Dutch football prediction using machine learning classifiers

Research paper business analytics

Abel Hijmans

Supervised by

Dr. S. Bhulai

Abstract

Sports betting is becoming more popular every year and more people are betting now than ever. With the growth of the betting market comes the growth of research done on match prediction. Research done in the 1950s has been the basis for match predictions up until the 1980s. Since then prediction techniques have shifted from distribution prediction towards a more modern data mining predicting. Using machine learning methods has been proved to be good analyzing methods for tournaments and league matches.

In this research paper multiple data mining techniques are analyzed and prediction results are compared to come to a good model for predicting matches of the Dutch football team (soccer team, in American English). Based on the prediction results of a random tree model, Naïve Bayes model and a k-nearest neighbor model one single model is chosen and results are looked at more in-depth. By using the best prediction model, one can see which variables of the dataset have little predictive power and which variables have a lot of predictive power. From the random tree model, which was the most predictive power, it is surprising to see that tactics of the coach hold little predictive value for predicting the end results of a match. To support this claim more research has to be done the prediction of the Dutch football team.

Preface

This research paper is compulsory for the Master Business Analytics at Vrije Universiteit Amsterdam. This paper is mandatory to be finished before the master internship starts and is a preparation for the project. The research paper must be conducted on a subject that has strong ties to the master Business Analytics and should therefore be lined to mathematics and/or artificial intelligence.

In this paper multiple data mining techniques are compared for the prediction of the outcomes of football games of the Dutch team. Based on those techniques multiple models are created and compared to come to the best model to predict match results.

Contents

Introduction

The Dutch football team (soccer team, in American English), unexpectedly missed the European championship of 2016. It was the first major tournament they missed since the world cup in 2002. In the qualification group the Netherlands were the big favorite against, among others, Iceland and Turkey. Losing against both teams meant that the Dutch did not qualify for the European championship and faith in the team was at an all-time low. Could these losses be designated to wrong tactics, bad team setup or was it bad luck that made the Netherlands lose. Many, so called, football experts gave their opinion about what went wrong, but could anyone have predicted these losses?

Much research is done on the prediction of football outcomes in leagues and tournaments but little research is done on the prediction of single matches of national teams. The reason for this is that much data is needed to make a good prediction and football teams' change quickly. These factors make predictions based on Poisson distributions (Moroney, 1956) hard and not very accurate. Other methods, such as classification, predict better on smaller datasets but also require more variables as input. The impact of coaches and decisions made by him and his staff seem to have a bigger impact on end results by the national team than in the league. This is due to the fact that losing once in the league has a smaller impact on the outcome of the league table than losing once in a world cup or even in a world cup qualifier.

Every coach has its own opinion about how to play and favors different type of players. For this research paper a dataset will be constructed that only has variables in it that are known beforehand. This is due to the fact that beforehand it is not certain with which team the opponent will play and how many fans will show up to a match. Therefore, the dataset which is used later to make predictions on, is so that it can be used by the coach to see what the impact of his decisions are.

This research will introduce three datamining models that have proven to have predictive powers on multiple datasets in general. These models will be explained, evaluated and tested on their predictive powers on a dataset that has 28 variables of the Dutch football team. In the end the following research question will be answered: How well is it possible to predict results for the Dutch national football team using data mining techniques?

Literature study

For quite a while scientists, businesses and gamblers have been trying to forecast game results in football for both tournaments and single matches. Many people made and make predictions just for the fun of it (a lot of football fans like to bet with their friends). Organized football gambling on the other hand has developed into a growing industry and is now worth billions (Keogh, 2013). As a result there is quite some literature on match prediction models. In this chapter an overview is presented of the most relevant research on match prediction. The last part of this chapter gives a small overview of general approaches in data mining.

In 1974 Hill (hill, 1974) showed that forecasts made before the season began had a significant positive correlation with end league table results. In his paper, he stated "this seems to indicate that football results are not pure chance (although there is obviously a considerable element of chance)." This means that in theory it should be possible to predict match results, up to a certain level, if enough data is available. In the 1950s a first analysis on football prediction was conducted. Moroney (Moroney, 1956) analyzed Poisson distributions and negative binomial distributions to predict the number of goals scored by both teams in a match. Both models proved to have analyzing strength that fit the results, which was confirmed by Benjamin (B. Benjamin, R.Charles, R. Pollard, 1971). Dixon (M.Dixon, G. Stuart, 1997) showed 30 years later, that using a Poisson distribution on a big dataset is a strong analytic tool to predict match results and can yield positive returns on the betting market. Key assumption is that goals scored by each team follow a Poisson distribution that has a parameter based on the 'attack rate', the 'defense rate' and a home effect. This model was based on research conducted earlier by Maher (Maher, 1982). He showed that using a Poisson distribution is the strongest tool to predict the goals scored by both teams and is favored over the negative binomial distribution or other distributions. The assumption that goals scored by each team follow a Poisson distribution is a very strong assumption and can be argued. The fluctuation in the scoring rate of teams only partly depend on time, and partly on which team is leading, which is not in line with the Poisson distribution theory as described by Moroney (M. Dixon, 1998). Furthermore Ridder shows that other factors have influence on the scoring rates of the team and cannot easily be captured by the attack or the defense rate (G.Ridder, 1194). Another variable that has an effect on goals scored is the international ranking of the football team (D. Dyte, S. Clarke, 2000).

