Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach

Research · September 2015

DOI: 10.13140/RG.2.1.1383.4729

2 authors, including:

Niek Tax
Eindhoven University of Technology
41 PUBLICATIONS **271** CITATIONS

Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach

Niek Tax and Yme Joustra

Abstract—This paper describes a public data based match prediction system for the Dutch Eredivisie. Candidate features was established through a structured literature study to identify factors with predictive value for match outcome. Model training was done on a self-made dataset from public sources, consisting of thirteen seasons of Dutch Eredivisie match data. Several combinations of dimensionality reduction techniques and classification algorithms have been tested on the public data training set in a structured way. The highest prediction accuracy on the public data feature set was achieved by using a combination of PCA (with 15% variance) with a Naive Bayes or Multilayer Perceptron classifier. Models have been created for a betting odds feature set and a hybrid feature set (the union of public data and betting odds features). McNemar's test showed no significant difference in accuracy of the highest accuracy hybrid feature set model and the highest accuracy betting odds features set model, but the results are still such to raise the assumption that a hybrid feature set of betting odds and public data is able to beat the bookmakers. Obtained results can be seen as a positive sign that it might be possible to engineer profitable betting decision support systems based on open data.

Index Terms—Sports, Data mining

1 Introduction & Research Questions

Forecasting of football matches has long been of interest to bookmakers determining odds for football matches as well as for punters trying to gain a football betting profit. With football being one of the most popular sports worldwide, a lot of data and statistics are recorded for the major competitions. The existing amount of football data has been a popular object of study within multiple scientific disciplines, including statistics and applied economics. Several studies have been conducted in football prediction, using either Machine Learning (e.g. [1], [2]) or statistical techniques (e.g. [3]). Besides forecasting, a lot of research has been conducted in the field of statistical identification factors of indication for football match outcome (e.g. [4], [5]). As the factors that were taken into account in predicting football match results differ between studies, there seems to be room for improvement in prediction accuracy. It is notable that studies describing machine learning match prediction systems hardly make any use of work on the factors influencing match results originating from the statistics or sport economics fields of science.

RQ1 *Which factors can be found for which studies have concluded them to be of predictive value for football match outcome?*

Football data is openly available on the web for most football competitions. This opens up the opportunity to built a football prediction system using only publicly available data. Aranda-Corral et al. [6] showed some factors to be unsuitable for use in automatic prediction systems due to being 1) hard to quantify (e.g. fan support) or 2) hard to retrieve automatically (e.g. injuries). Which bring us to the second research question.

RQ2 *Which factors described in literature are quantifiable and can be automatically retrieved for the Dutch football competition from publicly available data sources?*

Our research will take into account the latest thirteen seasons of the Dutch Eredivisie competition (2000-2013). A larger time scope is expected to have negative consequences for the number of automatically retrievable factors. We expect the amount of factors that are quantifiable and retrievable from public data sources to be resulting in a feature set size in the order tens to hundred, which is too large for our sample size in the order of thousands to prevent overfitting problems [7]. Several dimensionality reduction methods have been proposed through the years, of which the most common methods are described in survey articles by Molina et al [8] and Salappa et al

• *N. Tax and Y. Joustra are with the Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, Netherlands.*

[9]. It is to be researched which combination of dimensionality reduction techniques and machine learning techniques are best fitting in the context of Dutch football competition match prediction.

RQ3 *Which dimensionality reduction method and Machine Learning method results in the highest prediction accuracy in the Dutch football competition match result prediction task?*

Note that the third research question explicitly aims at finding the best prediction for the Dutch football competition instead of football competitions is general. We consider a prediction model created specifically for the Dutch football competition to be inapplicable to other football competitions, since different football competitions can be expected to have different standards and characteristics. E.g., Based on studies from Carmichael and Thomas [10] and Seçkin [11] combined we can conclude the home advantage effect to be different for the Turkish football competition (Super League) than for the English football competition (Premier League). Another factor limiting the applicability to other competitions is a difference in balance between football competitions, Goossens [12] demonstrated existence of balance differences between European football competitions. In a competition with higher balance, the chance of the underdog beating the title chaser will be higher.

Bookmakers have been offering odds on the various outcomes of a match for years, thereby they have already been creating match outcome prediction systems for years. Research [13], [14] has shown bookmaker odds to be a good predictor of match outcome, outperforming expert predictions. Goddard [5] remarks that bookmaker odds might partially be driven by insiders with access to privileged or last-minute information which is hard to capture in models, therefore it might be difficult to beat bookmaker predictions with a public data model. Bookmaker odds will be a good basis of evaluation for the usefulness of the public data prediction model, because of the high accuracy of bookmaker odds as a match outcome predictor. Therefore we formulate the fourth research question.

RQ4 *How does the public data model perform (in terms of prediction accuracy) compared to bookmaker odds?*

It might be the case that a public data model which includes bookmaker odds is able to outperform bookmaker odds, where a model only making use of public data is not. Therefore we will compare the prediction accuracy of a hybrid model, combining both public data and bookmaker odds into one model, to the prediction accuracy of the public data model and bookmaker odds model.

RQ5 *Is it possible increase performance (in terms of prediction accuracy) when combining betting odds and public data into a hybrid model, compared to a model using only one of both?*

2 METHODOLOGY

A structured literature study will be conducted to identify factors that might helpful to predict football matches. The structured literature study will be performed using the following set of queries: (”soccer” — ”football” — ”sport”) (”match” — ”result” — ”outcome”) (”prediction” — ”predicting” — ”forecast”) on the bibliographic databases Scopus and Web of Science. Articles found will be selected on whether their title or abstract contains the words *factor*, *feature*, *attribute* or a plural form or synonym of one of these words or contains a term from the fields of Machine Learning or statistics, as this often might suggest that either feature sets or dependencies are to be discussed. This process will yield an answer to our first research question.

For each factor that we will identify during the literature research we aim to retrieve quantified data from public web sources. We predefine the scope of our search for public web sources to the websites listed in Table 1 to be able to perform this search in a structured and consistent manner. This list of public web sources is composed by using the three most popular football websites in the Netherlands, combined with a website specialized in transfer market data and the website of the company holding the Dutch Eredivisie broadcasting rights. Based on these websites we can assess the feasibility of automatic retrieval of identified factors and therefore answer the second research question.

