Telecommunications and Informatics Engineering

Pattern Detection Applied to Soccer Results Forecast

Diogo Reis
diogo.reis@tecnico.ulisboa.pt

January 2018

Abstract

Soccer is not only recognized as the king of the sports, but also represents a large industry at an economic level. Despite this, the number of research papers applied to the sports betting market - soccer in specific - still have not reached the levels of, for example, the stock exchange market.

This paper proposes the development of a machine learning model on how to predict the outcome (home win, draw, away win) of a soccer match, based on statistic match information available in the pre-match. The information regarding each team will be analysed in terms of attack strength and defence strength separately, for the leading elite leagues in European soccer.

The developed model will be evaluated against other benchmark models - bet on the favourite, random walk and bet on the home team - using the average value of the odds offered at the main bookies.

Table of Contents

1 Introduction

In the past few years we have been assisting to a technological explosion that can be characterized in two different vectors. One is the exponential increase in the number of mobile devices with internet access, and the other is the growth of the volume and diversity of information stored today, accessible to all of us.

The importance given to information and its growth has never reached such high levels, and it is expected to continually increase. This information Age has motivated new areas of study such as Machine Learning and Data Science, that are currently in great evolution across all business sectors, and the sports betting market is not an exception.

This technological context contributes to the increase, year after year, of the transacted volume in the sports betting market. This results in additional interest to research works targeted to the theme. Despite the high similarity between the betting market and the stock market, the large difference in the transaction volume and information available clearly separates the number of research papers that we can find in both areas. Compared with other sports, the publication of research papers on soccer is below expectations, considering that it is the sport king of the world.

The annual *MIT Sloan Sports Analytics Conference* in Boston exposes the main sports research being done with focus on Basketball, Baseball, American Football, Ice Hockey, Tennis, and Soccer.

1.1 Motivation

A strong incentive in the study of sports-related analytics was the film *Moneyball* in 2011, about the book *"Moneyball: The Art of Winning an Unfair Game"*, from the author Michael Lewis, published in 2003. The book reports the 2002 season of the Oakland A's baseball team where, with a 41 million dollar budget, they were able to cope with the powerful 125 million dollar Yankees. The idea of creating a statistic analysis on a sport and take advantage of it has touched the minds of the gamblers community in general.

In Portugal, particularly, important steps have been taken to regulate the sports betting market. Since June 28, 2015 that the Decree-Law n.º 66/2015 [1] regulates the betting market through licenses issued by the certifying entity SRIJ (Service of Regulation and Inspection of Games).

After this regulation of the betting market in Portugal, the first entity to move forward with an offer for Portuguese betters was the *Santa Casa da Misericórdia*, through the game *Placard*, available only for face-to-face betting at authorized agents (Kiosks and stationery). The fact that *Placard* is only available for face-to-face betting, has stirred Portuguese culture in the betting market, thus drawing a diverse range of betters with very different gambling philosophies.

Next, three types of common betters will be presented, which will be used as a comparison reference for the model that will be defined in section 4.

[1] `http://www.srij.turismodeportugal.pt/fotos/editor2/legislacao/RJO%20Vers%C3%A3o%20inglesa_VF_1372015.pdf`

Non-risk Better This is a risk-averse better. He prefers to play safe and always bet on the most likely outcome.

Home Factor Better This is the old school style better, because he believes in the home-field advantage. Therefore, he is faithful to the strategy of always betting on the victory of the home team.

Random-walk Better This is the type of recreational better. This better does not make any analysis of the game, and may not even know the teams involved, but makes a bet for the pleasure of playing.

1.2 Objectives

This work aims the development of a betting model in the final result market - *HomeWin, Draw and AwayWin* - that can compete with the models currently followed by the bookmakers. In particular, it is intended to compete with the odds set by the *Betfair* exchange, where the price is a reflection of the supply and demand laws generated by betters. We will simulate *Betfair* exchange prices by averaging some bookmakers odds and removing the bookies margin. This model will be compared to the three types of betters defined in the previous section in order to have a real world comparison result.

The model will target the prediction of sports results where there is no clear favourite, since randomness in balanced games will be greater. This context of uncertainty will allow for a high number of opportunities to beat the market.

If a soccer match is an event with a well defined beginning and end, and with a component of unpredictability, why cannot be possible to do a statistical analysis that wins over the bookmaker? What information does the bookmaker have that I do not have?

2 Basic Concepts

In this section are introduced concepts related to sports betting and how bookmakers turn match probabilities into prices for their customers. The efficiency of the betting market will also be presented from the definition of Efficient Market Hypothesis (EMH). To conclude, we will also address the topic of Machine Learning in order to explore a possible inefficiency in the betting market through the detection of patterns in the data.

2.1 Soccer Probabilities

Three important concepts, which we will repeatedly address in game theory related to sports betting, are *probability*, *odd* and *expected value (EV)*.

