

Abstract

This master's thesis deals with the extent to which individual player performance can be predicted with the help of machine learning models. To answer this question, the current state of research in the field of sports analytics is first conducted through a literature review. Based on the resulting findings, an application is developed that automatically obtains, processes, and stores data from various sources. With the help of this data, different machine learning models are compared based on the CRISP-DM cycle. The influence of betting odds on the accuracy of the models is examined separately by dividing the models into a baseline group without betting odds and a treatment group with betting odds. These investigations show that betting odds improve the predictions of the models slightly. In comparing the machine learning models, an optimised random forest model achieved the most accurate forecasts and earned prizes in the experiment conducted on the test data set. The developed application is extended by this model. It is thus able to predict the individual player performance and send a line-up for the upcoming match day to the user based on these predictions.

Abstract German

Die vorliegende Masterarbeit beschäftigt sich mit der Frage, inwieweit, mit Hilfe von Machine Learning Modellen, individuelle Spielerleistungen vorhergesagt werden können. Um diese Frage beantworten zu können, wird zunächst der aktuelle Forschungsstand im Themenbereich des Sports Analytics zusammengefasst. Basierend auf den daraus hervorgehenden Erkenntnissen wird eine Applikation entwickelt, welche automatisiert Daten aus verschiedenen Quellen bezieht, verarbeitet und abspeichert. Mit Hilfe dieser Daten werden anschließend verschiedene Modelle des Machine Learnings, in Anlehnung an den CRISP-DM Kreislauf, miteinander verglichen. Dabei wird der Einfluss von Wettqouten auf die Genauigkeit der Modelle gesondert untersucht, in dem die Modelle in eine Basisgruppe ohne die Wettqouten und in eine manipulierte Gruppe mit Wettqouten eingeteilt werden. Aus diesen Untersuchungen geht hervor, dass Wettqouten die Vorhersagen der Modelle leicht verbessern. Im Vergleich der Machine Learning Modelle erzielte ein optimiertes Random Forest Modell die genauesten Vorhersagen und konnte im durchgeführten Experiment auf dem Testdatensatz Gewinne erzielen. Die entwickelte Anwendung wird um dieses Modell erweitert und ist dadurch in der Lage, zunächst die individuellen Spielerleistungen zu prognostizieren und auf Basis dieser Vorhersagen eine Aufstellung für den kommenden Spieltag an den Nutzer zu verschicken.

Contents

| | | |
|----------|---------------------------------------|-----------|
| 1 | Introduction | 1 |
| 1.1 | Motivation | 1 |
| 1.2 | Aim and Delimitation | 3 |
| 1.3 | Methodology | 4 |
| 2 | Foundations | 6 |
| 2.1 | Fantasy Leagues | 6 |
| 2.2 | SPITCH | 9 |
| 3 | Literature Review | 13 |
| 3.1 | Search Process | 14 |
| 3.2 | Concept Matrix and Criteria | 14 |
| 3.3 | Concepts | 17 |
| 3.3.1 | Features | 17 |
| 3.3.2 | Machine Learning Approaches | 19 |
| 3.3.3 | Betting Odds | 22 |
| 3.4 | Research Gap | 23 |
| 4 | Implementation | 24 |
| 4.1 | Business Context | 24 |
| 4.2 | Data | 30 |
| 4.2.1 | Procurement | 30 |
| 4.2.2 | Processment | 33 |
| 4.2.3 | Exploration | 37 |
| 4.3 | Models | 49 |
| 4.4 | Betting Odds Impact | 55 |
| 4.5 | Application | 57 |

| | |
|------------------------------------|-----------|
| 5 Conclusion | 59 |
| List of Figures | A |
| List of Tables | B |
| Bibliography | C |
| Decleration of Authenticity | G |

Chapter 1

Introduction

1.1 Motivation

'Data is the new oil.' – Clive Humby, 2006

What Clive Humby already recognised in 2006 is even more relevant today than ever before. Moreover, in the age of *Big Data*, the utilisation of this data is probably even more essential. The ability to gain insights by analysing this data and thereby anticipate the future is an art in itself and is still being vigorously explored in science today. More and more, efforts are made to predict the future with the help of this data and computing techniques such as *Artificial Intelligence* or *Machine Learning*.

At the same time, the *Big Data* age does not spare the world of sport. (cf. Rein & Memmert, 2016) In recent years, sports teams and associations have started to collect data during their games extensively. The teams analyse this data to keep improving, while the associations use the data to provide a more comprehensive gaming experience for their viewers. In almost every sport, this movement also led to a new sub-discipline of these sports: the Fantasy Leagues. Already started in 1962 for American football, this segment experienced a boom due to *Big Data*, as every sport could now be played as Fantasy League with an increasing complexity thanks to more accurate tracking possibilities.

Thereby, the technique of successfully analysing the data and predicting the future with machine learning models is enjoyable for all participants of the sport: the teams know where they can improve, the coach understands which tactics or players work best, and the supporters are even more involved in the sport. Above all, the Fantasy League organisers nowadays use these analyses to know how they have to design their game to remain as exciting as possible. Mainly because these models always try to look into the future, the idea of incorporating a metric that is supposed to represent the future is close at hand. An example of these metrics is a prediction market, which dares to look into the future with the help of swarm intelligence. There are different types of prediction markets in sports, ranging from spread betting to betting odds. This industry is even older than Fantasy Leagues, and not at all surprisingly, it is booming with the growth of data and analysis in sports. In fact, Fantasy Leagues are nothing more than more complex sports bets.

Although all mentioned areas are thriving, there still exists almost no scientific literature on their unification. This disregard leads to wasted potentials, as the stated benefits are numerous. Furthermore, the data required is already available and reliable. For these reasons, this thesis intends to examine one particular area of this unification to see how far machine learning can be used in this context to predict the future successfully.

1.2 Aim and Delimitation

This thesis aims to use the available data of a Fantasy Soccer League provider to create a machine learning model that predicts the individual performance of each soccer player for future matches as accurately as possible. Afterwards, based on this model, a team can be assembled and compared with other line-ups created by real-life opponents. For this purpose, different models will be examined that have proven successful in the existing literature. Therefore, a literature review will be conducted to overview the current state of research on these topics. Subsequently, the models will be complemented by betting odds to investigate whether betting odds provide a type of swarm intelligence and thus make the models more accurate. Finally, two questions are answered. Firstly: Is it possible to predict the future performance of players with the help of machine learning models? Furthermore secondly, does the influence of betting odds have a positive effect on these predictions?

However, the aim of this work goes beyond the mere investigation of different machine learning models. In addition, the predictions generated by the models are made available to users via an application implemented specifically for this purpose.

Every problem requires a different approach to the solution, and this is particularly true in the machine learning sphere. Although this thesis seeks to the best of one's knowledge to address the above questions, the solutions are not generally transferable. Each model, feature or statistic has been adapted to the exact use case at hand, although always attempting to remain as generalist as possible for similar problems. Furthermore, the result often only shows the steps that led to success. In the course of this research, many steps were taken, which did not lead to the desired results and, for this reason of comprehension, are not mentioned. In addition, reasonable simplifications had to be made when the reality was too complex to implement. These passages are explicitly mentioned in the text.

1.3 Methodology

Since the essential part of this work is the creation of various machine learning models and the procurement of the data to create them, this work reaches into the field of *Data Mining*. For this reason, the implementation method was based on an established procedure in this field: the *CRISP-DM* (Chapman et al., 2000) procedure. This process builds the foundation of Géron’s data mining process presented in his book ‘*Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*’, which offers a systematic approach to the type of problem at hand. This methodology is therefore used to systematically follow the data procurement process and the creation and evaluation of the different models.

Aligned with standard research designs, two types of models emerge in the data mining process. The first type of model is the *baseline* model, which will mainly use historical data from the Fantasy League provider for its predictions. The second type is the *treatment* model, which will additionally have access to the betting odds. With this design, a *t*-test can be used in the end to test whether betting odds have a statistically significant effect on the predictions. In parallel, Cohen’s D is calculated and used to determine the effect size.

The resulting models are compared on a test dataset experimentally. The experiment will pitch the baseline models against the treatment models and investigate whether the treatment models actually create better teams and thus perform better. For this purpose, standard metrics for the evaluation of machine learning models are used. Moreover, additional benchmarks are calculated, such as the optimal possible team or the rank in the player ranking, to investigate how such models perform against real players.

In this paper, reference is made to previous work in the literature. The current state of research is determined with the help of a literature review, which is conducted according to the established standards of Webster and Watson and vom Brocke et al., to ensure comprehensibility and reproducibility.

Following the methodologies above, the paper is divided into three main chapters. First, the basics of fantasy leagues are explained and then specified in more detail for the chosen provider *SPITCH*. Subsequently, the rules of *SPITCH* are explained shortly, limited to the essentials. In the next chapter, the literature review takes place, which presents the current state of research. In the final chapter, the data mining process takes place, where data is collected, processed and analysed, and models are created, optimised and compared with each other. This process results in two final models: a baseline and a treatment model. Finally, these two models are tested against each other in an experiment, and the effect size of betting odds is measured and evaluated.

Chapter 2

Foundations

2.1 Fantasy Leagues

Fantasy Leagues can look back on a history of over 60 years. Wilfred Winkenbach, a sports entrepreneur and enthusiast from the USA, designed a fantasy golf game in the 1950s. In this game, a team was made up of several golfers, and the team with the lowest swings in total won. Building on the success of this game, Winkenbach developed the first fantasy football league in 1962, which is similar to today's fantasy leagues. (cf. Green, 2014) This league consisted of 8 participants, friends or co-workers of Winkenbach, who met in a restaurant and wrote down their line-up for the coming season. The scoring system was kept very simple and was limited to the main events in a football game: touchdowns, field goals and interceptions. The simple reason was that each event had to be counted by hand by the game master. (cf. Fabiano, 2007) From this game, leagues quickly developed in other sports, such as baseball. One of the reasons this type of game first spread in the USA is the ease of assigning points to individual actions in the popular represented sports. For example, during an attack in American football, there are several plays, separated by pauses in which it can be assessed relatively clearly, for example, by the yards gained or lost, whether the play was successful. In soccer, on the other hand, there are fewer interruptions, plus unlike in baseball or American football, there are no intermediate milestones that can be reached between moves. These missing pauses lead to a more wild game, with difficult to evaluate actions. In addition, although there are roles within a soccer team, these roles, except the goalkeeper, are more strategic and do

not restrict the players in their playing actions. A defender can score goals or intercept passes just as well as a striker. In contrast, in American football or baseball, each player often has one single task per turn that can either succeed or not. All these factors did sports like soccer challenging, if not impossible, to implement as fantasy leagues.

However, with the advancement of modern image recognition technologies and player tracking devices such as two high-resolution cameras per playing side, these times are a thing of the past. (cf. Hoffmann, 2014) Nowadays, every event on a soccer pitch is automatically trackable and therefore offers the possibility to evaluate the performance of different players much more accurately. These advances allow fantasy soccer leagues to exist, as they can build their game on this data basis.

Nevertheless, the main goal of fantasy leagues is always the same for all sports: assemble a team that performs best. However, there are differences between the fantasy leagues in how this performance is evaluated. One major factor for this difference originates in the differences in the sport disciplines themselves. Despite that, even leagues in the same discipline can differ to create a unique selling point. Since this work is primarily focused on soccer, the following considerations are limited to fantasy soccer leagues. Furthermore, the decision was made to use the provider SPITCH, which only offered the first Bundesliga at the time of writing. That is why all comparisons are made concerning this game system.

The list of differences between the individual fantasy soccer leagues is long. Each provider of a league wants to have its unique selling point and highlights different tactical elements. It is not the aim of this thesis to show all the differences. This section merely serves to give a rough overview of the world of Fantasy Soccer Leagues and, in addition, to show that each game must be approached strategically differently and, as a result, different questions must be asked. The research in this paper, therefore, applies primarily to the game SPITCH. Nevertheless, this thesis aims to shed light on problems in as general a manner as possible and to be useful for research in similar areas.

A major difference in fantasy soccer is the national league in which the fantasy league is located. For example, there are providers for the English *Barclays Premier League*, the Italian *Serie A* and for the German *Bundesliga*, the latter being observed in this work. However, it is not only the selected national league that differentiates the various providers. Furthermore, a differentiation can be made between the availability of players. For example, one popular game mode exists where 20 fantasy managers compete against

each other, similar to the actual real-world competition. Each fantasy manager is assigned a team of random players. These own players can be traded with the other 19 teams for other players or money on a virtual transfer market. It is important to note that this game mode creates a segregated space, where the participants compete against each other continuously over an entire season. On top of that, each player exists **only once** and can only play for one fantasy manager's team at the time. In contrast, in SPITCH, any player can be bought and used by any fantasy manager. Furthermore, the participants do not play in a segregated space but with unlimited opponents. Further details and the general regulations are described in the following chapter.

2.2 SPITCH

This section intends to provide the necessary rules from SPITCH needed to understand the optimization problem at hand. Additionally, this section aims to outline the first approaches to a possible solution.

As already mentioned in the previous chapter, SPITCH is a provider for fantasy soccer leagues. (cf. SPITCH, 2021c) At the beginning of writing, SPITCH only provided competitions for the German *Bundesliga*. Until now, numerous national football associations from different countries joined. Furthermore, football managers can these days compete in various other game modes. This thesis solely deals with the traditional game mode for the German *Bundesliga*.

To counteract confusion that may arise, the following terms and their meaning in the context of this work, such as player or manager, are explained in more detail. Furthermore, each word is assigned a variable that will help understand the calculation of scores more quickly.