Public sources
www.vi.nl
www.elfvoetbal.nl
www.fcupdate.nl
www.transfermarkt.co.uk
www.foxsports.nl

TABLE 1
Search scope of public data sources

There are two distinct football match prediction tasks conceivable. One could predict the number of goals scored by each team, or one could forecast into the categories win/draw/lose. Goddard [3] showed that using the win/draw/lose approach allows higher accuracy compared to forecasting match outcome through predicting the number of goals scored by both teams, even when win/draw/lose forecasts are deduced from those predicted number of goals. Because the categorical version of the match prediction

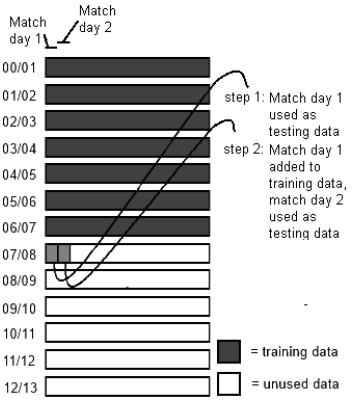

Fig. 1. Visualization of data used for training and testing in each step

task is sufficient to be used in a sport betting contest and because it allows a higher level of accuracy the categorical classification approach will be used.

Cross validation, the most common evaluation method in machine learning, cannot be used for the football classification task because of the temporal nature of the data. Using a cross validation evaluation method results in a problem that the classifier algorithm has an opportunity to learn from instances later in time compared to some of the test instances, therefore the classifier could have learned patterns in the data that it would not have been able to learn when chronological order was preserved. To avoid overestimating classification accuracy due to this problem we will use a purely retrodictive training approach. The football competition progresses through the year and after each match more training data becomes available. To include this new training data retraining the model is required after each new match if we want to evaluate a Machine Learning model taking into account the latest available data. To prevent classification of data points while only little data has been used for training during the evaluation process we use the first seven seasons of the thirteen retrieved seasons that we retrieved only for training, thus, not for evaluation.

Figure 1 visualizes the retrodictive approach used to evaluate the accuracy of different Machine Learning methods. It shows that for the first match of the eighth season, which is the first validation data point, the data points of the first seven seasons are used as training data. For each given match day in one of the seasons between 07/08 to 12/13 we use all the data points of all matches played in match days prior to this particular match as training data. To explain the evaluation procedure used to select the best performing model, say we start with *val = the first match of season 07/08* and *train = season 00/01 to 06/07*, the evaluation procedure consists of the following steps:

1) Train the model on *train*
2) Classify *test* with the model and check if prediction was correct
3) Adjust aggregated prediction accuracy
4) Add *test* to *train*
5) Set *test* to the following match in the set
6) If not at the end of the data set, go to step 1

Choices have to be made which dimensionality reduction techniques and classification techniques to include in the experiment, and which not to. Evaluating all existing dimensionality reduction and machine learning techniques is infeasible. The dimensionality reduction techniques and classification techniques that are to be evaluated with our retrodictive approach will be determined subsequent to the literature study, for which we have two reasons:

1) Which dimensionality reduction techniques and classification techniques can be expected to perform well depends on the characteristics of the set of candidate features
2) Related work found during the literature study performed to identify factors is expected to give insight in which dimensionality reduction and classification techniques worked well in related studies.

Bookmakers have long been doing match predicting with the purpose to calculate of their betting odds. Because of bookmakers long experience in the task and their commercial interest in the task, betting odds data will be used as baseline for evaluation of the to be created public data model. Odds data to be used for comparison consists of a combination of data from three bookmakers (Gamebookers, Bet&Win and Betbrain) and two forms (Asian Handicap and traditional 1x2 wagering). Asian Handicap betting is a form of betting that handicaps teams according to their form, therefore stronger teams must win with a certain goal difference for the punter to win the bet. The two bookmakers were used because of their popularity and because they complemented each other's missing data very well. Vlastakis et al [15] showed Asian handicap odds to be strongly interrelated to match outcome for the English Premier League in a nonlinear fashion.

McNemar's test [16] will be used to test whether or not the baseline or the public data based model significantly outperforms (on a 0.05 significance level) the other, therefore yielding the answer to the fourth research question.

For the fifth research question we create a hybrid model by adding betting odds features to the public data model, comparing their performance with McNemar's test (on a 0.05 significance level).

3 Factors

3.1 Previous performance in current season

3.1.1 Matches won, lost, ended in a draw

Miljkovic et al [17] showed the win/loss/draw ratios of basketball teams to be relevant features in predicting National Basketball League (NBA) basketball

matches. Aranda-Corral et al [6] modeled the currents seasons performance for a football classification task in one feature: points collected in previous matches.

3.1.2 Goals scored and conceded

Baio et al [1] proposed Gamma distribution mixture model for Italian Serie A match prediction, relying on the amount of goals scored and conceded of both teams. Miljkovic et al [17] used, amongst other features, the number of field goals scored and conceded and the number of 3-pointers scored and conceded for the prediction of basketball match outcome prediction. With feasible automatic retrieval of win/lose/draw percentages and average number of scored and conceded goals we add those features to the candidate feature set.

3.2 Performance in earlier encounters

Aranda-Corral et al [6] found the previous matches between the same teams to have medium to high correlation with the three match result categories in their study focusing on the Spanish national soccer league. Aranda-Corral et al for their prediction model decided to encode this into three categories: #wins, #draws and #losses for one of the two teams in current and all previous seasons. Hucaljuk and Rakipovic [18] also included the outcome of previous meetings between the two teams into his model for Champions League prediction.

3.3 Streaks

Heuer and Rubner [19] showed that losing streaks do negatively influence a team's winning probability, while winning streaks do not have any positive influence on winning probability. Goddard [5] has contradictory findings, concluding that losing streaks result in an increased winning probability and winning streaks result in a decreased winning probability. Constantinou et al [2] modeled form as one of the probabilistic directed acyclic graphical models created by football experts. Even though the helpfulness of streaks/form in predicting match outcomes is still under discussion, the ease of its automated retrieval makes it appropriate to add to the candidate feature set. In case the streak features turn out not to be helpful for match outcome prediction we can expect them to be filtered out by the dimensionality reduction methods.