The general method described above to predict match results are based on predicting the number of goals scored by both teams and then comparing those. Another method to analyze football results is to use a classification to directly predict a win, draw or lose of a team. Classification machine learning methods have been applied in the prediction in sport matches (R.Michalski, J. Carbonell, T. Mitchell, 2013). Byungo used Naïve Bayes as a machine learning method for the prediction of football matches (Byungho, 2008), while Kampakis (Kampakis, 2014) used random forest tree algorithms to analyze football outcomes and showed that these models are useful for the prediction of the outcome of football matches. Wilson (Wilson, 1995) used neural networks as a ranking method in American football team in his 1995 research. Huang used the same method as Wilson and showed that neural networks can be used to predict the European football cup of 2008 (Huang, 2010). These researches show that the use of classification in sports prediction have positive predicting results.

Most of the research that is done on outcomes of football matches is done on teams that play in a national league with many games in each season. National teams however, play a lot fewer games

compared to the teams in the national league. Goddard showed that using both methods on a dataset with 25 years of match results, there is no difference between the forecasting performances of score based prediction and end game result classification prediction (Goddard, 2005). Karlis showed that that using classification is a better predictor for end game results then Poisson distributions when looking at a small data set (Karlis, 2010). Therefore classification algorithms provide a better forecasting model when looking at the prediction of national teams, since the data available for these teams is quite scarce.

The data mining techniques used by the researches stated above are all well-known techniques and are widely used (Pujari, 2001). For supervised classification, different types of data mining techniques can be used and three widely used predictors, that are strong classification predictors, are described by Phyu (Phyu, 2009). Random tree induction models have the capability to break down complex classification problems into a collection of much simpler decisions, producing solutions that are easy to interpret. Random tree induction models proved to be successful in many different areas (Safavian, 1990). Naïve Bayes (or Bayesian networks) uses probability prediction as a classification tool and often classifies better than more sophisticated classifiers. The assumption for Naïve Bayes is that features are independent, which is generally a poor assumption. Bayesian predictions do yield good analyzing power on small and incomplete datasets (Uusitalo, 2007). Another frequently used classifier is the k-nearest neighbor classifier. The k-nearest neighbor classifier show high classification accuracy and high speed (Wu, 2008) , whilst results are easy to interpret.

From the literature study one can draw some conclusion that will form the base of the research conducted in this paper. Both goal scored prediction and classification prediction work well on big data sets. Since the Netherlands did not play enough games to consider the data set ‘big’, classification outperforms goal scoring prediction and is used to predict the outcome of the games played by the Netherlands. As classification models random tree model, Naïve Bayes and k-nearest neighbor classifiers are compared to see what model has the most predictive power.

Data

Introduction

The dataset used in this paper to forecast the end game results of the Dutch national football team had to be constructed using available data. To analyze the Dutch national football team as a complex network, Rob Kooij used the website www.voetbalstats.nl to gain information on some basic variables of the team (Kooij, 2008) .The dataset that was constructed for this research was based on the data found on the same website along with additional information found on www.fifa.com and www.goal.com on the starting squad of the Dutch football team.

The national team of the Netherlands played in total 773 games at the moment of writing, the first game being in 1905. Thus, the size of the dataset would consist at the most out of 773 data points. When using all data points the dataset is already considered to be small, but for prediction accuracy the dataset turns out to be even smaller. Surely matches played a long time ago hold little predictive value for the result of the current Dutch squad. Knorr-Helld (Knorr-Helld, 2000) stated that data on recent results are a better predictor than those based on results in the (distant) past. To increase predictive accuracy of the dataset only recent matches are used that can add predictive information to the dataset. For this paper the number of data points is set at 137, which corresponds with all matches played since 1st of January 2006 up until October 2016. The reason for choosing a time spawn of 10 years is that games from ten years ago still hold information about some players who played in the last game of October 2016, since they made their debut in 2006. Using only data that is less than 10 years old prevents the model from overfitting on data that hold little relevance for present day football.

Variables

The constructed dataset has various features that can all be placed in one of the following categories:

- Type of game
- Squad attributes
- Individual player attributes
- Team form

Type of game

In the ten years that the dataset covers the national team of the Netherlands played three different types of matches.
Friendlies, these were predominately used to test new tactics and players and to establish how strong the squad is in comparison to other teams. These matches matter only for the FIFA ranking and carry little value in general. Coaches prefer to use them to try out certain new tactics and accept the risk of losing, as do opponents. Therefore, results in friendlies may vary compared to other match types, even when other variables are similar.

Qualification games for big tournaments. These are often played against weaker opponents and there is not always an incentive to win. Nevertheless these types of matches are quite important, since enough of them need to be won to be allowed to participate in a major tournament. However, after

qualification is secured there is no more incentive to play your best and then these matches can, like friendlies, be used to try out new tactics.