Table 1: SPITCH Glossary

| Term | Variable | Meaning |
|---------|----------|--|
| Manager | M | Participants of SPITCH |
| Player | P | Real soccer player, i.e. Manuel Neuer |
| Value | V | Transfer market value of a Player P |
| Event | E | In-game events such as Goal, Pass, Unsuccessfull Pass etc. |
| Points | p | Points according to SPITCH points catalogue |
| Score | S | Sum of points p |
| Round | R | Game-Round, e.g. matchday |
| Line-up | L | Line-up consisting of 11 players P |

Like most fantasy leagues, the aim in the traditional game mode is to line up a team that performs best. Unlike most fantasy leagues, the managers M in a SPITCH competition only assemble a line-up L for the upcoming match day. So when planning the line-up, it is not necessary to think long-term for the entire season. A new line-up consisting of different players can be created for each round R . Each line-up consists of 11 out of 711 possible players P . Each player P_i , $\{i \mid i \in \{1, 2, \dots, 711\}\}$ has a score S_{PR} for each of the 34 rounds R_j , $\{j \mid j \in \{1, 2, \dots, 34\}\}$. For simplicity, as the rounds are separated and thus do not influence each other, the following declarations are all round-specific. The final line-up score S_L is the sum of 11 individual player scores S_{P_i} :

$$S_L = \sum_{i=1}^{11} S_{P_i} \quad (1)$$

This score S_L is used to create a ranking of managers M and therefore decides if the manager wins a prize or not. The individual player score S_{P_i} is calculated using the occurred events O during a match multiplied with their corresponding points p given by the SPITCH points catalogue. It exists a number of 33 different event types, such as pass, goal or tackle, therefore E_k , $\{k \mid k \in \{1, 2, \dots, 33\}\}$ applies. For example, a pass is granted two and a goal 200 points. For negative event types, such as a missed chance, negative points can also be awarded. (cf. SPITCH, 2021a) Hence, a player can have a negative score. The individual player score can be calculated using the following equation:

$$S_{P_i} = \sum_{k=1}^{33} O_k * p_k \quad (2)$$

Given equations (1) and (2), the final line-up score S_L can be calculated using:

$$S_L = \sum_{i=1}^{11} \sum_{k=1}^{33} O_{ik} * p_{ik} \quad (3)$$

The line-up allows nine players P per real-life club. (cf. SPITCH, 2021b) Furthermore, each player P_i has one of the following simplified positions: goalkeeper, defender, midfielder, attacker. As a result, for example, four players who, in reality, all play as right defenders can be lined up in SPITCH without any disadvantages. Players can not be lined up for

another position as their by SPITCH assigned simplified position. There is a selection of ten different formations that can be used to vary the number of defenders, midfielders, and attackers. However, this selection is limited to the relevant formations in reality, so there are only formations with a maximum of 5 players in one position form, except the goalkeeper position.

Each player P_i has a transfer market value V_i . As already explained in the previous chapter, in SPITCH, any player can be fielded by any manager. For this reason, the prices of the players are based on various factors, which, however, are not publicly available. These factors include how many managers draft this player, his historical performance, and his level of fame in reality. These values V exist to constrain the managers in their player choices. Since each manager only has a budget of €150m, he cannot exclusively field star players but must at the same time resort to more unknown players. This restriction turns the problem into a so-called **knapsack problem**. If the manager does not spend the budget completely, for example, by buying only inexpensive players, he will start the round with bonus points. The same applies vice versa if the budget is exceeded. This positive or negative score is called manager score S_M . The relation between the budget deviation Δ_{Budget} and manager score S_M is represented by the linear function:

$$S_M = \frac{\Delta_{Budget} \cdot 0.8}{100,000} = \Delta_{Budget} \cdot 0.8 \cdot 10^{-5} \quad (4)$$

Thereby, the factor $0.8 \cdot 10^{-5}$ is used by SPITCH as a balancing method. (cf. SPITCH, 2021b) For example, if the budget exceeds €10m, i.e., a budget of -€10m, the manager starts with $-\text{€}10\text{m} \cdot 0.8 \cdot 10^{-5} = -80$ manager score. Consequently, manager points do not exponentially increase or decrease the further one moves away from the budget threshold. For this reason, the transfer market value V of a player P can be converted and taken into account to his points p . For instance, a player P_1 with a transfer market value V_1 of €10m must therefore first score 80 points p to achieve a total positive score for the team. Since a linear relation can be established between the target value, the final line-up score S_L , and the weight of the transfer market values V , there is **no typical knapsack problem at hand**.

Since the transfer market value of a player V_i can be counted towards a player's individual score S_{P_i} , equation (2) can be supplemented by equation (4) to create the

adjusted player score:

$$S_{P_iM} = -V_i \cdot 0.8 \cdot 10^{-5} + \sum_{k=1}^{33} O_k * p_k \quad (5)$$

resulting in the following equation to calculate the **adjusted** final line-up score S_{LM} :

$$S_{LM} = S_L + S_M = \sum_{i=1}^{11} \sum_{k=1}^{33} O_{ik} * p_{ik} + \Delta_{\text{Budget}} \cdot 0.8 \cdot 10^{-5} \quad (6)$$

Each round, one player of the line-up can be appointed as captain, which results in his score getting doubled. The managers can participate for free or with a stake. The higher the stake, the higher the prize. The stake is graded according to so-called *pitches*, such as the *€2 pitch* or the *€30 pitch*. Only the participants of the individual pitches compete against each other. On the *free pitch*, the top 10 managers in the ranking, i.e., the ten managers with the highest adjusted final line-up scores S_{LM} , win. On all other pitches, the top 25% of managers, i.e., the upper quartile, win. Within these winning zones, the percentage of the prize won decreases exponentially. SPITCH does not publish the exact calculation of this decrease. The calculation of the price won will be addressed later in this work when evaluating the models in chapter 4.3.

Chapter 3

Literature Review

This chapter presents the current state of research in two different domains. The first is about predicting sporting events using machine learning. The latter examines sports betting with a particular focus on betting odds and how these can help to predict events in the future.

The established guidelines of vom Brocke et al. (2015) and Webster and Watson (2002), are used to determine the current state of research and respectively document the literature search process. As stated by Webster and Watson (2002), two types of literature reviews exist. This literature review belongs to the second type, which is, according to Webster and Watson, in general, shorter and where *'authors [...] tackle an emerging issue that would benefit from exposure to potential theoretical foundations'* (Webster & Watson, 2002, p. 14). First, as recommended by vom Brocke et al. (2015), the literature search process is documented as accurately as possible to facilitate future research on this topic. Then, the literature found is summarised in a concept matrix according to Webster and Watson (2002) and examined according to specially selected criteria. Based on this examination, research gaps are identified, and finally, the research question for this thesis is formulated.

3.1 Search Process

According to vom Brocke et al. (2015), in order to find relevant literature on the research areas dealt with, the topic is divided into separate concepts. These concepts help to find literature in scholarly databases using keyword search. The keywords searched for in this thesis were '*fantasy football*', '*machine learning*', '*prediction*' and '*betting odds*'. The keywords were entered in every existing combination to find articles that do not correspond to all keywords. Based on the research of Gusenbauer (2019), *Google Scholar* and *Microsoft Academic*, the most extensive academic search engines, were used for the literature search. When selecting the results from this search, attention is paid to the currently awarded VHB journal rankings (see e. V., 2015) for the sub-field of business informatics to ensure that the literature researched is of high quality. This ranking is chosen because it is well-known and accepted in the German research area. One journal that would be less considered following this ranking, but seems extremely relevant to the research in this thesis, is the *Journal of Quantitative Analysis in Sports* (JQAS). This journal gets published by the American Statistical Association (ASA), which according to themselves, '*is the world's largest community of statisticians*' (see "About ASA", n.d.). Using the papers from the JQAS and journals highly ranked by the VHB, the remaining literature is found using backward search and forward search suggested by Webster and Watson (2002).

3.2 Concept Matrix and Criteria

In the process mentioned above, 22 papers were examined and compared in a concept matrix (see table 2 on page 16) as required by Webster and Watson (2002). The concepts used to examine the papers will be briefly discussed from left to right in this paragraph.

The year of publication, the VHB ranking, and the distinction in which form the paper was published serve to evaluate the quality of the literature. That is to ensure that primarily the most recent papers in renowned peer-reviewed journals were analyzed. The sport discipline helps to notice similar approaches in different sports. While sports differ, some are more related than others. The main idea behind this is that there may be viable approaches from a similar sport that would have been unconsidered otherwise.

During the research, to the best of my knowledge, no publication was found which deals precisely with the problem at hand. For this reason, the research had to focus on similar approaches, objectives, or tasks. The solving approaches vary from more straightforward approaches such as mixed integer programming to more complex multi-hierarchical Bayesian models. Some publications used a combination of several methodologies, which are strongly dependent on the task to be solved. A distinction was therefore made between optimization and prediction tasks. Although almost all papers unanimously had the goal of setting up a team that would score as many points as possible, they came at the solution differently. The matrix distinguishes between publications that optimized only the team performance as a whole and those that predicted the performance for each individual player and then combined the best players into a team. At the same time, it investigated which papers relied on betting odds or another form of prediction markets. Lastly, the data used in each publication was analyzed. Due to the always different data, a generalized view was applied, which examines whether time-series data is used, whether the home advantage was taken into account and whether betting odds were used.

The articles are sorted by criteria in the following order: '*VBA Rank*', '*Machine Learning Approach*', '*Neural Network*', '*Individual Performance*', '*Betting Odds*'. This sorting ensures that the papers that are most similar to the thesis at hand and at the same time have a high VBA Rank are displayed first. For comparison purposes, the thesis at hand can be found at the bottom of the matrix. In this way, it can be quickly recognized that no publication deals precisely with the problem of the thesis. The paper that is closest to the topic is the paper by Landers and Duperrouzel (2017), even if it investigates football instead of soccer.

Table 2: Concept Matrix

| Paper | Published In | | | Solution Approach | | | | Task | | Solution Objective | | | Key Features | | | | |
|--------------------------------|--------------|----|-------|-------------------|------------|-----|-------------|------|----------|--------------------|------------|------------|--------------|---------|------------------|----------------|--------------|
| | Journal | CP | Other | VBA-Rank | Sport | MIP | ML-Approach | NN | Bayesian | Optimisation | Prediction | Individual | Team | Betting | Time-Series Data | Home-Advantage | Betting Odds |
| Landers and Duperrouzel (2017) | X | | | B | football | | X | | | X | X | X | | X | X | X | X |
| Lutz (2015) | | | X | n.R. | football | | X | X | | | X | X | | | X | | X |
| Deng and Zhong (2020) | X | | | n.R. | soccer | | X | X | | | X | | X | | X | X | X |
| Wheatcroft (2020) | X | | | JQAS | soccer | | X | | | X | X | X | X | X | X | X | X |
| Shah et al. (2021) | | X | | n.R. | soccer | | X | | | X | X | X | | X | X | X | X |
| Spann and Skiera (2009) | X | | | n.R. | soccer | | | | | X | X | | X | | X | | X |
| Anik et al. (2018) | | X | | A | cricket | | X | | | X | X | X | | | X | X | |
| Becker and Sun (2016) | X | | | JQAS | football | X | X | | | X | X | X | | | X | | |
| Goldstein et al. (2014) | | X | | n.R. | soccer | X | X | | | X | X | X | | | X | | |
| Egdi and Garby (2018) | X | | | JQAS | soccer | | | X | | X | X | X | | | X | | |
| Bonomo et al. (2014) | X | | | n.R. | soccer | X | | | | X | X | X | | | X | X | |
| Matthews et al. (2012) | | | | n.R. | soccer | | | X | | X | X | X | | | X | | |
| Skinner and Guy (2015) | X | | X | n.R. | basketball | | | X | | X | X | X | | | X | | |
| Pappalardo et al. (2019) | X | | | B | soccer | | X | X | | X | X | X | | | X | | |
| Yurko et al. (2019) | X | | | JQAS | football | | X | | X | X | X | X | | | X | | |
| Demediuk et al. (2021) | | X | | n.R. | e-sports | | X | | X | X | X | X | | | X | | |
| Edwards (2018) | | X | | n.R. | football | | | | | X | | X | X | | X | | |
| Karhnik et al. (2021) | | X | | C | cricket | X | X | X | | X | X | X | X | | X | | |
| Belien et al. (2017) | X | | | n.R. | cycling | X | | | | X | | X | X | | X | X | |
| Rein and Memmert (2016) | X | | | n.R. | soccer | | X | | X | | | | X | | | X | |
| Nevill and Holder (1999) | X | | | n.R. | soccer | | X | | | X | X | X | X | | X | X | |
| Thesis at Hand (2021) | | | X | n.R. | soccer | | X | X | | X | X | X | | X | X | X | X |

CP = Conference Proceeding, MIP = Mixed Integer Programming, ML-Approach = Machine Learning Approach, NN = Neural Network, Individual = Individual Performance, Team = Team Performance

3.3 Concepts

3.3.1 Features

The following paragraph summarises various concepts that have been frequently discussed in the presented literature. These topics are presented in alignment with the data mining process.