3.4 Managerial change

Table 2 provides an overview of studies on the effect of managerial change on team performance. Different studies conclude different effects for managerial change on team performance. Koning [20] and Balduck et al [4] hypothesize this effect to be dependent on the new managers characteristics and team characteristics, but do not provide what those characteristics might be. Even though research in the effects of managerial change fails to draw corresponding conclusions, the ease of retrieval of managerial change data makes it worthwhile to add managerial change information to the candidate feature set.

3.5 Home advantage

Palomino et al [23] and Carmichael and Thomas [10] showed that home advantage plays an important role in predicting match outcomes. There are however different findings on the role of home advantage, from studies by Carmichael and Thomas [10] and Seçkin [11] combined we can conclude the home advantage effect to be different for the Turkish football competition (Super League) than for the English football competition (Premier League). We take home advantage into account in the candidate feature set by adding home match performance (win/draw/loss percentages) features for the home playing team and away match performance features for the visiting team.

3.6 Matches with special importance

Goddard [5], [24] showed specific end-of-season matches, where a team still competes for participation in European competition for the next season or the prevention of relegation, to be a significant for match outcome.

3.7 Fatigue

Constantinou et al [2] use fatigue as one of the four features in his prediction model. Constantinou et al modeled fatigue as consisting of the toughness of the previous match, the number of days since the latest match, the number of regular first team players rested and the participation of first team players in national team matches.

3.8 National team players

Aranda-Corral et al [6] showed the presence of national team players in a team to be somewhat helpful in the prediction of match outcome for some national teams, but not for all. Aranda-Corral et al found these results in a study based on the Spanish La Liga. It is conceivable that the partitioning to relevant and non-relevant nationalities is conceivable to be different for the Dutch Eredivisie compared to the Spanish La Liga, because the level of both competitions differ. As the predicting value of the presence of national team players in a team is seems to be limited and because it hard to draw a strict non-arbitrary boundary between relevant and non-relevant nationalities for the Dutch Eredivisie we neglect the presence of national team players in the candidate feature set.

Study	Competition	Conclusion
De Paola and Scoppa [21]	Italian Serie A	Statistically insignificant positive effect
Balduck at al [4]	Belgian first, second and third division	Dependent on team and coach characteristics
Heuer et al [22]	German Bundesliga	No effect
Goddard [5]	English Premier League	Negative effect
Koning [20]	Dutch Eredivisie	Dependent on coach characteristics. Decrease in goals conceded.
Aranda-Corral et al [6]	Spanish La Liga	High correlation between the hiring of a new coach and match outcome

TABLE 2
Overview of studies on the effect of managerial change

3.9 Promotion to higher league

Aranda-Corral et al [6] found the participation of a team in a lower league in the to foregoing season to be of medium predictive value for match outcome for the Spanish La Liga. As the relation between participation in a lower league in the foregoing season and match outcome might also be present for the Dutch Eredivisie and automated retrieval of this feature is feasible we add this feature to the candidate feature set.

3.10 Expert predictions

Khazaal et al [25] showed that football expertise, age and gender did not have any impact on the accuracy of football match prognoses. As expert predictions are shown not of predictive value to match outcome and automated retrieval of expert predictions is hard we did not use expert predictions in the candidate feature set.

3.11 Football skills

Min et al [26] proposed the use skill grades on team level for offense, defense, possession and fatigue based on expert opinion. A likewise approach is inappropriate for a system making use of automated data retrieval, as the skill level rating task was performed manually. Ratings used in sport games such as the FIFA video game series might be an adequate substitute to a manually obtained expert opinion, but the difficulty in automated obtaining of historical information still makes it unsuitable. Football skills as perceived by football experts are not used as candidate feature.

3.12 Strategy

Palomino et al [23] used a game theoretic approach to show the influence of different playing strategies (e.g. attacking style, defending style) on match outcome. As the game strategy might change throughout the game and because the starting strategy is hard to automatically retrieve this factor is not suitable to include in the candidate feature set.

3.13 Travel Distance

Goddard [5] found the magnitude of home advantage in English league football to be dependent on the geographical distance the away team has to travel. Goddard [3] found the natural logarithm of the travel distance for the visiting team to be a useful feature for predicting English league football. One might argue this factor to be less relevant for the Dutch competition, where travel distances are only small. As travel distance is easily retrieved, they will be used in the candidate feature set.

3.14 Betting odds

Vlastakis et al [15] showed the size of Asian Handicap betting odds to be a significant predictor for both home and away team scores. An interesting finding was also that Asian Handicap odds appeared to be a better predictor for match outcome than regular 1X2 odds. Odachowski et al [27] conducted a machine learning experiment predicting football match results solely based on betting odds features on a data set of football matches from varying football competitions. Using only betting odds features showed to work well, with a prediction accuracy of 70%.

3.15 Club budgets

Aranda-Corral et al [6] proved the difference between team's budgets to be highly correlated with the result of a match using a Formal Concept Analysis (FCA), in a Spanish La Liga study. Aranda-Corral et al, taking into account only several past matches, concluded that club budgets are feasible to collect automatically. Because historical club budgets are hard to automatically retrieve, our validation methodology that takes into account several seasons of data makes the use of club budgets as candidate feature difficult.

3.16 Availability of key players

Constantinou et al [2] had football experts create a set of probabilistic directed acyclic graphical models on football match outcome. The availability of a team's key player(s) was drawn as part of this model by the football experts. We chose to represent this by the presence of the top scorer and top assister of a

team in the current season because the term *key player* as formulated by the experts is not objective and not measurable. As the impact of a team's the top scorer and top assister presence can be dependent on the team, we added an average goals (for top scorer) and average assists (for top assister) per match features to the candidate feature set.

4 CANDIDATE FEATURES

Based on the identified factors and their automatic retrieval possibilities (as described above) the following candidate features have been selected to form the candidate feature set. The candidate features listed in table 3 are retrieved for both home and visiting team.