Fair Odds in a Coin Flip To introduce the concepts of probability, odd and expected value (EV) in the complex sports betting market, we start from the simple example of tossing a balanced coin. Being the coin balanced, for a relatively high number of tosses, it is expected that about 50% of the times will come up head and another 50% of the times will come up tails. Since X is the 'tail' event, we can then say that one out of every two tosses it is expected to occur the event X.

We can define the probability of X from the Laplace rule for equiprobable events.

Definition 1. *Laplace probability*

$$probability = \frac{\#Favorable\ Cases}{\#Possible\ Cases} \tag{1}$$

$$probability\{tail\} = \frac{1}{2} = 0.5$$

If we want to turn event X into a betting event, the fair value to be paid to the better would be 2. This means that, if it is expected that the event will be held in two tosses, 2 is the fair price or fair odd.

The odd defined in different bookmakers may appear represented in different notations like decimal, fractional, percentage or American. Throughout this document we will use the typical notation in Europe, the decimal notation.

Being X the event *'Coming up tail when tossing a balanced coin'* which has a certain probability associated, then we define odd in decimal format.

Definition 2. *Decimal Odd*

$$Odd = \frac{1}{probability} \tag{2}$$

Returning to our tossing a coin event, the associated odd would be

$$Odd\{tail\} = \frac{1}{0.5} = 2$$

We can interpret the odd concept related to the event X as follows:

- **As a price** - The profit if the bet is won is $Bet\ Amount \times (Odd - 1)$
- **As a rate** - The event will occur once in Odd times

Regarding to the event *'Coming up tail when tossing a balanced coin'*, we can introduce the concept of expected value EV (indistinguishably also referred by $E[X]$) as the gains expected to achieve in the long run by repeatedly betting on event X.

Let X be a discrete random variable assuming the values $x_1, x_2, ...$ and $p(x)$ the probability function of X, $p(x_i) = P(X = x_i)$. We then define the expected value of X denoted by $E[X]$.

Definition 3. *Expected Value*

$$E[X] = \sum_{i=1}^{\infty} x_i P(X = x_i) \tag{3}$$

Let's then calculate the EV for bets of 10 € on event X.

Scenario 1 - Tail

$$\begin{aligned} Bet\ Result &= Bet\ Value \times (Odd - 1) \\ &= 10€ \times (2 - 1) \\ &= +10€ \end{aligned}$$

Scenario 2 - Head

$$\begin{aligned} Bet\ Result &= -Bet\ Value \\ &= -10€ \end{aligned}$$

Being the coin balanced, we have that 50 % of bets will be lost and 50 % will be won.

Let x_1 and x_2 be +10€ and -10€ respectively, both with 50% probability we have an expected value given by the following equation:

$$E[X] = +10€ \times 0.5 - 10€ \times 0.5 = 0€$$

Thus, it is trivial that when tossing a balanced coin, if we assign the odd of 2 both for the event head or tail, we have zero EV.

2.2 Bookmakers Odds

As described in the previous section, the amount paid by a bookmaker for a certain event is given by the inverse of the probability of that event occurring. This odd is called *fair value* or *fair price*. This is, fair value is the odd for an event that gives us a expected value of zero.

In the case of the bet on tail for tossing a coin, odd 2 is the value that in the long run would give an EV of zero. In terms of the market relative to the winner of a football game, there are three possible results: *HomeWin, Draw and AwayWin*. If all possible outcomes were equally likely to occur, we could assign a fair value of 3 (1/0.33) to each outcome.

Given the complexity involved in measuring the probability for the outcome of a football match, the odd defined for each match may have a *bias* associated with the game's own *fair value*. Thus, to safeguard that this bias is not against the bookmaker itself, the bookies define a price that gives them a profit margin, less dependent on the correct estimation of the true *fair value*.

For an outcome betting event, the Book Value [2] with a certain profit margin can then be defined as follows:

Definition 4. *Book Value*

$$Book\ Value = \frac{1}{Odd_{Home}} + \frac{1}{Odd_{Draw}} + \frac{1}{Odd_{Away}} \tag{4}$$

Let's imagine that for the match between *Liverpool FC* and *Chelsea FC* the following odds are set:

Liverpool FC	Draw	Chelsea FC
2.15	3.4	3.4

In this match the Book Value would be:

$$Book\ Value = \frac{1}{2.15} + \frac{1}{3.4} + \frac{1}{3.4} \quad = 1.05$$

The Book Value is then the value of the perfect book with an additional margin. This margin is the commission that bookie draws from the betting volume, either as a way to minimize its risk in the calculation of *fair value* or to directly increase its profits over all bets, regardless of the outcome of the game. Margin can be defined as follows:

Definition 5. *Margin*

$$\begin{aligned} Margin &= Book\ Value - Perfect\ Book \\ &= Book\ Value - 100\% \end{aligned} \tag{5}$$

Different methods of removing the margin for each of the outcomes are used by bookmakers. These are Equal margin distribution, Margin Weights Proportional, Odds Ratio, Logarithmic function.

