Expected goals in soccer

explaining match results using predictive analytics

Eggels, H.P.H.

Award date:
2016

Expected Goals in Soccer: Explaining Match Results using Predictive Analytics

Master Thesis

H.P.H. Eggels

Supervisors:
TU/e: dr. M. Pechenizkiy
TU/e: dr. R.J. Almeida
PSV: MSc. R. van Elk
PSV: dr. Luc van Agt

Eindhoven, March 2016

Abstract

Where data based decision making is taking over businesses and elite sports, elite soccer is lacking behind. In elite soccer, decisions are still often based on emotions and recent results. As results are, however, dependent on many aspects, the reasons for these results are currently unknown by the elite soccer clubs. In our study, a method is proposed to determine the expected winner of a match.

Since goals are rare in soccer, goal scoring opportunities are analyzed instead. By analyzing which team created the best goal scoring opportunities, a feeling can be created which team should have won the game. Therefore, it is important that the quality of goal scoring opportunities accurately reflect reality. Therefore, the proposed method ensures that the quality of a goal scoring opportunity is given as the probability of the goal scoring opportunity resulting in a goal. It is shown that these scores accurately match reality.

The quality scores of individual goal scoring opportunities are then aggregated to obtain an expected match outcome, which results in an expected winner. In little more than 50% of the cases, our method is able to determine the correct winner of a match. The majority of incorrect classified winners comes from close matches where a draw is predicted.

The quality scores of the proposed method can already be used by elite soccer clubs. First of all, these clubs can evaluate periods of time more objectively. Secondly, individual matches can be evaluated to evaluate the importance of major events during a match e.g. substitutions. Finally, the quality metrics can be used to determine the performance of players over time which can be used to adjust training programs or to perform player acquisition.

Contents

Chapter 1

Introduction

In sports, many people are conservative which makes revolutions in sports difficult. Currently, however, a revolution in sports is taking place. In this revolution, data plays a major role in the decision making of important decisions, e.g. buying players.

One of the most influential people in this revolution of data based decisions in sports is Bill James. Bill James is a statistician who has been studying Baseball since the 1970s, most often through the use of data. Among his contributions to baseball is a statistic to quantify player's contribution to scored runs: Runs created. Runs created correlates well to actual runs scored.

$$\text{Runs Created} = (\text{Hits} + \text{Walks}) * \frac{\text{Total Bases}}{\text{At Bats} + \text{Walks}}$$

The success of this statistic measure comes from the success story of Billy Beane. Beane, the general manager of a low-budget team, began applying James' principles. This resulted in the Oakland Athletics being competitive with teams with much higher budgets. This success story was captured by Michael Lewis in his book Moneyball [35]. In response to the successes of the Oakland Athletics, the major teams in the NFL also started hiring statisticians.

The revolution of statistics in baseball was also noticed in other sports, one of which being basketball. In the current days, basketball teams are using "Player Tracking" technologies to evaluate the efficiency of a team by analyzing player's movements during a match. In order to do so, the basketball players and the basketball are tracked with 25 Hz during matches [1].

Currently, data-based decisions are also more often made in soccer. The first origin of soccer analytics goes back to the 1930s when Charles Reep started analyzing soccer matches. During his career, he annotated around 2200 matches, which eventually lead to his article "Skill and Chance in Association Football" published in the Journal of the Royal Statistical Society in 1968 [8].

Reep was the first to collect that amount of data in soccer. His findings, however, were only of limited quality. Reep was too focussed on finding evidence for the way soccer should be played in his own vision. He, however, failed to see the drawbacks of this playing style [2].

Where soccer analytics go back a long way, the effective use of these analyses was limited for a long time. With the rising interest of analytics in other sports, soccer is currently catching up in the use of data-based decision making. Currently, the data is no longer collected by individuals such as Reep, but by entire companies such as Opta, Prozone, and StatDNA. The effort with which this data is collected also dramatically decreased. Where it took Reep about 80 hours to analyze a single game, currently data is collected by video cameras which track all the players in real time [8].

Besides the dramatic developments in the data collection methods, the data is also used for a wider variety of applications. Maybe one of the oldest applications of data analytics in soccer is in the betting industry where data is used to determine the odds of winning, losing, and drawing. Currently, however, it is possible to bet on almost everything, from the amount of corners for each side to the number of faults and yellow cards.

Furthermore, the data is extensively used by the soccer clubs themselves. Not very long ago, clubs had to exchange video files with each other to be able to analyze opponents. With the upcoming companies such as Prozone, this is not longer necessary and clubs have almost all matches of their opponent available. The rising use of data in soccer also raises some barriers. Anderson and Sally found many situations in which data could help analyzing matches, but where coaches were not willing to use the data since they would know better anyways [2].

Soccer clubs not only use data to analyze matches but also to analyze individual players. By analyzing individual players from both their own team and other teams, soccer teams are able to determine which players should be traded. Furthermore, they are able to find the best player from other teams to fill in a position at their own team. When Anderson and Sally used this approach to help a team in the transfer window, they got very positive feedback from the board. The manager, however, was not that enthusiastic about the data-driven approach over his own judgement [2].

At PSV, this data-driven revolution is also taking place. In the approach PSV is following, however, the projects originate from business-oriented demands. The demands for data-driven solutions mainly come from soccer trainers themselves. This way, soccer trainers can contribute to the way in which the projects are carried out and are more likely to effectively use the results.

1.1 Problem Statement

Being one of the best soccer clubs in the Netherlands, staying on top of the league is one of the most important goals at PSV. In soccer, however, achievements from last seasons do not count during the current season. This means that the performance of players and their coaches is analyzed based on achievements during a short period of time. In soccer, however, the results of teams (and thus the achievements) depend on many aspects. One of these aspects is luck. A team can perform way better than their opponent but still end up drawing a match or even losing.

Since clubs in soccer are so achievement oriented and achievements do only count for a short period of time, the performance of players and coaches can go from extremely good to extremely bad in just a week time. The achievements of a team are, however, not only dependent on the skills of players and coaches but also in other aspects (including luck). Especially since goals (which determine the result of a match) are so rare in soccer. Due to the many aspects influencing the result of a match it is not only difficult to pinpoint the one responsible for a particular achievement, it is already difficult to understand why a match resulted in a particular way. It is, therefore, the goal of this thesis to provide a more objective way to get an understanding of a result in a match.

Due to the many aspects influencing the result of a match, it is hard to get a complete understanding of the result of a match. This thesis, therefore, focuses on getting an understanding of a result of a match by analyzing the quality and the quantity of goal scoring opportunities related to the result of a match. This leads to the following research question.

> "How can we quantify the quality of a goal scoring opportunity created by a team?"

By answering this question, not only the quality of a created goal scoring opportunity can be determined, but also the performance of the player in that situation can be determined. The shot of a player results in a goal or not. Comparing this to the expected quality of the goal scoring opportunity leads to insights in the performance of the shot of that player. Therefore, this research also answers the following question:

> "How can we quantify the value of a shot given the scoring opportunity?"

1.2 Methodology

The size of data of both the tactical data and the physical data provided by PSV ask for data mining solutions. Data mining is extracting useful information from large datasets and databases.

Data mining, therefore, lies in the intersection of statistics, machine learning, data management, and databases, pattern recognition, artificial intelligence and much more [9].

One of the major data mining frameworks is the Cross-Industry Standard Process for Data Mining (CRIPS-DM)[45]. This data mining framework is used as the generic framework for this thesis. A graphical representation of this model is provided in Figure 1.1.

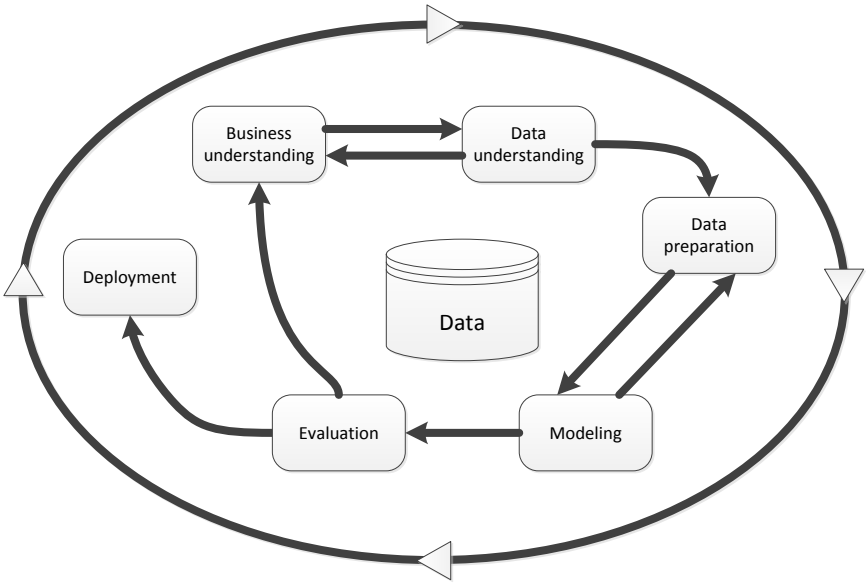


Figure 1.1: Graphical representation of the CRISP-DM methodology [45]

Below is some more elaboration on the steps provided in the CRISP-DM methodology by Shearer:

1. **Business Understanding:** Focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives.
2. **Data Understanding:** First collection of the data after which the familiarity with the data is increased such that data quality is identified and first insights are discovered.
3. **Data Preparation:** Covers all activities to construct the final dataset.
4. **Modeling:** Selection of modeling techniques and calibrating the parameters of these techniques to optimal values.
5. **Evaluation:** Review the model's construction to be certain that it properly achieves the business objectives.
6. **Deployment:** Organizing the gained knowledge in a presentable way such that the customer can use it.

1.3 Related Work

In literature, many studies have attempted to predict the result of soccer matches before the match actually started. Various perspectives have been used to tackle this problem. A common perspective to look at this problem is the prediction of soccer matches from a betting perspective [12, 33].

Furthermore, both statistical approaches and machine learning approaches have been undertaken to predict soccer match outcomes. Statistical approaches often consider goals scored by a

team as Poisson processes, hereby predicting the probability of a team to win a game [30, 24]. Other statistical approaches have studies the relationship between possession time and team's performance [29].

In machine learning, many complex systems have been proposed to predict the winner of a match. An ensemble of k-nn predictors is proposed by Hoekstra [26]. Dobravec notes the importance for non-expert based features [15]. Therefore, he proposed a method of predicting match outcomes using these latent features. Other attempts have been used by analyzing the density of network graphs in order to predict match outcomes [19].

The approaches taken in these papers are all dependent on data available before the match. More insights might be generated when including data generated during a match. Kerr uses data generated during a match to predict match outcomes [31]. Therefore, Kerr trains Logistic Regression on features extracted during the entire match e.g. possession time.

Lucey et al. [36], however, take a different approach. They determine different features (e.g. distance to the goal), to determine the quality of individual goal scoring opportunities. They call this metric the **Expected Goal Value (EGV)**. The Expected Goal Value is determined by applying Logistic Regression to the extracted features.

Similarly to Lucey et al. a quality metric is determined for individual goal scoring opportunities. Contrary to Lucey et al., however, in our approach, it is ensured that the quality metric can be interpreted as the probability that a goal scoring opportunity results in a goal. Therefore, it is important that the predicted scores match reality accurately. Furthermore, the quality metric in our approach is used to estimate match results.

1.4 Main Results

In this thesis, a metric is proposed which determines the quality of a scoring opportunity, i.e. the probability of a scoring opportunity to result in a goal. Aggregating this metric, called the Expected Goals, for matches leads to expected match results. It is shown that Expected Goals accurately represent the probability of a scoring opportunity resulting in a goal. Furthermore, it is shown that the Expected Goals leads to valuable insights into match results.

In close matches, the aggregated Expected Goals most often determine the expected winner of a match incorrectly. This is, however, as expected as goals are rare in soccer. Therefore, one single scoring opportunity resulting in a goal could make the difference between one team winning or the other team winning. The results of this thesis have also been published in Eggels et al [21].

1.5 Outline of this Thesis

The main structure of this thesis is based on the CRISP-DM methodology. The first chapter already briefly introduced the problem. Secondly, according to the business understanding phase of the CRISP-DM methodology, the preliminary plan of the data mining tasks is defined in Chapter 2. The datasets and a general understanding of the data are provided in Chapter 3. Modeling techniques are selected and applied to the data in Chapter 4. The model is evaluated in Chapter 5. According to the final step of the CRISP-DM methodology, the gained knowledge is visualized in Chapter 6. To show the possible applications of the expected goals model for soccer clubs, a case study is performed with the data about FC Barcelona. Here, data from FC Barcelona is used instead of PSV due to confidentiality. Finally, the thesis is completed with a conclusion and recommendations for future work.

Chapter 2

Predictive Analytics for Characterizing Scoring Opportunities

This chapter translates the given business problem into a predictive analytics framework, including a set of data mining tasks. It is important to make this translation in the early stages of this thesis since these choices determine the interpretation of the eventual results. First of all, the category of data mining tasks in which the business problem can be categorized is discussed. Secondly, the appropriate techniques of this category are discussed. Then, a technique (calibration) is introduced to improve these techniques. Some of the characteristics of the model are then elaborated on. Of these characteristics, the need for a range interval is discussed in Section 2.4. Then, the influence of the bias, variance, and the irreducible error and their implications are elaborated on. The eventual interpretation of the obtained scores are discussed in Section 2.6. Finally, the chapter is finished with a conclusion.

2.1 Data Mining Tasks

Data mining can be categorized into multiple different types. These tasks differ mainly based on the types of objectives that a person wants to achieve with the data mining application. A classification of data mining tasks is provided below [20].

- **Exploratory Data Analysis (EDA):** The goal of EDA is to explore the data without any clear ideas of the aspects that we are looking for. EDA becomes much more difficult when the dimensions of the data are increasing since it is more difficult to visualize this data. An example of EDA is research by Becker, Eick, and Wilks who created a set of intricate spatial displays for visualization of time-varying long-distance telephone network patterns [3]
- **Descriptive Modelling:** The goal of Descriptive modeling is to describe the complete data set. One could, for example, divide the data into different clusters (clustering or segmentation) or describe the relationships between different variables (dependency analysis). Clustering techniques have been used to analyze the long-term climate variability in the upper atmosphere of the Earth's Northern hemisphere [11]
- **Predictive Modeling: Classification and Regression:** Aims to build a model that permits the value of one variable to be predicted from the known variables. The difference between classification and regression is based on the output of both methods. Where classification results in a categorical variable, regression results in numerical output. Predictive models have, for example, been used to develop a system which tracks the characteristics of all unique telephone numbers in the United States [13]

- **Discovering Patterns and Rules:** Besides techniques which build models, there also exist data mining techniques who are concerned with pattern detection. An example of such techniques is the task in which combinations of items that occur frequently in transaction databases are found [20]
- **Retrival by Content:** Consists of finding patterns from the data which are previously defined by the user. This kind of tasks is most commonly used for text and image data sets. Search engine Google, for example, uses this kind of retrieval methods to locate documents on the Web [6]

The objective of this thesis is, as stated in Section 1.1, to determine the quality of a goal scoring opportunity. The quality of such a goal scoring opportunity is determined based on input variables. Therefore, the problem of this thesis asks for predictive modeling techniques. More specifically, the problem of this thesis asks for a numerical output in which the quality of a goal scoring opportunity is provided. Therefore, one could argue that regression techniques best suit this problem. As we see later in this chapter, however, there also exists methods in which classification techniques can result in numerical output. Furthermore, the provided data set exists of binary target variables (goal or no goal) and therefore, classification techniques suit this problem best.

Other approaches could, however, still be useful with the current data set. Exploratory Data Analysis and Descriptive modeling could, for example, be used to to get a better understanding of the data.

2.2 Classification

Classification is the task of learning a target function f that maps each attribute set x to one of the predefined class labels y [49]. In the case of the expected goal model the class labels are goal or no goal. Therefore, $y_i \in \{\text{goal}, \text{no goal}\}$.

In classification problems, no single best classification algorithm exists which is better than all the other available classification algorithms [51]. One way to determine which algorithm to choose is to determine the best classification algorithm for a given problem by cross-validation [32]. By using cross-validation, multiple classification algorithms are trained and performance metrics are calculated accordingly. Based on these metrics, the best model can then be selected. For this thesis, four different algorithms are used: Logistic Regression, Decision trees, Random Forest, and Ada Boost. Implementations of these algorithms by scikit-learn are used [42]. The random forest and the Ada-boost algorithm are examples of ensemble learners. Ensemble methods use multiple learning algorithms to obtain better predictions than could have been obtained from the individual learning algorithms [44].