The first concept discussed is the *preprocessing of the data*. In order to achieve optimal results, the data mining process must adjust the data in advance without compromising its validity. Many of the authors tackle the problem that few exceptional players outperform the average players. These outliers are firstly hard to predict and secondly degrade the prediction accuracy. To solve this problem, the authors used various techniques to boost their prediction accuracy. For example, Landers and Duperrouzel (2017) developed a calculated threshold that players must reach at least to be included in the analysis. All players below this threshold are sorted out. At this point, it should be mentioned that it is crucial to choose a threshold instead of a point range, as in this case, the players with the highest points are not omitted. This approach can only be applied if data from previous games are available. Egidi and Gabry (2018), Lutz (2015), Yurko et al. (2019) focus in their papers on what to do if this data is not available. Yurko et al. (2019) state that one major problem they could not solve that negatively influences the team performance is the uncertainty of players appearing in the lineup due to unpredictable events with no evidential data like injuries. Lutz (2015) investigates the case of new players who joined at the beginning of the season ('new joiners'), as these players naturally do not have previous game data. He suggests taking the mean points of all players on a similar position in this case. (cf. Lutz, 2015, p. 3) In contrast to that, Egidi and Gabry (2018) take a different approach. In their paper, they compare two different solutions to this problem. On the first try, they put the expected points to zero, and on the second try, they guessed the points in a calculated range. In both cases, the processed points from the player often were too low to be considered for their starting lineup. Nevertheless, they find out that the second approach is more precise and improves their models overall. Furthermore, Egidi and Gabry (2018) discover that simplifying the data, if more details do not add value, increase their models' accuracy as well. This is similar to the approach Deng and Zhong (2020) take. In their studies, they use the *Kaggle European Soccer Database*, a table with

a total of 144 attributes, wherefrom they only carefully select 28 attributes to improve the model.

This links to the second concept, the *feature selection*. As mentioned, Deng and Zhong (2020), Egidi and Gabry (2018) reduce the attributes fed to their model to increase accuracy. In his paper, Lutz (2015) examines precisely the question of if and how far the number of attributes must be limited. He proceeds in three different ways. First, he does not exclude any features, figuring out that this approach is the least accurate. Secondly, he selects the features manually according to his assessment. Lastly, he chooses a more analytical path: *Recursive Feature Elimination with Cross Validation* (RFECV). This method '*recursively eliminates features and checks if the regression method's results improve by cross-validating.*' (Lutz, 2015, p. 4) This calculated approach yields the highest prediction accuracy. This method in combination with *univariate selection* was used by Anik et al. (2018) as well. One key feature that is discovered in this way is the position of the players. Demediuk et al. (2021), Egidi and Gabry (2018), Lutz (2015) all increased their accuracy by modeling each position separately. Similar to Lutz' second manual approach, Deng and Zhong also select their features based on their perception and note that '*sufficient background knowledge of the practical application is essential.*' (Deng & Zhong, 2020, p. 4). That confirms the discovery of Rein and Memmert, who claims that at the current state of research, '*most [Machine Learning] soccer analyses are performed by computer scientist research groups with little apparent involvement by sports scientists.*' (Rein & Memmert, 2016, p. 6).

From these researches could be inferred that it is beneficial to interview sports experts on their opinion on essential features if manual feature selection is used. However, if this is not possible, feature selection algorithms should be applied. In addition, the models could be even further improved by omitting players and features that offer little added value for the predictions. Each position should thereby be modelled separately. Finally, missing data can be dealt with in three ways: setting it to zero, giving it a mean value from similar players, and estimating it accurately. The latter is promising the most success.

3.3.2 Machine Learning Approaches

Once the feature selection process is complete, the next step, respectively the third concept, is to select the right *machine learning approach*. First, as in the concept matrix, a distinction must be made between optimization and prediction tasks. Only two different methodologies were chosen for the optimisation task: either brute force optimisation (Landers & Duperrouzel, 2017) or mixed-integer programming (Becker & Sun, 2016; Beliën et al., 2017; Bonomo et al., 2014; Edwards, 2018; Matthews et al., 2012). Which of these two methods is used depends primarily on how much computing power is required for the previous task.

For the prediction task, a variety of methods are used that range from simple linear regression to more complex feed-forward deep neural networks. In the following, the focus is limited to the three most frequently used and most promising methods: *Gradient Boosted Decision Trees (GBDT)* (Deng & Zhong, 2020; Landers & Duperrouzel, 2017), *Random Forest* (Demediuk et al., 2021; Deng & Zhong, 2020; Karthik et al., 2021; Shah et al., 2021) and *Deep Neural Networks (DNN)* (Deng & Zhong, 2020; Karthik et al., 2021; Landers & Duperrouzel, 2017; Lutz, 2015; Skinner & Guy, 2015). In the paper of Deng and Zhong (2020), all of the previously mentioned methods are used and compared. However, only a 'simple' Decision Tree model is used instead of a GBDT model. As a criterion for their Decision Tree, they use the 'information entropy', which is '*a mathematical measure of the degree of randomness in a set of data, with greater randomness implying higher entropy and greater predictability implying lower entropy.*' (Deng & Zhong, 2020, p. 4) They note that while Decision Tree models compute faster and require fewer data processing than Random Forest models, for this reason, they are more prone to over-fitting as the number of data increases, making them less accurate in general. The DNN they create consists of '*5 fully-[connected] dense layers, five activation layers, two dropout layers and one batch normalization layer.*' (Deng & Zhong, 2020, p. 4) They apply the *Softmax* function to transform the data and use *sparse categorical cross entropy* as base for the model's loss. The batch size is set to 32, and the model trains 500 epochs. According to their research regarding prediction accuracy, the DNN is the most accurate (0.99), followed by the Decision Tree model (0.91) and the Random Forest model (0.84), with the DNN probably being over-fitted with an accuracy of 0.99.

Landers and Duperrouzel (2017) use a GBDT model in their studies in which they want to predict the individual player performance for each player in the *National Football*

League (NFL). They test the team their model predicts against 300.000 randomly selected teams and achieve the highest scores in five of eleven weeks. In addition, they let their model predict the 100 best team constellations and thereby manage to get into the profit range of the 20th percentile in 68% of the cases. Unfortunately, they do not provide further information on their model but again emphasize the variety of well-thought and self-engineered features they use. Furthermore, like Deng and Zhong, they also agree to the straightforward implementation of Decision Trees, as it is not necessary to normalize or scale the features. (cf. Landers & Duperrouzel, 2017, p. 6)

In the researches from Shah et al. (2021), they attest the Random Forest model to produce the best results for their problem. They compare four different approaches to calculate the expected rate of goals. The calculations are based on: previous goals, expected goals from prediction markets, Linear Regression and Random Forest. To compare their models, they use the *Brier Score*, 'a score function that helps determine the accuracy of any probabilistic model.' (Shah et al., 2021, p. 7) Another implementation of the Random Forest algorithm is used by Demediuk et al. (2021), who calculate a so-called '*Performance Index (PI)*' for each player during a e-sport game of Dota2. Here, the algorithm is used to predict the chance of winning the game based on real-time in-game data. This, later on, helps the final calculation of the PI. Depending on the length of the game, the Random Forest model predicts the correct winner with an accuracy of 0.55 to 0.8.

Of all the methods used in the literature reviewed, DNNs are the most commonly used. Similar to Deng and Zhong (2020), Karthik et al. (2021) benchmark their feed-forward DNN against Machine Learning algorithms like K-Nearest Neighbours (KNN) or Random Forest. Although their DNN, with an accuracy between 0.88 and 0.94, does not appear to be over-fitted, it also outperforms all Machine Learning models by a margin of at least 0.08. Their DNNs input layer has one neuron for each feature fed to the classifier. '*The model consisted of three hidden layers, each with 64, 32 and 16 neurons, respectively. Finally, the output layer consisted of 7 neurons [...]. A learning rate of 0.3 is used for training 500 epochs. A categorical cross-entropy loss function with sigmoid activation functions in hidden layers and a softmax activation function in the output layer is used for training the classifier. The basic hyperparameters [...] were empirically optimized using the grid search approach.*' (Karthik et al., 2021, p. 7) In contrast to this more complex DNN, Lutz (2015) uses a DNN with only one hidden layer and compares it to his results with *Support Vector Regression (SVR)*. The DNN with the best accuracy trained 50 epochs has 50 hidden units and uses the Sigmoid squashing function. This straightforward DNN

already outperforms his SVR model slightly. In his conclusion, he states that DNNs with multiple hidden layers could provide increased accuracy. (cf. Lutz, 2015, p. 5)

In summary, it can be concluded from the methods analyzed that there is no algorithm in this area of research that predominantly offers the best prediction accuracies. Instead, various methods must be experimented with and adapted precisely to the problem at hand. Nevertheless, the relevant literature shows methods that promise more success than others, which should be specially addressed for this reason. These methods include Decision Trees and Deep Neuronal Networks.

3.3.3 Betting Odds

The final concept discussed in this section is the influence of *betting odds* in current researches in this area. The impact of values that want to predict the future can already be observed in the literature reviewed. Landers and Duperrouzel (2017) for example, want to predict the winning team against the spread, based on historical spread betting data. Shah et al. (2021) use a metric called 'expected goals', which indicates how many goals a soccer team will score according to the participating bettors. Although Deng and Zhong (2020) do not explain any further how they use or process the betting odds in their dataset, they claim to feed them to their models. In his paper, Wheatcroft (2020) investigates the overreaction of soccer betting odds in mismatch to the underlying reality. He, therefore, explains ubiquitous biases in sports betting, like the home-underdog bias and the contrary away-favourite bias. Although studies from Nevill and Holder (1999) show that the home advantage does exist, many of the papers discuss how many influences the home advantage has. (Bonomo et al., 2014; Deng & Zhong, 2020; Landers & Duperrouzel, 2017; Shah et al., 2021) In his studies, Wheatcroft (2020) shows that there is a tendency to overestimate the influence of the home advantage. Furthermore, he defines a nominal statistic called '*combined odds distribution*' (COD) which indicates '*the performance relative to expectations of a team in its previous matches*' (Wheatcroft, 2020, p. 4). If the COD is above 0.5, the team performed better than predicted by the odds and vice versa. From this statistic, he infers that the hot-hand bias exists in soccer betting odds, where an event seems to have a higher probability if it occurred recently in the past. In addition, he explains that even though algorithms are generally considered not to be biased prone, they still are created by biased humans.

In their studies, Spann and Skiera (2009) compare three different forecast methods, namely prediction markets, tipsters and betting odds. They discover that prediction markets and betting odds are more accurate than expert opinion. This discovery is in contrast to the results of Goldstein et al. (2014), who figured out that the prediction accuracy of smaller, smarter crowds tends to be higher than the general swarm intelligence. However, to obtain the highest prediction accuracy according to Spann and Skiera (2009), all three forecasting methods should be combined.

In the end, future predicting values such as spread bets, betting odds or prediction markets have already been used in the literature but not yet as planned in this work. Furthermore, it could be shown that almost every form of using these values promises general success.

In addition, it was proven that even though these values are not free of biases, they still offer added value in forecasting.

3.4 Research Gap

In the following concluding section of this chapter, as requested by Webster and Watson (2002), the research gap found and the research question resulting from it will be addressed. Some of the authors in the presented literature state that *'investigations into these [fantasy sports] games in the academic literature are virtually nonexistent'* (Landers & Duperrouzel, 2017, p. 1) or *'the prediction of Fantasy Football has barely been studied'* (Lutz, 2015, p. 1). Rein and Memmert (2016) even claim that all main characteristics of big data implementations are highly relevant and provide specific solutions to address tactical analytics in elite soccer. In addition to that, as already described in the previous paragraph, future predicting values, especially betting odds, seem to offer a significant value in increasing prediction accuracy. Although Deng and Zhong (2020) include betting odds in their studies, it is only one of many features, and they only aim to predict team performance instead of individual player performances. In contrast, this thesis closely observes this influence, taking the results of Wheatcroft (2020) and Goldstein et al. (2014) into account.

The central research question is based on the question: *'How accurately can individual soccer player performances be predicted using historical data?'*. After this question is resolved; the central research question can be answered, which is:

'How accurately can individual soccer player performances be predicted using historical data and betting odds?'

Chapter 4

Implementation

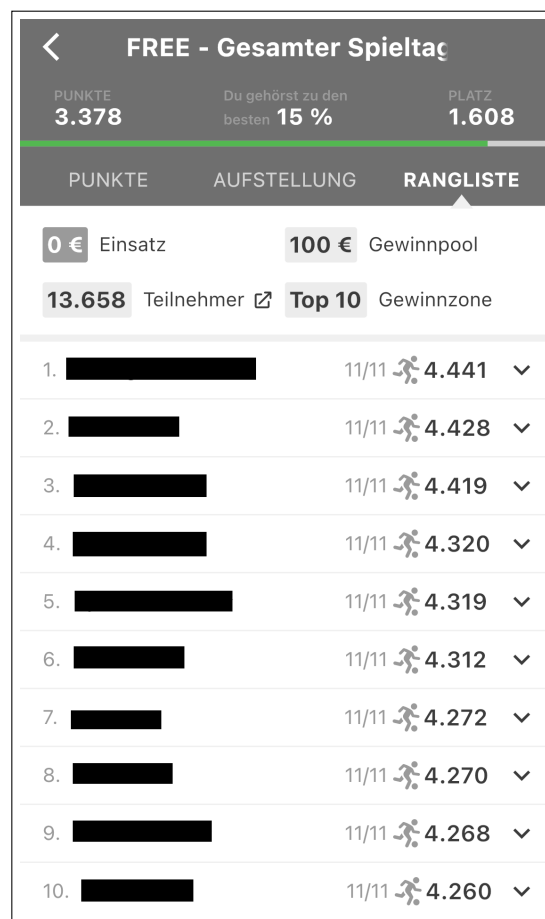
As already mentioned in chapter 2, the task Fantasy League managers face is to set up an optimal line-up. To find out this optimal set-up, indirect anticipation of the future is always useful. As can be seen from the literature review in chapter 3, many attempts have already been made to make this anticipation no longer manual and based on random factors. Quite the contrary, through the combination of big data and sport (Rein & Memmert, 2016, cf.), it is now possible to use the data collected in the past to make automated predictions for the future. A promising approach to this type of challenge is machine learning. In order to approach this machine learning problem methodically, the book *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* by Géron is used as a guideline. The implemented methodology results in a *Python Jupyter Notebook*, in which it is analysed whether and with which models this problem can be solved. In the first step, Géron advises to think fundamentally and to answer basic questions about the problem. This is done in the following section.