Candidate feature	Data type
1. Team id	Scalar
2. Avg goals scored per match this season	Scalar
3. Avg goals conceded per match this season	Scalar
4. Result of previous match	Nominal
5. Result of two matches ago	Nominal
6. Result of three matches ago	Nominal
7. Result of four matches ago	Nominal
8. Result of five matches ago	Nominal
9. Team was in a lower league previous year	Boolean
10. Number of matches coached by current coach	Scalar
11. Team hired new coach during previous month	Boolean
12. Top-scorer suspended or injured	Boolean
13. Top-assist suspended or injured	Boolean
14. Avg goals scored by top-scorer	Scalar
15. Avg assists given by top-assist	Scalar
16. Days since previous match	scalar
17. Percentage of wins this season	scalar
18. Percentage of lose this season	scalar
19. Percentage of draw this season	scalar

TABLE 3
Candidate features retrieved for both teams

Several candidate features are not retrieved once per team, but are only retrieved for the home playing team (listed in Table 4) or for the visiting team (listed in Table 5).

Candidate feature	Data type
20. % of wins in home matches	Scalar
21. % of draws in home matches	Scalar
22. % of losses in home matches	Scalar

TABLE 4
Candidate features retrieved for home team

Candidate feature	Data type
23. % of wins in away matches	Scalar
24. % of draws in away matches	Scalar
25. % of losses in away matches	Scalar

TABLE 5
Candidate features retrieved for visiting team

Some candidate features are retrieved only for the current home team and visiting team playing. Those candidate features are listed in Table 6. In this table we mean by 'same ground' the combination team1 - team2 and we mean by 'both grounds' the combination team1 - team2 and team2 - team1, in terms of home team - visiting team. These features are encodings of the earlier encounters factor as described the factors section, to which we added our own extra features for only those earlier encounters in which the current home team was also the home team. Those extra, same ground earlier encounter features, were inspired by our own hypothesis that the correlation of the earlier encounter features to match outcome might be dependent on the ground that the match is played on.

Candidate feature	Data type
26. % of earlier encounters won by home team - same ground	Scalar
27. % of earlier encounters won by visiting team - same ground	Scalar
28. % of earlier encounters resulted in a draw - same ground	Scalar
29. % of earlier encounters won by home team - both grounds	Scalar
30. % of earlier encounters won by visiting team - both grounds	Scalar
31. % of earlier encounters resulted in a draw - both grounds	Scalar
32. Distance between home city and away city	Scalar

TABLE 6
Candidate features retrieved for combination of home team and visiting team

The final candidate features are the odds features: home win, away win, draw and Asian handicap odds. Those candidate features are listed in Table 7.

Candidate feature	Data type
33. Gamebookers home win odds	Scalar
34. Gamebookers draw odds	Scalar
35. Gamebookers away win odds	Scalar
36. Bet&Win home win odds	Scalar
37. Bet&Win draw odds	Scalar
38. Bet&Win away win odds	Scalar
39. Gamebookers Asian handicap home team odds	Scalar
40. Gamebookers Asian handicap away team odds	Scalar
41. Gamebookers size of handicap (home team)	Scalar
42. Betbrain maximum Asian handicap home team odds	Scalar
43. Betbrain average Asian handicap home team odds	Scalar
44. Betbrain maximum Asian handicap away team odds	Scalar
45. Betbrain average Asian handicap away team odds	Scalar

TABLE 7
Candidate features retrieved for combination of home team and visiting team

5 Machine Learning techniques

5.1 Dimensionality reduction techniques

Based on results from experiments on the effect of different dimensionality reduction approaches on classification accuracy for several data sets ([8], [9]), the following dimensionality reduction approaches are chosen to include in the experiment.

- Principle Component Analysis
- Sequential Forward Selection
- ReliefF [28]

The following dimensionality reduction approaches not encountered in the literature study will additionally be included in the experiment.

- Correlation-based Feature Subset Selection [29]

5.2 Classification techniques

In Table 8 one can see that there is a dichotomy in research, with one some studies using features originating from previous matches and characteristics and other studies using features based on ratings of experts. We can consider the feature sets of the studies using data from previous matches and the current match (Miljkovic et al [17], Hucaljuk & Rakipovic [18], Tsakonas et al [30] and McCabe & Trevathan [31]) to be more similar to our candidate feature set described in Tables 3, 4, 5 and 6. Bayesian Networks seems to be an often used technique in the related work, but only seems to perform well on feature sets using on expert rated features, as only one study with features similar to our candidate feature set decided to use Bayesian Networks (Hucaljuk & Trevathan [18]) and they concluded it not to be one the best performing models. The following list of classifiers showed to work well in the related work (references below to related work providing evidence) using comparable feature sets.

- Naive Bayes [17]
- LogitBoost (with Decision Stump) [18]
- Neural Network (including Multilayer Perceptron) [18], [31]
- Random Forest [18]
- Genetic Programming [30]

Although we earlier claimed the cross validation evaluation method to have a risk overestimating prediction accuracy we can use it to obtain a rough estimate of the performance of classification algorithms. With the WEKA [32] machine learning toolkit we conducted a small indicative experiment in which we run all classification algorithms with their default parameters on the dataset in which we used all candidate features in a 10-fold cross validation experimental setting. All classification algorithms performing above a manually selected threshold of 52% classification accuracy are presented below. The classification algorithms which are not considered generally known in the Machine Learning field hold references to studies describing the algorithms.

- Random Forest
- CHIRP [33]
- FURIA [34]
- DTNB [35]
- Logistic regression
- Decision tree (J48)
- Hyper Pipes

In the experiment we will evaluate the classification performance of the union of classification algorithms that worked well in our 10-fold cross validation experiment and classification algorithms found to work well according to literature. Genetic programming and logistic regression showed to execute infeasibly slow using the retrodictive evaluation method and are therefore omitted from further experiments. The final list of classification algorithms to be evaluated with the retrodictive evaluation method is presented below.