Logistic Regression Estimation of the probability of occurrence of an event as a function of a relatively large number of variables [50].

Decision Tree is a classifier expressed as a recursive partition of the instance space [37].

Random Forest is a set of multiple decision trees where the predicted class is determined from the mode of the classes [25]. Random forests are able to correct for the poor generalization of decision trees [22]

Ada Boost starts by fitting a classifier on the original data. Then, additional copies of the classifier are fitted on the data where the weights of incorrectly classified instances such that the new classifiers focus on more difficult cases [18]. For this thesis, decision trees are used as the underlying classifier for Ada Boost.

In order to determine the best of these algorithms, multiple performance metrics can be used which all have their own benefit. Some of these metrics are computed with the help of the confusion matrix (Table 2.1). These metrics are listed below [47].

Table 2.1: Confusion matrix

		Actual	
		Positive (Goal)	Negative (Non-Goal)
Predicted	Positive (Goal)	True Positive (TP)	False Positive (FP)
	Negative (Non-Goal)	False Negative (FN)	True Negative (TN)

- **Precision:** $\frac{TP}{TP+FP}$
- **Recall:** $\frac{TP}{TP+FN}$
- **F-score:** $\frac{\text{Precsion} \cdot \text{Recall}}{\text{Precsion}+\text{Recall}}$
- **Area under the ROC Curve (AUC):** $\frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$

A typical way of evaluating classifiers is using the 0-1 loss function, where zero error is given to the values which the classifier predicted correctly and 1 to the incorrect predictions. The objective of this thesis, however, is not to predict all the samples correctly but to score different goal scoring opportunities. Therefore, it is more important to rank different scoring opportunities relative to each other. Therefore, the Area Under the Curve (AUC) seems the best performance metric. The AUC can be interpreted as the probability that the classifier assigns a higher score to a random positive example than to a random negative example [4]. The AUC, therefore, gives insights in the way in which the classifier is able to rank better scoring opportunities indeed higher. Performance metrics resulting from the 0-1 loss function do, however, still provide value as they give insights into how much the classifier is indeed able to give high scores to good scoring opportunities. Therefore, the precision, recall, and F-score are also provided in the evaluation of classifiers.

2.3 Calibration

As previously mentioned in Section 2.1, classification outputs can be represented as a probability. This is most often done by computing the class membership probabilities. Class membership probabilities can be interpreted as the confidence of a classifier of a sample belonging to a certain class. To obtain better probabilities, these probabilities are re-calibrated. Two main techniques exist to map the model predictions to posterior probabilities: Platt Calibration and Isotonic Regression. Since we are dealing with a two class problem, these methods can easily be applied to the current problem.

Platt Calibration: Platt proposed to transform SVM predictions to posterior probabilities by passing them through a sigmoid function [43].

Isotonic Regression: A more general approach is proposed by Zadrozny and Elkan who used Isotonic regression to calibrate predictions from SVMs, Naive Bayes, Boosted Bayes, and decision trees [53, 54]. It is more general in the sense that the mapping function is no longer a sigmoid function but only needs to be monotonically increasing [40].

Niculescu-Mizil and Caruana show that Platt scaling outperforms Isotonic regression when the data set is relatively small. When the size of the data set, however, increases (1000 samples or more) Isotonic regression outperforms Platt scaling [40]. Since the data set provided consists of more than 1000 samples, Isotonic regression is used further on.

The performance of the calibrated model differs from the performance of the non-calibrated classifier. For the calibrated classifier, however, the main objective is to provide good estimates of the actual probabilities. The performance of the calibrated classifier can, therefore, be determined by comparing the calibrated probabilities to actual probabilities. Since no actual probabilities are provided, bins with similarly calibrated probabilities are created. The actual ratio of goals is then

determined for those bins and plotted against the mean probabilities of the bins. This method is similar to the method proposed by Niculescu-Mizil and Caruana [40]. Furthermore, the brier score proposed by Brier is calculated [5]. This score corresponds in many ways to the mean squared error and thus provides the error of the probabilities to the classification problem.

2.4 Confidence Interval

Since many factors influence the quality of goal scoring opportunities, it is likely that the quality of goal scoring opportunities has a high variance. A single point prediction would therefore not provide enough information to draw valid conclusions. Therefore, it is important to provide a range prediction instead of a single point prediction. Therefore, a confidence interval is required. The boundaries of the confidence interval are given by:

$$\left(x - t^* \frac{s}{\sqrt{n}}, x + t^* \frac{s}{\sqrt{n}}\right), \tag{2.1}$$

where x is the point prediction, t comes from the student-distribution, s is the estimated standard deviation, and n the number of samples (in our case 1).

Equation 2.1 shows that the standard deviation is required to compute the prediction interval. Since all these factors play a role in determining the quality of a goal attempt, it would, however, make no sense to compute the standard deviation of the complete data set. It would seem better to compute the standard deviation of similar samples. Two techniques were discussed to compute the standard deviation of similar samples.

n-Nearest Neighbours finds the n Nearest Neighbours of the sample. The standard deviation could then be computed over these n samples.

Clustering techniques cluster similar samples together. The inner-standard deviation of these clusters could then be computed as the standard deviation of the points.

To compute the standard deviation using n-Nearest Neighbours, the Nearest Neighbour algorithm should have to compute the nearest neighbours of each individual sample. This would be very resource intensive. By using clustering algorithms, however, only the standard deviation of the cluster has to be computed. Therefore, clustering is used instead of Nearest Neighbours. Gaussian mixture models is a clustering technique which is commonly used for kernel density estimation. Therefore, Gaussian mixture models are used to calculate the standard deviation of clusters.

2.5 Bias, Variance & Irreducible Error

Before the implications of Bias, Variance & the Irreducible Error can be discussed, firstly, the theory of these concepts has to be discussed. Tan defines Bias, Variance, and the Irreducible error as follows [49]:

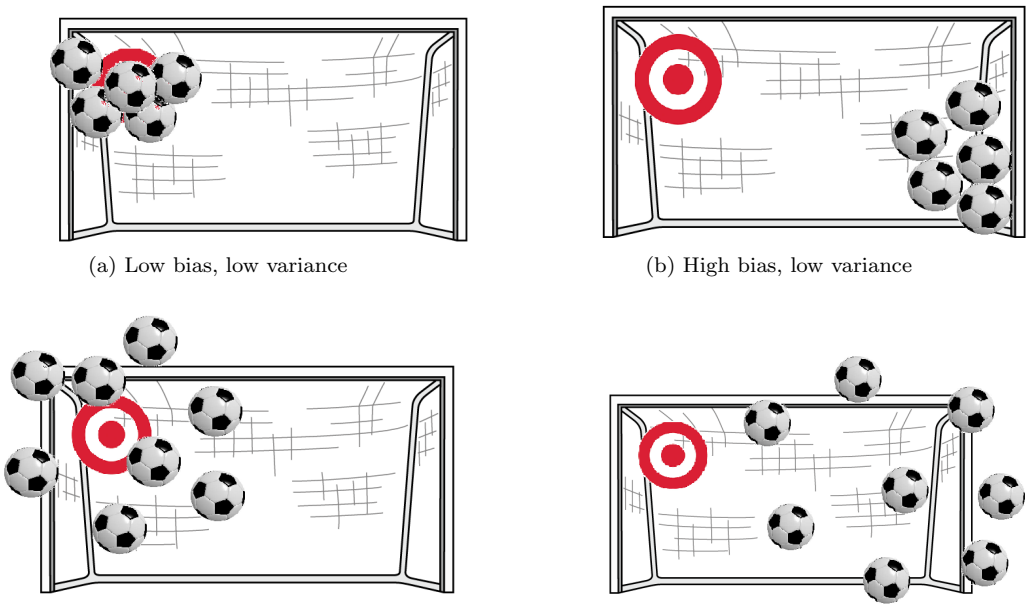
Bias is the average distance between the actual value and the predicted value.

Variance is the deviation between a single prediction x and the average prediction $\bar{x}$.

Irreducible Error (Noise) the term that cannot be reduced to any model.

Variance and Bias are best explained graphically in Figure 2.1. Figure 2.1 shows that for bias, the soccer balls are indeed close to each other. They are, however, not close to the target. In the case of high variance, the mean of the points indeed lies close to the target, the individual attempts, however, do not hit the target.

Now the concepts of Bias, Variance & Irreducible Error are clear, the implications of these concepts to the stated problem can be discussed. During this discussion, it is important to keep



(a) Low bias, low variance

(b) High bias, low variance

(c) Low bias, high variance

(d) High bias, High variance

Figure 2.1: Visual representation of Bias and Variance

the goal of this thesis in mind: Analyze goal scoring opportunities created by a given team with given players.

In each predictive modeling task, ideally one would create a model with zero bias and zero variance. Relating such a model to Figure 2.1 would lead to a situation in which the target would almost always be hit (Figure 2.1a), except for external factors influencing the shot (the irreducible error). In terms of classification, the classifier would always be able to assign the proper class label to each case. In most real world applications this is, however, impossible. Therefore, the goal is to minimize the bias and variance of the predictive model.

In predictive modeling, the input of the predictive model is one of the main aspects which determine the bias, variance, and the irreducible error of the model. To make this clear, lets go back to the situation in Figure 2.1 in which a player aims at a target in a goal. Without any knowledge of the situation, it is hard to determine the likelihood of the player hitting the target. With prior knowledge such as the distance of the player to the target, this would be somewhat easier. Therefore, one could argue that collecting features of the situation of the player could lead to a reasonable estimate of the likelihood that the player hits the target.

Now lets consider the situation in which a player with poor shooting accuracy attempts to score. Due to the player characteristics the outcome of the goal attempt, given the situation, would differ significantly from this player. The predictive model is, however, not able to anticipate to these fluctuations and would, therefore, have a higher error. The goal attempts of a player with good shooting accuracy, however, would fluctuate less and are therefore more predictable. A predictive model trained on only the players with good shooting accuracy would, therefore, have more predictive power. Therefore, one could argue that the predictive model should be trained on only the players with good scoring accuracy.

But, what if the original player with the poor shooting skills would try to hit the target again? Would the predictive model trained on the best players still be able to accurately determine the likelihood of the player hitting the target? Intuitively, this would still provide some insights since the situation in which the players find themselves did not change. The likelihood of the poor

player hitting the target would, however, be significantly lower than the excellent player hitting the target. Therefore, it seems important to take the quality of the player into account as well. Since lower quality players are less predictable, the predictive model would probably lead to a less optimal model (higher variance). The practical meaning of such a model would, however, be limited since it would be only applicable to the better players.

Now lets have a look at the situation in which the player, instead of hitting the target, attempts to score a goal. To make the situation more realistic, a goalkeeper is standing in the goal trying to prevent the player from scoring and thus aims to stop the goal attempt. In this case, the likelihood of the player scoring does not only depend on the situation and the player quality anymore but also of the quality of the goalkeeper. If an excellent goalkeeper would be on the line, the likelihood of the goalkeeper stopping the goal attempt would be higher than when a poor goalkeeper is on the line. Therefore, the quality of the opposing goalkeeper is included in the model.

The same reasoning could be applied to the other defenders who are trying to stop the attacker from scoring. The quality of a goal scoring opportunity is, however, determined at the moment at which the player attempts to score. At this moment, the position of the defenders on the field is, however, already given. The only action the defenders could perform at this moment is blocking the goal attempt. Intuitively, there, however, seems to be no significant difference in player quality regarding the blocking of a goal attempt. Therefore, it seems useless to include defender quality as input for the predictive model.

2.6 Interpretation

After calibration is applied to the classifier, the classifier scores can be interpreted as posterior probabilities. Let p_i be the probability that a goal is scored from a goal scoring opportunity i. The goal scoring opportunities can then be modelled as a Bernoulli random variable $y_i \sim Ber(p_i)$. Then, the expected number of goals in a match of n goal scoring opportunities is equal to:

$$E[\#Goals] = E\left[\sum_{i=1}^{n} y_i\right] = \sum_{i=1}^{n} E[y_i] = \sum_{i=1}^{n} p_i \tag{2.2}$$

2.7 Conclusion

In this chapter, the choice of predictive modeling, and more specifically classification, is justified. Furthermore, different classification algorithms are discussed. These classification algorithms are trained based on the Area Under the Curve (AUC). This performance metric ensures that better goal scoring opportunities indeed score higher. Since the classification algorithms only provide a confidence of belonging to a particular class, a calibration step is added. By calibrating the classification algorithms, more realistic probabilities are obtained. Due to the background of the problem, it is important to provide a range prediction to single predictions. Finally, the importance of adding player quality data to the input and the interpretation of the final scores were discussed.

Chapter 3

Soccer Match Data

In order to apply the methods discussed in Chapter 2, a good understanding of the data is required. Therefore, this chapter elaborated on the available data. First of all, the three different data sources are discussed. Three aspects of the given data sources are therefore discussed. First of all, the methods to collect the data are discussed. Secondly, the general format of the obtained data is discussed. Finally, for each of the data sources (potential) data quality issues are discussed. After the introduction of the data sources, this chapter shows how these data sources are combined to provide more valuable insights. Finally, this chapter is finished with a conclusion.

3.1 Tactical Data

ORTEC is one of the largest providers of advanced planning and optimization solutions and services. ORTEC consists of three business units: ORTEC Logistics Solutions, ORTEC Consulting, ORTEC Living Data. ORTEC Sports is part of part of ORTEC Living data. ORTEC Sports tries to provide sports teams and individuals with analysis which allows fine-tuning individual and team achievements [41].

Besides the analysis provided by ORTEC, sports teams can also use ORTEC data to do their own analysis. Further on in this thesis, the ORTEC data is referred to as tactical data. Tactical data consists of important events which occurred during a match. These events include passes, goal attempts, saves, fouls cards and much more.

3.1.1 Data Collection

ORTEC data is collected by connecting to the ORTEC API. Firstly, the available competitions have to be determined from the API. Then, each of the competitions consists of several editions. These editions correspond to the seasons for which data is available. Each of these editions has several phases. These phases often exist of final-semi final-etc. for knock-out competitions. For normal competitions, often a regular competition is provided. Some of these competitions also have playoffs, e.g. Dutch Eredivisie. Finally, the phases consist of the data of matches.

In terms of ORTEC, matches are a set of events. These events in itself have related events. Take for example a goal attempt. This goal attempt could only occur due to some previous events, e.g. a pass. The pass previous to this goal attempt is the related event and therefore stored as a related event. The pass in itself, however, is also the main event and would, therefore, occur twice in the data if all related events are extracted from the API. Another example is the situation in which a player commits a foul. This foul could result in the player getting a card. The card is then the related event of the foul. The card, however, is not stored as an event on its own. When collecting the ORTEC data, one has to be careful with related events. Extracting all of the related events would result in many events occurring multiple times in the data. Extracting none of the

related events, however, would result in missing data. Currently, only cards (yellow and red) are included as related events. The structure of the ORTEC API is provided in Figure 3.1.

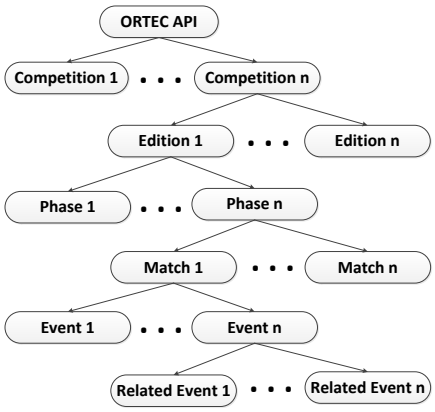


Figure 3.1: Structure of the ORTEC API

The events resulting from the ORTEC API are provided in JSON format. Since collecting all the data takes time, the JSON data is stored in CSV files for later use.

3.1.2 General Format

As mentioned in the introduction of this chapter, tactical data consists of important events which occurred during a match. Each of these events contains attributes to provide additional value. The different attributes collected from the ORTEC API are provided below.

- **Tijd (Time stamp):** The time of the match in milliseconds (ms)
- **Helft (Half):** Categorical variable which states whether the match is in the first (Helft = 1) or the second half (Helft = 2).
- **Effectiviteit (Effectiveness):** The effectiveness of a moment in the match. The effectiveness is represented with a number between one and five. The effectiveness is, however, legacy and is no longer used by ORTEC anymore. Therefore, information about the meaning of these numbers is not provided.
- **Categorie (Category):** The category of a moment specifies the type of moment in the match e.g. pass, throw in, goal attempt, save, etc.
- **Speler (Player):** The player who performed the specific moment
- **Team:** The team of which the player belonged to
- **Attribuut (Attribute):** The attribute provides some more information about the specific moment. The attribute states for example with which part of the body the action was performed. The attribute is also legacy and therefore no longer used by ORTEC. For this project, the attribute is only used to determine with which part of the body the action was performed.
- **Definitie (Definition):** The definition provides some more in-depth information about the moment. The definition provides information about the direction of the action, whether it was successful etc.
- **Wedstrijd (Match):** The match in which the specific moment was performed.
- **Ronde (Round):** The round of the league in which the match of the moment was performed.

