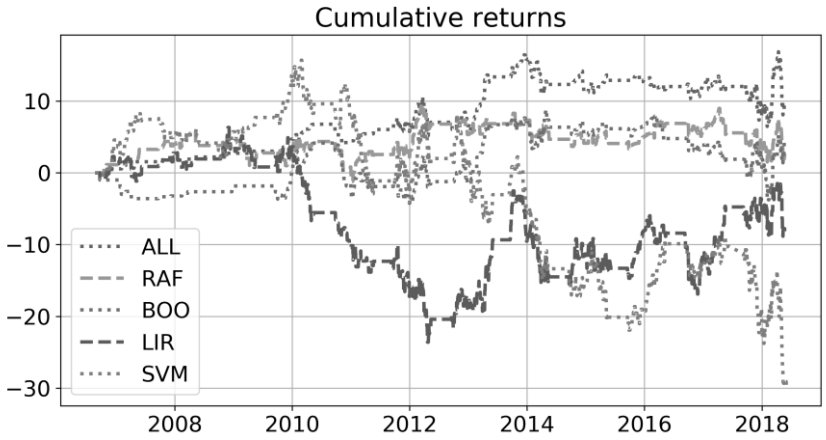

Figure 1. Cumulative returns of RAF, BOO, LIR, SVM, and ALL. from football seasons 2006/2007 to 2017/2018. Source: Stübinger et al. (2020).

3. Implications for the efficiency of the market

Market Efficiency states that all publicly available data is manifested in fair betting odds (up to a surcharge added to the odds by the bookmakers). This implies that no systematic outperformance should be possible.

Fama (1970) introduces the concept in the context of financial markets (weak form) which states abovementioned hypothesis. Lately, football betting has been getting more interest in research as there is a vast amount of publicly available data – both for players/teams statistics and various bookmakers odds. The latter should reflect all the information that can be derived from aforementioned data if the efficiency hypothesis holds true.

3.1. Inefficiencies in football odds

Angelini & Angelis (2018)[1] give an excellent introduction on challenging the efficiency hypothesis in the area of football. They analyze 11 major European football leagues covering 11 years of data and revealing inefficiencies in three of the leagues. However, the analysis is based on econometrical models, only, whereas Stübinger et al. (2020) cover the application of state-of-the-art machine learning technique with a strong emphasis on the prediction of the correct outcome of a game rather than modeling the mechanics between the influencing factors. Stübinger et al. (2020) point out diminishing profitability of

[1] See Angelini & Angelis (2018) and references therin.

systematic betting approaches over time, which can be explained by improved modeling techniques and the improved availability of processing power and data. In this section, we link the aforementioned results of Stübinger et al. (2020) to the efficiency hypothesis.

Efficiency in this context is always referring to betting odds including all available information. Thus, an information asymmetry would lead to systematic wins on average when used to identify 'weak' betting odds (of either the bookmaker or the betting person). On the other hand, when applying betting strategies that use only limited amount of information (compared to the odds), one can expect systematic losses over time. This is in line with the results of Stübinger et al (2020), where strategies that use no or little information (random-betting or favoring the home team) never gain any positive returns. However, using todays computation power and data, one can generate an historical information advantage in times where the odds did not reflect that knowledge which results in positive returns for earlier periods.

In older times, the process of reflecting believes about the outcome of a game using betting odds was in the hand of individual bookmaker companies. As the amount of information kept on growing, providers of betting-odds-as-a-service[2] started entering the market who offer real-time information before and during sport events leveraging enormous data collection tools and modelling power. Also, by providing the major bookmaker companies with similar betting odds, arbitrage effects are no longer available.

3.2. Efficiency and Machine Learning

The results of Stübinger et al. (2020) state that it would have been highly profitable to use machine learning based betting strategies in early 2000s. However, in those times neither the data nor the computation power has been available as easily as it is of today. By choosing the data source and the applied ML algorithms constant over time, we observe diminishing returns when using the predictions of a football game to bet against the bookmakers. We think that this is strongly related to the increased ability of odd providers to collect and process data together with more sophisticated modeling techniques. Whereas in older times a single, well informed person with knowledge and experience of analyzing games in his favorite league could have had an overview on all available data by screening the newspaper or relevant internet articles, nowadays in times of social media[3] and openly accessible datasets[4] this can only be achieved using algorithms and cloud computing. The information lead of bookmakers that is reflected by more accurate odds makes it impossible

[2] Such as https://www.betradar.com/ or https://txodds.net/.

3 E.g, Schumaker et al (2016) using Twitter data.

[4] https://datahub.io/collections/football

for individuals to beat the market in the long run, which is a strong confirmation for the market efficiency hypothesis.