Tournament matches, either for the European Cup or the World Cup. Winning one of these tournaments is a big dream for many professional players. Most games at a major tournament are based on a knockout system: losing means you are out. Both teams send in their best squad and there is no room for errors. Some information about the match type and outcome is shown in Table 1 below.

type of match	played	lost	draw	won
Friendly	60	12	19	29
Qualification	52	8	5	39
Tournament	25	6	3	16

Table 1: Type of games played by the Netherlands and results

Squad attributes

For the squad attributes there are in total 9 variables which are known beforehand by the coach and can be used for analyzing match results. These attributes are: number of defenders/midfielders/attackers, number of matches already played by the defenders/midfielders/attackers, average age of the various players, goals scored by the team and the percentage of defenders that is left footed. Reason for adding this last variable is that coaches quite often select defenders based on their strongest foot- whilst the other variables have a big impact on the tactic set by the coach.
Some of the statistics used in the dataset are based on the matches already played. For instance, the feature goals in team is the total number of goals scored by the team, previous to the match they are about to play. The same goes for the average games defenders/midfielders/attackers (the average number of games played by individual players previous to the match being analyzed). Other features, like the percentage of left footed defenders, are based on the starting eleven send in by the coach.

	Minimum	Maximum	Average
Number of defenders	3	5	4,1
Number of midfielders	1	5	3,1
Number of attackers	2	4	2,8
Average games defenders	15	64,3	31,8
Average games midfielders	2	94,5	47,3
Average games attackers	1	86,5	40,6
Goals in team	1	154	72,0
Average age in team	20,8	28,9	26,2
Percentage left footed	0	0,8	0,5

Table 2: Team attributes of the Netherlands

From Table 2 one can see that the most noticeable are the minimum average number of games of defenders, midfielders, attackers and the number of goals scored by the team. All these minimum values are derived from two single matches. One was played on 11th of August in 2010. A friendly match against Ukraine, and the first match played by the national team after the lost World Cup final against Spain. Many internationals preferred a longer holiday, to regain mental focus after the blow of losing the final

over practicing against Ukraine. The other outlier, played on the first of June 2016, was a practice match against Poland, in which the number of goals scored by the team was extremely low. This match was the first game played after it became clear that the Netherlands did not qualify for the European championship of 2016. When arranging *Goals_in_team* and *games_played_by defenders/midfielders/attackers* from low to high one can see that extreme low and extreme high values are rare, as can be seen in Figure 1.

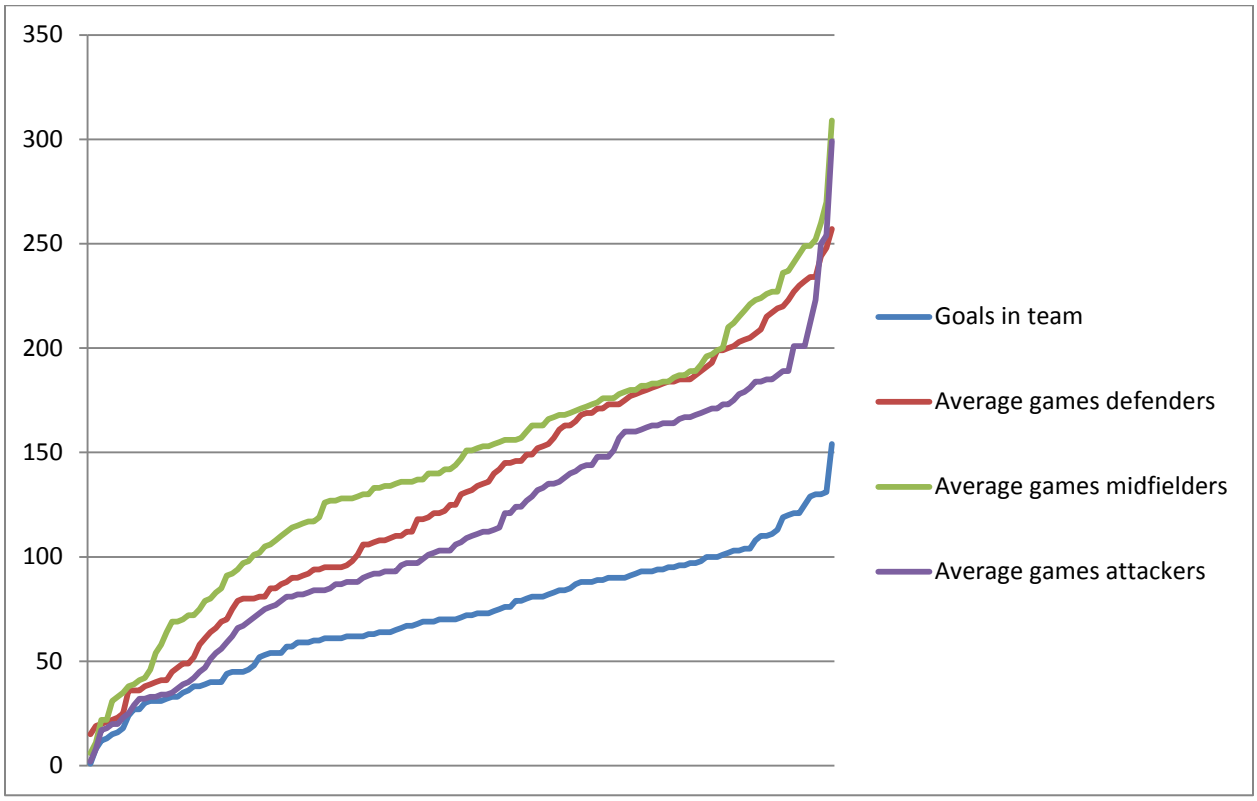

Figure 1: Plot of team variables

To reduce overfitting done by the model on the dataset, outliers such as extreme low or high values are assumed to be non-existent and are replaced by a maximum or minimum value. In Table 3 below is shown what those values are and how much data points are replaced. By replacing outliers with the new maximum and minimum values, the model is less likely to over fit.