- **Location X:** The horizontal location of the action.
- **Location Y:** The vertical location of the action.

A sample of the ORTEC data set is provided in Table 3.1.

Table 3.1: A small sample of the ORTEC data

Time stamp	Half	Effectivity	Category	Player	Team	···
162012	1	3	defending action	Joshua Brenet	psv	···
164930	1	3	defending action	Adam Maher	psv	···
164930	1	3	foul	Karim El Ahmadi	feyenoord	···
164930	1	3	attacking action	Karim El Ahmadi	feyenoord	···
183352	1	3	indirect free kick	Andrs Guardado	psv	···
184426	1	3	pass	Adam Maher	psv	···
185982	1	3	pass	Andrs Guardado	psv	···

···	Attribute	Definition	···
···	clearance ‖	isClearance ‖	···
···	duel touched ‖ body ‖	isDuelPart ‖ isDuelWonByDefender ‖ isDuelStanding ‖	···
···		isDuelPart ‖ isDuelStanding ‖	···
···	body ‖ duel untouched ‖	isPossessionLoss ‖ isDuelPart ‖ isDuelStanding ‖	···
···	left foot ‖	isPossessionGain ‖	···
···	right foot ‖ direct ‖	isPassCompleted ‖ isPassBackward ‖	···
···	left foot ‖	isPassCompleted ‖ isPassForward ‖	···

···	Match	Round	Date	Location X	Location Y
···	psv - feyenoord	4	30-8-2015	13.187	31.864
···	psv - feyenoord	4	30-8-2015	23.862	13.476
···	psv - feyenoord	4	30-8-2015	76.336	85.480
···	psv - feyenoord	4	30-8-2015	76.572	88.131
···	psv - feyenoord	4	30-8-2015	23.626	21.033
···	psv - feyenoord	4	30-8-2015	30.141	21.788
···	psv - feyenoord	4	30-8-2015	19.309	19.270

3.1.3 Data Quality

At ORTEC sports, match data is collected real-time. Since employees need time to react, the time stamp of the ORTEC data is not likely to match the actual time stamp. When matching time stamps with other data sources, this could lead to potential mismatches. Section 3.4.2 elaborates on how this potential danger is solved. It is, however, important to note that the reaction time is not constant for all the events during a match. Therefore, it is impossible to match the time stamps of the match exactly for all the events.

Similarly, the ORTEC employee has to define the location at which the event occurred manually as well. Therefore, the location is not likely to match the exact location of the field. Section 3.4.3 elaborates on the difficulties of matching locations.

3.2 Spatiotemporal Data

Inmotio is a software package developed by Abatec AG and TNO. From 2012 on, Inmotio is part of Inmotiotec GmbH located in Austria. Inmotiotec tracks players using Radio-frequency identification (RFID). This allows far better tracking of players compared to GPS tracking. Inmotio is mainly used to quantify the physical load of players. Recently, the tracking data is also used to

perform tactical analyses, such as quantification of pass options, transition moments, crosses, and ball pressure [27].

During matches, however, the RFID technology cannot be used due to legislation. Therefore, during matches players are tracked with the use of cameras. This, however, leads to quality issues which are addressed in Section 3.2.4

3.2.1 Data Collection

Inmotio makes data collection extremely easy. Within their software, where the normal analyses are performed, Inmotio provides an export option. This option is used to export the data to CSV format.

3.2.2 General Format

Tracking of players during matches is performed with a frequency of 10Hz. This means that the exported files consist of 26 (22 players, 3 referees and the ball) lines each 100ms. Each of these lines consists of the following attributes.

- **Time stamp:** The time of the match in milliseconds (ms)
- **X:** The horizontal position of the specific player on the field.
- **Y:** The vertical position on the field
- **Marker Name:** The match is divided into six quarters with each quarter their own Marker Name. A list of the complete set of Marker Names is provided in Table 3.2
- **Snelheid (Speed):** The speed of the specific object at the given time in meters per second (m/s)
- **Acceleration:** The acceleration of the specific object at the given time in square meters per second (m^2/s)
- **ID:** The ID of an object during this specific match. Objects are not necessarily the same for objects over different matches.
- **Naam (Name):** The name of the object. This could be a player name, the ball or the referee.

Table 3.2: Different versions of marker names. In this table, the marker names are already in lower cases and spaces have been removed

Start first half	Start second half
s1	s2
start	start2
start1	

A sample of the Inmotio data is provided in Table 3.3.

Table 3.3: A small sample of the Inmotio data

Time stamp	X	Y	Marker Name	Speed	Acceleration	ID	Player
164700	33.941	-27.691	Start1	5.35	-4.77	17	Adam Maher
164800	34.308	-28.122	Start1	4.81	-6.02	17	Adam Maher
164900	34.624	-28.446	Start1	4.16	-7.01	17	Adam Maher
165000	34.888	-28.664	Start1	3.43	-7.05	17	Adam Maher
165100	35.107	-28.796	Start1	2.78	-5.9	17	Adam Maher

3.2.3 Data 2014-2015

As mentioned earlier, there are some significant changes between the Inmotio data of season 2014-2015 and the Inmotio data of season 2015-2016. The main difference between the two is that in season 2014-2015, there is no ID which identifies the ID number of a player during a match. Due to this, we are forced to identify a player by name only.

Furthermore, there are some differences in column names. Where in season 2015-2016 Dutch names are used for the Name (Naam) and the speed (Snelheid) of the player, in season 2014-2015 the English names Name and Speed are used. Besides these changes, the Marker names had minor changes. In season 2015-2016, the Marker Names were "Start1" and "Start2" with capitalization or without and with or without a space between start and the number. In season 2014-2015 this notation is also used for some of the matches but some other matches also use a slightly different notation with for example "s2" or "45" for the start for the second half. The complete set of changes in marker names can be seen in Table ...

3.2.4 Data Quality

Tracking players with the use of cameras lead to some (potential) data quality issues. First of all, when players are outside the range of the cameras, the camera loses the player. When the player comes back into play, someone has to manually select the player again. This could result in some time in which the player is not tracked.

Furthermore, when multiple players are very close to each other on the field, players could be switched by the cameras. When this happens, this is not easily seen from the data since the player still has a location. Situations, where players are very close on the field, occur many times during a match. Take for example a corner kick, in which many players are in the same area of the field. This leads to many (potential) situations in which data quality issues could occur.

To handle the potential dangers of camera tracking, employees are situated behind the Inmotio software to correct possible mistakes real time. Since the employees do, however, need some reaction time, the locations could still be incorrect for a small period of time. Another possible solution is to apply Data Quality Assessment (DQA) on the data. Data Quality Assessment is the process of evaluating data in order to determine whether the data meets the required quality. Data Quality Assessment is, however, a time-consuming task and therefore expensive.

3.3 Player Data

No insights in player quality could be extracted from the tactical data nor the spatiotemporal data. Therefore, other ways of expressing player quality have to be used to determine the quality of a player. Currently, broadcasting companies are using game related player data originating from Football manager as the data source for their match analysis [48]. Furthermore, statistical companies such as Prozone use the same data to enrich their data [7]. Currently, there is a general agreement that the game Football Manager provides player data of good quality that can be used in soccer analytics. The data from the game Football Manager is, however, not publicly available. The choice is therefore made to switch to a different game of which the data is publicly available. After some research, data originating from the soccer game FIFA is found which is used in this thesis as a measure of estimating player quality [46]

3.3.1 Data Collection

As mentioned in the introduction of this chapter, data from the game FIFA is used to measure player quality. This data is publicly available on the web but could, however, not easily be downloaded [46]. Therefore, web scraping techniques are used to scrape the data from the web. The downside of these techniques is that, when websites change, the web scraper has to be modified accordingly.

The web scraper starts from a base URL. In this base URL, the main competitions of which data should be collected can be selected. Then, the web scraper starts by selecting a single team, of which the individual players are scraped. This process is repeated until all the teams of all the competitions are finished. Then, the web scraper continues the process for other seasons if specified by the user. The process if graphically represented by Figure 3.2

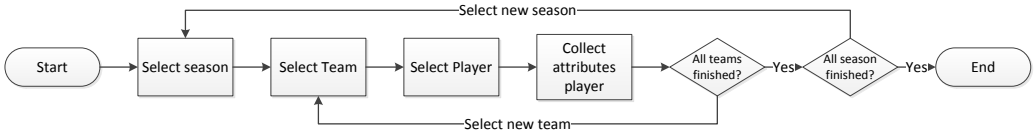


Figure 3.2: Process of scraping the player data

The scraped data is stored in CSV files such that the data does not have to be collected every time.

3.3.2 General Format

As described in the data collection phase, player data of different seasons is collected. Therefore, a player could appear more than once in the data. In order to be able to select the right player, the player's name, current club and the season of which the data are collected are stored. Furthermore, FIFA stores 33 player attributes of each player. These attributes all contain values ranging from 0 (very poor) to 100 (perfect). A sample of the player data is provided in table 3.4

Table 3.4: A small sample of the FIFA data

Player	Team	Finishing	Heading	···	GK Diving	GK Reflexes
Jeroen Zoet	PSV	12	19	···	79	84
Mats Hummels	Borussia Dortmund	55	90	···	15	6
Daniel Alves da Silva	FC Barcelona	60	71	···	5	7
Georginio Wijnaldum	Newcastle United	76	72	···	16	6
Memphis Depay	Manchester United	73	55	···	8	10

3.3.3 Data Quality

Players are evaluated by EA Sports, the manufacturer of the FIFA games. It is, however, unknown, which criteria EA Sports uses to evaluate the attributes of the different players. Many public discussions are held every year on which players are overrated or underrated by EA Sports. The quality of the scores of the player attributes is therefore at least questionable.

In 2014, the Daily Mirror researched the quality of the player attributes [14]. Therefore, they studied three different player attributes: Passing, Tackling, and finishing. To illustrate the results, the results from Passing and Tackling are provided in Table 3.5. The Daily Mirror compared the ratings provided by EA Sports with the actual performance of a player. Then, they ranked the players on both and show the difference of the top performing players in the Barclay Premier League.

Table 3.5 shows that the differences in player ranking can become large. It is, however, important to note that these are all the high performing players of the league. Differences in the ranking of the players are not very significant and therefore, differences in ranking the players are expected. Besides that, it is not known on which basis the players are ranked. It could, for example, mean that the difficulty of a pass or tackle results also influences the score.

Table 3.5: Analysis of the player data quality (Source: Daily Mirror [14])

Passing			
Name	Rating	Accuracy	Difference
Barry	85	86.8	0
Coutinho	85	80.6	6
Fernandinho	85	88.3	-3
Fletcher	85	89.1	-6
Milner	85	84	3
Nasri	85	91.5	-9
Oscar	85	83.5	4
van Persie	85	76.7	7
Allen	86	86.8	5
Carrick	86	88.6	0
Gerrard	86	86	7
Hazard	86	83.3	10
Yaya Toure	86	90.1	-2
Arteta	87	92.1	-2
Britton	87	90.3	0
Mata	87	88.6	3
Ozil	88	88	8
Silva	89	88.2	8

Tackling			
Name	Rating	Success	Difference
Cahill	85	79.4	-7
Canas	85	58.3	5
Clichy	85	77.2	-6
Evra	85	68.4	3
Terry	85	62.1	4
Zabaleta	85	76.7	-4
Agger	86	74.1	3
Ivanovic	86	83.1	-4
Koscielny	86	75.9	1
Mertesacker	86	70.7	5
Cole	87	69.2	9
Ferdinand	87	90	-2
Medel	87	77.2	1
Vidic	89	75.9	5
Kompany	90	70.8	9

In conclusion, the results show that the rankings from EA Sports are questionable but the results do not show that there are serious data quality issues. Therefore, the attributes ranked by EA Sports are used in this thesis.

3.4 Merging the Data

To get the most interesting data set, both data sets have to be combined. This section elaborates on how both data sets are combined and which issues were encountered during the combination of both data sets[1]. To combine the data, the tactical data is used as the main file. This means that both the spatiotemporal data and the player data are matched to the tactical data. Firstly, different names exist in the tactical data, player data and spatiotemporal data for a given player. Therefore, player names have to be matched.

The tactical data and the spatiotemporal data start the match at slightly different moments. Therefore, timestamps have to be matched. Finally, the spatial data is different for both the tactical data and the spatiotemporal data. The selection of spatial data is therefore described in Section 3.4.3.

3.4.1 Player Name

In order to be able to merge the different data sources, the right player should be matched. Since player names are, however, not consistent among the different data sources, a mapping should take place. Both, the names of the player data and the names of the spatiotemporal data are mapped to the tactical data.

There is a wide variability of reasons why names are not consistent. First of all, player names have to be inserted manually in the Inmotio system. Therefore, sometimes typos could occur. Secondly, differences in encoding lead to different player names. There are, for example, no accents in the spatiotemporal data and the player data, where there are accents in the tactical data. Thirdly, differences in capitalization occur. Finally, some data sources use more middle names of players than other or sometimes even use nicknames for players. This makes it hard to map the player names.

In order to still map most of the player names all accents are removed. Furthermore, all player names are put to lower cases. Still many different player names do, however, occur. Therefore, the similarity of two player names has to be determined to select the most similar one. Therefore, the Levenshtein distance is computed. The Levenshtein distance computes the number of characters in one string that has to be changed to get to the desired string [34].

Table 3.6: Example of different names and corresponding Levenshtein distance

Name ORTEC	Name Inmotio	Levenshtein distance
Adam Maher	Adam Maher	0
Luuk de Jong	Luuk De Jong	1
Denys Oliynyk	Denys Oliinyk	1
Jerry st. Juste	Jeremiah st. Juste	5
Anthony Limbombe Ekanga	Anthony Limbombe	7

Since many player names occur and some might be similar. the number of player names from one data source to which the player name is compared should be minimized. When mapping the player names from the spatiotemporal data, the player names are therefore matched for each individual match. By mapping the player names for each individual match, the maximum number of player names of the tactical data to which the player name can be matched equals 28 (22 starting players and at most 3 substitutions for each team).

To limit the number of player names to be matched with the player data, the players are matched to the players of a particular team in a season. Since, however, the team names are also inconsistent among the data sources, there might not always be a direct match. Mapping the team names, however, seems more difficult since there are many teams with short names and it might be more beneficial to map teams to short teams (lower number of incorrect characters).

[1] Merging of the data sets has been conducted in cooperation with Kees Hendriks, who performed his graduation at the same time [23]

3.4.2 Timestamp

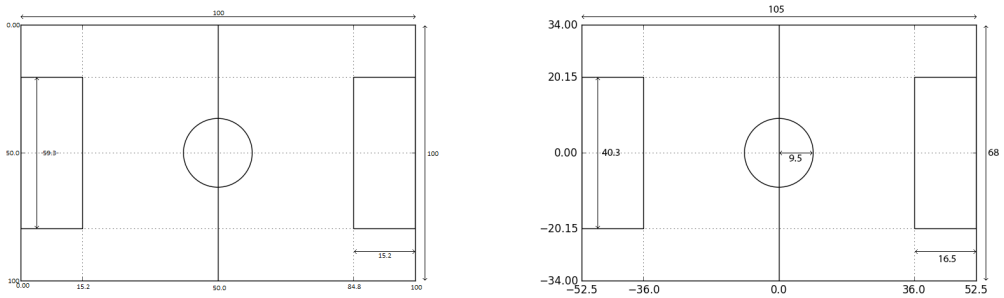
The timestamps of the tactical data and the spatiotemporal data are slightly different due to different starting moments of the match. Therefore, the timestamps have to be modified. By subtracting the start of the match, both data sources start the match at $t = 0$. By ensuring that both data sources start the matches as $t = 0$, timestamps could be matches immediately.