[2] The Book Value is the sum of all odds of an event. When this sum is 100 %, we have a Perfect Book

2.3 Efficient Market Hypothesis

The Efficient Market Hypothesis was originated in the 1960s thanks to the work of economist Eugene Fama [5]. Eugene divided the market's efficiency into Weak-form, Semi-Strong and Strong-Form of EMH. The Weak-Form EMH notes that the market is efficient to the extent that current prices include all of its past. In other words, it states that the price of the past is already reflected in the current price, in a way that this last can not be predicted with greater precision than the current price. Semi-strong form covers Weak-form and adds that the current price quickly adjusts to new information that you can not profitably take advantage of. Strong-form, incorporates both Weak-Form and Semi-Strong forms of EMH and additionally reports that no type of public or private information allows you to take advantage of the present prices, i.e. these prices are already a reflection of that same information.

Compared with the betting market, the market efficiency has a direct parallelism in its three forms, but since in this work we will only use the prices available at the beginning of the game, the temporal oscillations obtained by new information regarding the game will be discarded. Therefore, we will work with the market efficiency in the Weak-Form.

If the betting market respects Weak-Form EMH, then it will not be possible to achieve profits systematically from betting systems based on the history of games analysed, since the price of the game at present will already reflect all this information itself, and thus the profits will be null in the long run. In order to test the efficiency of the betting market, we will use the prices available for the games of the past, and betting systems such as those defined in the section 1 tend to balance, leaning to zero profit in the long run. This applies to prices after the removal of the margin from the bookmaker.

2.4 Machine Learning

Machine learning is a field of study that includes knowledge of statistics, computer science and domain knowledge. The machine learning programming consists of learning from data itself, in opposition to the traditional rule-based programming.

The first use of the term machine learning dates back to 1959 by the author Arthur Samuel [9].

Definition 6. *Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed.*

Machine learning can be divided into two large groups, supervised and unsupervised learning. The prediction of the sports result *Home Win, Draw, Away Win* is inserted in the category of supervised learning, where the model will be fed previously with a set of features related to a soccer game and labeled with the outcome of that same game, in order to predict a categorial data class.

Definition 7. *Features are business domain attributes, which strongly characterize the prediction class. Features must be informative, discriminating and independent of each other.*

The big challenge is to create a model that can learn from the input dataset with high prediction accuracy and at the same time prevent overfitting. This will allow that this model can be applied to unseen data.

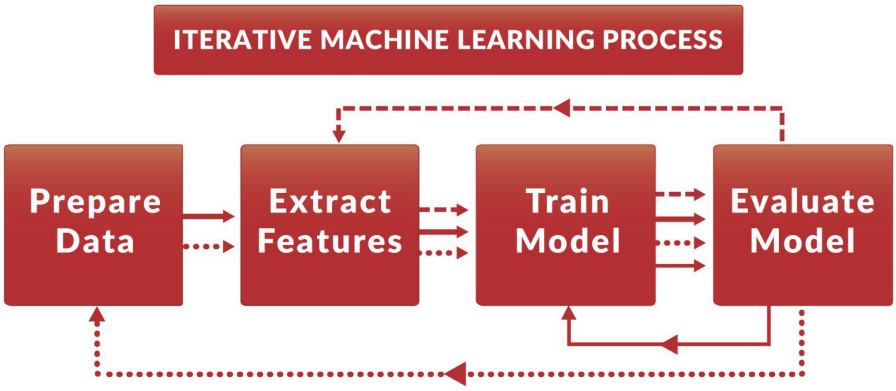

Fig. 1. Machine Learning Process

As presented by figure 1 machine learning is an iterative process that includes the Prepare Data, Extract Features, Train Model and Evaluate Model blocks. The Prepare Data phase includes the whole process of extracting and processing the data. In the phase of Extract Features or features engineering, is where the data is transformed from its original format, to raw data that contains characteristics that help to successfully classify our prediction class. It is followed by the training phase of the model with our dataset, and later by the evaluation of this same model. Further details regarding the evaluation of the model will be presented in section 5. The 12 lessons of Pedro Domingos [4] will be taken into account during the machine learning process.

3 Related Work

The evolution of the works applied to soccer has several divisions that include the study of the game style of the team, prevention of injuries, scouting, prediction of the exact result or match outcome, among others. *Beating the bookies* is the focus of this work, that consists in exploring possible market inefficiencies that may exist in specific contexts such as balanced soccer matches. There are many factors that may cause inefficiency in the sports betting market, such as poorly calculated odds by some bookmakers, betting volumes that force a variation in prices, or purposely misaligned prices to attract betters. The effectiveness of the betting market is being challenged by statistic models in an attempt of predicting the outcome of a game better than the bookmakers, and by the creation of machine learning algorithms that analyse a number of descriptive characteristics trying to increase the rate of success against the implicit probability that the bookies offers.

Avery et al. [1] explored the market inefficiency caused by 'noise traders', i.e. gamblers that bet on past winners, follow the advice of pseudo-experts, and bet on teams with prestige. In their work they adopted a betting system designed to exploit the sentiment-induced mispricing that resulted in a profitable strategy.