	New minimum value	New maximum value	number of replacements
Goals in team	12	131	4
Average games defenders	20	250	3
Average games midfielders	20	260	4
Average games attackers	15	250	4

Table 3: New minimum and maximum values of team variables of the Netherlands

Individual player attributes

Variables attributed to individual players can be used to analyze a team's performance even though football is a team sport. When individual players have extreme qualities, it is possible that morale in the team rises or nerves are calmed, resulting in a better performance. On www.goal.com qualities of every player are described and rated by an expert. Although these qualities can have a huge impact on the game, they are not facts but rather personal opinions or judgements. Opinion on certain skills of well-known football players may not vary much but experts do have very different opinions on lesser known players (comparisons of Slory's rating by experts on www.goal.com differ hugely from that on other sites). Therefore, in this section not every player is looked at and rated individually but we focus only on two categories of players within a team – namely the captain and the star player. In theory they can have the biggest influence on morale. A measurement for captain qualities is the number of games played in the Dutch national squad by the captain. The idea behind this is that an experienced captain knowns how to coach the team and knows what is needed to win a game. Other measurements concerning the captain, for instance, verbal communication skills, are too subjective to be taken into account in this analysis. A star player in a team might be highly motivational for the team since good players can give the opponent a hard time and might even hold an extra 'fear factor'. Who is or is not a star player is of course very subjective and opinions about who is the best player might vary from one person to another. The general rule, however, is that better players are more expensive and thus the star player is the one with the highest transfer value. This used to be a good indicator for the quality of a player until the last couple of years. Now it has become a bit biased since transfer fees have risen a lot lately. Still it remains up to now the best indicator for player qualities. Some descriptive statistics about the individual player ratings are shown in Table 4.

	Minimum	Maximum	Average
Games of captain	4	128	84,7
Value of star player	2	35	24,1

Table 4: Statistics of individual player ratings

From Figure 2 one can deduce that there are only a few outliers for the *value of star player* while *games of captain* can be described as an outlier if the number of games is less than twenty. Therefore, the minimum value for star player value is set at 12 (changing three data points) and number of games played by the captain is set to be a minimum of 20 (changing 7 data points).

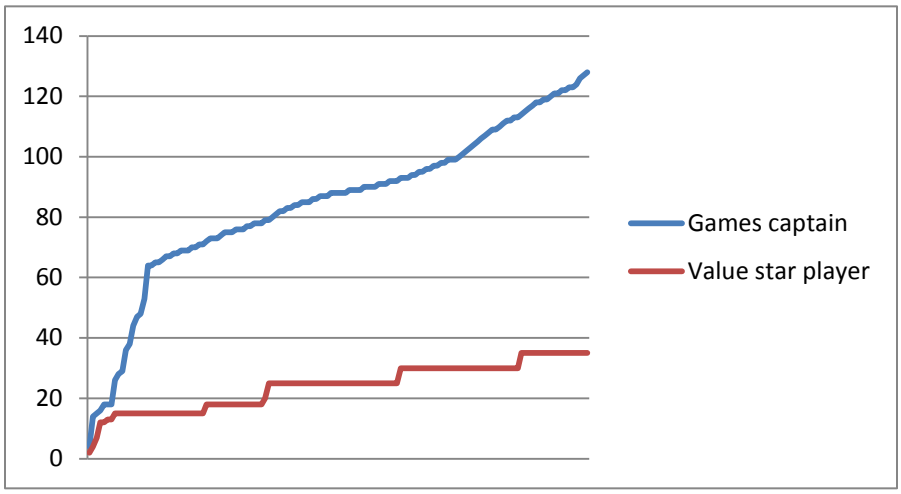

Figure 2: Plot of individual player ratings

Team form

The last variable taken into account in the dataset is the shape both teams are in. This is done by looking at the independent list of the FIFA that ranks teams based on results obtained in the past. This list shows the ranking of all nations that are members of the FIFA; number one is the best team, number 207 the worst; the list is updated every month and gives an easy to interpret number corresponding with the form the team is in. Team forms of both the Netherlands and the opponent are shown in Table 5.

	Best	Worst	Average
Rank Netherland	1	26	7
Rank opponent	1	207	45

Table 5: Team form ratings

As one can see the average ranking of the Netherlands in the past ten years was a lot better than the ranking of the average opponent. The vast majority of opponents the Netherlands faced were teams ranked in the top 100 or ranked lower. Here we assume that countries not in the top 125 are all at an equal (and equally bad) level and therefore, the worst ranking an opponent can get (or the Dutch team, if the ranking of the Netherlands drops dramatically) is capped to be 125. At the moment of writing number 125 is North Korea while the country on the bottom of the list is Tonga. Thus, it is assumed that both countries are on the same, equally bad, level. By capping, a total of 8 database points are changed to reduce overfitting done by the model.