There is, however, one more difference between the tactical data and the spatiotemporal data. The tactical data restarts counting the second half, where the spatiotemporal data continues counting throughout the half time break. This can, however, be solved easily. By subtracting the start of the second half, the second half also starts at t=0. Since this could (potentially) lead to issues in data manipulations, later on, the choice is made to start the second half at sixty minutes after the start of the game. The correct time stamp could, therefore, be determined with equation 3.1

$$t_{new} = \begin{cases} t - t_{\text{Start}}, & \text{if first half} \\ t - t_{\text{Start second half}} + 60 * 60000, & \text{otherwise} \end{cases} \tag{3.1}$$

3.4.3 Spatial Data

Spatial data of the tactical and the spatiotemporal data do not have the same dimensions. First of all, the origin of the tactical data lies in the upper left corner of the field, where the origin of the spatiotemporal data lies in the middle of the field. Secondly, the units used by both data sources are different. The spatiotemporal data uses meters to measure the distance on the field. The tactical data, however, uses values between 0 and 100 to measure this distance. The effects of the use of different units are provided in Figure 3.3.



(a) Tactical data

(b) Spatiotemporal data

Figure 3.3: Dimensions of the fields of the tactical data and the spatiotemporal data

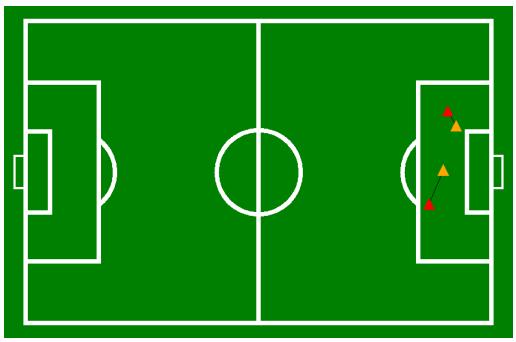
There is, however, one more difference in measuring the location of players on the field. In the tactical data, teams always play from left to right, in the spatiotemporal data, however, teams play as they played actually during the match (one-half from left to right and the other half from right to left). To overcome this issue, firstly, the goalkeeper (player with the highest absolute value) at the beginning of the match was selected. With the assumption that the goalkeeper is always on his own half, the direction of play could be determined from that team and thus the location could be adapted.

Sometimes, however, the goalkeeper does not have a location at the beginning of the match. Then, a striker could have the most absolute value and the wrong side of the field could be selected. To exclude these cases, the most absolute player for each half is selected. Since the goalkeeper is the most absolute player during a half, the direction of play can be determined and thus the

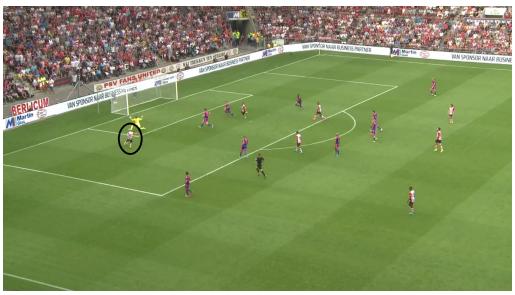
location of the players can be modified. Equation 3.2 provides the conditional equation for the first half. Here, $s_{i,t}$ is the location of player i at time t and n_i is the number of measurements for player i. When modifying the location of the players during the second half, the only difference is $t > 3300000$ instead of $t < 3300000$

$$s_{i,t} = \begin{cases} s_{i,t}, & \text{if} \min\left(\sum_{i=0}^{10} \frac{\sum_{t=0}^{t<3300000} s_{i,t}}{n_i}\right) < \max\left(\sum_{i=0}^{10} \frac{\sum_{t=0}^{t<3300000} s_{i,t}}{n_i}\right) \\ s_{i,t} * -1, & \text{otherwise} \end{cases} \tag{3.2}$$

After the locations of the players have been adjusted such that they always play from left to right, the difference of dimensions can be accounted for. The quality of the locations of the both data sources has to be determined in order to select one. Figure 3.4 shows the locations of the tactical data and the spatiotemporal data of three goal scoring opportunities in a match.

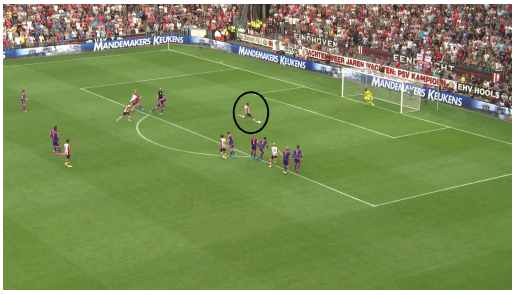


(a) The locations of which the goals are scored according to the tactical data (yellow) and spatiotemporal data (red)



(b) A video snapshot of the first goal

(c) A video snapshot of the second goal



(d) A video snapshot of the third goal

Figure 3.4: Comparison of the spatial data from ORTEC with the Inmotio data

As can be seen from the data, the locations are quite similar. The main difference in location comes from moment c where the spatiotemporal data seems to be better. The spatiotemporal data, however, has issues with tracking players which could lead to missing data or incorrect values that are way off (Section 3.2.4). Furthermore, more data is available from the tactical data source. Therefore, the locations of the players are extracted from the tactical data source.

An important note to make here is that the tactical data does not provide data about surrounding players and that therefore, data from surrounding players is still extracted from the spatiotemporal data. The location of the player attempting to score and his surrounding players, therefore, comes from different data sources. This could lead to small differences when comparing these locations. Figure 3.4, however, already showed that these locations are very similar.

3.5 Conclusion

In this chapter, the three different data sources are introduced. With the introduction of the data sources, the data collection methods, the general format, and (potential) data quality issues are discussed. Finally, this chapter elaborated on the methods applied to combine the data sources. In order to combine the data sources, adjustments in player names, timestamp and location had to be made.

Chapter 4

Modeling

This chapter elaborates on the modeling steps which have to be taken to actually obtain the predictive model. First of all, features are extracted from the three data sources. These features are then prepared to ensure that the model is trained properly. Furthermore, the class imbalance is discussed and a solution is proposed.

4.1 Feature Extraction

The raw data discussed in Chapter 3 is used to determine some key features which are used as input variables for the model. The features which are collected from the data are divided into features based solely on ORTEC data and features based on ORTEC data as well as Inmotio data. This distinction is made since the ORTEC data includes more matches as the Inmotio data does.

4.1.1 Tactical Features

The features that can be extracted from the tactical data can be divided into multiple categories. First of all, features regarding the physical location of the player can be determined. Secondly, the context of the goal scoring opportunity influences the quality of the goal attempt. Then, the action previous to the goal attempt is determined. The current score could influence the mental state of the players and could, therefore, influence the quality of the goal scoring opportunity. Finally, some additional features are extracted.

Spatial Features

One of the most important aspects of the ORTEC data is the location on the field. The location of the field from which an attacker is trying to score seems to be an important variable in determining the chance of a goal from a given goal scoring opportunity. There are various ways in which the location of the field can be included in a model. Lucey et al. for example determine a probability distribution of shot locations resulting in a goal [36]. During this thesis, however, a different approach is used. From the (x, y) coordinates, two features are determined: (1) The distance to the goal and (2) the angle to the goal. Three different cases can be selected in order to calculate these features. The attacker can be standing on the left-hand side of the goal (2) right in front of the goal or (3) on the right-hand side of the goal. The formula of all these cases are presented in Table 4.1. A geographical representation of the three scenarios is also given in Figure 4.1

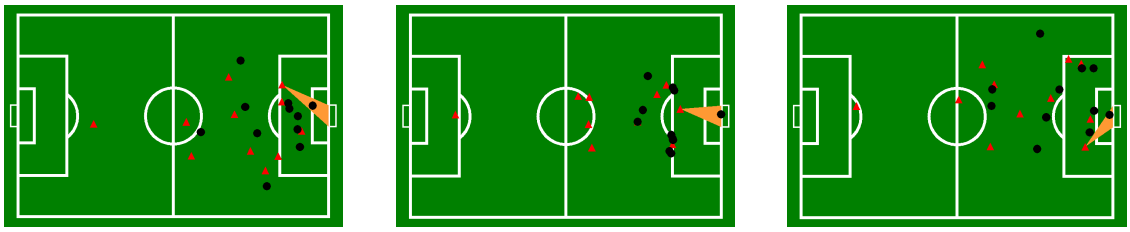
Where angle1 $= \frac{\tan^{-1}(\mathrm{GW}/2+y)}{\Delta x}$ and $\Delta x = \frac{\mathrm{FW}}{2} - x$

Context

Besides the location of the goal attempt, the ORTEC data also includes information about the context of a goal attempt. Like Lucey et al. six different contexts are used: Open-play, Counter,

Table 4.1: Formula to calculate the distance to the goal and the angle to the goal

Relative position	Distance to the goal	Angle to the goal
Left side	$\sqrt{(\frac{\text{FW}}{2} - x)^2 - (\frac{\text{GW}}{2} - y)^2}$	$\text{angle1} - \tan^{-1} \frac{\|y\| - \text{GW}/2}{\Delta x}$
In front	$\frac{\text{FW}}{2} - x$	$\text{angle1} - \tan^{-1} \frac{\text{GW}/2 - \|y\|}{\Delta x}$
Right side	$\sqrt{(\frac{\text{FW}}{2} - x)^2 - (\frac{-\text{GW}}{2} - y)^2}$	$\text{angle1} - \tan^{-1} \frac{\|y\| - \text{GW}/2}{\Delta x}$



(a) Scenario 1: Left side (b) Scenario 2: In front (c) Scenario 3: Right side

Figure 4.1: Different scenarios for which the distance to the goal and the angle to the goal have to be computed

Corner, Penalty, Direct free kick, and Indirect free kick.

To determine the context of a goal scoring opportunity, the contexts have to be defined. The definitions of the different contexts are given below.

- **Open play:** Goal attempt created from a series of passes on the opposition's half of the field.
- **Counter:** Fast break-away after gaining possession on own half.
- **Corner:** Set piece from the corner of the field when the ball passed the goal line, is not counted as a goal and is last touched by a defender
- **Penalty:** Set piece after a foul of the opposition in own box. A penalty is always taken 11 meters from the goal
- **Direct free kick:** Set piece after a foul of the opposition outside the own box which is taken directly on goal.
- **Indirect free kick:** Set piece after a foul of the opposition outside the own box which is not taken directly on goal.

Now the different contexts are defined, these contexts can be extracted from the data. For both penalty and direct free kicks, this is very straightforward since they are both directly encoded by ORTEC. For the other contexts, however, some more work has to be done. In case the context of the goal scoring opportunity is not one of those, we have to look at previous events from the data. Here, we look at the last fifteen seconds before the event. It can, however, occur that there have not been any events during the last fifteen seconds e.g. there was an injury treatment. In this case, the last three events are used.

For the previous events, the first of the following is used:

- If the event was a corner: context $\rightarrow$ Corner
- If the event was an indirect free kick: context $\rightarrow$ Indirect free kick

- If the event was possession gain on the first 15 meters of the opposition's half: context → Counter

It is, however, very difficult to define open play situations. Therefore, the context is encoded as open-play when none of the previous events occurs in the selected previous events.

Origin of Goal Scoring Opportunity

For some contexts, the context itself might not be accurate enough. Therefore, additional information about the origin of a chance is created from the data. A distinction is made between a dribble, rebound, cross pass, long pass, and possession gain.

- **Dribble:** The attacker dribbled and thus took on a defender just before shooting.
- **Rebound:** The attacker picks up the ball after it reflects from the bar, is blocked by a defender, or saved by the goalkeeper.
- **Cross pass:** The attacker attempts to score from a cross given from the side of the field.
- **Long pass:** A scoring opportunity created after a pass longer than 30 meters
- **Possession gain:** The attacker gains possession by taking the ball from a defender or the goalkeeper.

To extract these features from the ORTEC data, the last 10 events before the goal scoring opportunity are investigated. The first from the above events which occurred before the goal scoring opportunity is selected and the corresponding feature is extracted.

Current Score

Another feature which can be extracted from the ORTEC data is the current score line. The current score line could eventually represent the defensiveness of the oppositional team. The effect of this feature is best explained by some scenarios. Teams, who are behind, for example, are more likely to attack which means that more of their players are attacking. This automatically results in fewer players defending their own goal. Which leads to fewer defenders and thus more space for the opposition. Furthermore, more attackers leads to more passing options and thus more choices for both attacker and defender to choose from which could affect the quality of goal scoring opportunities.

Degree of Difficulty

Even when all the above features are included in the model, the quality of goal scoring opportunities can still not be defined properly. Lets take, for example, a cross from open-play where the team of the attacker is 1-0 behind. Intuitively, the quality of the goal scoring opportunity is not the same if the ball is above the ground compared to the same situation in which the ball is on the ground. Since there is no z value of the ball of the moment on which the attacker shoots, the best we can do is to select the high attribute as encoded by ORTEC. This results in a categorical variable where the ball is high or not.

Furthermore, the quality of the same goal scoring opportunity depends on with which part of the body the attacker attempts to score. If the attacker, for example, attempts to score with his head aiming the ball is more difficult than when the attacker attempts to score with his foot. Besides the increase in accuracy, attempting to score with the foot also results in higher power and thus less time for a goalkeeper to respond to the attempt.

4.1.2 Player Quality Features

As discussed in Section 2.5, two features from the player quality data are included for the model. These features are the quality of the player attempting to score and the quality of the goalkeeper. The player quality data contains several attributes which could be of importance for both players. These attributes are provided in Table 4.2.

Table 4.2: Selected attributes of the player and the Goal Keeper from the player data

Player attributes	Goal Keeper attributes
Penalties	GK Diving
Free Kick Accuracy	GK Handling
Heading Accuracy	GK Positioning
Long Shots	GK Reflexes
Finishing	

Table 4.2 shows different player attributes of the players from the Player data. Not all the attributes of player are, however, useful in a particular situation. Therefore, different scenarios are created for each of the player attributes. An overview of the player attributes used as a feature for the different scenarios is provided in Table 4.3.

Table 4.3: Mapping of the player attributes to given situations

Scenario	Player attributes
Penalty	Penalties
Direct Free Kick	Free Kick Accuracy
Header	Heading Accuracy
Long shot (Distance $> 16m$)	Long Shots
Other situations	Finishing

For the Goal Keeper, it is, however, difficult to determine the importance of the attributes in a given situation. At the time of the shot, for example, the location of the ball relative to the goalkeeper when passing the goalkeeper is not known yet. Therefore, it is impossible to decide whether the goalkeeper should dive or not (and thus if diving is important in this situation). Therefore the mean of the selected Goal Keeper attributes is used as a feature of the goalkeeper.

4.1.3 Spatiotemporal Features

As seen in the last Section 4.1.1 and Section 4.1.2, many features can be extracted which influence the quality of a scoring opportunity. Even more features can, however, be extracted when the spatiotemporal data is included as well since the spatiotemporal data includes data about other players on the field where this data is not available in the tactical data. By matching the goal scoring opportunity as discussed in Chapter 3, the locations of all the players at the time of the goal scoring opportunity are known. From the positions of the players, some more features can be generated.

Not all players are, however, participating in the play at the time of the goal scoring opportunity. Therefore, only the relevant have to be selected from the data. For this thesis, the relevant players are the players who are standing in the area between the attacker and the two goal posts. These players are determined by in polygon calculations where the players have a defendable radius of $1.3m$.

Intuitively, it is harder to score a goal when the area between the attacker and the goal is denser. Thus, it is more difficult to score when there are more players in line of the goal and the attacker. Not all the players standing between the attacker and the goal are, however, the same. Where defenders try to block the ball, attackers would probably try to avoid the ball. This is a major difference when extracting different features. Furthermore, the goalkeeper, who also tries

to save (block) ball, defends a larger area than the defender. Therefore, the number of attackers and the number of defenders between the goal and the attacker are extracted as features.

Whether or not the goalkeeper is standing between the attacker and the goal could be included as a categorical variable as well. Not all the situations in which a goalkeeper is between the attacker and the goal are, however, the same. In situations where the goalkeeper is very close to the attacker, it is very difficult to shoot the ball passed the goalkeeper. On the other hand, when the goalkeeper is standing far from the attacker, the goalkeeper has more time to react to the goal attempt and is thus more likely to stop the goal attempt. Since the distance of the goalkeeper to the attacker seems to play an important role, the choice is made to include the Euclidean distance of the goalkeeper to the attacker as a feature instead of whether or not the goalkeeper is in line with the attacker and the goal. A special case exists, however, when the goalkeeper is not in line with the goal and the attacker. In this case, the maximum distance of the goalkeeper to the goal from the whole data is used as this distance.

The same logic from the goalkeeper can be applied to the case of the defender. On one side, it is harder to shoot passed the defender when the defender is standing close, on the other side, the defender has more time to respond to the goal attempt when he is standing further away. Therefore, besides the number of defenders standing in line of the goal, the Euclidean distance of the defender to the attacker is also extracted as a feature.