Arbitrage techniques have already revealed that the betting market includes inefficiencies that can be exploited. Vlastakis et al. [13] revealed that, although arbitrage opportunities appear to be rare, they are highly profitable, yielding returns between 12% and 200%. Through the wisdom theory of the crowd, authors of *Beating the bookies with their own numbers* [7] applied the same principle to the different odds that some bookmakers offer, using the average of all the bookies as true price, exploring the prices that deviated from that average. Their success has led them to be blocked by bookies with 'discriminatory practices against successful clients'.

Constantinou et al. [3] proposes Probability Score Classification in order to take into account the probability distribution and intrinsically the associated implied odds. A probabilistic bayes network model was presented that was used to generate the English Premier League (EPL) match forecasts during season 2010/11 with good results. Authors of *The market for English Premier League odds* [6] aimed to provide a probabilistic methodology for calibrating real-time market odds for the evolution of the score difference for a soccer game. Rather than directly using game information, they used real-time market odds to calibrate a Skellam model to provide a forecast of the final result.

As stated in the work of Steffen Smolka [10], the quantity of information available for free is limited, thus the phase of event engineering will require additional attention. The work is focused on EPL 2016/17, using offensive and defensive team performances features applied on Support Vector Machine and Neural Network models. Though they only evaluated the models based on accuracy measures.

As proved in Xu's research [14] the number of yellow cards applied to the home team in the last game is a feature to be taken into account. Based on the results of the probit regression model, bookmakers' odds for EPL 2006-

2007 are effective forecasts of soccer match outcomes, but could be improved by incorporating the effect of the number of yellow cards the home team receives in its last match.

Some works have achieved good results in terms of a prediction accuracy in the match outcome market, but were not analysed from profit point of view, i.e., betting prices were not taken into account. Vaidya et al. [12] used data regarding ten years from the EPL, on logistic regression, random forrest and naive bayes models. They have achived better results by using the three models combined in a voting system.

On the other hand there are works with a greater focus on creating betting systems in order to maximize profit, and with less focus on forecast accuracy. Zaan's thesis [15] demonstrated that staking the full optimal Kelly criterion maximizes the expected profit but at the cost of an enormous increase in variability. The results achieved by Moya et al. [8] offer compelling evidence that a finely tuned sports betting system involving a solid selection process and optimized staking has the potential to produce large profits with a limited initial bankroll after a relatively short amount of time.

Instead of testing the model with a dynamic betting system [11], which adjusts the value of the bet according to risk or probability of it, our work will follow a simple betting system so that the results obtained are clearly consequences of the model and not the system itself. The state of the art in the betting market, somewhat limited when compared to other studies, provides a good starting point for our solution that will be presented in the next section.

4 Solution

In order to contribute to the work presented in the previous section, this chapter details a solution that intends to take a step forward in the study of predicting the outcome of a soccer match. First, the main requirements of the solution are presented, followed by a description of the architectural proposal. Finally, the planning for the implementation of the solution is presented.

4.1 Requirements

To create guiding lines for the implementation of the project, as well as its evaluation, three requirements and three non-requirements were raised.

In the developed solution it should:

- Be possible to include/exclude the bookmakers margin from match odds.
- Be possible to filter the matches to include in the model by their minimum odd value.
- Be evaluated in terms of Accuracy and Profit.

The developed solution should not:

- Include data that is not freely accessible and of public knowledge.
- Be evaluated by the complex betting/banking system.
- Be tested on real bets due to legal restrictions that differ from country to country.

4.2 Architecture

The typical architecture of a machine learning project was adopted to implement the solution. In figure 2 can be observed the main modules that include Historical Data, Feature Engineering, Split Datasets, Model Creation and Evaluation. Next, these are presented with a brief explanation of each one.

Historical Data The dataset used [3] includes data from 11 different countries, but special focus will be given to the five European elite leagues (*Bundesliga I,La Liga, Ligue 1,Premier League and Serie A*). More detailed information on the dataset will be presented in section 4.3.

Feature Engineering Features extracted directly from the dataset will be used, as well as new synthetic features that will be created from the data. This is the module where the business logic will be implemented.