Extra variables

Based on the four types of variables new variables are built. An example is the difference between the current ranking and the average ranking. These new variables are based on the average value of a variable and compared to the value in every match. These types of variables give extra information on how the variables in the team are compared to the average and should provide extra insights.
Including those new variables the total dataset consists of 137 rows, each with 29 variables, no missing values and one dependent variable: the number of points obtained in that game.

Models

As described in the literature study, there are multiple machine learning techniques that can be used for classification. In this paper we will look at few techniques and compare their results. In this chapter three models discussed by Phyu (Phyu, 2009) are analyzed. This should, eventually, enable us to choose the method, out of the three models described by Phyu, with the best predictive power as the final forecast model.

GBM

Generalized boosted models (GBM) is a classification that uses a mixture of weak predictors like decision trees, that when combined form a strong prediction model. The GBM model predicts the variable *Points_obtained_in_game* by looking at the other variables in the dataset. The model focusses on the number of points that will be obtained according to the given variable values. The model gives a probability that the team will win, draw or lose. This is achieved by putting one or more constraints on each node in the decision tree. At this node two 'branches' connect a new next node where new constraints are set. The number of nodes in the model must be given as a parameter. In order to obtain the best prediction, multiple number of nodes can be tested. Too few nodes leads to a model where not enough decisions can be made and thus the outcome may be random, while too many nodes leads to overfitting of the data. The end nodes represent the values in the variable that needs to be predicted. The model makes multiple 'trees' to see what tree best fits the data. Fitting the data is done in the following way:
The GBM model minimizes the mean squared error given by $(\hat{y}-y)^2$ where $\hat{y}$ is the predicted value. The model uses $\hat{y}=F(x)$ where F(x) is a model based on the training)set. The GBM assumes that there is a non-perfect model, F_m at each stage m, $1 \leq m \leq M$ that can be improved by constructing a new model that adds an estimator h to the previous model such that $F_{m+1} = F_m(x)+h(x)$. The GBM model fits h to the residual $y-F_m(x)$ such that $Fm+1$ corrects F_m, such that the residual becomes smaller.

Naïve Bayes

Naïve Bayes classification uses probability classifiers to predict outcomes based on the maximum-likelihood. The model assumes that all variables in the dataset used to predict the variable *Points_obtained_in_game* are independent. The classifying model is based on the assumption that the value of a feature in the dataset is not dependent of the value of other features in the dataset. Naïve Bayes first looks at the class that needs to be predicted and then considers the probability that some features have the given (fixed) values. The classification is done by looking at the highest probability that the given case falls in one of the classes and is then assigned to the class with the highest probability. So in our testing_set the model could be looking at the following chances:

$$win = \frac{P(win) * P(Age|win) * P(attackers|win) * P(Home|win)}{normalization\ constant}$$

This is a simplified example but gives an idea of how this classifier works. In the given dataset some features cannot be used. This is due to the fact the variables are not independent while the model does assume so. Therefore, only the original variables were used and the newly constructed variables were left out when testing this model.

K-nearest neighbor

The k-nearest neighbor model uses the space of the independent variables to classify an unknown instance. This space is n-dimensional, where n is the number of variables. The algorithm classifies the instance by 'votes' of its neighbors. This implies that the algorithm looks for the k nearest neighbors and counts how many of these neighbors are in each class. The unknown instance gets classified as the class with the most neighbors. The closest neighbors are those whom have the least distance towards the unknown instance, the distance for continuous variables is calculated the following way:

$$Distance(x, y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

For factor variables the distance is either 0 or 1. The distance is 0 when both factor variables have the same value, while the distance is 1 if the factor variables have a different value. A problem that arises often with k-NN classifications is the difference in measurement. When difference in age, measured in years, is big this will most likely not much influence the distance when another variable is money earned in a year.

Results

To gain insight in how well each method described in the previous chapter scores, the number of games predicted incorrectly is used as evaluation score. The better the predictive power of each model the lower the number of wrong predictions. For each model, the total dataset was hustled and then at random 80% of the data were put in a so called training_set, while the remaining 20% is put in a testing_set. Reason for setting the training_set at 80% is that the Dutch football has won approximately 70% of the games played the last 10 years. By setting making the training_set larger it can be that the testing_set only consists off won games, while if the training_set was made smaller it can be possible that the whole training_set consists of games that have been won.
The games predicted are those in the testing_set. For each model this is done a 1000 times. Cross validation is used to come to the best prediction model for classifying the testing_set. Results from the model were on each run the average of this 1,000. On every run the dataset was randomly hustled again and a new training_set and testing_set were chosen. Every run constituted of different games to train the model on and different games to predict. The score assigned to each model is the average percentage of games predicted incorrectly.
In this chapter adjustments that needed to be made to the whole dataset are discussed along with the results of each model. Afterwards, the results of the best prediction model are discussed in more depth.