The features derived from the spatiotemporal data are visualized in Figure 4.2.

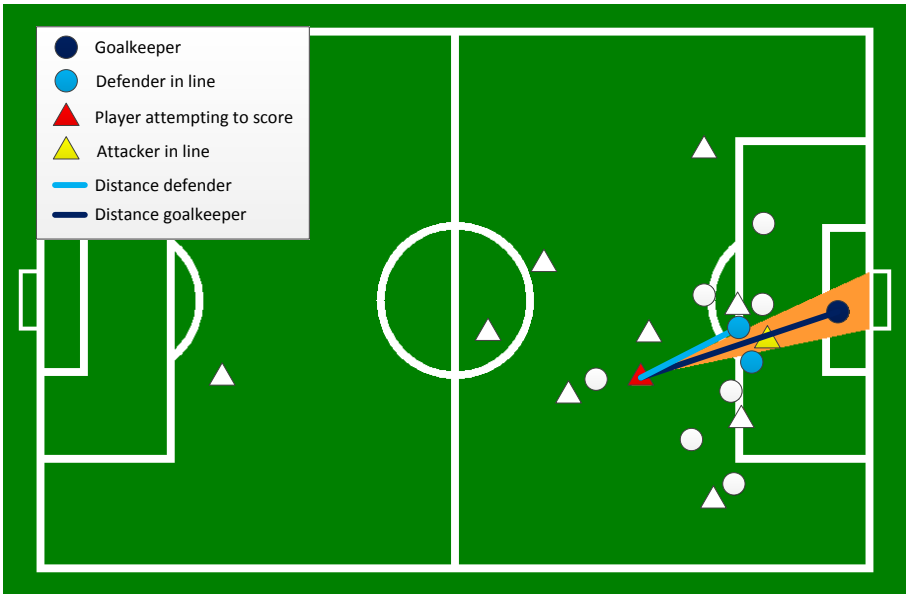


Figure 4.2: Visualization of the features derived from the spatiotemporal data.

4.2 Data Preparation

For numerical values, the classification algorithms implemented by scikit-learn work best within the range $[0, 1]$ [42]. Therefore, min-max normalization is used to transform the features in such a way that they lie within this range. The min-max normalization is implemented by scikit-learn [42]. The formula for min-max normalization for a range of values of a feature (X) is given in Equation 4.1, where X_{min} is the minimum value of the feature, X_{max} is the maximum value of the feature and min and max are respectively the minimum and maximum value of the desired range.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} * (max - min) + min \tag{4.1}$$

Not all features are, however, within order. Take for example the feature context. Here, it is difficult to say whether a corner is necessarily better than a free kick. These variables are so-called categorical variables. These variables are set to dummy variables such they can be included in the classification algorithms. This would, for example, lead to the case where a variable says whether it is a corner (1) or not (0) and another variable saying whether it is a free kick (1) or not (0). An overview of the categorical variables and the numerical variables is provided in Table 4.4.

Table 4.4: Distinction between numerical and categorical variables

	Numerical	Categorical
Dist to goal	Number of attackers in line	Part of body
Angle to goal	Number of defenders in line	Originates from
Player quality	Distance nearest defender in line	Current score line
Goal keeper quality	Distance goal keeper	Context
		High

4.3 Class Imbalance

Since goals are such rare events in soccer, there is a large imbalance of goals versus non-goals in the data. Since this imbalance would largely influence the performance of the model, regularization has to be applied to the data in order to get proper performance. Various techniques exist to overcome this issue. The major approaches are listed below [39].

- Balancing
- Generating synthetic instances
- Ensemble learners
- AUC ranking/ cost-sensitive classification
- One-class classification setup

For this thesis, a combination of balancing (under-sampling the majority class) and generating synthetic instances is used. Chawla et al. show that this is a good method to deal with imbalanced data [10]. Both methods are briefly discussed below.

4.3.1 Generating Synthetic Instances

Generating synthetic instances is different from previous over-sampling techniques since it does not simply duplicate samples from the minority class but generates synthetic class samples by interpolating between existing instances. One of the existing algorithms to generate synthetic instances is the SMOTE algorithms [10].

The minority class is over-sampled by taking the minority class samples and by creating synthetic samples for each of them. Depending on the amount of over-sampling required, neighbors of the k Nearest Neighbours are randomly chosen. Interpolating on all the attributes of these neighbors leads to new synthetic samples.

4.3.2 Under-Sampling the Majority Class

In normal under-sampling, random samples from the majority class are removed from the data set. Yen and Lee show, however, that a clustering-based under-sampling method leads to better performance of the eventual training data [52].

Their method uses k-Means Clustering to cluster the samples together. For each of these clusters, a number of samples are removed from the data set to get the desired number of samples in the actual data set. By applying this method, the training data represents the original data better.

4.4 Conclusion

In this chapter, the features extracted from the different data sources are provided. An overview of these features is provided in Table 4.5. Furthermore, the preparation of these features is discussed. Here, the distinction was made between the preparation for numerical features using min-max normalization and the casting of categorical features to dummy variables. Finally, a method of solving the class-imbalance problem is proposed where a combination of under and over-sampling is used.

Table 4.5: Features for the data sources

ORTEC	FIFA	Inmotio
Context	Player quality	Number of attackers in line
Part of body	Goal keeper quality	Number of defenders in line
Dist to goal		Distance nearest defender in line
Angle to goal		Distance goal keeper
Originates from		
Current score		
High		

Chapter 5

Evaluation

This chapter elaborates on the performance evaluation of the models provided in Chapter 4. Multiple techniques are used in this chapter to evaluate the performance of the predictive model. First of all, the performance metrics for predictive modeling are discussed in Section 5.1. Secondly, the performance of the calibration step is evaluated in Section 5.2.

The true business value of the models is, however, generated by determining the likelihood of a goal scoring opportunity. The generic performance metrics can only determine the performance of the model based on the data itself. Therefore, to further evaluate the performance of the model, the performance of the model is determined by conducting an eye test with a business expert.

The goal of this thesis is to explain match results based on expected goals. Therefore, the relation to the expected goals and the match outcomes is evaluated in Section 5.4. Finally, the chapter ends with a conclusion.

5.1 Performance Metrics

The performance of the classification algorithms shows to what extent the algorithm was able to differentiate the goal attempts resulting in goals from goal attempts which did not result in goals. Therefore, this section shows to what extent the classification algorithm was able to rank situations according to their quality.

5.1.1 Cross Validation

To properly test the predictive model, the model should be trained on previously unseen data. To lower the variability of the performance metrics, this method is repeatedly executed. Averaging these performance scores leads to the eventual performance of the model. This model validation technique is called cross-validation.

For this thesis, Stratified 10-Fold cross-validation is used. With Stratified 10-Fold cross-validation, the data is split into ten different folds. Since Stratified cross validation is used, the proportion of classes is similar over the different folds. Each of the ten folds is the unseen data (test data) exactly once. The other nine folds are, at that time, the data on which the predictive model is trained.

Since the classification algorithms have a predefined parameter setting, inner k-fold cross validation is used to perform parameter selection. This means that the training data is split into ten folds. Each of these folds is then the validation set once, while the predictive model is trained on the other folds. Based on the scores of the validation sets, an optimal parameter setting is selected. The process of the cross-validation step is provided in Figure 5.1. The parameters evaluated for the different classifiers are provided in Table 5.1.

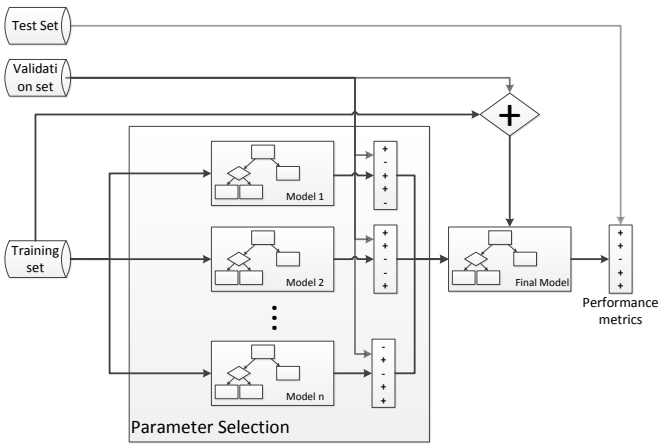


Figure 5.1: Visualization of the model (parameter) selection and estimation of generalization performance of the selected (final) model

Table 5.1: Parameter settings for the different classifiers

Classifier	Parameter	Setting 1	Setting 2	Setting 3	Setting 4
Logistic Regression	C	$1e^3$	$5e^3$	$1e^4$	$5e^4$
	Solver	Liblinear	Sag	Newton-cg	Libfgs
Random Forrest	Max Depth	None	3	5	
	Bootstrap	True	False		
	Criterion	Gini	Entropy		
	# Estimators	5	10	20	30
Decision Tree	Criterion	Gini	Entropy		
	Max Depth	None	3	5	
Ada-boost	Learning Rate	0.5	1	1.5	
	# Estimators	5	10	20	30

5.1.2 Performance Metrics before Calibration

As mentioned in Section 2.2, the AUC is used to select the best predictive model. The AUC is, therefore, also used to optimize parameter settings as explained in Section 5.1.1. Besides the AUC, other metrics could, however, show more insights in the performance of the model. The Performance metrics of the different classification algorithms are provided in Table 5.2.

Table 5.2: Performance metrics of the model.

Model	Splitted	Precision	Recall	F-score	AUC
Random Forrest	no	0.703	0.812	0.745	0.735
Decision Tree	no	0.705	0.680	0.687	0.707
Logistic Regression	yes	0.745	0.589	0.632	0.707
Logistic Regression	no	0.756	0.678	0.708	0.728
ADA-boost	no	0.626	0.773	0.691	0.682

Table 5.2 shows that the Random Forest classifier outperforms, or at least equals, the other classification algorithms on all performance metrics. Furthermore, the table shows that splitting the data for different context does not lead to a superior classifier.

5.1.3 Performance Metrics after Calibration

After calibrating the classifier, the classifier returns different scores. These scores, obviously, lead to different performance metrics. These performance metrics are provided in Table 5.3.

Table 5.3: Performance metrics of the model.

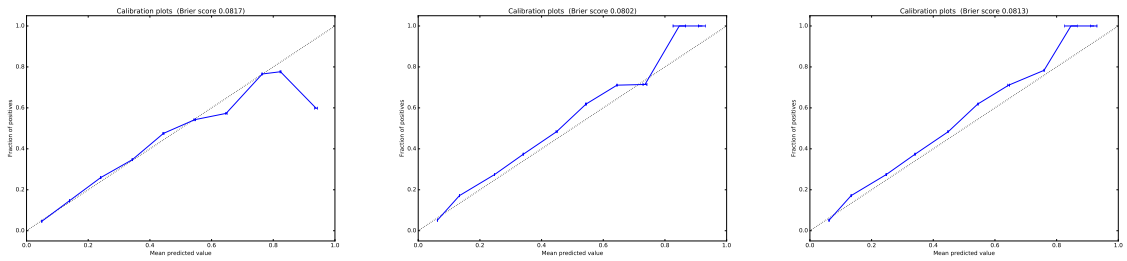
Model	Splitted	Precision	Recall	F-score	AUC
Random Forrest	no	0.771	0.104	0.187	0.775
Decision Tree	no	0.661	0.169	0.269	0.777
Logistic Regression	no	0.696	0.155	0.254	0.785
ADA-boost	no	0.526	0.013	0.022	0.692

Table 5.3 shows that, after calibration, the performance metrics lower. Especially the recall and F-score are significantly lower than before calibration. The lowering of the recall can be explained by the definition recall. The lowering of the F-score then results from the lower recall.

In Section 2.2, recall was already defined as the number of correctly predicted goals divided by the number of actual goals. Since the scores resulting from the calibrated classifier are much lower, less goal scoring opportunities are predicted to result in a goal. This also corresponds to the real life case, where there are only a few goal scoring opportunities which result in goals more than half of the time. Intuitively, it is, therefore, correct that the recall is low.

5.2 Reliability Graph

To ensure that the class membership probabilities can be interpreted as posterior probabilities, the calibration step is performed. The successfulness of calibration can be determined with the use of a reliability graph. Three different reliability graphs are presented in Figure 5.2. Firstly, a reliability graph is provided with all the data. As can be seen in Figure 5.2, the reliability graph of the complete data drops at the end below the optimal line. Since the goal attempts which have high scores are mainly penalties, the effect of penalties on the reliability graph is determined. Therefore, Figure 5.2 also contains the reliability graphs in case the penalties are left out or when penalties are set to a constant value. Research already showed that scoring a penalty (excluding the rebound since this leads to a second goal attempt) typically has a probability to result in a goal of 0.76 [38].



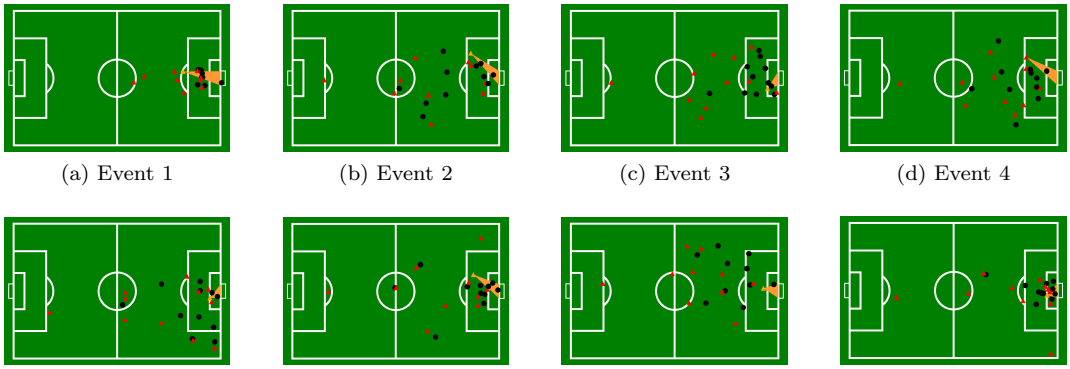
(a) Original (b) Without penalties (c) Corrected penalties

Figure 5.2: Calibration plots.

As can be seen from Figure 5.2 the reliability graphs all follow the optimal line closely. This indicates that the predicted values follow the actual values. Therefore, the calibrated class membership scores can be interpreted as probabilities.

5.3 Eye Test

It is important to have a predictive model with reasonable performance. In order to be base decisions based on the expected goal attempts, the expected goals for individual goal scoring opportunities should seem reasonable as well. Therefore, this section provides some examples of goal attempts with their corresponding expected goals. These examples are provided in Figure 5.3. The corresponding features and expected goal values are provided in Table 5.4 and Table 5.5.



(a) Event 1 (b) Event 2 (c) Event 3 (d) Event 4

(e) Event 5 (f) Event 6 (g) Event 7 (h) Event 8

Figure 5.3: Examples of goal scoring opportunities

Table 5.4: Features and posterior probabilities for the events 1-4

Feature	Event 1	Event 2	Event 3	Event 4
Context	Free-Kick	Counter	Open-play	Counter
Part of body	Foot	Foot	Foot	Head
Dist to goal	20.314	17.078	5.842	6.824
Angle to goal	0.368	0.289	0.726	1.075
Originates from	Pass	Pass	Rebound	Pass
Current score	In front	In front	Draw	Behind
Player quality	54	54	77	92
GK quality	61	77		77
Goal	Yes	Yes	Yes	Yes
Probability	0.099	0.074	0.445	0.310
95% Confidence Interval	[0, 0.227]	[0, 0.287]	[0.094, 0.796]	[0.188, 0.432]

Table 5.5: Features and posterior probabilities for the events 5-8

Feature	Event 5	Event 6	Event 7	Event 8
Context	Indirect Free-Kick	Corner	Counter	Corner
Part of body	Foot	Foot	Foot	Head
Dist to goal	5.531	14.040	9.216	5.605
Angle to goal	0.961	0.394	0.869	1.075
Originates from	Long Pass	Dribble	Pass	Long Pass
Current score	In front	Draw	In front	Draw
Player quality	75	77	77	68
GK quality	77	66.25	61	77
Goal	No	No	No	No
Probability	0.344	0.070	0.418	0.230
95% Confidence Interval	[0.216, 0.472]	[0, 0.382]	[0.205, 0.631]	[0.128, 0.332]

5.4 Match Outcomes

The goal of this thesis is to explain match results based on the analysis of expected goals. Therefore, an important business question to solve is if the expected goals are indeed representative of the match results. Therefore, Table 5.6 provides the number of matches that are correctly predicted with the use of the expected goals model. Furthermore, the predicted results were at most one goal off is shown, the number of matches in which the result was correct (win team 1, draw, win team 2). Finally, the Mean Squared Error (MSE) for the number of goals for a team per match is provided and the MSE for the number of goals per match is provided.