[3] `http://www.football-data.co.uk/data.php`

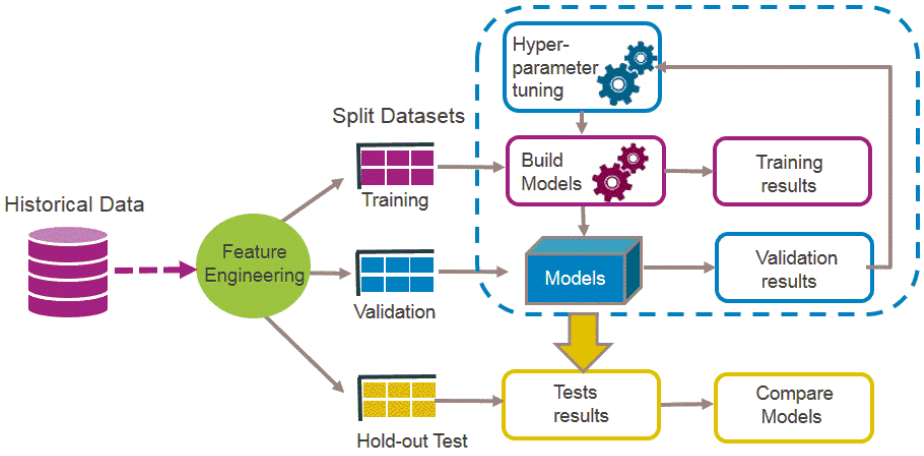

Fig. 2. Solution Architecture

Split Datasets Although the volume of data is limited when compared to other applications of machine learning these days, the literature suggests a clear separation between training and testing dataset. If it is possible, the validation of the model will be done in the separate validation dataset.

Model Creation This module is generally seen as a black-box, since the complexity and diversity of machine learning models is high. The focus is on achieving a balance between improving the accuracy of the model, while avoiding overfitting the training dataset.

Evaluation This module is responsible for making all the work done measurable. Further details can be found in section 5.

4.3 Solution Description

In this subsection each module of the solution architecture presented in figure 2 will be discussed in further detail. An iterative implementation of the whole machine learning process will be conducted, consisting in two interactions called Phase 1 and Phase 2.

Historical Data The dataset used to support the project will be separated in two phases. The first one will allow to raise possible problems about limiting the volume of data or features that are not very informative, or that are very correlated between them.

The second phase will focus on achieving a compromise between adding data volume (more leagues or seasons) or adding new features, taking into account that the overhead associated with cleaning and data handling.

Next is presented a brief description of the possible data to include in the model for each phase:

- **Phase 1**
  - Leagues: Bundesliga I,La Liga, Ligue 1,Premier League and Serie A
  - Season: 2007/08 to 2016/17
  - Odds: Average odds from Bet365, Blue Square, Bet&Win, Gamebookers, Interwetten, Ladbrokes, Pinnacle, Sporting Odds, Sportingbet, Stan James, Stanleybet, VC Bet, William Hill.
  - Fields: More detailed information regarding the Match Info, Match Statistics and Odds is found on appendix A.

- **Phase 2**
  - Include more Seasons
  - Include more Leagues
  - Include Elo Ratings
  - Include Fatigue data

Feature Engineering This module will be responsible for the success of the entire model. According to other works [10,12] each team will be separated in a defensive and offensive form, but all features will be normalized to values from zero to one. This allows that in any matchday, the value of the feature can reflect the relative strength of the feature of that team, compared to the total of its championship or team cluster in which is inserted. The Home Field Advantage factor will also be taken into account [2] as well as the yellow cards as suggested in Xu's work [14].

- **Phase 1**
  - Offensive and Defensive form
  - Offensive and Defensive goals
  - Relative classification
  - Cards flag
- **Phase 2**
  - Elo Rating factor
  - Fatigue
  - Sliding window metrics
  - League Related Info

Split Datasets All data splitting will be done randomly, to avoid training the model with sequential matches or matches of a single outcome class. In a first phase the data separation will be done at 70% for training and 30% for testing. In a second phase the validation dataset will be included and the separation will change to 60% - 20% - 20% for training, validation and testing respectively.

Model Creation Different models will be tested in order to realize in practice the significant differences between them. In the first phase will be tested the most common algorithms, which in the second phase will be pruned to optimize their results.

Classification Algorithms

- Naive Bayes
- Logistic Regression
- Support Vector Machines
- Decision Trees
- Neural Networks

Evaluation The model evaluation should contemplate both magnitudes, a good performance of accuracy and profit, comparing with other benchmark models through systems of two different betting systems.

In both phases the best performance will be sought in the two strands listed above. The difference between phases 1 and 2 will be in the subsets of data used. Further details on the evaluation of the system will be addressed in the 5 section.

5 Evaluation Method

Special importance will be given to the evaluation component of this project and different tests will be conducted in order that the achieved conclusions are according to the data itself. The evaluation will be done combining accuracy and profit measures to produce more consistent results.

If part of the process of machine learning is a black-box, sometimes complex to read for the analyst, it is also the point where we will translate our analysis into measurable results that can be interpreted by the human being.

Below are presented the metrics under which the model will be evaluated, together with the list of benchmark betting systems for comparison, and some data related questions.

5.1 Accuracy Metrics

In order to evaluate the quality of the model predictions different metrics will be used.

Accuracy

Simply measures how often the classifier makes the correct prediction. It's the ratio between the number of correct predictions and the total number of predictions. It is calculated from the following equation:

$$Accuracy = \frac{\#Correct}{\#Predictions} \tag{6}$$

Confusion Matrix

A confusion matrix of binary classification is a two by two table, formed by counting of the number of the four outcomes of a binary classifier, i.e. *true positives*, *true negatives*, *false positives* and *false negatives*.