GBM

In the GBM model some parameters still had to be determined to minimize the mean squared error, namely the number of trees, the interaction depth and the shrinkage. The number of trees is the parameter on which the model bases the number of trees that are produced in each run. More trees reduce the errors in the training_set but may lead to overfitting. The interaction depth is the number of nodes in the tree, while the shrinkage is the learning rate. Shrinkage reduces the impact of each additional added tree. Choosing a smaller shrinkage will increase the running time of the program but may lead to a better forecast. The parameters for training the model were set at:

$$Number\ of\ trees = 1500$$
$$Interaction\ depth = 5$$
$$Shrinkage = 0.1$$

For every run the model is fine-tuned so that the best number of trees is chosen. This is done by pruning the model within the loop so that the number of trees is different in each run. On average the number of trees is 10. When choosing a GBM where the number of trees is the average, the parameters are as follows:

$$Number\ of\ trees = 10$$
$$Interaction\ depth = 5$$
$$Shrinkage = 0.1$$

The results from using GBM as a prediction method with the number of trees pruned every run, is as follows:

average_games_wrong	average_percentage_games_wrong
11.14	39.78

Results from the GBM model where to parameters are the average, give the following results:

average_games_wrong	average_percentage_games_wrong
11.62	41.5

Naïve Bayes

For the Naïve Bayes classifier to work the explanatory variables in the dataset need to be independent. Therefore, when using this model to analyze the results of the Dutch national football squad, all dependent variables had to be removed. The dataset is shrunk and limited to only 17 variables. This makes the model quite weak but does fulfill the requirements. When using this very (!) small dataset results are as follows:

average_total_wrong	average_percentage_wrong
14.4	0.51

The literature states that in general GBM models outperform naïve Bayes models. Still on average the results vary quite a lot, predicting on average almost three more games wrong. Therefore, the model is also tested using the whole dataset which has variables that are co-dependent to see what the differences in results are.

average_total_wrong	average_percentage_wrong
16.2	0.58

As one can see the results of the prediction using the whole testing_set, become a lot worse with almost 60% of the games predicted incorrectly. This is in line with the theory and thus it seems not useful to use the whole dataset when using naïve Bayes as a predictive model.

K-nearest neighbor

For the k-nearest neighbor predictive model to work one has to specify to how many neighbors the model must look in order to classify the unknown input and the dataset has to be normalized. Normalization of the variables in the dataset is needed to make sure that all variables have the same impact on the distance. This is due to the fact that some variables (number of attackers) have a low maximum value, while others (ranking of the opponent) have a much bigger value and thus have a bigger impact on the distance. Normalization is done by setting each lowest value of the variable at value 0, the biggest value of the variables is set at 1 and the rest in between is given a value between 0 and 1 based on its original value. After the normalization is done, a run is done where k runs from 1 to 35 to see at which k the error made by the model is the smallest. To rule out the factor luck/coincidence for the classification model, a 1,000 runs are done for every k and the errors are stored. The plot of the errors against the k is shown below in Figure 3.

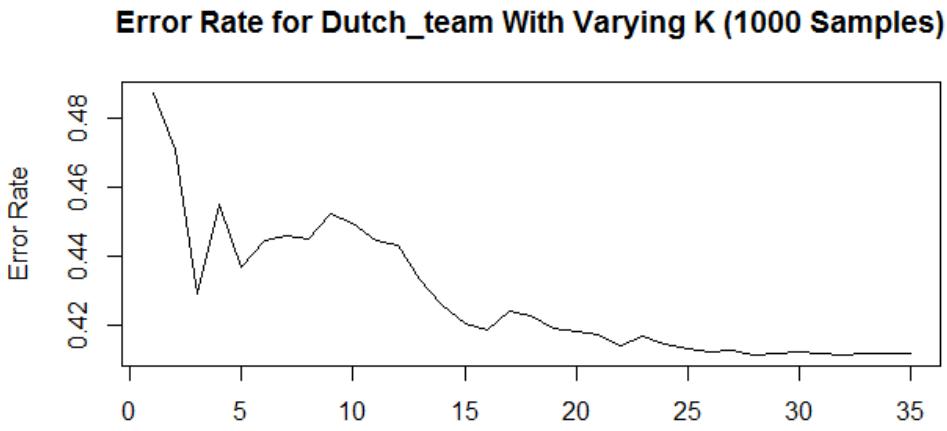

Figure 3: Errors ratings of k-nearest neighbor using different k

From the plot one can see that the minimum amount of error is achieved with a k equal to 32. Errors are not reduced when k is bigger or smaller than this 32. This k is used in the k-nearest neighbor model for classification. Running this model a 1000 times gives the following results:

```
Average_games_wrong                    average_percentage_wrong
              11.57                                       41.32
```

Results (2)

In Table 6 the results of each method are shortly summed up again.

Method	Percentage predicted correctly
Random tree	60.22%
Naïve Bayes	49.00%
K-nearest neighbor	58.68%

Table6: prediction results of the three used machine learning classifiers

Based on the predictive powers of each model, the GBM model is looked at into more detail to see what variables hold predictive powers.

GBM-results

The GBM method is the method with the best predictive score. Based on the dataset it is possible to predict 61.22 % of the matches correctly using this method, which is a better predictive score than when other methods are used. Therefore, some more other characteristics of the predictions made are shown below. In Figure 4 the number of wrong predictions by the GBM model is shown.