Table 5.6: Evaluation of match outcomes according to the expected goals model

League	Season	#Matches	#Correct	#1 Goal	#Result	MSE Match
1.Bundesliga	2015-2016	2	1	1	1	1.655
Premier League	2015-2016	1	0	0	0	2.289
Champions League	2015-2016	133	37	96	74	1.956
Eredivisie	2013-2014	322	86	221	175	2.556
	2014-2015	322	81	222	177	2.395
	2015-2016	322	84	222	178	2.284
Jupiler League	2013-2014	380	95	239	202	2.902
	2014-2015	379	94	255	215	2.517
	2015-2016	342	88	208	185	2.762
KNVB Beker	2015-2016	26	6	17	16	2.601
Ligue 1	2013-2014	380	105	257	183	2.106
	2014-2015	380	106	289	189	2.172
	2015-2016	380	101	272	198	2.153
Primeira Liga	2014-2015	280	92	202	165	1.948
	2015-2016	231	58	150	123	2.426
Primera Division	2013-2014	380	118	271	220	2.471
	2014-2015	380	112	254	207	2.153
	2015-2016	380	102	267	201	2.385
Total		5020	1366	3443	2709	2.366

What stands out from Table 5.6 is that in only 1366 of the 5020 matches, the exact score of the match was predicted based on the expected goals. If, however, one goal difference is accepted, 3443 of the 5020 matches have correctly predicted scores. Therefore, it seems that the expected goals model is, in most cases, almost correct. The MSE Match strengthens this statement. The MSE match shows that the average MSE of the result of a match is 2.366. Therefore, the average number

of goals predicted difference goals of both teams differs $\sqrt{2.366} \approx 1.538$ from the actual difference in goals, which means that many matches would be in the range of only one goal difference.

Furthermore, Table 5.6 shows that the results of match outcomes are not very different across leagues or seasons.The only exceptions are the 1.Bundesliga and the Barclays Premier League, where too few matches were evaluated. In the other cases, about 1 out of four matches the actual score is predicted correctly, and the number of matches in which the score was at most one goal off is about 2.5 times as high. Furthermore, the MSE values of the number of goals scored per team and the difference in goals scored do not differ much from the mean. Therefore, one could conclude that the expected goals model is generalizable across different leagues and different seasons. In the further analysis, results of different leagues and seasons could, therefore, be aggregated.

So far, just the exact results are examined. More interestingly, maybe, is how often the expected goals model predicted the correct winner. This is given by the number of correct results in Table 5.6. Obviously, the number of correctly predicted matches is higher than the correctly predicted scores. What stands out, however, that the number of correctly predicted matches is not close to the number of scores predicted correctly where one goal difference was allowed. This shows that games where the model is one goal off in the match, this one goal also influences the result of the match. To evaluate in which cases the one goal difference most often influences the result, the problem of predicting the winner of a match is defined as a three class problem where either Team 1 wins, Team 2 wins or the game ends in a draw. The confusion matrix of the three-class problem is provided in Table 5.7.

Table 5.7: Confusion matrix of the three class problem of predicting the winner of a match

		Actual			
		Win 1	Draw	Win 2	Total
Predicted	Win 1	1079	329	210	1618
	Draw	599	524	591	1714
	Win 2	220	362	1106	1688
Total		1898	1215	1907	5020

Table 5.7 shows that most of the incorrect classified match outcomes originate from predicted draws. Predicted draws are, most likely, games which were very tight. Table 5.6 already showed that in many cases, the model was only one goal off. In tight games, one goal off means that the result of the match is predicted incorrectly. This was most likely the case of the predicted draws. To dive even further in the predicted match outcomes, the most common actual match results and predicted match results are provided in Table 5.8.

Table 5.8: Most common results and the predicted result

		Actual								
		2-0	2-1	3-0	2-2	0-0	1-1	1-0	3-1	Total
Predicted	1-1	193	222	68	61	170	216	374	62	1469
	2-1	182	201	121	53	41	86	159	131	1189
	1-2	37	85	10	44	26	97	85	33	476
	3-1	48	46	42	16	7	20	15	42	382
	2-2	21	50	9	27	3	29	24	31	278
	2-0	71	16	36	0	11	20	38	23	259
	1-0	36	19	12	1	47	13	75	6	221
	3-0	21	9	20	0	1	4	13	11	126
Total		630	705	347	246	363	553	840	372	5020

Table 5.7 already showed that the predicted goals model is most often incorrect when the model predicts a draw. Therefore, it is most interesting to look at the predicted draws in Table 5.8. Lets take for example the predicted score of 1-1. What stands out that in 596 of the 1469 cases,

the actual result was a close win with at most one goal off (374 cases of 1-0 and 222 cases of 2-1). Furthermore, in 447 of the 1469 cases, the result was correct, which is a similar ratio as in Table 5.7. Of these 447 cases, however, the model predicted another score in 231 of the cases. This could indicate that the expected goals at which the expected goal model assigns a goal to a team has to be tweaked. The value at which the model assigns a goal to a team is called the threshold. The result of the threshold value on the predicted results is provided in Figure 5.4.

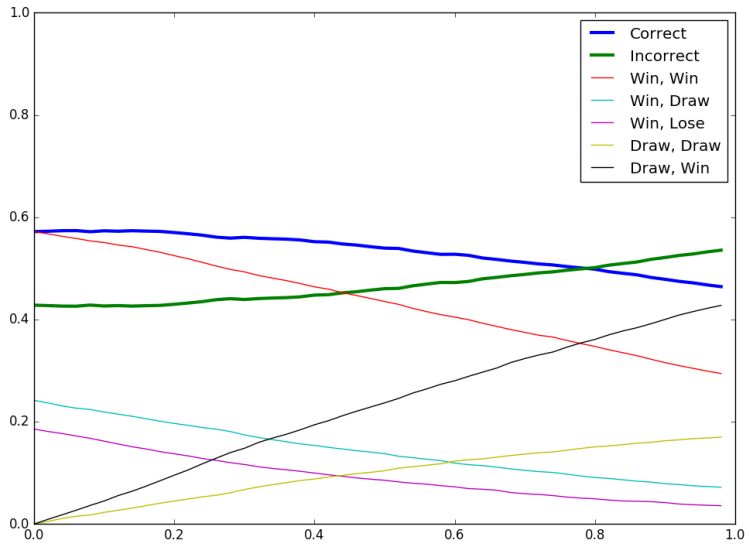


Figure 5.4: Influence of the threshold on the obtained results

Figure 5.4 shows that the highest number of correctly predicted match results is obtained with a threshold of around 0.2. When doing so, the number of correctly predicted wins is very high. The correctly predicted draws, however, are very low. Since the correctly predicted number of draws was low already the choice is made to stick with the threshold value at 0.5.

5.5 Conclusion

In this chapter, the performance of the predictive model is discussed. At first, the performance metrics of the different classification algorithms were evaluated. The random forest classifier is clearly outperforming the other classification algorithms. Then, the performance of the calibration is determined. Reliability graphs showed that the predicted scores follow the optimal line closely. Therefore, the conclusion could be drawn that the calibrated class membership probabilities could be interpreted as actual probabilities.

To determine to which extent the proposed expected goals model fits the business problem, the scores for individual goal attempts were evaluated. Here, it was also seen that the confidence intervals are very wide. Therefore, no valid conclusions on a single goal attempt can be made. Finally, the match outcomes as predicted by the expected goals model are determined and evaluated to the actual results of matches. The expected goals model is limited in predicting the match result, especially in tight games.

Chapter 6

Case Study: FC Barcelona

The main goal of this thesis is to evaluate matches over a period of time to base strategic decisions on. This is, however, only a part of the possible applications of the expected goals as introduced in this thesis. The different applications for soccer clubs can be classified into three different categories: analysis of a period of time, a single match, and a particular player. These three different categories are discussed in this chapter.

Due to confidentiality, data of PSV cannot be shared publicly. Therefore, the analyses in this chapter are performed on different soccer clubs. More specifically, the soccer club FC Barcelona is used for this case study. The visualizations, however, are similar to the case of PSV.

6.1 Season Analysis

Strategic decisions in soccer clubs are often based on subjective reasoning over objective reasoning. By evaluating the results over a given period of time, the strategic decisions could be based on an objective evaluation. To evaluate periods of time, Expected Goals is used to evaluate these periods of time in three different ways. First of all, the expected goals scored and conceded are given of all the matches in the selected period. Secondly, The player performance could be evaluated over the same period of time. Finally, the effectiveness during one match could be evaluated over a period of time. By evaluating match effectiveness during a match over a period of time, weak periods of time could be analyzed in more detail.

6.1.1 Expected Goals during a Season

First of all, periods of time (in this case a season), can be analyzed by soccer clubs. Analyzing the performance during a season allows clubs to analyze the "objective" performance of their club. Decisions such as a change of tactic, different players, or even firing of a coach can then be made based on this more objective performance. Figure 6.1 provides an overview of the expected performance over a season. The bars in this plot represent the expected goals scored (red) and conceded (blue) of the club of choice. The winner of a match can be determined by looking at the white bar, which represents the expected difference in goals of the match. Furthermore, the actual number of goals scored and conceded are given as a number in the histogram. This allows the analyst to easily compare the expected result of a match to the actual result of a match.

The purpose of this figure is to be able to analyze weak periods of time during a season. Since individual matches are almost undependable of each other (excluding injuries and suspensions) the matches are displayed as discrete events. Bar charts are able to visualize these discrete events over a given period of time [16]. To provide even more information in the bar chart, some of the pre-attentive attributes are used [17]. Pre-attentive attributes are attributes in a visualization that can easily be analyzed by the human brain.

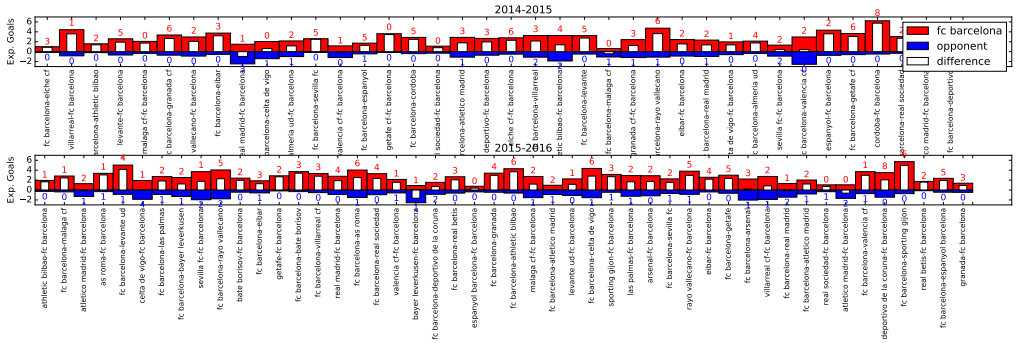


Figure 6.1: Expected goals by and conceded by FC Barcelona

For this visualization, the pre-attentive attribute color is used to make a clear distinction between the selected team, the opposing team and the difference between these different information types.

Another visualization of the performance over time is in the form of a league table. The league table is often used to present the ranking of teams during the league. This league table, however, can also be created with expected match outcomes. The expected league table could then be compared to the actual league table to check the expected performance over time versus the actual performance. The league table, however, does not provide insights in which games are expected to be lost. Figure 6.1 could provide insights in individual matches. The expected league table of the Primera Division at the end of season 2015/2016 is provided in Table 6.1.

Table 6.1: Expected league table Primera Division 2015/2016

Rank	Team	Matches	Wins	Draws	Lost	Points	GS	GA	GD
1	fc barcelona	38	33	5	0	104	104	33	71
2	real madrid	38	30	4	4	94	89	45	44
3	atletico madrid	38	27	6	5	87	60	31	29
4	sevilla fc	38	16	12	10	60	66	53	13
5	celta de vigo	38	14	13	11	55	54	55	-1
6	malaga cf	38	12	17	9	53	48	47	1
7	athletic bilbao	38	12	16	10	52	51	47	4
8	eibar	38	14	10	14	52	52	53	-1
9	real sociedad	38	11	16	11	49	47	49	-2
10	rayo vallecano	38	11	16	11	49	58	61	-3
11	espanyol barcelona	38	11	13	14	46	46	57	-11
12	villarreal cf	38	9	14	15	41	40	52	-12
13	las palmas	38	11	7	20	40	46	57	-11
14	deportivo de la coruna	38	8	14	16	38	41	53	-12
15	valencia cf	38	9	10	18	37	47	62	-15
16	getafe	38	9	10	19	37	45	62	-17
17	real betis	38	6	16	16	34	41	56	-15
18	sporting gijon	38	6	14	18	32	43	66	-23
19	granada	38	5	16	17	31	42	58	-16
20	levante ud	38	6	11	21	29	41	64	-23

6.1.2 Player Performance during a Season

Players who score often in a season often get a lot of attention of better clubs. These players, however, are very expensive and often prove to be not good enough for these higher ranked teams.

The expected number of goals of the players in a league can be used to evaluate players who did not have an exceptional season in terms of actual goals but did have a good season in terms of exceptional goals. Identifying these players could provide a good way of determining good players who are relatively cheap compared to the more popular players. Figure 6.2 shows a way of determining these players.

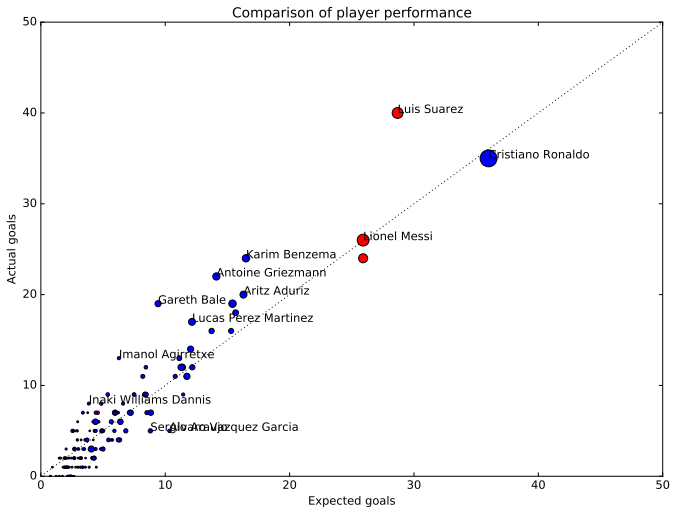


Figure 6.2: Expected goals by and conceded by FC Barcelona

The Expected Goals over a season are here plotted with the Actual Goals scored during that same season. In order to easily see the difference between the distributions of the Expected Goals and the Actual Goals, a Q-Q plot is used [28]. As one would expect, the players roughly follow the diagonal, which means that the Expected Goals of the players come close to the actual goals. In this plot, some more visualization technique is used. The size of the circles (one of the pre-attentive attributes [17]) is used to display the number of goal scoring opportunities that the players needed in order to receive the Expected Goals and the Actual Goals. Furthermore, the color red is used to select the players of a particular team (in this case FC Barcelona).

6.1.3 Team Effectiveness during a Match

Soccer clubs are also interested in the differences in performance during time periods in a match. Therefore, a time series of the Expected Goals and the Actual Goals is plotted of the matches in a season are plotted. Since there is too much information for the human brain to easily process this information, the average goals and expected goals are also plotted for the same time series.