4.1 Business Context

Based on the rules explained in section 2.2, two main findings emerge. The first point is the main objective of the game: assemble a team of 11 players that will score as many points as possible on the upcoming match day. The second point is that the constraint placed on the players, the budget, can be converted to a player's total score. Regarding the first point, it is necessary to distinguish at what point the main goal of the game is

achieved. There are three different angles to approach this: if only the problem itself is considered, the goal would be to put together the team that scores highest. Secondly, from a game perspective, it would be enough to field the best team of all competitors, i.e., to place first in the final ranking. Lastly, a purely economic objective would be to put together a team that makes a profit by ending in the profit zone at the end of the game.

Past rankings show that even the first place of a matchday never achieves the whole number of points. Instead, the top ranks only reach 80 to 85% of the maximum possible score. This paragraph aims to present an example for this claim. Therefore, the following screenshots from the 8th matchday of the 2021/22 Bundesliga season are evaluated. The first screenshot shows the top 10 ranking for this matchday. This screenshot indicates that the scores of the top managers differ only slightly.



| FREE - Gesamter Spieltag | | | | | |
|--------------------------|--|--------------------------|--|-----------|---|
| PUNKTE | | Du gehörst zu den besten | | PLATZ | |
| 3.378 | | 15 % | | 1.608 | |
| PUNKTE | | AUFSTELLUNG | | RANGLISTE | |
| 0 € Einsatz | | 100 € Gewinnpool | | | |
| 13.658 Teilnehmer | | Top 10 Gewinnzone | | | |
| 1. | | 11/11 | | 4.441 | ▼ |
| 2. | | 11/11 | | 4.428 | ▼ |
| 3. | | 11/11 | | 4.419 | ▼ |
| 4. | | 11/11 | | 4.320 | ▼ |
| 5. | | 11/11 | | 4.319 | ▼ |
| 6. | | 11/11 | | 4.312 | ▼ |
| 7. | | 11/11 | | 4.272 | ▼ |
| 8. | | 11/11 | | 4.270 | ▼ |
| 9. | | 11/11 | | 4.268 | ▼ |
| 10. | | 11/11 | | 4.260 | ▼ |

Figure 1: Top 10 Ranking FREE Pitch, 8th matchday 2021/22

Additionally, the following screenshots show the line-ups of the best three managers for this matchday.



Figure 2: Line-Ups of Top 3 Managers from 8th matchday 2021/22

A line-up can already be assembled from the players of the best three managers, which is far better than rank one on this matchday. The best line-up from the presented players would have been:

Formation: 5-2-3

Captain: Gnabry

Lewandowski (586), Kramarić (476), Müller (419)

Gnabry (1312), Hofmann (521)

Oxford (494), Ginter (478), Mavropanos (472), Elvedi (381), Schlotterbeck (348)

Bredlow (385)

Budget needed: €198.2m (-385.6 Manager Score)

Total Score: **5,486**

Therefore, rank 1 in the competition reached 81% ($4,441 \div 5,486$) of the score, which confirms the statement made earlier. Even though this is only one example, similar proportions of the scores can be seen on any other matchday.

Furthermore, it is important to notice that there are no well-known managers who regularly achieve top ranks. According to the observations made so far, it seems challenging to field the best possible team, let alone to achieve first place regularly. A probable explanation for this is the strong influence of luck due to things that are not predictable, such as injuries. Therefore, it is much more realistic to pursue the last-mentioned goal, to end in the winning zone regularly. The previous rankings reveal that the score needed to reach the winning area is lower, between 60 and 65% of the maximum possible score. For these reasons, this work aims to write a model that regularly achieves 65% of the best possible score.

Regarding the second point, the best possible score is achieved by the line-up with the highest adjusted final line-up score S_{LM} . To create this line-up, the eleven players with the highest adjusted player score S_{P_iM} , which can be placed in one of the available formations, must be found. In addition, the player with the highest player score S_{P_i} must be appointed as captain. As the score of a player of the upcoming match day can be seen as a label for the prediction, this is **supervised machine learning**. Additionally, as the target to predict is a numerical value, this is a **regression problem**. Since the transfer market values and all possible line-ups are already known in advance, only the final score of each player has to be predicted. From another point of view, if the prediction of the final points of each player had an accuracy of 100%, the best line-up could be calculated automatically. Therefore, it can be concluded that the model, which is supposed to achieve a regular 65% of the best possible score, can be divided into two parts. In the first step, a machine learning model predicts the score for each player. In the second step, the best line-up is calculated based on these scores. Since no machine learning model is needed for the second step, the goal of the work can be specified further. The updated goal is to write a model that predicts upcoming individual player performances as accurately as possible to enable the system to achieve a line-up score of 65% of the best possible score.

The literature review shows that in the context of predicting individual player performances, future predicting metrics such as betting odds have not yet been investigated. In other contexts, these kinds of metrics have already helped to improve predictions. (cf. Landers & Duperrouzel, 2017) Moreover, these values are generally considered to have a high potential in this regard. (cf. Goldstein et al., 2014; Wheatcroft, 2020) For this reason, the influence of betting odds will be examined more closely in this thesis. To investigate this influence, the machine learning models will first predict the players' final

scores without the betting odds features. These models are called **baseline** models. Then, the models will again predict the final scores of the players with the betting odds features. These models are called **treatment** models. Finally, it can be examined which models provided the more accurate predictions and consequently which of the models compiled the better line-up as a result.

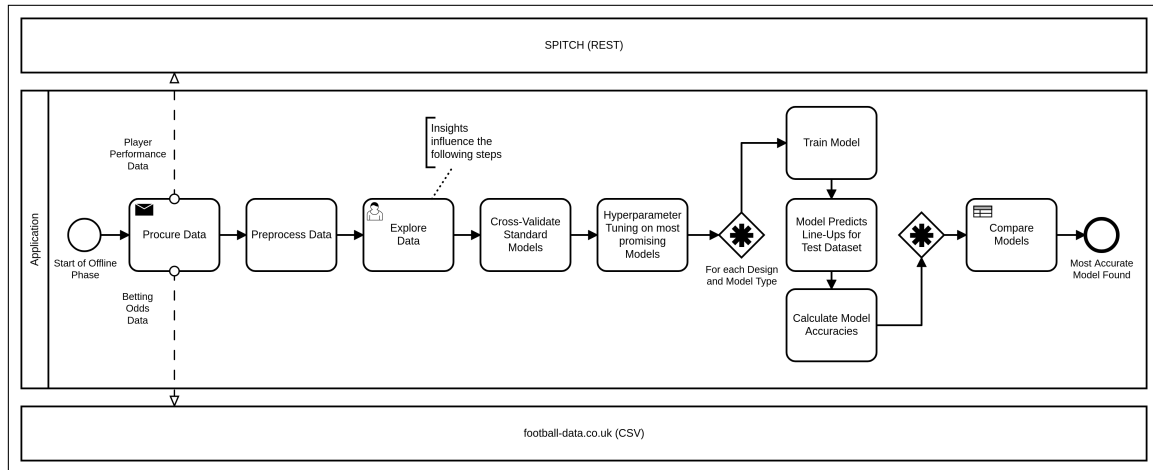


Figure 3: Offline Phase Process

From these investigations now presented, a machine learning model will emerge that provides the most accurate predictions. This model is then implemented in a tool that, prior to a matchday, collects the current and required data and feeds it to the model. Based on the model's predictions, the best line-up for the upcoming match day is created and provided to the end-user. The creation of the tool is therefore divided into two phases: an offline (see figure 3) and an online phase (see figure 4). First, the offline phase takes place, in which the model with the most accurate predictions is found. This requires a lot of historical data. The tool can then be used in the online phase. In this phase, the next predicted line-up can be queried. Current data is used for the prediction. In *Big Data* terms, the models in the offline phase are so-called batch processing models, while the model in the online phase is a stream processing model.

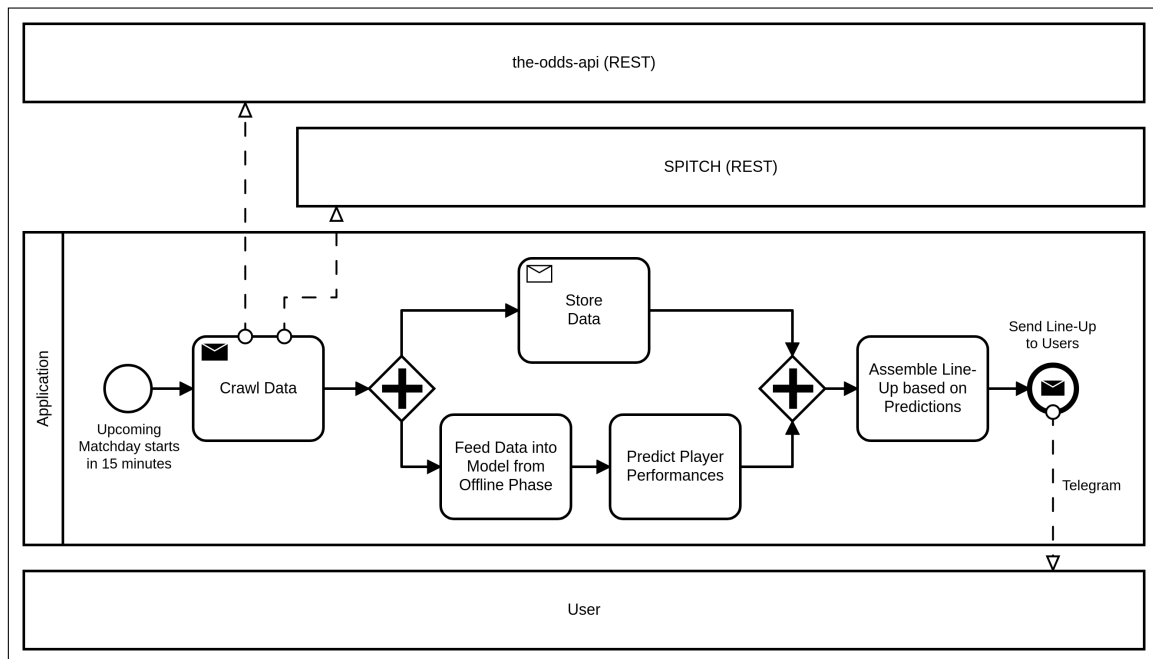


Figure 4: Online Phase Process

As already mentioned at the beginning of this section, SPITCH is a competition in which, on the one hand, even the best-placed players take turns and, on the other, almost never achieve the highest possible score. These two attributes are characteristics of gambling. Furthermore, football is complex and has many influencing factors that cannot all be taken into account. The betting odds try to take some of these factors into consideration, which is why they should probably achieve better results. However, it may also be that the bettors' assessment of bias is inferior. For all these reasons, it can be assumed that it will probably be difficult to achieve the set target.

4.2 Data

The cornerstone of any machine learning project is the underlying data. In the following chapter, the three essential steps in dealing with data that took place during the thesis are highlighted. First, the data was obtained from various sources. Then the data was processed so that it could be brought into relation with each other. Finally, the data was analysed to gain insight into the characteristics of the data and to draw implications for the machine learning models.

4.2.1 Procurement

As mentioned in the previous section, two types of models will be compared: baseline and treatment models. While the baseline models only use the data provided by SPITCH, the treatment models additionally use betting odds from other sources. For this reason, the data needed for this project is divided into two groups: the SPITCH data and the betting odds. Furthermore, it is stated in the previous section that the final tool is divided into two phases: a batch processing offline phase and a stream processing online phase. These classifications end in four different sources needed to implement the tool: SPITCH data and betting odds for each of the online and offline stages.

The tool's architecture is designed after the microservice principles. The microservices are described throughout the *Implementation* chapters, and the final architecture is again summarised in the last section of this chapter. The tool and with it each associated microservice is launched on a virtual private server (VPS) hosted on *Hetzner*. (see Hetzner, 2021) The core of the architecture is the PostgreSQL *Database* microservice, which stores all of the data mentioned in the following paragraphs.

SPITCH Data

The SPITCH data derives from the same source regardless of the ongoing tool's phase. The data is crawled directly from their API-endpoint <https://api.spitch.live> from the *Crawler* microservice. The *Crawler* therefore executes a script in a manually definable schedule. Additionally, the *Proxy* microservice is running. The requests sent from the *Crawler* are

piped over the *Proxy*. The *Proxy* then sends the request via a scaling number of different IP addresses (nodes) within the *Tor* network. This procedure bypasses the rate limits of the recipient's firewall, as the requests can no longer be clearly assigned to a single IP address, allowing the *Crawler* to send multiple requests per second. However, with programming ethics in mind, the requests per second still were set to a relatively small number to avoid over floating the endpoint.

SPITCH data refers to player-specific and event-based data. Player-specific data includes tables such as the *Player* table with information about their position, transfer market value, names, and team membership. The player data was crawled from the endpoint <https://api.spitch.live/contestants>. This endpoint sends all players and teams from the SPITCH database in the *JavaScript Object Notation* (JSON) format. The JSON is received, converted, and stored in two separate tables *Player* and *Team* in the own *Database*. This procedure took place once in the offline phase and takes place before each matchday in the online phase. The latter ensures that the player and team data is up to date for the model's predictions. During the conversion of the request, the transfer market value of each player is stored in a different table *Market Value* with an additional period of validity. In data management or warehousing terms, the table *Market Value* is implemented as *Slowly Changing Dimension Type 2*. This modification was made because the transfer market value is a feature of the players that changes frequently but whose history is helpful for later investigations. For this reason, information would be lost if the transfer market value were merely overwritten on update.