- Naive Bayes
- LogitBoost (with Decision Stump)
- Neural Network (Multilayer Perceptron)
- Random Forest
- CHIRP
- FURIA
- DTNB
- Decision tree (J48)
- Hyper Pipes

6 Feature retrieval

We have built a web scraper using Java and JSoup for retrieving the right data for the features. The data of the odds features, see also Table 7 have been retrieved from Football-data[1]. Furthermore, the features 10,11,16 in Table 3, that provide information about the coach of a team and information about matches played in other competitions like the Champions League or European League have been retrieved from Transfermarkt.de [2]. The distance, feature 32 in Table 6 between two football teams has been determined by retrieving the distance between the cities of the teams. This distance is retrieved from Google Maps[3], where the two cities have been used as 'To' and 'From' parameters. The first distance provided by googlemaps is used as distance. Finally, the remaining features which is the public data, has been retrieved from Elf Voetbal[4]. One note has to be made about the data of the features 12 and 13 from Table 3. The values of these features have been determined by checking the data of the match that needs to be classified on Elfvoetbal. However, the matches we retrieved from this website have already been played and normally this data is still unknown on this website until after the match. When classifying an new not yet played match, the values of these features 12 and 13 can be determined using data from FCupdate[5], which shows if a player is suspended or injured for the to be played match.

1. http://www.football-data.co.uk
2. http://www.transfermarkt.co.uk
3. https://maps.google.nl
4. http://www.elfvoetbal.nl
5. http://www.fcupdate.nl

Study	Good classifiers	Poor classifiers	Features
Miljkovic et al* [17]	Naive Bayes		Results in current season, streaks, avg goals scored/conceded
Joseph et al [36]	Bayesian Network	MC4 Naive Bayes k-NN	Expert judgment of team quality
Hucaljuk & Rakipovic [18]	LogitBoost Neural Network Random Forest	Bayesian Network k-NN Naive Bayes	Streaks, earlier encounters, ranking, number of injured players, avg goals scored
Constantinou et al [2]	Bayesian Network		Points obtained and subjective information by experts
Baio & Blangiardo [1]	Poisson model	Bayesian Network	Goals scored by home and away team
Min et al [26]	Bayesian Network		Expert rated features
Tsakonas et al [30]	Genetic programming		Difference in number of injured and banned players, difference in scored goals in last 5 matches, rank difference, host factor, goal difference over 10 season
McCabe & Trevathan* [31]	Multilayer Perceptron		Point for and against, win-loss record, home and away Performance, performance in previous n games, ranking, location, player availability

TABLE 8
Related work using Machine Learning approaches to predictive sports modeling

7 Results & Discussion

The results are split in two parts. The first part discusses and evaluates for the models public data, betting odds and hybrid the prediction accuracy of the classifiers in combination with dimensionality reduction algorithms. The second part discusses the results of the McNemar's test.

7.1 Machine Learning Techniques Evaluation

Each subsection explains the results by referring to Tables or Appendices. The Tables show the accuracy of different combinations of classifiers and dimensionality reduction algorithms. Combinations which are marked with '—' are not performed for one of the following reasons:

1) Calculating the accuracy takes too long (longer than one day);
2) The classifier already contains a dimensionality reduction algorithm (in case of Random Forest).

Cells marked light gray show the highest accuracy measured for each dimensionality reduction algorithm. The dark gray cell shows the highest accuracy seen on all dimensionality reduction/classifier combinations. The second column indicates the classification model parameters, except for the Multilayer Perceptron classifier. The training process of the Multilayer Perceptron takes a lot of time for each match, therefore we have retrained to classifier once per season, once per half season and per 10 matches. This approach deviates from the evaluation methodology as described in the methodology section in such a way that the Multilayer Perceptron is expected to show lower results than it would have shown on the described evaluation methodology.

7.1.1 Public data model

An interesting observation from Table 11 is that ReliefF is outperformed by the other dimensionality reduction methods for all classifiers and its accuracy is even lower than not using any dimensionality reduction method at all for most classifiers. The low performance of ReliefF is remarkable given the fact that it was designed specifically for attribute sets with strong mutual dependencies, a characteristic that does apply to our candidate feature set. The Hyper Pipes classifier showed a highest accuracy of 48.810%, which is remarkably low as it already performed above the 51% threshold on the 10-fold CV experiment. It might be the case that Hyper Pipes benefits more than other algorithms from the possibility of taking into account future data that 10-fold CV allows for this classification task, due to the nature of the algorithm. The low accuracy of Hyper Pipes is not surprising however given that the algorithm is designed to handle very high dimensional data (hundreds of features of more) and very sparse data [37].

The highest performance was seen using Naive Bayes (combined with PCA with 3 PC's) or a Multilayer Perceptron (combined with PCA with 3 or 7 PC's). The the strong performance of the Multilayer Perceptron might not be surprising given the ability of MLP's to learn more complex structures in data. The high prediction accuracy of the Naive Bayes classifier is more remarkable, given the assumption of independence between features that the algorithm relies on. We hypothesize the dependence between features in the candidate feature set to be canceled out by applying the PCA with only three principle components. This is not the only occurrence in literature where a very simple learning algorithm performs really well on a sport match prediction task, Miljkovic et al [17] managed to predict the correct winner of NBA basketball matches in about 67% of instances also by using a Naive Bayes classifier. Aranda-Corral et al [6] managed to classify Spanish Soccer League matches into the categories win/draw/lose with 59.74% accuracy using only public data. This does not automatically

imply that the same prediction accuracy could be achieved for the Dutch Eredivisie, because of possible differences in balance and predictability between the competitions.

Tables 14, 15, 16 and 17 show the confusion matrices of the four best performing public data models. The confusion matrices show the difficulty of predicting a draw result, the possibility of a draw result is simply neglected by all four models. The low prior of a draw result (about 22%) partly explains this behavior of the prediction models, in addition we note that the features in our candidate feature set all seem to hint either to a home win of an away win where no feature seem to have much predictive value for the draw class.