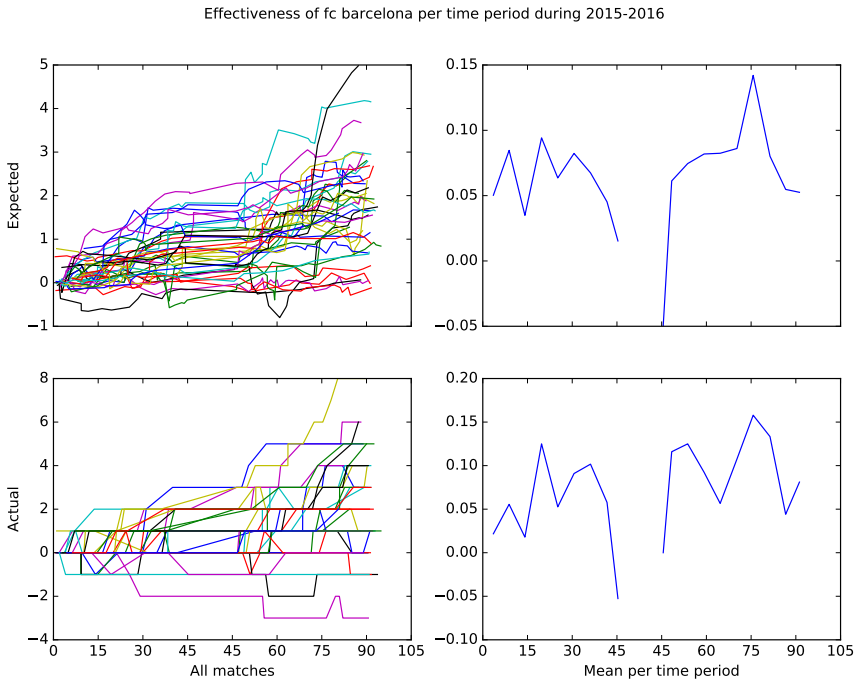


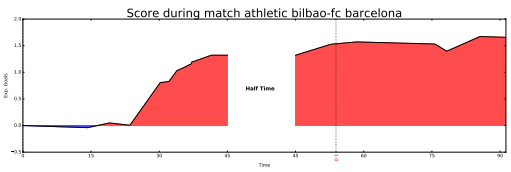
Figure 6.3: Evaluation of effectiveness over time periods during a particular season

6.2 Match Analysis

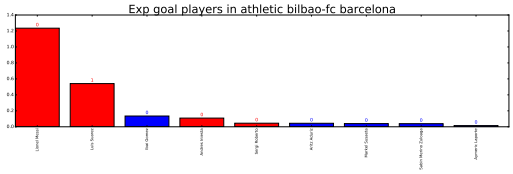
A more in-depth analysis of a single match could also provide valuable information about the performance of a single match. Two visualizations for match analysis are discussed in this section.

First of all, a time series of a match could be plottes just as Figure 6.3. Plotting this time series in one figure, effects of substitutions, goals scored or conceded and other major events could be analyzed in more detail by analyzing those events to the plot. In Figure 6.4, the goals scored and conceded are visualized. It can be seen that the match was in balance in the beginning of the match. Then, FC Barcelona (in red) creates some good scoring opportunities. After this moment, the match is in balance again (Expected goal is almost constant). Where FC Barcelona was not able to score during one of these major scoring opportunities, they did score later on in the match with a minor scoring opportunity. The pre-attentive attribute color is used in this plot to show the team who is expected to be in front [17]. The color filled beneath the plot is Red when the selected team (in this case FC Barcelona) has the upper hand or blue if the opponent has the upper hand based on Expected Goals.

Furthermore, the Expected Goals of individual players can be showed in a bar chart. This allows clubs to determine the most valuable player during a match based on created goal scoring opportunities and compare the Expected Goal scoring opportunities to actual goals scored. In this match, Lionel Messi created the best scoring opportunities. He was, however, not able to finish one of these scoring opportunities. His teammate Luis Suares, however, was able to score a goal with less optimal scoring opportunities. The pre-attentive attribute color is used to show to which team a player belongs, i.e. red for FC Barcelona and blue otherwise.



(a) The continuing of one match



(b) Expected goals of all players during one match

Figure 6.4: Match analysis figures

6.3 Player Analysis

Since the Expected Goals are determined based on individual goal attempts, the performance of players could also be analyzed with the use of Expected Goals. Expected Goals could be analyzed to get more insights into the performance of that player relative to the performance of similar players. Secondly, the favorite and most effective side of a field of a player could be determined. Thirdly, the favorite body parts of players could be determined. Lastly, the performance of a player over a season could provide insights into periods of time in which a player performs well/ poor.

6.3.1 Comparison of Players

Based on the player characteristics in the Player Data (Section 3.3) and a given player, the most similar players could be determined. The determination of the most similar players is performed with the use of a Nearest Neighbour classifier. This classifier determines the players closest to the given player using the Euclidean distance. The most similar players of FC Barcelona striker Lionel Messi are provided in Table 6.2. Table 6.2 also provides the three best attributes of Lionel Messi (as given by FIFA) and for the most similar players.

Table 6.2: Similar players of Lionel Messi

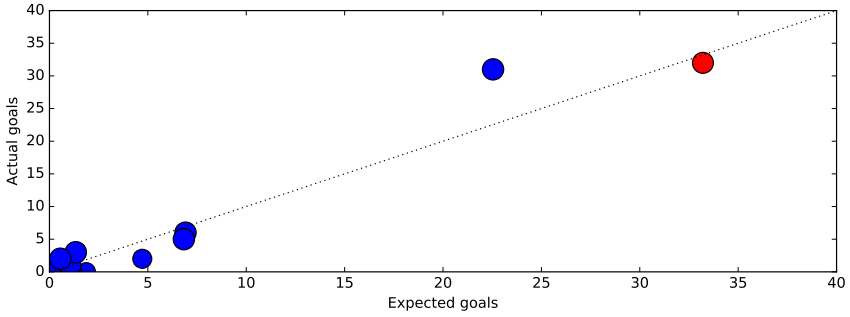
Player	Team	Ball Control	Dribbling	Acceleration
Lionel Messi	FC Barcelona	96	96	95
Arjen Robben	FC Bayern Munich	90	93	90
Paulo Dybala Rillo	Juventus	90	89	90
Eden Hazard	Chelsea	90	94	93
Neymar da Silva Santos Jr.	FC Barcelona	93	94	91
Sergio Aguero	Manchester City	89	89	92
Franck Ribery	FC Bayern Munich	90	90	86
Jonas Goncalves Oliveira	SL Benfica	88	84	76
Lorenzo Insigne	Napoli	88	87	93
Yevgen Konoplyanka	Sevilla FC	85	86	93
Carlos Vela	Real Sociedad	81	82	87
Antoine Griezmann	Atletico Madrid	84	86	86
Juan Manuel Mata Garcia	Manchester United	88	85	78
Marco Reus	Borussia Dortmund	85	86	89
Antonio Di Natale	Udinese	90	82	84
Julian Draxler	VFL Wolfsburg	85	87	78

Table 6.2 shows that Lionel Messi scores high on the attributes ball control, dribbling, and acceleration. The similar players, selected by the Nearest Neighbour algorithm, score high on these attributes as well

The Expected Goals and the actual goals of these similar players are plotted with the use of a

Q-Q plot in Figure 6.5. Similar to Figure 6.2, using the Q-Q plot [28], one could determine which player performs better than expected or performs worse than expected. If a player performs as expected, it will lay on the diagonal of the plot.

To ensure that the selected player, in this case, Lionel Messi, easily stands out from the others, the pre-attentive attribute color is used [17]. Furthermore, the size of the circles is used to show the similarity between the players. The similarity score is a result from the Nearest Neighbour algorithm.



Team	Player	Matches	Exp Success	Success
FC Bayern Munich	Arjen Robben	2.0	0	2
Juventus	Paulo Dybala	5.0	1	1
Chelsea	Eden Hazard	8.0	1	0
FC Barcelona	Neymar Jr.	0.0	0	0
Manchester City	Sergio Aguero	8.0	4	2
FC Bayern Munich	Franck Ribery	6.0	1	0
SL Benfica	Jonas Oliveira	0.0	0	0
Napoli	Lorenzo Insigne	0.0	0	0
Sevilla FC	Yevgen Konoplyanka	0.0	6	6
Real Sociedad	Carlos Vela	29.0	6	5
Atletico Madrid	Antoine Griezmann	44.0	22	31
Manchester United	Juan Garcia	0.0	1	1
Borussia Dortmund	Marco Reus	0.0	0	0
Udinese	Antonio Natale	0.0	0	0
VfL Wolfsburg	Julian Draxler	7.0	1	3

Figure 6.5: Comparison of similar players of a specific player

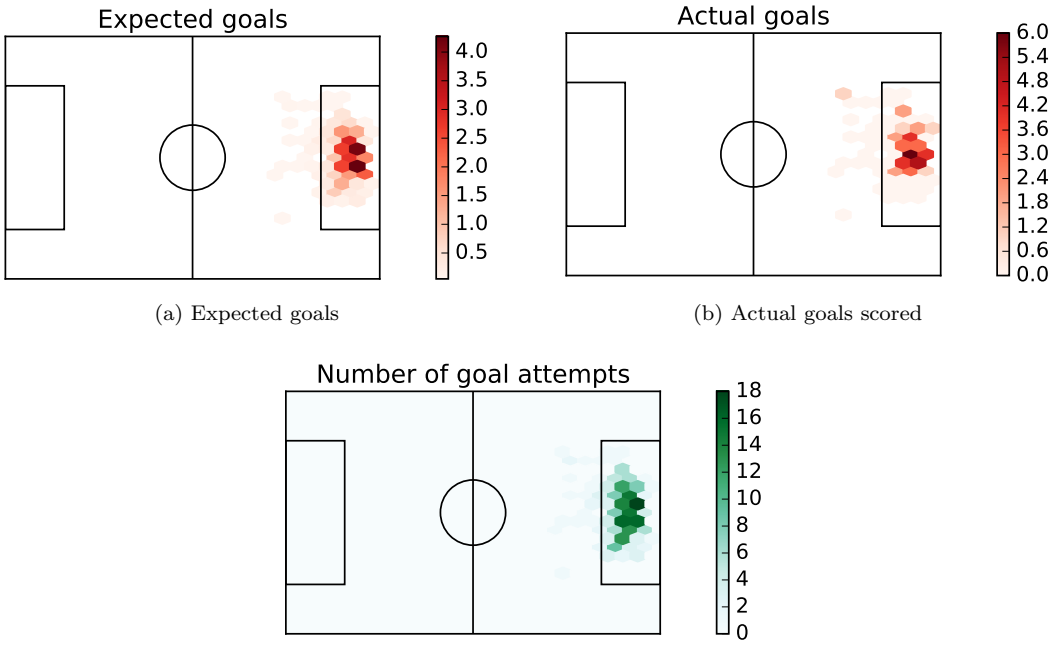
6.3.2 Analyzing Shot Location

Since the location of which the goal scoring opportunities are attempted, another player characteristic can be evaluated: the location of goal attempts. Analyzing the location of goal attempts of players could lead to interesting insights for own players but also for players from opposing teams. When analyzing own players, the most efficient locations can be determined. Furthermore, inefficient locations can be avoided in match situations or the individual training could be adjusted to make these locations more efficient.

Analyzing opposing players, however, could provide useful information in match preparation. Furthermore, scouts could use this information of opposing teams to select the correct player for a team. This is best illustrated with an example. Typically, wingers can be classified into two categories. The first category in which the winger crosses the ball in order to provide a scoring opportunity for other players. In the second category, the winger dribbles to the center of the pitch and attempts to score himself. An analysis of the location from which players shoot on goal could help in classifying the difference in these players.

In order to be able to fully analyze the favorite parts of the field, three different views have to be created. First of all, the expected goals are provided for different parts of the field. To compare the expected goals to the actual goals, the actual goals have to be provided for the different locations of the field as well. Finally, in order to determine whether players are shooting with goal

scoring opportunities of high quality, the number of goal scoring opportunities are provided for the different parts of the field as well. The three different views are provided in Figure 6.6.



(a) Expected goals

(b) Actual goals scored

(c) Number of goal attemts

Figure 6.6: Analysis of goal attempt locations of a specific player

In the design of Figure 6.6 numerous aspects have been considered. First of all, expected goals, actual goals, and the goal attempts displayed graphically on the field. In order to achieve this, a Choropleth map is used. A Choropleth map is a map in which aggregated data is visualized on geographical locations [28]. Furthermore, to allow easy processing by the human brain, the colors of the expected goals and the actual goals are both red while the color for the number of goal attempts is green. The difference in colors is chosen in this way since expected goals and actual goals can be compared easily (numbers should match more or less) while a more careful comparison is necessary with the number of goal attempts.

6.3.3 Analyzing Body Parts

Similarly to the analysis of the location of goal scoring opportunities, the use of different body parts can be evaluated. Three different parts of the body are present in the tactical data: head, body, foot. Analyzing the use and effectiveness of these body parts could provide more insights into the player characteristics. These insights could then be used to adjust training programs, to prepare matches, and for scouting. Figure 6.7 is a visualization of the use of these body parts.

Instead of using the Choropleth map as a visualization technique, since there are not that many different body parts which can be used, Graduated Symbol Maps are used. Graduated Symbol Maps place symbols on an underlying map, which allows for higher dimensional visualizations. In Figure 6.7 the color transparency show the effectiveness of the body parts. Here, the effectiveness is determined for each of the body parts. Theoretically, all circles could have a value of 1 in which case the player scored all his goals. Furthermore, the size of the circles is used to visualize the relative number of goal attempts of the different body parts. Since the number of goal attempts is already visualized in the figure, the third plot (as given in Figure 6.6) is no longer necessary.

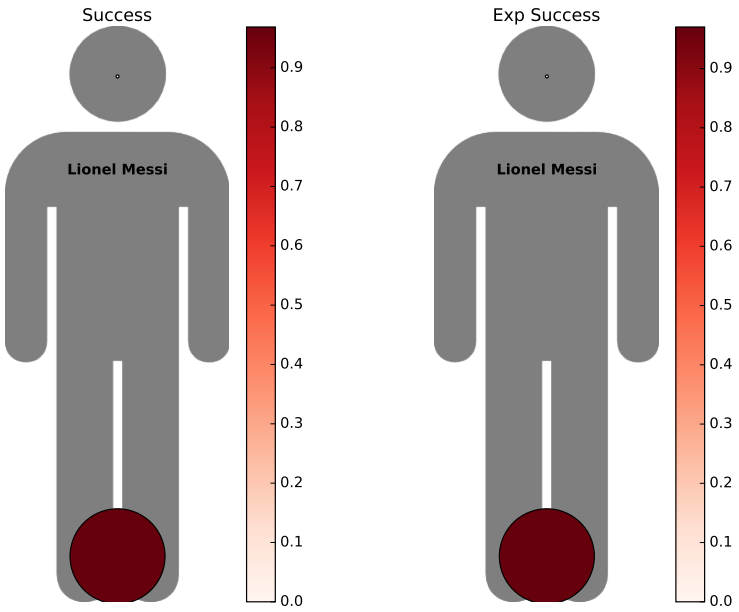


Figure 6.7: Analysis of favourite body parts of a specific player

6.3.4 Performance during a Season

As discussed in Section 6.1.1, teams can have stronger and weaker periods during a season. Similarly to teams, player performance can fluctuate during a season as well. To analyze whether these fluctuations are time dependent, the expected goals can be provided over a period of time. Since matches are discrete events over time, a bar chart is used to visualize the performance of a player over a season. This bar chart is provided in Figure 6.8.

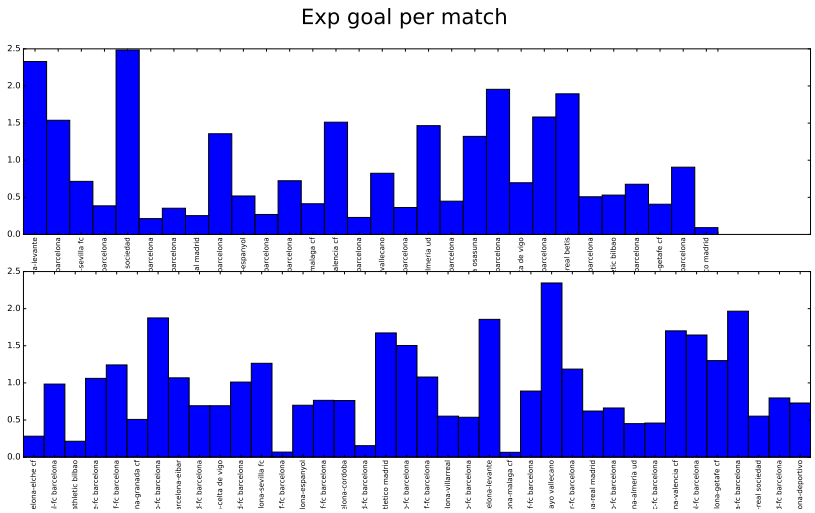


Figure 6.8: Expected goal of a player over a season

Figure 6.8, however, lacks information about the number of goal attempts a player needed to

get the expected goals. In order to provide this kind of information, the ratio of the expected goals of a player are represented as the expected goal per goal attempt. The expected goals per goal attempt of Lionel Messi are provided in Figure 6.9.