However, the core data is event-specific, stored in the *Event* table. Each event on the pitch during a match is stored in this table, represented by one row. Therefore, one row in the table interprets like: '*Player A has performed event B in minute C on matchday D.*' i.e. '*Manuel Neuer played a pass in the 35th minute of the 6th matchday.*' This data can only be gathered matchday- and player-specific, using the corresponding endpoints following the structure:

https://api.spitch.live/matchdays/matchday_id/players/player_id/events

Consequently, the *Crawler* first has to get all the player and matchday identifiers (id) to gather all the data provided by these endpoints. The tool is able to get all the player identifiers using the method described above in the player paragraph. Simultaneously, the *Crawler* has to get all the matchday identifiers by requesting the endpoint

<https://api.spitch.live/matchdays> and storing the received information in the table *Matchday*. If player and matchday identifiers are provided, the event data can be queried. The service then takes two extra checks before starting the crawling. First, it checks for the latest matchday in the *Event* table to prevent duplicate queries. This step is particularly essential in the online phase, as the tool is executed regularly on different occasions. Second, it proves which of the matchdays in the table *Matchday* have already occurred by taking the current timestamp and comparing it with the timestamps of the matchdays. This checkup ensures that no endpoints from matchdays that lay in the future are getting requested, leading to empty responses. After these two checks are made, the *Crawler* first gets and then stores all the available event data in the *Database*.

It is important to notice that SPITCH most likely uses an official API as well to gather this information, like <https://www.api-football.com/> for instance. For further studies, it would be beneficial to investigate to what extend the data from SPITCH matches with the data from such APIs since they are easier to request and provide much more data, which reaches far into the past. At the same time, however, it must be noted that more data could also lead to less accurate results, as the development of individual players over the years represents an additional and very complex influence. Only the connection between player data and betting odds could be examined more closely. In both cases, the amount of data now available is likely to be sufficient.

The following table serves as an overview of the various SPITCH-based tables in the *Database*, explaining how many records are included and how much memory they occupy.

Betting Odds

The betting odds data in this thesis are obtained in two different ways. Here, a distinction is made between the online and offline phases. In the offline phase, a comma-separated values (CSV) file is taken from the following website:

<https://www.football-data.co.uk/germanym.php>. In this CSV file are all betting odds for each game of a season for different betting odds providers. Furthermore, columns like the average betting odds from all providers are added. In the online phase, the data is taken from this API: <https://the-odds-api.com/>. In this phase, care is taken to request the odds as late as possible, as they often change before the start of the matchday given

to the most recent information. Here, no average betting odds column is provided and therefore needs to be calculated itself by the *Crawler* microservice. This data gets stored in the *Odds* Table of the *Database*. Each row thereby represents the winning, draw, and losing odds for a team for a matchday.

4.2.2 Processment

After the data has been obtained in the previous section, it is processed and merged in this section. In the offline phase, these are the first steps in the Jupyter Notebook, while in the online phase, these steps take place in the *Prediction* microservice.

In the database query, the player-specific tables are joined with the *Event* table. This join converts the identifier columns into human-readable and understandable columns. For example, the name, club, and position are obtained from the player identifier. After all the event data in the database has been fetched and made comprehensible, this data must first be supplemented because the API only returns events that have indeed happened on the pitch. However, it could also be interesting to know that a player **did not** play, for example, a single pass on a matchday. This information is currently not yet available in the data set and will be added in the following step. Therefore, the points catalog from SPITCH (see SPITCH, 2021a) has to be loaded, which contains all events that can happen in SPITCH. After that, the events can be grouped by player and matchday. For each group, the points catalog gets iterated over. If there exists no corresponding entry in the group for an event type, an entry is added with the occurrence of zero. After this step, the event data is ready to use.

In the Jupyter Notebook, the betting odds data gets read from the CSV file mentioned in the previous section. From the data obtained, only the relevant columns get extracted, which include the date, name of the home and away team, and the average of the betting odds. Each row in the dataset is equal to a match between two teams. In order to assign the correct match day to these matches, the date column must first be converted into a *datetime* object. Now the *Matchday* table in the *Database* can be used to check to which matchday the date belongs. After this step, the rows can be split, resulting in two rows, with each row being equal to betting odds for one team for one matchday. Therefore, the odds for the home team are renamed in odds to win for the home team and vice versa

for the away team. Additionally, the column *is_home* is added to investigate the **home advantage** mentioned often in the literature review.

In the online phase, the data was already written to the database in this format by the *Crawler*. This is why the data only needs to be read from the database by giving the upcoming matchday.

After the event data and the betting odds have been converted into the desired form, they can be merged. The key with which the two data sets must be linked contains the matchday and the team. Thereby the following problem occurs: while the team data obtained by SPITCH is in German format, the betting odds data is in International format. Since the names of the teams in the respective data sets do not match, each International formatted name must first be assigned its German equivalent. Furthermore, as it is best practice to always join via identifiers, the German identifier from the database is assigned to each international name. The Python library *smart-match* is used to achieve this. Smart-match calculates the similarity between two entered strings using various methods such as the *Levenshtein distance* or the *Smith-Waterman* algorithm. All methods were evaluated and compared with each other, and finally, the *Smith-Waterman* algorithm was chosen, as it produced by far the highest similarity values. The algorithm iterates over each international name in the betting odds dataset, checking the similarity with each German name using the *Smith-Waterman* formula and permanently storing the name and the identifier for the highest similarity. In this way, each row in the betting odds dataset has a team identifier assigned, which makes it possible to merge the two datasets.

With the help of the SPITCH points catalog and the individual events for each player, the player score can now be calculated for each match day. It is important to mention that the transfer market value of a player is not deducted from the score, as the models should only concentrate on the points actually achieved. In this case, the transfer market values would only blur the scores. To calculate each player's individual score, according to equation (2) on page 10, each number of appearances is multiplied by the corresponding point value per event, and these products are then summed. In the resulting dataset, each row represents the player's score for a matchday together with his position, team, and betting odds.

Through this, an additional feature can be implemented utilizing the individual player score: the player *performance trend*. This feature takes the recent scores of a player and

calculates a trend score based on them. How differently the historical values had to be weighted was investigated in the course of the thesis. Several methods were compared: the Simple Moving Average (SMA), the Cumulative Moving Average (CMA), and the Exponential Moving Average (EMA). In addition, different parameters were used for the SMA and EMA: the SMA was calculated with three, five, and ten historical values and the EMA with an α of 0.1, 0.3, and 0.5. The SMA sums up the n past values and divides the total by n , resulting in each past value having the same weight. The EMA, however, weights more recent values by the exponential factor α . From these investigations, the EMA with an α of 0.5, also known as the half-life, performed best in predicting future performances merely based on recent values. Therefore, the player performance trend column is created by calculating the EMA with an α of 0.5 based on the individual player score column.

According to Géron’s methodology and the intention of this work, several machine learning models are compared with each other. For this reason, the data now available must be made equally feasible for each model. Because of this, all numerical features must be normalized and all categorical features encoded. The features *player performance trend* and all betting odds-specific columns are normalized using a technique called *Min-Max-Scaling*. With this technique, each feature X is scaled in the range between zero and one using the following formula:

$$X_{\text{SC}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (7)$$

The categorical features *team* and *position* are encoded using the technique *One-Hot-Encoding*. This technique creates a boolean column for each instance of a categorical feature. Thus, for example, the position column becomes four columns according to the pattern *is_goalkeeper*, *is_defender*, and so on. By converting the data in this way, the final data set put into the models has 31 columns.

After the data has been merged and made human-readable in the previous paragraphs, a distinction must be made between two different data sets. The first data set contains all event types with the number of occurrences for each player and each matchday. The second data set, in contrast, only includes the calculated final score for each player for each matchday. Thus, there are two ways in which machine learning models could be

used to predict player performance. The first option would be to predict for each player the number of occurrences for each event type. Then the final score can be calculated based on the predicted values. The second option is to predict the final score directly. Since a proportional relationship between the events and the final score can be established through the points catalog, only the influence of the features is decisive for which variant is chosen. To further explain the variables, the following figure 5 shows the dependent and independent variables and their connections to each other. Thereby, the influences of the added features are marked with a *plus* for a strongly predicted effect and a *tilde* for a moderate impact.

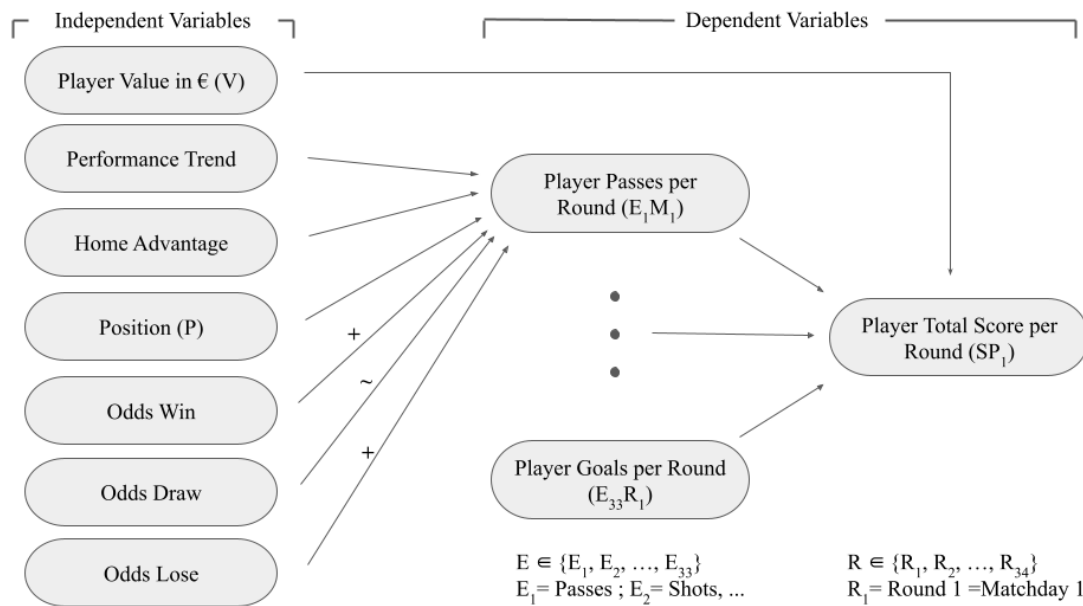


Figure 5: Effects between Dependent and Independent Variables

4.2.3 Exploration

As already mentioned in the previous section, two data sets exist which can be interpreted. First, this section examines all features that are present in both datasets. Finally, the different features are examined to finally decide on a dataset to be used for the models.

The following table describes all the features that appear throughout the datasets, their description, and datatype. In *Python*, *Strings* are stored as objects.

Table 3: Overview of Features and their Datatype

| Feature | Datatype | Description |
|-------------------|----------|---|
| name | object | name of a player |
| position | object | position of a player, i.e. goalkeeper, attacker |
| matchday | int64 | matchday where the results happened, i.e. 1, 2, 34 |
| team_name | object | name of the team in which the player competes |
| is_home | bool | boolean indicating whether the game was played at home or not |
| odds_win | float64 | odds for the team to win |
| odds_draw | float64 | odds for the team to draw |
| odds_lose | float64 | odds for the team to lose |
| event_type | object | event type, i.e. goal, pass, unsuccessfulTackle |
| occurrences | int64 | number of occurrences of event_type for this matchday |
| performance_trend | int64 | recent performance trend for a player |
| score | int64 | individual player score S_{P_i} |

Each feature is first analyzed individually and then in combination with each other. This analysis is mainly done in the order of the table, starting with the dataset containing event types and their occurrences.

Name

The name of a player is synonymous with the player itself in this context. As can be seen from the following figure, there aren't identical numbers of entries for each player. This deviation should not be the case because, in the previous section, an entry was assigned to each player for each matchday and event type. The reason for this difference is that there are entries for the event types **player_on** and **player_off** in the data set, which is not found in the points catalog from SPITCH and has no influence on the score. For this reason, these events can be ignored.

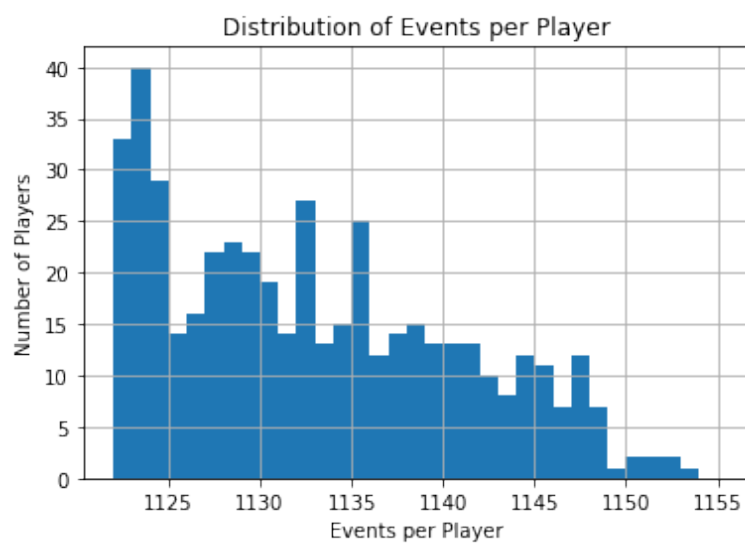


Figure 6: Distribution of Events per Player

After deleting the two event types mentioned above from the data set, there are 1,122 rows for each of the 467 players. This is equivalent to 33 different event types on 34 matchdays. Therefore, the data set has a total of 1,122 rows times 467 players, i.e. 523,974 entries.

Matchday, Event Type, and Home Advantage

Accordingly to the calculation from the previous section, for each matchday, there are 33 event types times 467 players, i.e., 15,411 rows, and for each event type, there are 34 matchdays times 467 players, 15,878 entries. Since in the German *Bundesliga* every team plays every team twice, once at home and once away, there are precisely the same number of rows for matches played at home and matches played away.