Table 1 presents the confusion matrix, where in the columns are the predicted classes and the rows are the actual true classes.

	Positive	**Negative**
Positive	True Positive	False Negative
Negative	False Positive	True Negative

Table 1. Confusion Matrix

Precision and Recall

The precision score quantifies the ability of a classifier to not label a negative example as positive. The precision score can be interpreted as the probability that a positive prediction made by the classifier is positive. Precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. The score is in the range [0,1] with 0 being the worst, and 1 being perfect.

The precision and recall scores can be defined as:

$$precision = \frac{\#true\ positives}{\#true\ positives + \#false\ positives} \tag{7}$$

$$recall = \frac{\#true\ positives}{\#true\ positives + \#false\ negative} \tag{8}$$

F1 Measure

The F1-score is a single metric that combines both precision and recall via their harmonic mean:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

5.2 Profit Metrics

Regarding profit evaluation, two simple betting strategies will be adopted, fixed-size and fixed-profit presented next.

Fixed-Size

$$Stake_i = Constant\ bet\ value \tag{10}$$

Fixed-Profit

$$Stake_i = \frac{profit}{odd - 1} \tag{11}$$

The choice of simple betting strategies aims to minimize the impact of this component, since banking management is out of the scope of this work.

5.3 Benchmark Models

The performance of the created model will be compared to the three betting systems defined in section 1, so that all results will be easily comparable with real betting cases, as follows:

- Non-risk Better
- Home Factor Better
- Random-walk Better

5.4 Questioning the Data

Much of the analytic insights results from questioning the data. The following questions will guide us through the testing process for the defined models:

- What is the impact of training the model with all the championships at once, or one at a time?
- How does the model behave if we restrict ourselves to more balanced matches where uncertainty is greater?
- If we remove the implicit uncertainty in the opening and closing season matches do we have an impact on the results?

After answering the previous questions, sure that many others will emerge, which will provide good insights to the behaviour of the created model, as well as to the definition of new features to incorporate.

6 Conclusion

The soccer industry is undoubtedly growing, maximizing the volume traded in the betting market. Several works in the attempt to beat the models created by the bookmakers have been developed and, although they can not demonstrate totally consistent results, they provide good starting points. A gap in the state of the art still exists in terms of the evaluation of the results, either because they are limited to simple tests of accuracy of the models or because they introduce complex betting strategies or bankroll management that do not allow us to draw conclusions from the model itself. This work aims to exploit betting inefficiencies, present in particular scenarios like balanced matches, matches with reduced book value, isolated championships, etc. This is motivated by the fact that a global market efficiency does not mean that this efficiency is maintained under specific business conditions.

6.1 Expected Results

With the different test conditions that will be applied in conjunction with a new feature creation, it is expected to identify consistent patterns in the betting market and, above all, to detect inefficiencies in the market. In particular, it is expected to create a model that performs better than *Non-risk, Home factor* and *Random-walk* betters.

6.2 Preliminary Results

As starting point, the five elite leagues of European football were exploited for the 2012/13 seasons until 2016/17 in the subset of balanced matches, i.e. with a minimum odd greater than 2.3, results in 2027 soccer matches.

As shown in figure 3 a bias of 3% (61.80 euros) in terms of the tie market can be retrieved within our 2027 matches sample.

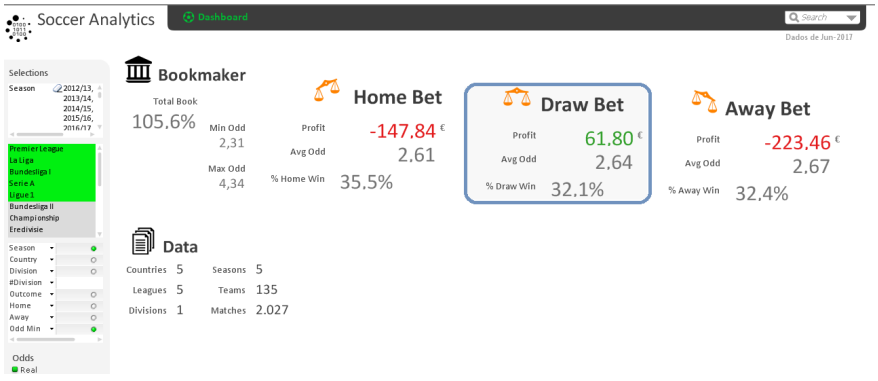

Fig. 3. Draw bias

With direct exploration it was possible to identify a 3% inefficiency in the market, thus it is expected that using machine learning techniques it is possible to identify more cases that allows to create a consistent strategy within the complex and competitive betting market.

6.3 Schedule

Figure 4 presents the *Gantt* calendar planning the three major components, i.e. Writing MSc Thesis, Phase 1 and Phase 2. As mention in section 4 the same tasks will be performed during both phases , but with different levels of detail.