3.3. Bookmakers perspective

Naturally, the view of the bookmakers plays an important role in the context of football betting. Provider determine the offered odds on the basis of a two-stage procedure. First, they estimate the probabilities of the outcomes home win, draw, and away win. The methods applied here vary between naive baseline approaches and highly complex machine learning models. The sum of the three probabilities is always 1. The fair betting odds would be the inverse of the respective probabilities. Second, bookmakers usually diminish the fair odd value by their risk premium, i.e., a certain amount to cover their costs in the long run. Consequently, we can calculate this key figure by determining the difference between the sum of the inverse of the actually offered betting odds and 1. A higher risk premium is tantamount to a more inefficient market environment as the bookmakers need to hedge their earnings due to the increased uncertainty of the outcome of a game. Figure 1 presents the average risk premium of the online bookmaker Bet365, one of the leading betting providers with around 23 million customers, from 2006 to 2018. We observe a decrease of the risk premium from around 10 percent in 2006 to around 5 percent in 2018. This fact is not surprising since more and more betting providers are entering the market and the information available is increasing. This picture is very similar to the increasing market efficiency as we know it very well from the stock market environment.

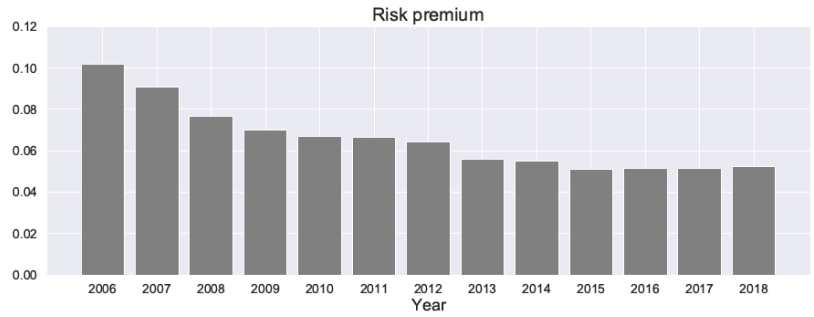

Figure 2. Average risk premium of the bookmaker Bet365 from 2006 to 2018.

Additionally, we want to provide some insights on the distribution of the risk premium across different type of football teams. Figure 3 displays the relation of average points and average risk premium for the teams of the Primera Division from 2006 to 2018. Top teams like Real Madrid and Barcelona achieves around 2.3 points per match, bad teams like Granada and Gijon around 1 point per match. We observe an average risk premium of approximately 6.2 percent for the top teams as well as the bad teams. Teams of average

quality in terms of long-run performance like Zaragoza and Mallorca possess a higher risk premium. We may conclude that is much more difficult to predict the outcome of matches involving teams of average quality.

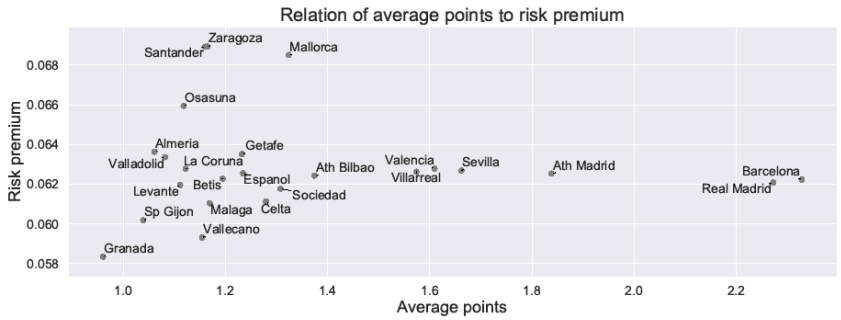

Figure 3. Relation of average points and average risk premium for the teams of the Primera Division from 2006 to 2018.

4. Outlook

In this paper we revisited the results of Stübinger et al (2020) from the perspective of the market efficiency hypothesis. We come to the conclusion, that diminishing returns over time placing bets based on signals generated by machine learning algorithms confirm that all available information is reflected in fair betting odds in recent times. The fact that using available information cannot systematically outperform the bookmakers confirms the weak efficiency hypothesis by Fama (1970) in recent times.

References

Angelini, G., & Angelis, L., (2019). *Efficiency of online football betting markets*. International Journal of Forecasting, 35(22), 712-721.

Fama, E. (1970). *Efficient Capital Markets: A review of theory and empirical work.* The Journal of Finance, 25(2), 383-417.

Schumaker, R. P., Jarmoszko , A. T., & Labedz, C. S. (2016). *Predicting wins and spread in the Premier League using a sentiment analysis of Twitter*. Decision Support Systems, 88, 76-84,

Stübinger, J., & Knoll, J. (2018). *Beat the Bookmaker – Winning football bets with machine learning (best application paper).* In International Conference on Innovative Techniques and Applications of Artificial Intelligence (pp. 219-233). Springer, Cham.

Stübinger, J., Mangold, B., & Knoll, J. (2020). *Machine learning in football betting: Prediction of match results based on player characteristics*. Applied Sciences, 10(1), 46.