Position

From the figure 7 below, it can be seen that there are four different positions in the dataset, with the midfielder being the most common position with 193 entries out of 467 (41.3%) and goalkeeper being the rarest position with 35 entries (7.5%). This distribution is reasonable given that most formations have more midfielders than any other position, and there is always only one goalkeeper.

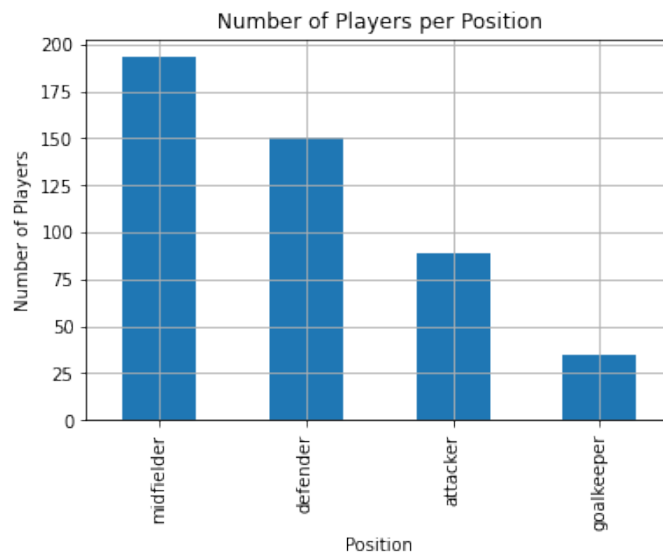


Figure 7: Number of Players per Position

Team Name

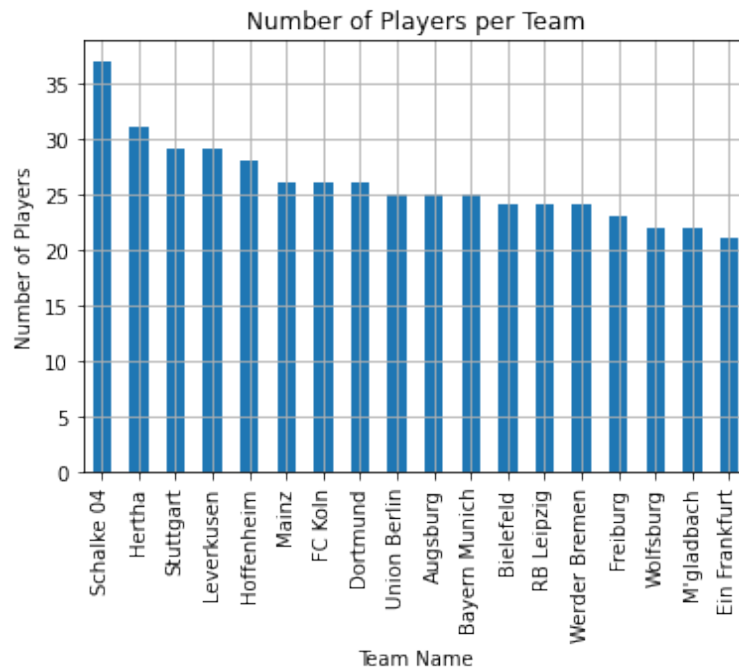


Figure 8: Number of Players per Team

The figure shows that Eintracht Frankfurt has the fewest players with 21. Schalke 04 has the most players with 37. The average number of players per team is 25.

Betting Odds

Many different betting odds systems exist. The betting odds used in this thesis follow the most popular model. In this model, the betting odds represent the factor by which the stake is multiplied if the event on which the bet was placed occurs. Through this system, bets can be placed on any game outcome with a chance of winning. The more probable the outcome, the lower the factor. This factor is influenced by various aspects such as the assumed playing strength of the teams, but also how many bettors have already bet on this outcome. Since every bookmaker has its own methods for this calculation, an average of 13 different platforms is chosen. This betting odds model has the advantage that it is lucrative for the bookmaker in every case, as he chooses the factors in such a

way that no matter what the outcome of the match, the stakes paid in exceed the profit to be cashed out. For this reason, these betting odds should be assessed with caution. However, they represent a relatively general and, above all, up-to-date picture of the teams' probabilities of winning. The figure below shows that the betting odds are mostly in the range between two and four for winning or losing. Anything above or below that range indicates a one-sided game. Here, 24.69 has been the maximum value, and the corresponding counterpart 1.09 was the minimum value. While the range for a draw is smaller, the factor itself is higher in general.

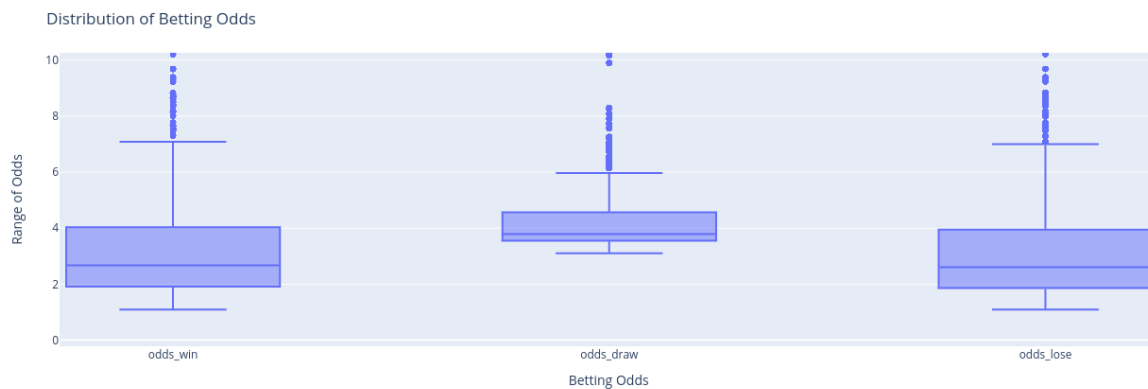


Figure 9: Distribution of Betting Odds

Occurrences

The bar chart below shows that most of the rows have a value of 0 (**87.3%**). The frequency monotonically decreases as one moves further away from zero. Therefore, the main task of machine learning models would be to predict situations where the result are not zero, which leads to a classification problem instead of the initial regression problem. As announced at the beginning of the section, a final decision has to be made in favor of one data set. In addition to the figure below, further investigations with this dataset, which cannot be presented in this thesis due to the large volume, showed that its use is unsuitable. For this reason, it was decided to use the data set that contains the calculated scores.

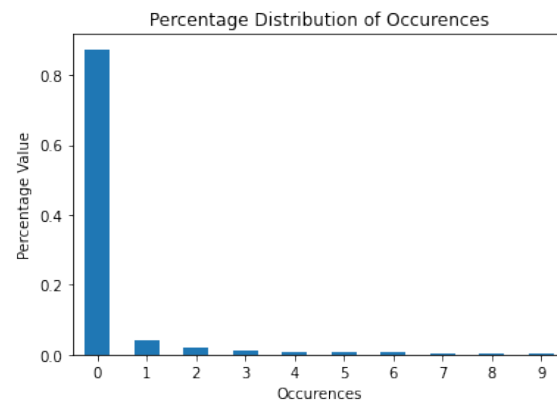


Figure 10: Percentage Distribution of Occurrences

Score

An interesting statistic in relation to the score, especially about the significance of the performance trend, is the autocorrelation. The autocorrelation indicates the extent to which previous values are predictive of future values of the same feature. Figure 11 shows an autocorrelation between 0.65 and 0.7 for the lags between -10 and 10. Additionally, the correlation slowly decreases as one moves further away from the actual value. This small decrease indicates that more current values give better predictions than past values. This conclusion confirms the choice of exponential moving averages for the calculation of the performance trend. Overall, an autocorrelation of 0.7 indicates a medium to strong relationship between past and current values.

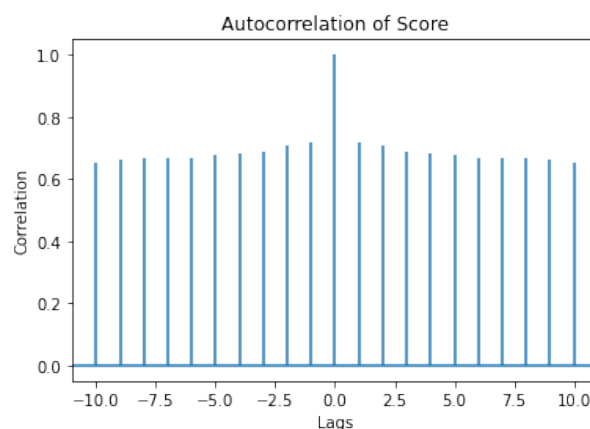


Figure 11: Autocorrelation of Score

Looking at the score, two groups of players are particularly interesting. The first group consists of players who rarely score many points and are consequently not drafted by many managers. The second group consists of players who score relatively confidently. The players in this group are, therefore, probably the most expensive because many managers draft them. Two different approaches are taken in the analysis of the data to separate the players in these two groups. To begin with, the players who belong to the first group will be examined by looking for the players with the highest average of points per game. Secondly, it is examined which players have scored the most points overall over the entire season. These are the players in the second group.

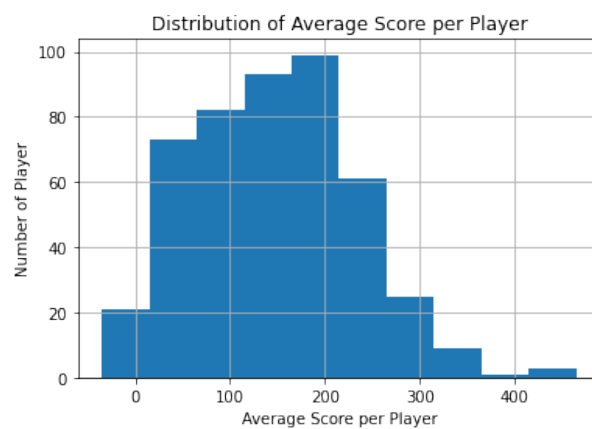


Figure 12: Distribution of Average Score per Player

Figure 12 shows the distribution of the average scores of all players. This figure indicates that most players receive an average score of around 200. There exists a small group of players who achieve an average score of over 300. Interestingly, more players achieve more than 400 points than players who achieve between 350 and 400 points.

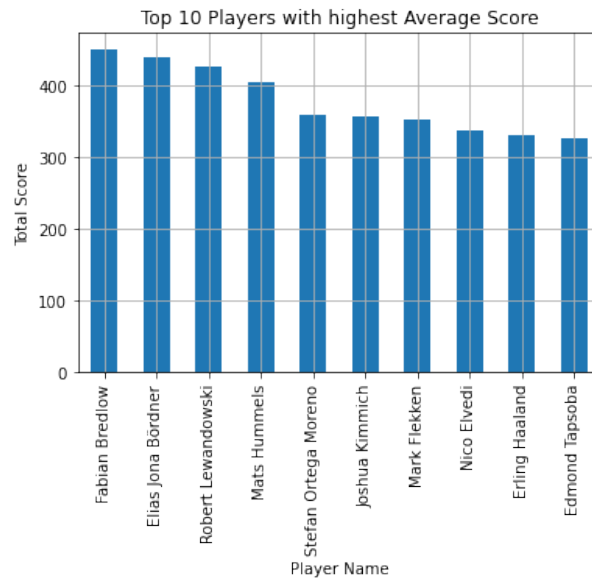


Figure 13: Top 10 Players with highest Average Score

The above figure displays the ten players with the best average score. Surprisingly, many players in this group are not well-known and therefore inexpensive. These are the players who need to be ideally lined up by the model because, on the one hand, they score a lot of points and, on the other, their low transfer market value means they don't have to earn a lot of points to be valuable. Nevertheless, there are also very well-known players, such as *Mats Hummels* or *Robert Lewandowski*. With these players, the models have to weigh up whether to put them in the line-up or not, as their high transfer market value means that they have to earn a lot of points first in order to recoup the manager points they have cost. These players are especially worth choosing as captain because the points scored are doubled, but their transfer market value deduction is not.

Regarding the second group of players, figure 14 shows that there are many players who have scored between zero and 1,000 points throughout the season. Given this and the prior observations, it can be concluded that many players in the data set are rarely fielded, but when they are fielded, they still achieve an average score.

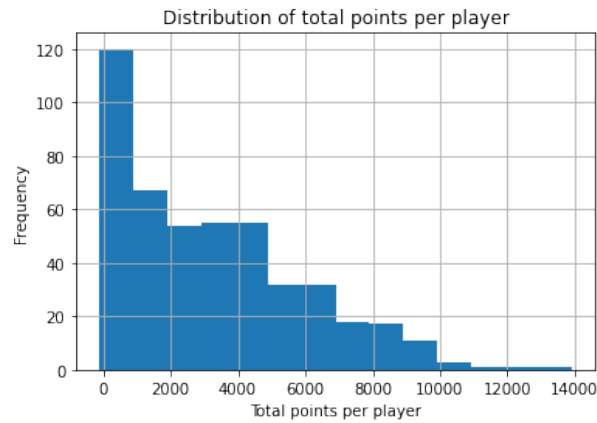


Figure 14: Distribution of Total Score per Player

On the other end of the axis, there is only a small number of extraordinary players who scored over 10,000 points throughout the season. These players have two characteristics: they are fielded often and score confidently. The following figure shows the Top 10 players regarding the total score over the entire season.