The Historical Data task comprehends the extraction of data and cleaning operation. In the second phase new sources will be included.

Feature Engineering is the task associated to the creation of features directly from the fields present in source data, as well as changes made on these same fields.

Split Datasets is where we separate the data into training, validation and testing datasets, in order to train, evaluate and test the model with different data.

The Model Creation task is where different classification models will be created and compared to each other during the Evaluation task.

Test and Debug are the tasks intrinsic to any development project, where bugs that arise during development are fixed.

The Writing MSc Thesis task will be performed throughout all development stages,in order to document the steps taken and the reasons why key decisions were made.

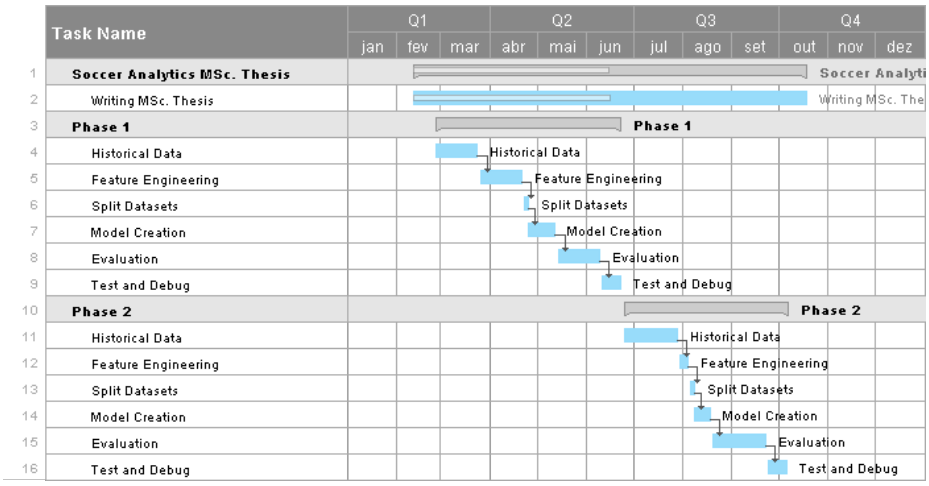

Fig. 4. Gantt Calendar

References

1. Avery, C., Chevalier, J.: Identifying investor sentiment from price paths: The case of football betting. The Journal of Business 72(4), 493–521 (1999)
2. Carmichael, F., Thomas, D.: Home-field effect and team performance: Evidence from english premiership football. Journal of Sports Economics 6(3), 264–281 (2005), https://doi.org/10.1177/1527002504266154
3. Constantinou, A.C., Fenton, N.E., Neil, M.: Pi-football: A bayesian network model for forecasting association football match outcomes. Know.-Based Syst. 36, 322–339 (Dec 2012), http://dx.doi.org/10.1016/j.knosys.2012.07.008
4. Domingos, P.: A few useful things to know about machine learning. Commun. ACM 55(10), 78–87 (Oct 2012), http://doi.acm.org/10.1145/2347736.2347755
5. Fama, E.: Efficient capital markets: A review of theory and empirical work. Journal of Finance 25(2), 383–417 (1970), https://EconPapers.repec.org/RePEc:bla:jfinan:v:25:y:1970:i:2:p:383-417
6. Feng, G., Nicholas, P., Jianeng, X.: The market for english premier league (epl) odds. Journal of Quantitative Analysis in Sports 12(4), 167–178 (2016), https://EconPapers.repec.org/RePEc:bpj:jqsprt:v:12:y:2016:i:4:p:167-178:n:1
7. Kaunitz, L., Zhong, S., Kreiner, J.: Beating the bookies with their own numbers - and how the online sports betting market is rigged. ArXiv e-prints (Oct 2017)
8. Moya, F.E., Gill, P.S.: Statistical methodology for profitable sports gambling. In: NA (2012)
9. Samuel, A.L.: Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 3(3), 210–229 (Jul 1959), http://dx.doi.org/10.1147/rd.33.0210
10. Smolka, S.: Beating the bookies: Predicting the outcome of soccer games. NA p. 5 (7 2016)
11. Stenerud, S.G.: A Study on Soccer Prediction using Goals and Shots on Target. Master's thesis, norwegian university of science and technology (6 2015)
12. Vaidya, S., Sanghavi, H., Gevaria, K.: Football match winner prediction. International Journal of Computer Applications 154(3), 31–33 (Nov 2016), http://www.ijcaonline.org/archives/volume154/number3/26474-2016912066
13. Vlastakis, N., Dotsis, G., Markellos, R.N.: How efficient is the European football betting market? Evidence from arbitrage and trading strategies. Journal of Forecasting 28(5), 426–444 (2009), https://ideas.repec.org/a/jof/jforec/v28y2009i5p426-444.html
14. Xu, J.S.: A look into the efficiency of bookmakers odds as forecasts in the case of english premier league. NA p. 27 (5 2011)
15. Zaan, T.v.d.T.: Predicting the outcome of soccer matches in order to make money with betting. Master's thesis, Erasmus University Rotterdam (Mar 2017), http://hdl.handle.net/2105/37404

A Appendix

The dataset used in phase 1 was extrated from `www.football-data.co.uk` and contains information related to the match itself, additional match statistics and bookmakers prices for the event match outcome.