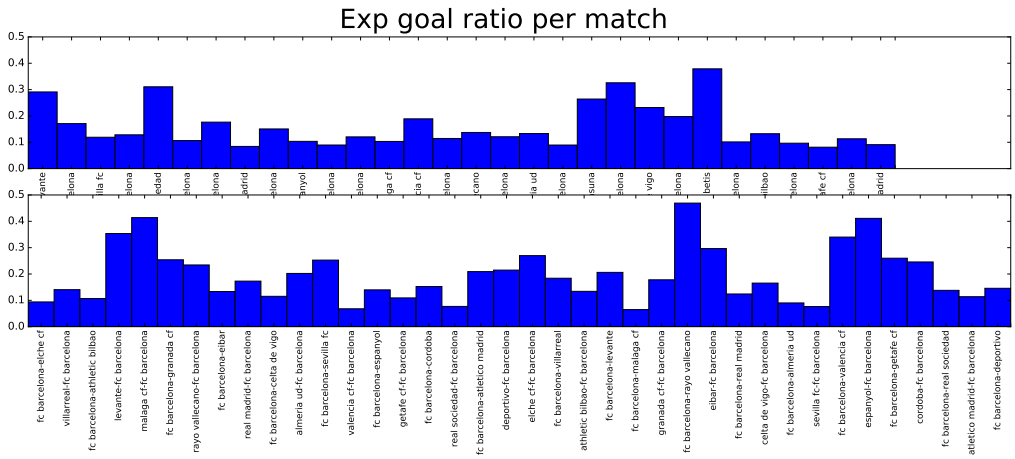


Figure 6.9: Ratio of expected goals of a player over a season

6.4 Graphical User Interface

To allow for easy use of the visualizations described in this chapter, a Graphical User Interface (GUI) is created [1]. The GUI allows the user to easily switch between different leagues, seasons, teams, matches, and players. The visualizations can then be viewed in the GUI and stored to pdf for later use. Furthermore, a multi-paged report can be generated in pdf format for a season, a match, or a player. The report in combination with the GUI allows the technical staff to share reports. A screenshot of the GUI is provided in Figure 6.10.

[1] The GUI is created in cooperation with Kees Hendriks, another student graduating at PSV. The different tabs at the top of the GUI represent different projects performed at PSV. Therefore, other students can easily add their projects [23].

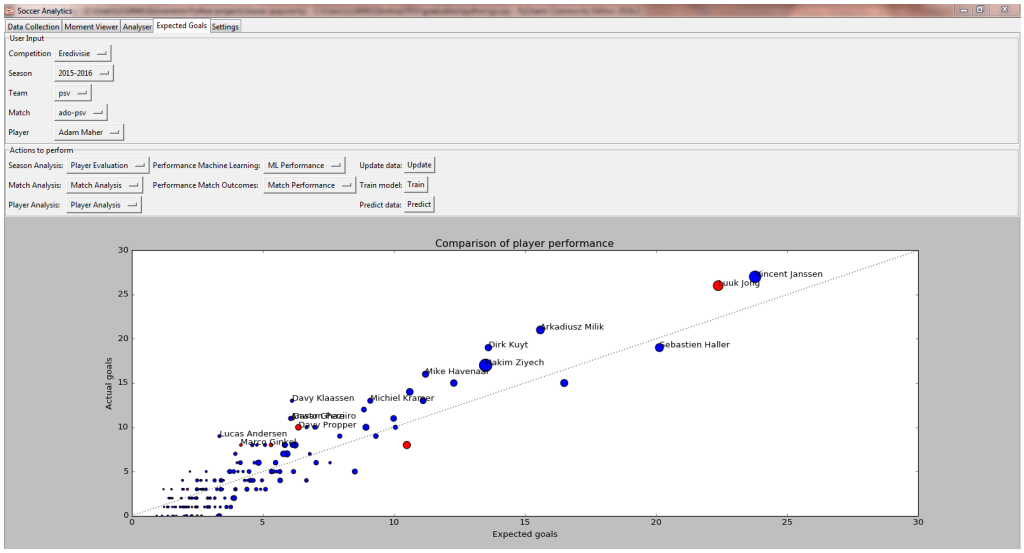


Figure 6.10: Screenshot of the Graphical User Interface

Chapter 7

Conclusion

In this chapter, the thesis is concluded. Therefore, the main contributions of the thesis are discussed. Furthermore, limitations of the proposed model and future work are discussed.

7.1 Main Contributions

This thesis provides a new metric (Expected Goals) to evaluate the quality of a goal scoring opportunity. In order to achieve the expected goals, classifiers are trained to rank the scoring opportunities, i.e. better scoring opportunities score higher. Therefore features should be extracted from the situation of the scoring opportunity. Since the probability of the scoring opportunity resulting in a goal also depends on both the player attempting to score and the goalkeeper trying to prevent this, these features are added to the features which describe the scoring opportunity. Since the scores from derived from the classifiers do not accurately match the actual probability of resulting in a goal, the classifiers are calibrated. Calibration ensures that the scores can be interpreted as probabilities.

The quality of individual scoring opportunities should be interpreted with care since the standard deviation (and thus the confidence interval) of the probability is relatively high. Due to the law of large numbers, scoring opportunities can, however, still be aggregated. Aggregating the expected goals leads to expected match outcomes. The performance of the expected goals model is also evaluated regarding the expected match outcomes. Here it is seen that especially when the expected goal model predicts a draw, the predictions are often incorrect. A possible explanation of this is that in the cases of a predicted draw, the match was tight and the game could have gone either way with only one goal difference.

Finally, this thesis proposed some approaches which can be taken by teams to use the expected goals model to improve decision-making. These approaches can be divided into three different categories: Season analysis, Match analysis, and player analysis. Different visualizations are proposed to visualize different useful approaches.

7.2 Limitations & Future work

In this thesis, four different classification algorithms with different parameter settings have been evaluated. Since not all algorithms perform the same, even more classification algorithms with different parameter settings could be tested. Furthermore, more features could be added as input for the classifiers. In the current setting, random forest performs best, but this might change when more algorithms and features are added.

A limitation of the current expected goals model is that the quality of individual goal scoring opportunities should be interpreted with care due to the relatively large confidence interval. The variance of between scoring opportunities is partly explained by the problem itself since players make different decisions in similar scoring opportunities. Future work could, however, still focus

on reducing the variance between similar scoring opportunities. More and more data is getting available as the time continues since more matches are played. More data could reduce the variance of the model. Furthermore, feature selection could be performed to select only the most important features.

Even more insights in match results could be obtained by adding actions before the actual goal attempt. This could be used to determine the best possible action to perform at a given moment for a given player. By doing this, crosses which did not get touched would also be included in the expected goals model. Furthermore, choices made by individual players could be evaluated. This could then be used by trainers for training purposes. An example of such a situation in which the best possible decision could be determined is provided in Figure 7.1

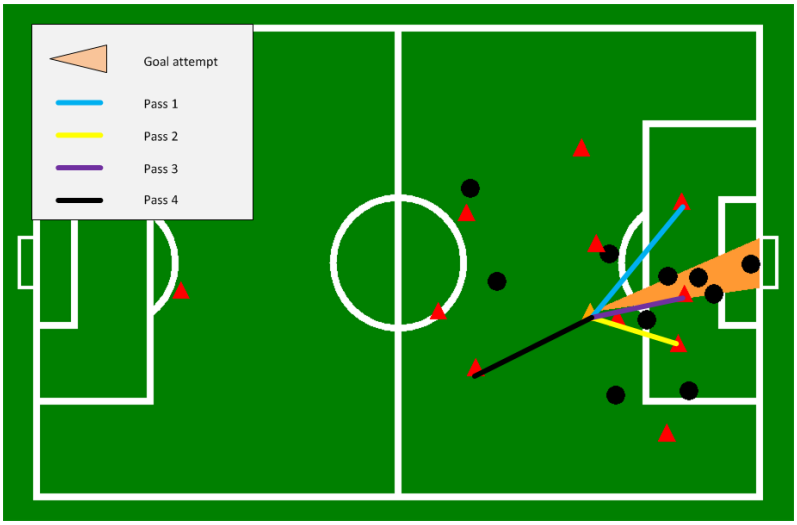


Figure 7.1: Determination of best possible option of a player at a given moment

In Figure 7.1 the player has five intuitive options (4 passes and a goal attempt). Currently, the expected goals model is only able to determine the probability that the goal attempt will result in a goal. In other words, the expected goals model is able to determine the quality of the choice to shoot on goal. The model is, namely, trained on data in which a player already chose to shoot on goal. Passing the ball to one of his teammates could, however, lead to a better Expected Goals. Intuitively, when pass 1 succeeds of Figure 7.1, the goal attempt created for the receiving player. Since it seems quite possible that this pass will succeed this, intuitively, seems to be the best decision.

Creating a model which determines the best decisions at a given time, however, leads to new challenges. First of all, the assumption that players are not moving is no longer reasonable. Since players could move in all possible directions, it is hard to predict where they are going. If, however, only one action is investigated (as in Figure 7.1) the position of the players could be determined by the current speed, acceleration, and direction of movement. The speed and acceleration are already given in the spatiotemporal data. The direction of movement can be determined by, for example, looking at the direction of movement of the last 500 milliseconds. Furthermore, the probability of success has to be predicted for multiple types of events. Similar approaches as presented in this thesis could, however, be followed to determine the probability of success for other events.

Currently, the predictive model is trained on the complete population of players with player characteristics. As discussed in Section 2.5, this leads to the desired Expected Goals, and thus expected performance. By training the predictive model, the model could provide insights into different types of performance. When the predictive model is, for example, trained on only the best players in the population, the model would indicate desired performance (closest to optimal performance). Furthermore, the model could be trained on only the players of a league to provide

insights into the typical performance of players in that league. This could give insights into the differences in leagues.

Besides improvements in the expected goal model itself, the expected goals model opens new possibilities for further research. Some of the possibilities can already be seen from visualizations as shown in Chapter 6. First of all, the influence of events on the outcome of matches could be determined. A first approach is provided with Figure 6.4a. A more quantitative approach would, however, provide insights into a more general influence.

Secondly, the expected goals could be used in player acquisition. In this thesis, a first approach for getting similar players is provided in Section 6.3.1. Elaboration on this approach could lead to interesting insight which improves decision making of player acquisition. The expected goals could be included to find players who are performing better than expected.

Bibliography

[1] SportVU Player Tracking, howpublished = http://www.stats.com/sportvu/sportvu-basketball-media/, note = Accessed: 2016-03-24. 1

[2] Chris Anderson and David Sally. *The numbers game: why everything you know about football is wrong*. Penguin UK, 2013. 1, 2

[3] Richard A. Becker, Stephen G. Eick, and Allan R. Wilks. Visualizing Network Data. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):16–28, 1995. 5

[4] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997. 7

[5] Glen W Brier. Verification of forecasts expersses in terms of probaility. *Monthly Weather Review*, 78(1):1–3, 1950. 8

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh International World-Wide Web Conference*, pages 107–117, Brisbane, Australia, 1998. 6

[7] J.J. Bull. Sky sports use football manager database to profile players in real life. *The Telegraph*, 2015. 15

[8] C . Reep, B . Benjamin. Skill and Chance in Association Football. *Journal of the Royal Statistical Society .*, 131(4):581–585, 1968. 1

[9] Soumen Chakrabarti, Martin Ester, Usama Fayyad, Johannes Gehrke, Jiawei Han, Shinichi Morishita, Gregory Piatetsky-Shapiro, and Wei Wang. Data mining curriculum: A proposal (version 1.0). *Intensive Working Group of ACM SIGKDD Curriculum Committee*, page 140, 2006. 3

[10] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 28

[11] Xinhua Cheng and John M. Wallace. Cluster Analysis of the Northern Hemisphere Wintertime 500-hPa Height Field: Spatial Patterns, 1993. 5

[12] Anthony C. Constantinou, Norman E. Fenton, and Martin Neil. Pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, 36(February 2016):322–339, 2012. 3

[13] Corinna Cortes and Daryl Pregibon. Giga-mining. In *KDD*, pages 174–178, 1998. 5

[14] Daily Mirror. FIFA ratings vs real life: How does the video game measure up to actual Premier League stats? http://www.mirror.co.uk/sport/football/news/fifa-ratings-vs-real-life-4027925, 2014. Online; accessed 2 August 2016. 16, 17

[15] Stefan Dobravec. Predicting sports results using latent features: A case study. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on*, pages 1267–1272. IEEE, 2015. 4

[16] Robert Donnelly and W Michael Kelley. *The humongous book of Statistics Problems*. Penguin, 2009. 39

[17] Stephen Few. Tapping the power of visual perception. *Visual Business Intelligence Newsletter*, 2004. 39, 41, 42, 44

[18] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995. 6

[19] Thomas U Grund. Network structure and team performance: The case of english premier league soccer teams. *Social Networks*, 34(4):682–690, 2012. 4

[20] David Hand, David Hand, Heikki Mannila, Heikki Mannila, Padhraic Smyth, and Padhraic Smyth. *Principles of data mining*, volume 30. 2001. 5, 6

[21] Ruud van Elk Harm Eggels, Mykola Pechenizkiy. Explaining soccer match outcomes with goal scoring opportunities predictive analytics. In *MLSA*, Riva del Garda, Italy, 2016. ECML/PKDD. 4

[22] Trevor J.. Hastie, Robert John Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2011. 6

[23] Kees Hendriks. PSV football data analysis: An elaborate analysis on what possession loss caauses and what causes possession loss. Master's thesis, University of Technology Eindhoven, the Netherlands, 2016. 18, 47

[24] Andreas Heuer, Christian Mueller, and Oliver Rubner. Soccer: is scoring goals a predictable Poissonian process? *EuroPhysics Letters*, 89(3):2–6, 2010. 4

[25] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995. 6

[26] Vincent Hoekstra, Pieter Bison, and Guszti Eiben. Predicting football results with an evolutionary ensemble classifier. page 68, 2012. 4

[27] Inmotio. Football. `http://www.inmotio.eu/en-GB/34/football.html`, 2016. Online; accessed 5 August 2016. 14

[28] Heer Jeffrey, Bostock Michael, and Ogievetsky VADIM. A tour through the visualization zoo. *Communications of the ACM*, 53(6):56–67, 2010. 41, 44, 45

[29] P. D. Jones, N. James, and S. D. Mellallieu. Possession as a performance indicator in footbal. *International Journal of Performance Analysis of Sport*, 4(February 2016):98–102, 2004. 4

[30] Dimitris Karlis and Ioannis Ntzoufras. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003. 4

[31] Matthew George Soeryadjaya Kerr. *Applying machine learning to event data in soccer*. PhD thesis, Massachusetts Institute of Technology, 2015. 4

[32] R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International Joint Conference on Artificial Intelligence*, number 0, pages 0–6, 1995. 6

[33] Helge Langseth. Beating the bookie: A look at statistical models for prediction of football matches. In *SCAI*, pages 165–174, 2013. 3

[34] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966. 18

[35] Michael Lewis. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004. 1

[36] Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, and Iain Matthews. "Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data. In *Proc. 9th Annual MIT Sloan Sports Analytics Conference*, pages 1–9, 2015. 4, 23

[37] Oded Maimon and Lior Rokach. *Data mining and knowledge discovery handbook*, volume 2. Springer, 2005. 6

[38] Tim McGarry and Ian M Franks. On winning the penalty shoot-out in soccer. *Journal of Sports Sciences*, 18(6):401–409, 2000. 33

[39] RA Mollineda, R Alejo, and JM Sotoca. The class imbalance problem in pattern classification and learning. In *II Congreso Español de Informática (CEDI 2007). ISBN*, pages 978–84, 2007. 28

[40] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning ICML 05*, (1999):625–632, 2005. 7, 8

[41] ORTEC. Performance Analytics. `http://ortecsports.com/performance-analytics/`, 2016. Online; accessed 5 August 2016. 11

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 6, 27

[43] J.C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 7

[44] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010. 6

[45] Colin Shearer, Hugh J Watson, Daryl G Grecich, Larissa Moss, Sid Adelman, Katherine Hammer, and Stacey a Herdlein. The CRIS-DM model: The New Blueprint for Data Mining. *Journal of Data Warehousing14*, 5(4):13–22, 2000. 3

[46] Sofifa. Players. `http://sofifa.com/players`, 2016. Online; accessed 5 August 2016. 15

[47] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009. 6

[48] K Stuart. Why clubs are using football manager as a real-life scouting tool. *The Guardian*, 4:20–58, 2014. 15

[49] Pang-Ning Tan et al. *Introduction to data mining*. Pearson Education India, 2006. 6, 8

[50] Strother H Walker and David B Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179, 1967. 6

[51] David H Wolpert. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7):1341–1390, 1996. 6

[52] Show-Jane Yen and Yue-Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, 2009. 29

[53] Bianca Zadrozny and C Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *Icml*, pages 1–8, 2001. 7

[54] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining KDD 02*, pages 694–699, 2002. 7