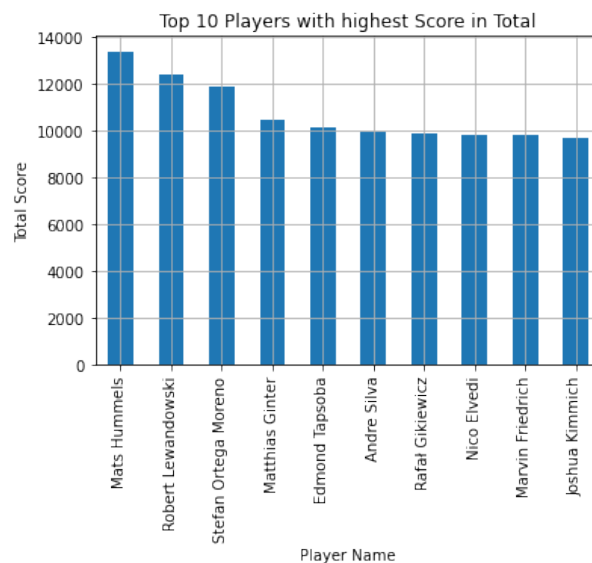


Figure 15: Top 10 Players with highest Total Score

In contrast to the previous group, this top 10 contains many well-known players, once again including *Mats Hummels* and *Robert Lewandowski* at the front positions.

Correlation between position and score

Interestingly, all positions are represented in the previous figure 15. For this reason, further insights into the correlation between the position and the scores are provided in this paragraph. The figure 16 shows that goalkeepers score the most points on average, followed by defenders and midfielders. Attackers score the fewest points. This finding is counterintuitive as goals are scored highest in SPITCH and most goals are scored by attackers. It can be deduced from this finding that it often depends on many small events, such as passing or taking the ball away, to achieve a high score.

Furthermore, conclusions for the formation can be drawn from this diagram. For example, if goalkeepers score the most points on average, it might be a good strategy to choose the goalkeeper as captain. In addition, a defensive formation should be chosen, as defenders and midfielders score more points on average than attackers.

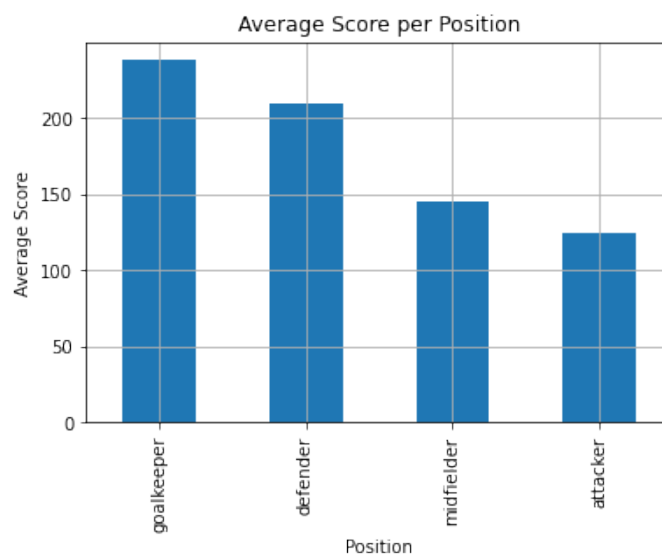


Figure 16: Average Score per Position

The following two data explorations are carried out exclusively with the Betting Odds CSV data set (Football-Data, n.d.) and should therefore be considered separately from the previous data set.

Correlation between Betting Odds and Match Result

This section aims to examine the influence betting odds have in the prediction of future player performance. For this reason, the match result is taken from the CSV in addition to the betting odds. First, the percentage of cases in which betting odds predicted the correct final result is examined. Therefore, the delta between the odds for the home and away team is calculated. From the previous investigations of the betting odds, it is known that when the odds for a draw are highest, the odds for the home and away team usually differ by a maximum of 0.5. For this reason, a draw is expected for a delta of less than 0.5. Otherwise, a win is expected for the team with lower odds. When this procedure is applied to the entire data set, the betting odds predict the correct outcome **53.6%** of the time. This result is a remarkable 21% difference to the probability of one-third if one had to guess the result. Thus, this study concludes that betting odds are indicative of which team is likely to win.

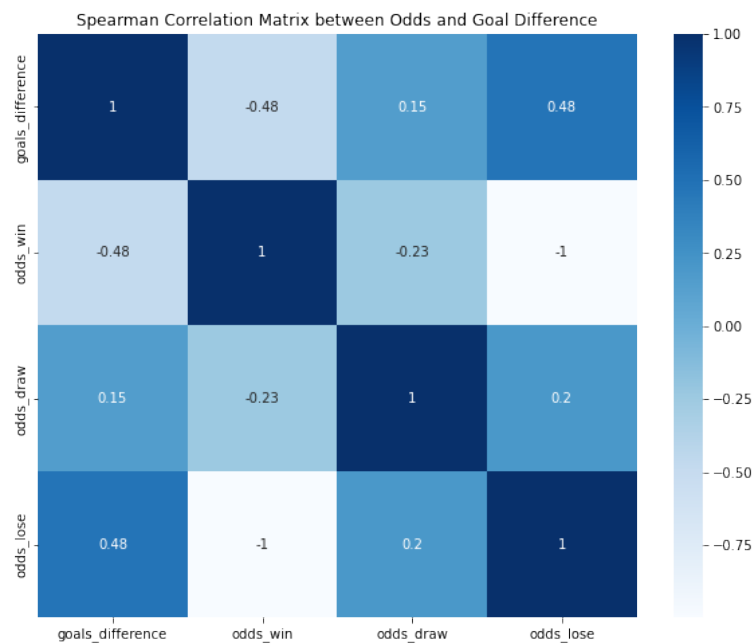


Figure 17: Spearman Correlation Matrix between Odds and Goal Difference

However, the game of SPITCH is more complex than just predicting the winning team. It would also be interesting to know whether betting odds can also predict the level of victory or defeat. Because teams that are expected to win higher than others consequently

score more goals, which leads to higher scores. In order to prove this influence, the final goal difference is computed. Afterward, the correlation between this variable and the betting odds is calculated using *Spearman's rank correlation coefficient*. The correlation matrix in figure 17 indicates a medium correlation of 0.48 or -0.48 respectively between the goal difference and the odds for winning or losing. Accordingly, the betting odds not only allow to predict which team is likely to win but also to estimate with a certain degree of accuracy how clear the result will be.

Home advantage

The literature review showed that home advantage had been used in many models written to predict sports outcomes. In the data set used for the models in this thesis, a column regarding the home advantage was also added. The following figure shows that home advantage also played a role in the 2020/2021 season. From the perspective of the home team, 42% of the matches were won and 31% lost.

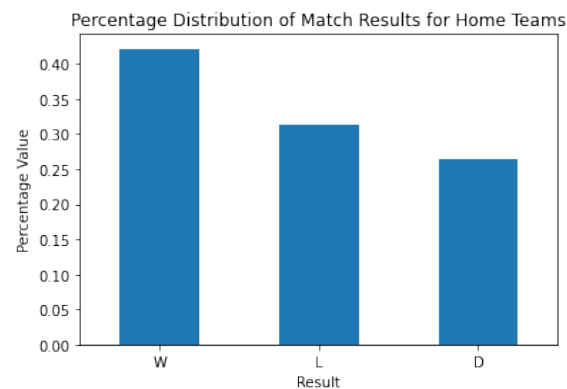


Figure 18: Percentage Distribution of Match Results for Home Teams

4.3 Models

One crucial step is the splitting of the dataset into test and training data. According to best practices, the splits can range from 60% to 40% until 80% to 20%. In this thesis, the test dataset contains the last nine matchdays of the 2020/2021 Bundesliga season, i.e., the matchdays 26 to 34. This distribution represents a 70:30 split.

Before the models can be implemented, a metric has to be found for comparing the models. For machine learning problems, it is helpful to define a metric beyond the usual metrics that make the models' predictions comparable to the desired outcome. Reflecting on the considerations from the Business Context chapter 4.1, the realistic goal of the models should be to predict players' performances accurately enough to regularly produce a line-up that achieves a score of 65% of the best possible score. By reaching this goal, the models would be able to win a prize now and then. For this reason, the following two own metrics are created. The first metric measures the percentage of points achieved by the model compared to the best possible score. The second metric takes the first metric to calculate the prize money earned in case of reaching the winning zone.

In order for the first metric to be measured, the best line-up must be calculated for each matchday in the test dataset. Therefore, the achieved score from each player for each matchday is taken and decreased by their current transfer market value according to equation (5) to get the **adjusted player score** S_{P_iM} . Using these values, the best line-up can be assembled by taking the corresponding number of best players for each position for each available line-up and adding up their scores. The resulting sum is the highest possible score that can be achieved for each formation. The formation with the most points is the best possible line-up, which corresponds to 100% for the first metric. The percentage of points for the team set up by the model is then calculated by dividing the resulting score by the best possible score. This metric is called the **percentage of best possible score** $S_{LM\%}$.

For the second metric, as SPITCH, unfortunately, does not give any insight into the exact calculation of the prize money within the profit zone, a formula has to be created using the available data, which represents this calculation as closely as possible. Due to the lack of data, conscientious simplifications have to be made. Like already mentioned in the chapter *Business Context*, the best ranks usually achieve 80 to 85% of the maximum

possible score. Accordingly, in the simplified calculation of the price, the first place is won as soon as $S_{LM\%}$ is 80% or above. Every percent less means one place further down in the ranking. Thus, the *Rank* can be calculated from $S_{LM\%}$ using the equation:

$$Rank = \frac{0.81 - S_{LM\%}}{0.01} \quad (8)$$

The corresponding prize to each rank can be seen on the screenshot from *SPITCH* in figure 19.



Figure 19: Percentage of Prize per Rank for the €30 Pitch

This distribution applies to all *pitches* with stakes. The following function can be derived using the data from this screenshot, which calculates the share of the profit $Profit_{\%}$ from the associated rank $Rank$:

$$Profit_{\%} = 27.56 * e^{-0.72 * Rank} + 1.3 \quad (9)$$

The prize money consists of the players' stakes and therefore is the product from participants and stake. For the evaluation, the models attend the €2 pitch with 1,000 participants. Thus, the prize money equals €2,000 each round.

Before the machine learning models are created, two *dummy* models are defined. These models serve as comparisons and represent alternative strategies. The first model randomly assembles a line-up and is called *guess*. The second model fields the best line-up from the previous matchday for the upcoming matchday and is called *previous best*. The table below presents the results of these models, ordered by their performance on *Mean $S_{LM\%}$* :

Table 4: Dummy Models Results

| Design | Model | Mean $S_{LM\%}$ | MAE of \hat{S}_{P_i} | Prize Money | # Prizes Won |
|--------|---------------|-----------------|------------------------|-------------|--------------|
| dummy | previous best | 0.4133 | 197.45 | -€18 | 0 |
| dummy | guess | 0.3511 | 197.18 | -€18 | 0 |

The results show that the model *previous best* performs with a mean of 41.33% of the best possible score and therefore performs better than the model *guess*. This result is no surprise, considering the strong autocorrelation of the player scores shown in the Data Exploration section 4.2.3, indicating that players who recently performed well will most likely perform well in the future. However, although the model has set up better line-ups, it has a higher mean absolute error per player (MAE of \hat{S}_{P_i}). Both models fail to win a prize, losing nine times €2, which equals a net loss of €18.

The results of the dummy models are now compared with the machine learning models available in the *scikit-learn* Python package. The selection of regression models follows the results of the literature review. For this reason, these models are *Multiple Linear Regression*, *Logistic Regression*, *Decision Trees*, *Random Forests*, *K-Nearest Neighbors*, and *Support Vector Machines*. Each model exists twice, once in *baseline* and once in *treatment* design. The treatment models have access to the betting odds data while the baseline models don't. All models are cross-validated with five stratified folds on the training dataset to minimize the bias in the accuracy scores of the models. The following table shows the results of the cross-validation, sorted by the MAE of \hat{S}_{P_i} .

Table 5: Cross-Validation Results

| Design | Model | MAE of \hat{S}_{P_i} |
|-----------|----------------------------|------------------------|
| treatment | Multiple Linear Regression | 99.32 |
| baseline | Multiple Linear Regression | 99.73 |
| treatment | Random Forest Regression | 102.95 |
| treatment | K-Nearest Neighbors | 105.41 |
| baseline | K-Nearest Neighbors | 106.38 |
| baseline | Random Forest Regression | 107.55 |
| treatment | Logistic Regression | 117.82 |
| baseline | Logistic Regression | 118.52 |
| treatment | Support Vector Machines | 120.06 |
| baseline | Support Vector Machines | 120.52 |
| baseline | Decision Tree Regression | 132.91 |
| treatment | Decision Tree Regression | 135.91 |
| dummy | guess | 197.18 |
| dummy | previous best | 197.45 |

Three findings emerge from the results of the cross-validation. The first finding is that five of the six machine learning models provide more accurate predictions when accessing the betting odds. The second finding is that even the worst-performing machine learning model has a much smaller MAE than the dummy models. Therefore, machine learning algorithms seem to offer a predictive value in general. The last finding is that *Multiple Linear Regression* and *Random Forest Regression* are the two most accurate machine learning models. However, the models so far have only used their default parameters. Through a process called **hyperparameter tuning**, the models can improve even further. This procedure is applied in the next paragraph to the two previously mentioned most accurate models.

Multiple Linear Regression and *Random Forest Regression* differ in their complexity. While the former has only three Boolean parameters that can be optimized, *Random Forest Regression* has more than ten parameters, most of which are numerical. Furthermore, two out of three of the linear regression parameters are unsuitable for optimization in this

case. The parameters *normalize* and *positive* are dropped since the data has already been standardized, and there are features in the dataset such as betting odds where one feature at least must be negative. The only optimization parameter remaining is *fit_intercept*, which allows the regression to be calculated with or without an interception. This parameter did not make the models more precise.