7.1.2 Betting odds model

Table 12 shows the results of training prediction models based only on the betting odds features. The two classifiers performing with the lowest accuracy were Hyper Pipes and Naive Bayes. Hyper Pipes being optimized for sparse very high dimensionality data explains the poor performance, as the betting odds data consists of thirteen features. Naive Bayes assumes independence between features. Betting odds features can be expected to be very highly correlated, hence, big differences between bookmaker odds for the same match would enable punters to arbitrage. The high correlation between betting odds features might be causative to the poor prediction performance of Naive Bayes.

The highest accuracy is 55.297% using classifier FURIA with 10 folds as parameter, meaning that one tenth of the data is used as pruning set to prevent the algorithm from creating learning rules which lead to overfitting. This accuracy is higher than the accuracies predicted using only public data, which might indicate that betting odds are partly based factors not included in our public data model.

7.1.3 Hybrid model

Table 13 shows the prediction accuracies using the hybrid model. While comparing the accuracies of Table 11 with Table 13 one can see that the combination of using betting odds an public data improves the prediction model. ReliefF shows to be part of the highest accuracy model when used in combination with LogitBoost. Because the betting odds features are expected to have very high mutual dependencies (more than public data) and ReliefF is specifically designed to handle features of high mutual dependence this result was to be expected. The strong prediction performance of LogitBoost is in line with Hucaljuk & Rakipovic [18] findings on their study on Champions League prediction models.

The LogitBoost/ReliefF confusionmatrix shown in Table 18, shows that the hybrid model still has difficulties of predicting draw outcomes. In comparison with the public data models, the hybrid model has slightly higher prediction numbers for the draw category, possibly this is due to the betting odds for a draw outcome which is a feature with predictive value for draw which the public data feature set lacked.

7.2 Statistical model evaluation

McNemar's test [16] gives insight in the statistical significance of differences between predictive models. The fourth research questions concerns whether it is possible to outperform betting odds predictions by using public data. To answer this question we apply McNemar's test to the highest accuracy (as listed in Table 11) public data model found (Naive Bayes model on three principle components) and the highest accuracy (as listed in Table 12) betting model (10-fold FURIA model).

7.2.1 Public data model and Betting odds model comparison

Let the highest accuracy public data model be m1, the highest accuracy betting odds model be m2 and let the function a represent model accuracy. The following hypotheses can be stated:

$H_0 : a(m1) = a(m2)$
$H_a : a(m1) \neq a(m2)$

Betting odds ->Public data	Correct	Wrong
Correct	908	104
Wrong	115	723

TABLE 9
Contingency table for betting odds model and public data model comparison

Applying McNemar's test on the data of Table 9 results in the following chi-square statistic: $\chi^2 = 0.5525$ $p\text{-}value = 0.4573$

The chi-square statistic has one degree of freedom which allows us to calculate the above stated corresponding p-value. As the p-value of 0.4573 is larger than 0.05, we cannot reject the null hypothesis that the public data model performs equally well as the betting odds model. Therefore we can say that we cannot statistically differentiate between the performance of the public data model and the betting odds model.

7.2.2 Betting odds model and hybrid model comparison

Now let m3 be the highest accuracy hybrid model. The following hypotheses can be stated corresponding to the fifth research question.

$H_0 : a(m2) = a(m3)$
$H_a : a(m2) \neq a(m3)$

Betting odds ->Combined	Correct	Wrong
Correct	964	73
Wrong	59	754

TABLE 10
Contingency table for betting odds model and hybrid model comparison

Applying McNemar's test on the data of Table 10 results in the following chi-square and p-value statistics:

$\chi^2 = 1.4848$
$p\text{-}value = 0.2230$

As the p-value of 0.2230 is larger than 0.05, therefore we cannot reject the null hypothesis that the combined model using betting odds and public data performs equally well as the betting odds model. This implies that we cannot statistically differentiate between the performance of the public data model and the betting odds model.

8 Conclusions

Because the first two research questions serve mainly to be able answering of research questions three in a structured way, the conclusion section will focus on research questions three, four and five.

RQ3 *Which dimensionality reduction method and Machine Learning method results in the highest prediction accuracy in the Dutch football competition match result prediction task?*

The highest accuracy for the public data model was seen when the Naive Bayes or Multilayer Perceptron classifier was used in combination with a Principle Components Analysis (with 3 or 7 Principle Components), which achieved an accuracy of 54.702%. Spann et al [14] showed it to possible to design a profitable betting strategy for a prediction model which was less accurate than our public data model (53.98%). Although this study was not carried out on the German Bundesliga instead of the Dutch Eredivisie it gives a positive sign of the possibility of designing a profitable betting strategy based on this model.

RQ4 *How does the public data model perform (in terms of prediction accuracy) compared to bookmaker odds?*

The highest accuracy measured using betting odds features was 55.297% when the FURIA classifier was used with 10 folds as parameter. This accuracy is only slightly higher than the predicting accuracy of the public data model. Using McNemar's test we did not find any statistical difference in accuracy between the public data model and the betting odds model.

RQ5 *Is it possible increase performance (in terms of prediction accuracy) when combining betting odds and public data into a hybrid model, compared to a model using only one of both?*

A combination of LogitBoost and ReliefF scored the highest accuracy on the hybrid model which makes use of public data as well as betting odds features. This hybrid model achieved an accuracy of 56.054%, therefore we can conclude to increase the combined accuracy of the bookmakers by over 0.7 percentage points by including public data into the model. Although McNemar's test showed no statistical difference between the betting odds model and the hybrid model, the results are still such to raise the assumption that a model based on public data and betting odds combined is able to beat the bookmakers. Overall we can see the results of this study as a positive sign that it might be possible to engineer profitable betting decision support systems based on non-confidential information.

9 Future work

This work can be succeeded by experiments concerning betting strategies to evaluate whether prediction models described in this study can be used to gain profit. Another subject of further study could be our hypothesis that the prediction models described in this paper will show increasing accuracy trends from the beginning of season to the end of season, as each seasons tends to have its own characteristics that classification models need to for to get accustomed to. Combining both topics, it could be the case that a betting strategy only betting on matches from a particular game day in the season will be more beneficial than a basic betting strategy betting on all matches in the season. Further research is needed to validate or contradict these hypotheses.