A full description of the included fields is listed below.

Match Info

$$
\begin{aligned}
Div &= LeagueDivision \\
Date &= MatchDate(dd/mm/yy) \\
HomeTeam &= HomeTeam \\
AwayTeam &= AwayTeam \\
FTHG &= FullTimeHomeTeamGoals \\
FTAG &= FullTimeAwayTeamGoals \\
FTR &= FullTimeResult(H = HomeWin, D = Draw, A = AwayWin) \\
HTHG &= HalfTimeHomeTeamGoals \\
HTAG &= HalfTimeAwayTeamGoals \\
HTR &= HalfTimeResult(H = HomeWin, D = Draw, A = AwayWin)
\end{aligned}
$$

Match Statistics

$$
\begin{aligned}
HS &= HomeTeamShots \\
AS &= AwayTeamShots \\
HST &= HomeTeamShotsonTarget \\
AST &= AwayTeamShotsonTarget \\
HC &= HomeTeamCorners \\
AC &= AwayTeamCorners \\
HF &= HomeTeamFoulsCommitted \\
AF &= AwayTeamFoulsCommitted \\
HY &= HomeTeamYellowCards \\
AY &= AwayTeamYellowCards \\
HR &= HomeTeamRedCards \\
AR &= AwayTeamRedCards
\end{aligned}
$$

Odds

$$
\begin{aligned}
BbAvH &= Betbrainaveragehomewinodds \\
BbAvD &= Betbrainaveragedrawwinodds \\
BbAvA &= Betbrainaverageawaywinodds \\
BbAv{>}2.5 &= Betbrainaverageover2.5goals \\
BbAv{<}2.5 &= Betbrainaverageunder2.5goals
\end{aligned}
$$

B Glossary

Bookmaker The bookmaker, also called the bookie or simply 'the house', it refers to the business or organization that provides an odds market for sporting events, with prices available for all possible outcomes. A 'book' is simply the full record of all betting transactions made with the bettors for a particular event.

Event This refers to the specific sporting event. Examples of events are India vs Sri Lanka playing the final of the Cricket World Cup or Real Madrid playing against Barcelona in the Spanish Soccer League.

Market A betting market is a type of betting proposition with two or more possible outcomes. The result of the match (home win, away win, or draw), the number of goals scored (two or less goals, three or more), or the time of the first goal are a few examples of different markets for a single sporting event.

Bank The total amount of money a gambler has to place bets on sporting events.

Stake The amount of money being risked in a single bet.

Fair odds The odds that would be offered if the sum of the probabilities for all possible outcomes were exactly 1 (100%). For example, supposing we had a market with three possible outcomes A, B, C with probabilities of success P(A) = 0.5, P(B) = 0.4 and P(C) = 0.1, the fair odds would be 2.00, 2.50, and 10.00 respectively, which are just the inverse of the estimated probabilities.

Overround Also called vigorish (or vig for short) in American sports betting, the overround is a measure of the bookmaker's edge over the gambler. The bookmaker will never offer fair odds on a market. In practice, the payout offered on each selection will be reduced, which in turn increases the reflected probability of an event. When odds have been adjusted in this way the sum of the probabilities for all events will exceed 1 (100%). The overround is the amount by which the sum of all probabilities exceeds 100% and it is the bookmaker's profit margin. For example, if we had a market with two possible outcomes A, B, where P(A) = P(B) = 0.5, the fair odds on each selection would be 2.00. However, the bookmaker may offer payouts of 1.85 on each selection. The corresponding probabilities for each selection are now 1/1.85 = 0.5405, and the sum of the probabilities for all outcomes is 0.5405 x 2 = 1.081. The overround is 8.1%, and for every $100 paid out by gamblers the bookmaker expects to make a profit of 8.1 dollars, assuming that there are balanced bets on both A and B.

Pick The selection among all the possible outcomes on which the gambler is placing the bet.

Result The actual outcome of the event. If the pick and the result are the same the gambler wins the bet and is paid an amount equal to his stake times the odds offered on the selection. If the result is different from the pick the gambler loses his entire stake.

Profit The amount of money additional to the original stake that the gambler receives when the bet is won. Bookmakers sometimes use the term Winnings, but this term refers to the amount of money paid back including the original wager, which is somewhat misleading. It is preferable to speak about the profit made in a bet instead of the winnings of a wager.

Yield A measure of the profitability of a series of bets, it is calculated as the sum of the profit made from all the placed bets divided by the sum of the money staked in all bets, usually expressed as a percentage. For example, if after 10 bets of \$1 each there is a net profit of \$1.50, the yield is $(1.5/10) = 0.15$, i.e. 15%.