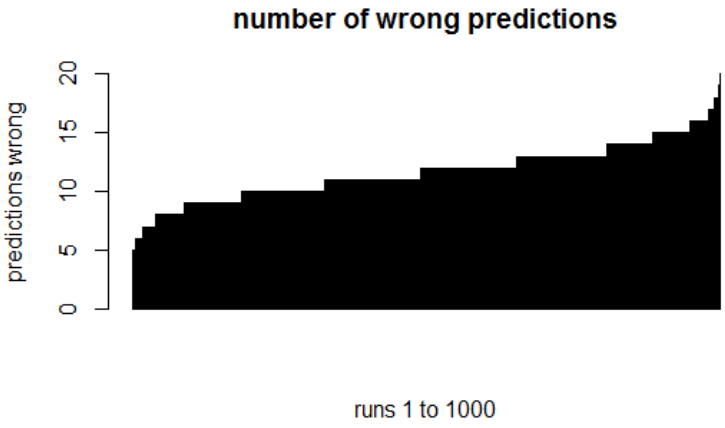

Figure 4: Number of predictions wrong predicted by GBM model sorted from low to high

In Table 7 extra statistics from the predictions made by the GBM model are shown. From the table one can see that the number of predictions made wrong by the model vary between 5 and 20, while the standard deviation of the percentage predicted wrong is 8.54%.

Statistic	Value
Average number of games wrong	11,14
Average percentage of games wrong	39,78
Standard deviation percentage of games wrong	8,54
Minimum number of games predicted wrong	5
Maximum number of games predicted wrong	20

Table 7: Statistics from the GBM model prediction

By doing a Sharipo-Wilk normality test we tested if the nul-hypothesis H_o :the number of games predicted incorrectly follows a normal distribution, can be rejected. Results of the Sharipo-wilk normality test are as follow:

```
Shapiro-Wilk normality test

data:  sorted_predictions
W = 0.98324, p-value = 2.615e-09
```

The p-value of the Sharipo-Wilk normality test is less than 5% and thus it is not possible to state that the number of games predicted wrong by the GBM model follow a normal distribution. The q-q plot shown in Figure 5 approves the results found by the Sharipo-Wilk normality test. One can see that the black dots are not nicely spread out over the red line showing that the number of games predicted wrong indeed do not follow a normal distribution. Therefore, we cannot construct a 95% confidence interval, for the percentage of games predicted incorrectly, that carries any value.

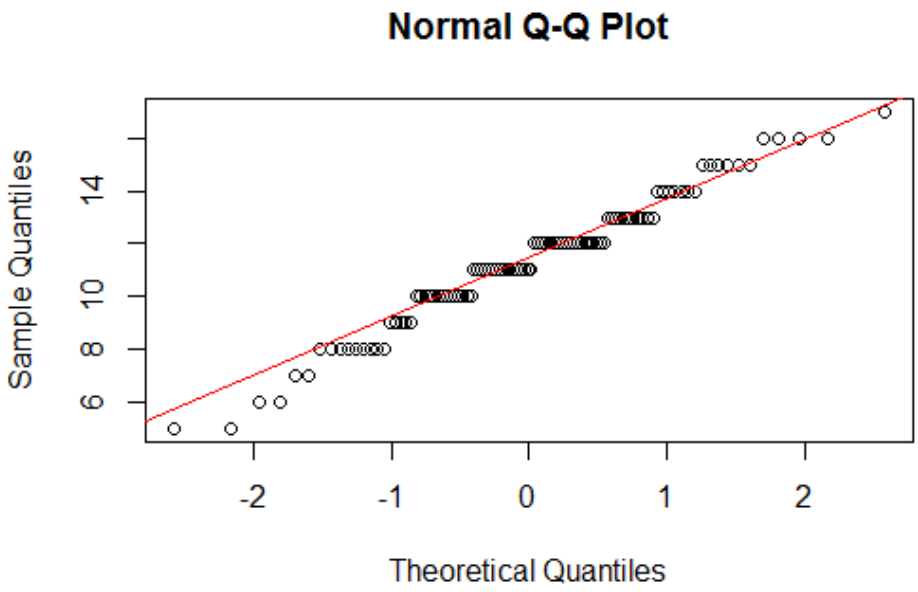

Figure 5: q-q normal plot for number of games predicted wrong

From the GBM model we can check the predictive value of the variables. When checking which of those have the highest and lowest impact on the predictions we got some interesting results. In Figure 6 the plot of all variables against it relative influence is shown.