APPENDIX A
RESULT TABLES

Dimensionality reduction algorithm ->		ReliefF			All Features	CfsSubsetEval	PrincipleComponents		
Classifier:		1 Attributes	5 Attributes	10 Attributes	51 Attributes	15 Attributes	3 PC's	5 PC's	7 PC's
CHIRP		48.648%	53.405%	54.324%	—	54.324%	53.51%	54.594%	54.432%
DTNB		50.432%	51.405%	52.594%	—	52.702%	53.405%	53.405%	53.405%
FURIA	3 folds	48.648%	48.486%	50.162%	54.270%	54.432%	53.891%	54.270%	53.837%
	10 folds	48.648%	48.486%	51.081%	54.054%	52.702%	53.729%	54.324%	53.567%
HyperPipes		48.648%	48.648%	48.648%	48.756%	48.648%	48.810%	48.702%	48.594%
J48		50.432%	51.621%	53.783%	51.297%	53.081%	54.486%	54.432%	54.270%
NaiveBayes		50.432%	54.216%	53.351%	51.945%	53.621%	54.702%	54.540%	54.108%
Multilayer Perceptron	Season	50.486%	48.324%	—	—	—	54.702%	54.324%	54.702%
	Half season	50.702%	49.081%	—	—	—	54.702%	54.270%	54.594%
	10 matches	50.756%	49.405%	—	—	—	54.000%	54.054%	53.837%
RandomForest	10 trees	50.648%	—	—	51.027%	—	48.972%	—	—
	50 trees	50.432%	—	—	53.730%	—	—	—	—
	250 trees	50.486%	—	—	54.108%	—	—	—	—
	500 trees	50.432%	—	—	53.676%	—	—	—	—
LogitBoost		49.567%	52.378%	53.189%	53.513%	53.621%	54.324	54.540%	53.621%

TABLE 11
Prediction accuracies for the public data model

		Betting odds
Classifier:		
CHIRP		54.000%
DTNB		54.108%
FURIA	3 folds	54.540%
	10 folds	55.297%
HyperPipes		50.216%
J48		53.729%
NaiveBayes		47.459%
Multilayer Perceptron	Season	54.756%
	Half season	54.756%
	10 matches	54.594%
RandomForest	10 trees	52.540%
	50 trees	53.351%
	250 trees	—
	500 trees	—
LogitBoost		54.864%

TABLE 12
Prediction accuracies for the betting odds model

Dimensionality reduction algorithm ->		ReliefF		All Features	CfsSubsetEval	PrincipleComponents		
Classifier:		5 Attributes	10 Attributes	65 Attributes	18 Attributes	3 PC's	5 PC's	7 PC's
CHIRP		53.405%		—	55.297%	54.216%	53.567%	54.378%
DTNB		54.000%	53.621%	—	53.567%	55.135%	55.135%	55.027%
FURIA	3 folds	54.054%	54.540%	—	54.702%	55.297%	54.756%	54.918%
	10 folds	54.054%	54.540%	—		54.864%	54.702%	54.702%
J48		54.162%	53.891%	53.513%	53.567%	55.027%	54.702%	54.486%
NaiveBayes		50.756%	51.513%	50.810%	53.459%	54.648%	54.162%	53.675%
MultiLayer Perceptron	Season	51.837%	—	—	52.432%	54.972%	54.972%	55.135%
	Half season	49.081%	—	—	51.891%	54.918%	54.756%	55.081%
	10 matches	—	—	—	51.081%			
RandomForest	10 trees	—	—	51.459%	—	48.324%	50.810%	48.810%
	50 trees	—	—		—	48.540%	51.351%	52.756%
	250 trees	—	—		—	49.675%	51.945%	52.648%
	500 trees	—	—		—	49.081%	52.108%	53.135%
LogitBoost		56.054%	55.675%	55.189%	55.567%	54.702%	54.810%	54.648%

TABLE 13
Prediction accuracies for the hybrid model

APPENDIX B
CONFUSION MATRICES

	Predicted HOME	Predicted AWAY	Predicted DRAW
ACTUAL HOME	754	143	3
ACTUAL AWAY	266	256	3
ACTUAL DRAW	293	130	2

TABLE 14
Confusion matrix: classifier NaiveBayes using PCA (3PC's)

	Predicted HOME	Predicted AWAY	Predicted DRAW
ACTUAL HOME	751	149	0
ACTUAL AWAY	264	261	0
ACTUAL DRAW	291	134	0

TABLE 15
Confusion matrix: classifier Multilayer perceptron season based using PCA (3PC's)

	Predicted HOME	Predicted AWAY	Predicted DRAW
ACTUAL HOME	751	149	0
ACTUAL AWAY	264	261	0
ACTUAL DRAW	291	134	0

TABLE 16
Confusion matrix: classifier Multilayer perceptron half season based using PCA (3PC's)

	Predicted HOME	Predicted AWAY	Predicted DRAW
ACTUAL HOME	759	141	0
ACTUAL AWAY	272	253	0
ACTUAL DRAW	293	132	0

TABLE 17
Confusion matrix: classifier Multilayer perceptron season based using PCA (7PC's)

	Predicted HOME	Predicted AWAY	Predicted DRAW
ACTUAL HOME	764	132	4
ACTUAL AWAY	254	265	6
ACTUAL DRAW	283	134	8

TABLE 18
Confusion matrix: classifier LogitBoost using ReliefF 5 attributes

REFERENCES

[1] G. Baio and M. Blangiardo, "Bayesian hierarchical model for the prediction of football results," *Journal of Applied Statistics*, vol. 37, no. 2, pp. 253–264, 2010.

[2] A. C. Constantinou, N. E. Fenton, and M. Neil, "pi-football: A bayesian network model for forecasting association football match outcomes," *Knowledge-Based Systems*, 2012.

[3] J. Goddard, "Regression models for forecasting goals and match results in association football," *International Journal of Forecasting*, vol. 21, no. 2, pp. 331 – 340, 2005.

[4] A. L. Balduck, A. Prinzie, and M. Buelens, "The effectiveness of coach turnover and the effect on home team advantage, team quality and team ranking," *Journal of Applied Statistics*, vol. 37, no. 4, pp. 679–689, 2010.