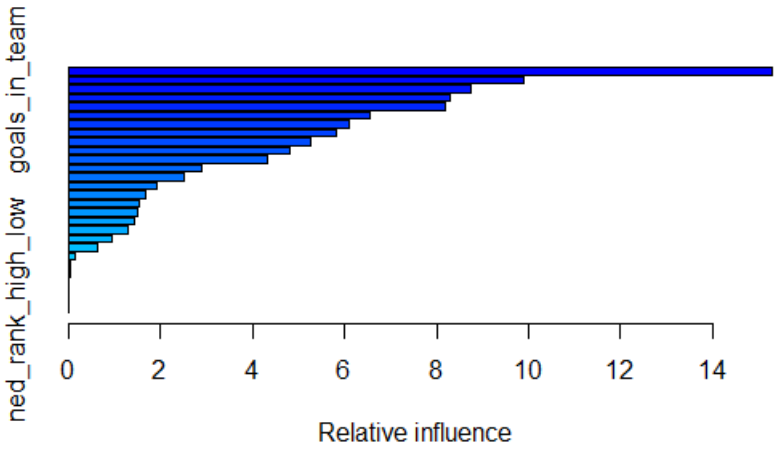

Figure 6: relative influence of all variables in dataset

From all the variables there are four that have zero influence on the decision tree namely:

- number_of_defenders
- average_caps_defenders_team_compared_to_total
- most_expensive_player_price_compared_to_average
- ranking_Netherlands_high_or_low_in_comparison

The three variables with the highest and lowest impact are in the Table 8 below.

Lowest impact	Highest impact
Number of midfielders	Difference in rank between the teams
Number of games played by captain compared to total	Percentage of defenders that are left footed
Ranking opponent high or low	Number of goals in the team

Table 8: Highest and lowest variable influence

Conclusion

When using a GBM classification method it is possible to predict 60,22% of the games of the Netherlands, on average, correctly. The other 39,78% of the games played are predicted incorrectly. Using the GBM model to predict does yield better results compared to the prediction results of Naïve Bayes and the k-nearest neighbor models. Using Naïve Bayes or the k-nearest neighbor models to predict the outcomes of the Netherlands football team resulted in a prediction score that varied between 42% and 58,62%.

From the model it follows that the best prediction the GBM model made is with an accuracy of 82.15% correct, while the worst prediction made by the GBM model is only 28.57% of the games predicted correctly. These results are based on a dataset that only includes information about the Dutch team and no information about the opponent except for their FIFA-ranking. From the variables in the dataset, tactical variables do not have much predictive value while the difference on the FIFA ranking is the variable that has the most predictive value. To improve the results of the random tree model more variables should be added that are time consuming to implement. Variables that can be added based on the opponent for instance are all the information from the last game that team played. Result, the number of defenders, the number of goals they scored and many more. Variables that have not been included about the Dutch team that can be added are distance run by each player the last game, last game played for their club, percentage of games won when certain setup was played, etc. Adding those can increase the predictive power of the model but would be very time consuming. Assumptions that have been made are that weather conditions do not have an impact on the results of matches played. This assumption can be further investigated to improve prediction results

Discussion

To improve the prediction results on the matches of the Dutch football team more research will be needed. To improve the prediction results research on other data mining techniques like neural networks should be conducted along with other methods to see if these methods give better prediction results. To further improve the research more data should be collected along with more variables that are known on forehand. Variables that can be included are the last game the opponent played and more statistics about the Dutch team's last game.

Bibliography

B. Benjamin, R.Charles, R. Pollard. (1971). Skill and chance in ball games. J*ournal of the Royal Statistical Sociaety*, 623-629.

Byungho. (2008). A compound framework for sports results prediction: A football case Study. *Knowledge-Based systems*, 551-562.

D. Dyte, S. Clarke. (2000). A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research society*, 993-998.

G.Ridder. (1194). Down to ten: estimating the effect of a red card in soccer. J*ournal of the American Statistical association*, 1124-1127.

Goddard, J. (2005). Regression models for forecasting goals and amtch results in association football. *International journal of forecasting*, 331-340.

Hill, I. (1974). Association football and statistical inference. *Applied statistics*, 203-208.

Huang, K. (2010). A neural network method for prediction of 2006 world cup football game. *The 2010 international Joint Conference on Nerual Networks*.

Kampakis. (2014). Using Twitter to predict football outcomes. *arXiv*.

Karlis, D. (2010). Robust fitting of football predictions models. *IMA Journal of Management Mathematics*.

Keogh, F. a. (2013). F*ootball betting-the global gambling industry worth billions.* retrieved november.

Knorr-Helld, L. (2000). Dynamic rating of sports teams. *Journal of the Royal Statistical Society*, 261-276.

Kooij, R. (2008). Het Nederlands elftal als complex netwerk. *Nieuws archief voor wiskunde*, 50-55.

M. Dixon, M. R. (1998). A birth process model for association football matches. *Journal of the royal statistical society*, 523-538.

M.Dixon, G. Stuart. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society*, 265-280.

Maher, M. (1982). Modelling assocation football scores. *Statistica Neerlandica*, 109-118.

Moroney, M. (1956). Facts from figures. *Facts from figures*, 472.

Phyu, T. (2009). Survey of classification techniques in data mining. *Proceeding of the INternational MultiConference of Engineers and Computer Scientists*.

Pujari, A. (2001). Data mining techniques.

R.Michalski, J. Carbonell, T. Mitchell. (2013). Machine learning: An artificial intellegence approach. *Springer Science & business Media*.

Safavian, R. (1990). A survey of decision tree classifier methodology. *a survey of decision tree classifier methodology*.

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological modelling*, 312.

Wilson, R. (1995). Ranking college football teams: A neural network approach. *interfaces*.

Wu, X. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 1-37.