[5] J. Goddard, "Who wins the football?" *Significance*, vol. 3, no. 1, pp. 16–19, 2006.

[6] G. A. Aranda-Corral, J. Borrego-Díaz, and J. Galán-Páez, "Complex concept lattices for simulating human prediction in sport," *Journal of Systems Science and Complexity*, vol. 26, no. 1, pp. 117–136, 2013.

[7] D. Hawkins, "The problem of overfitting," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[8] L. C. Molina, L. Belanche, and À. Nebot, "Feature selection algorithms: A survey and experimental evaluation," in *Proceedings of the 2002 IEEE International Conferenceon Data Mining*. IEEE, 2002, pp. 306–313.

[9] A. Salappa, M. Doumpos, and C. Zopounidis, "Feature selection algorithms in classification problems: An experimental evaluation," *Optimisation Methods and Software*, vol. 22, no. 1, pp. 199–212, 2007.

[10] F. Carmichael and D. Thomas, "Home-field effect and team performance evidence from english premiership football," *Journal of Sports Economics*, vol. 6, no. 3, pp. 264–281, 2005.

[11] A. Seçkin, "Home advantage in association football: Evidence from turkish super league," in *Proceedings of the EcoMoD2006 Conference*, 2006.

[12] K. Goossens, "Competitive balance in european football: Comparison by adapting measures: National measure of seasonal imbalance and top 3," University of Antwerp, Faculty of Applied Economics, Working Papers, Dec. 2005.

[13] D. Forrest, J. Goddard, and R. Simmons, "Odds-setters as forecasters: The case of english football," *International Journal of Forecasting*, vol. 21, no. 3, pp. 551 – 564, 2005.

[14] M. Spann and B. Skiera, "Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters," *Journal of Forecasting*, vol. 28, no. 1, pp. 55–72, 2009.

[15] N. Vlastakis, G. Dotsis, and R. N. Markellos, "Nonlinear modelling of European football scores using support vector machines," *Applied Economics*, vol. 40, no. 1, pp. 111–118, 2008.

[16] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[17] D. Miljkovic, L. Gajic, A. Kovacevic, and Z. Konjovic, "The use of data mining for basketball matches outcomes prediction," in *Proceedings of the 8th International Symposium on Intelligent Systems and Informatics*. IEEE, 2010, pp. 309–312.

[18] J. Hucaljuk and A. Rakipovic, "Predicting football scores using machine learning techniques," in *MIPRO, 2011 Proceedings of the 34th International Convention*. IEEE, 2011, pp. 1623–1627.

[19] A. Heuer and O. Rubner, "Fitness, chance, and myths: an objective view on soccer results," *The European Physical Journal B*, vol. 67, no. 3, pp. 445–458, 2009.

[20] R. H. Koning, "An econometric evaluation of the effect of firing a coach on team performance," *Applied Economics*, vol. 35, no. 5, pp. 555–564, 2003.

[21] M. De Paola and V. Scoppa, "The effects of managerial turnover evidence from coach dismissals in italian soccer teams," *Journal of Sports Economics*, vol. 13, no. 2, pp. 152–168, 2012.

[22] A. Heuer, C. Müller, O. Rubner, N. Hagemann, and B. Strauss, "Usefulness of dismissing and changing the coach in professional soccer," *PloS one*, vol. 6, no. 3, p. e17664, 2011.

[23] F. Palomino, L. Rigotti, and A. Rustichini, "Skill, strategy and passion: An empirical analysis of soccer," in *Proceedings of 8th World Congress of the Econometric Society*, 2000, pp. 11–16.

[24] J. Goddard and I. Asimakopoulos, "Forecasting football results and the efficiency of fixed-odds betting," *Journal of Forecasting*, vol. 23, no. 1, pp. 51–66, 2004.
[25] Y. Khazaal, A. Chatton, J. Billieux, L. Bizzini, G. Monney, E. Fresard, G. Thorens, G. Bondolfi, N. El-Guebaly, D. Zullino, and R. Khan, "Effects of expertise on football betting," *Substance Abuse Treatment Prevention and Policy*, vol. 7, May 2012.
[26] B. Min, J. Kim, C. Choe, H. Eom, and R. B. McKay, "A compound framework for sports results prediction: A football case study," *Knowledge-Based Systems*, vol. 21, no. 7, pp. 551 – 562, 2008.
[27] K. Odachowski and J. Grekow, "Using bookmaker odds to predict the final result of football matches," in *Knowledge Engineering, Machine Learning and Lattice Computing with Applications*. Springer, 2013, pp. 196–205.
[28] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Ninth International Workshop on Machine Learning*, D. H. Sleeman and P. Edwards, Eds. Morgan Kaufmann, 1992, pp. 249–256.
[29] M. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1999.
[30] A. Tsakonas, G. Dounias, S. Shtovba, and V. Vivdyuk, "Soft computing-based result prediction of football games," in *The 1st International Conference on Inductive Modelling*, 2002.
[31] A. McCabe and J. Trevathan, "Artificial intelligence in sports prediction," in *Proceedings of the 5th International Conference on Information Technology: New Generations*, 2008, pp. 1194–1197.
[32] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18.
[33] L. Wilkinson, A. Anand, and D. T. Nhon, "Chirp: a new classifier based on composite hypercubes on iterated random projections." in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 6–14.
[34] J. C. Huehn and E. Huellermeier, "Furia: An algorithm for unordered fuzzy rule induction," *Data Mining and Knowledge Discovery*, 2009.
[35] M. Hall and E. Frank, "Combining naive bayes and decision tables," in *Proceedings of the 21st Florida Artificial Intelligence Society Conference*. AAAI press, 2008, pp. 318–319.
[36] A. Joseph, N. Fenton, and M. Neil, "Predicting football results using bayesian nets and other machine learning techniques," *Knowledge-Based Systems*, vol. 19, no. 7, pp. 544–553, 2006.
[37] J. Eisenstein and R. Davis, "Visual and linguistic information in gesture classification," in *Proceedings of ACM SIGGRAPH 2007*. ACM, 2007, p. 15.