The hyperparameter tuning of the Random Forest Regression Model happens with the six most influential parameters. Since four of these parameters can be any natural number, the Python packages called *RandomizedSearchCV* and *GridSearchCV* are used. Each of these packages takes a list of possible parameter settings and fits the models with different combinations from these parameter settings using cross-validation. *GridSearchCV* takes all parameters and tests every possible combination, whereas *RandomizedSearchCV* takes number ranges and randomly tests a pre-defined quantity of combination. Therefore, *RandomizedSearchCV* narrows down the optimal parameter settings, and *GridSearchCV* finds the optimal parameters in the narrowed down area afterward. Through this procedure, the following optimal parameters for the Random Forest Regression are found, resulting in the hyperparameter-tuned model:

```
RandomForestRegressor(max_depth=13,max_features='sqrt',bootstrap=True,
min_samples_leaf = 9, min_samples_split = 2, n_estimators = 300)
```

This model reaches the lowest MAE of **98.14** and thus predicts most accurately.

After all the models have been created, fitted, and cross-validated, they can finally be used on the test dataset. The following table presents the results, ordered by their performance on *Mean S_{LM}* :

Table 6: Models' Results on the Test Dataset

| Design | Model | Mean $S_{LM}\%$ | MAE of \hat{S}_{P_i} | Prize Money | # Prizes Won |
|-----------|--------------------------------|-----------------|------------------------|-------------|--------------|
| treatment | Tuned Random Forest Regression | 0.5189 | 104.56 | €33.86 | 2 |
| baseline | Tuned Random Forest Regression | 0.4922 | 104.92 | -€18.00 | 0 |
| treatment | Multiple Linear Regression | 0.4911 | 105.28 | €8.01 | 1 |
| baseline | Random Forest Regression | 0.4733 | 114.06 | -€18.00 | 0 |
| treatment | Random Forest Regression | 0.4700 | 108.81 | -€18.00 | 0 |
| baseline | Multiple Linear Regression | 0.4667 | 105.56 | -€18.00 | 0 |
| baseline | K-Nearest Neighbors | 0.4489 | 111.81 | -€18.00 | 0 |
| treatment | K-Nearest Neighbors | 0.4344 | 113.11 | -€18.00 | 0 |
| dummy | previous best | 0.4133 | 197.45 | -€18.00 | 0 |
| baseline | Decision Tree Regression | 0.3922 | 141.60 | -€18.00 | 0 |
| treatment | Logistic Regression | 0.3867 | 123.85 | -€18.00 | 0 |
| baseline | Logistic Regression | 0.3811 | 123.95 | -€18.00 | 0 |
| treatment | Decision Tree Regression | 0.3556 | 145.67 | -€18.00 | 0 |
| dummy | guess | 0.3511 | 197.18 | -€18.00 | 0 |
| treatment | Support Vector Machines | 0.3400 | 126.94 | -€18.00 | 0 |
| baseline | Support Vector Machines | 0.3344 | 127.16 | -€18.00 | 0 |

The table shows that the **hyperparameter tuned random forest regression model** in the treatment design, i.e., with the betting odds as features, made the **most accurate predictions** on the test dataset, with an average of just under **52%** of the points from the best possible line-up. The model is 16% or 11% respectively, better than the dummy models. On average, the model misjudged each player by 104.5 points in the test and is thus 6 points worse than on the training dataset. If the winning zone would start at 60% of the best possible score, the model would have **won a prize two times** and brought in total prize money of **33.86€**. This profit is possible because the standard deviation of the best model is 7.6%, which indicates that it is quite realistic for the model to reach more than 60% in nine tries.

Interestingly, although the multiple linear regression model in the treatment design made less accurate predictions on average than the random forest regression model in the baseline design, it still won a prize. This means that despite poorer individual predictions, it lined up a better team, probably because of luck.

Furthermore, it is noticeable that on the test dataset, only four out of seven models performed better in the treatment design than in the baseline design. This could indicate

that the models were overfitted on the betting odds in the training dataset. However, the influence of the betting odds will be examined in detail in the next section.

4.4 Betting Odds Impact

The results of the previous section suggest that betting odds have a positive influence on the predictive accuracy of the models. In this section, a two-sided t -test will be used to check whether this assumption is correct. With the research question on page 23 in mind, the hypothesis H_1 is:

'The accuracies of a model that predicts individual player performance increase under the influence of betting odds.'

Therefore, the null hypothesis H_0 is:

'The accuracy of a model that predicts individual player performance does not increase under the influence of betting odds.'

The data used for the t -test are two result series from the hyperparameter tuned Random Forest Regressor, one from the baseline and one from the treatment design. The absolute difference between the line-up and the best possible score for each matchday in the test dataset is chosen. Otherwise, the data would not be normalized considering the variances in the scores per matchdays. For the t -test, the Python package *researchpy* is used. Table 7 on page 56 shows the results.

Table 7: Two-Sided t-Test Results

| Metric | Result |
|------------------------|--------|
| Degrees of freedom | 16 |
| t | 0.5928 |
| Two side test p value | 0.5212 |
| Difference < 0 p value | 0.7192 |
| Difference > 0 p value | 0.2808 |
| Cohen's d | 0.3130 |
| Hedge's g | 0.2661 |
| Glass's delta | 0.2665 |
| Pearson's r | 0.1631 |

The two main metrics to focus on are *Cohen's d*, which indicates the strength of the effect, and the *p-value*, which shows how statistically significant the sample is. With a *Cohen's d* of **0.313**, the effect can be categorized as **small**. Furthermore, a 58.7% chance exists that a model that uses betting odds will assemble a line-up with a higher score than a model without them, and 62.2% of the line-ups assembled by the treatment models will be above the mean of the baseline models. (cf. Magnusson, 2021) However, a *p-value* of 0.5212 indicates that the sample used for calculating is not statistically significant. The reason for this is probably the too-small sample of only nine match days. Therefore, with a usual significance level α of 0.05, the null hypothesis H_0 can not be rejected. Nevertheless, in future studies with a larger dataset, the implied effect could be further investigated.

4.5 Application

This last section of the chapter aims to present the entire application that was developed in the course of this thesis. The source code for the application, as well as the Jupyter Notebook in which the data mining part was implemented, can be found as a *Git* repository on *GitHub* under the following link:

<https://github.com/jakobheine/fm-analytics/tree/v.1.0.0>

As already mentioned at the beginning of this chapter, the application is implemented in a microservice architecture. All services are created as *Docker* containers, which makes them easy to deploy. The entire tool is deployed on a virtual private server (VPS) hosted on Hetzner. (Hetzner, 2021) The following figure shows the architecture and how the microservices interact with each other.

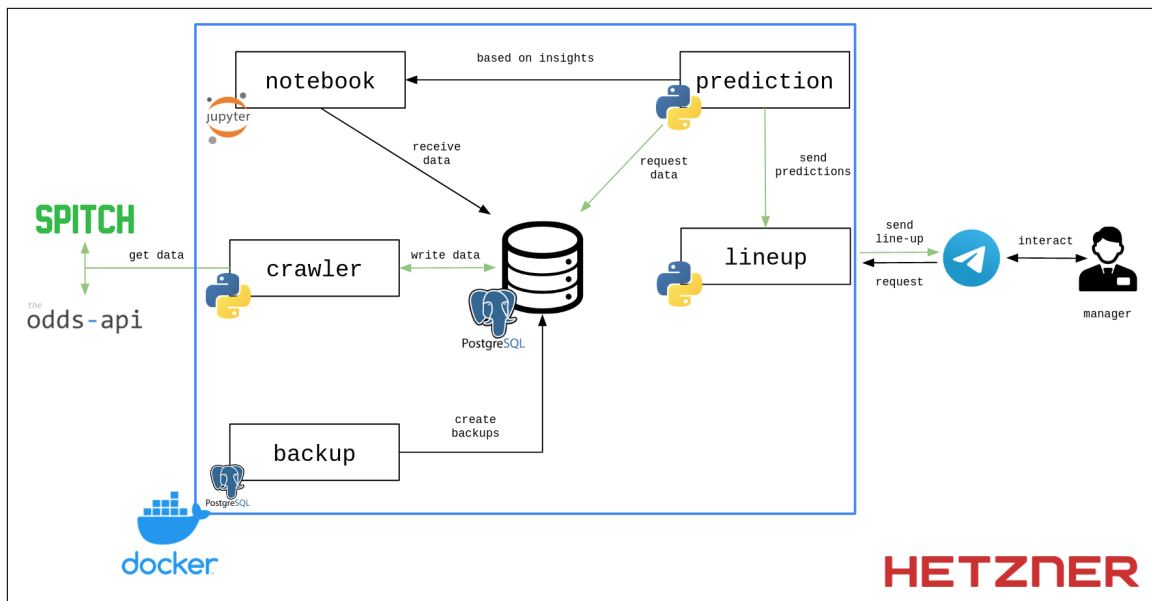


Figure 20: Microservice-Architecture of the Application

The green arrows indicate the primary process. In this process, the *Crawler* collects the current data from *SPITCH* and *the odds-api* via the *Proxy* and stores it in the *Database*. This data is then queried by the *Prediction* service, which predicts the individual scores of the players based on the findings from the *Jupyter Notebook* and forwards them to

the *Lineup* service. The *Lineup* service creates the optimal line-up from the predicted scores and the transfer market values, which it then sends to the end-users via a *Telegram Bot*. This process starts 15 minutes before the kick-off of a matchday so that the data is as up-to-date as possible, but the managers still have enough time to line up the team predicted by the tool. In addition, the manager can start the process himself at any time. Furthermore, the *Backup* service creates a backup of the database every 24 hours.

To deploy the application, only a server with *Docker*, *Docker-Compose*, and *Git* installed is needed. After the *Github* repository has been cloned, the application can be started with the command: `docker-compose up`.

The following screenshot shows the messages the application is sending to its users.

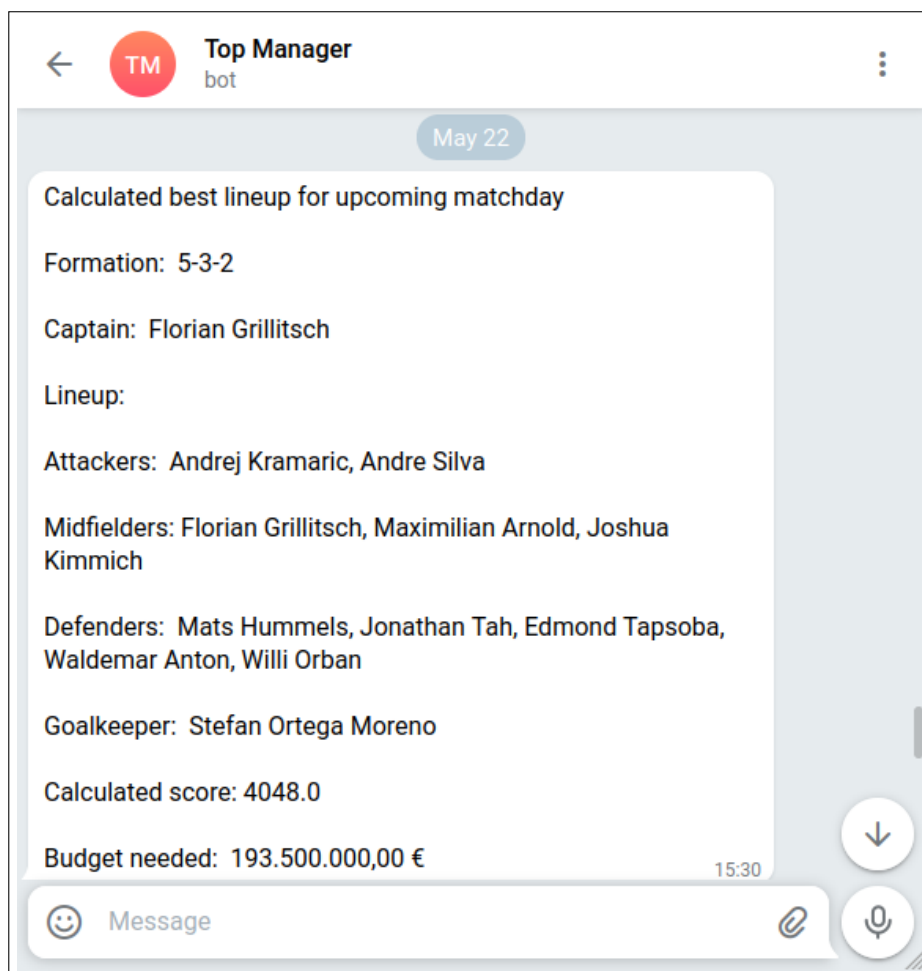


Figure 21: Screenshot of the Applications Messages

Chapter 5

Conclusion

Fantasy Leagues flourish under the influence of the *Big Data* age. The immense growth in data and statistics makes these games more complex than ever before, yet this growth offers a wide range of improvement potentials. In this thesis, it was investigated to what extent this data can be used in machine learning models which predict future player performances. Therefore, the fundamental research question asked was: *'How accurately can individual soccer player performances be predicted using historical data?'*

To answer this question, three steps had to be taken. The first step to answering this question was a literature review to overview possible solution approaches. This review revealed that no research has yet been done on precisely this kind of problem. Nevertheless, related research was found and applied, resulting in multiple feature and model proposals. One of these features was the betting odds, which raised the question of whether betting odds affect the predictions positively. Thus, the central research question could be updated into *'How accurately can individual soccer player performances be predicted using historical data and betting odds?'*

In the second step, the question at hand was translated into a machine learning problem. Through this translation, the goal could be set to write a model that regularly earns a prize in a Fantasy League competition by reaching at least a line-up score of 65% of the best possible score. In order to write this model, a data basis had to be established first. An application was written to procure, process and, store the data.

In the final step, using the data, several models, with and without betting odds, were compared with each other. From the comparison, two types of machine learning models emerged that performed best: *Random Forest* and *Multiple Linear Regression*. Hence, the thesis proves that these two types of models, which have been shown to solve similar problems in the past, also performed well on the issue at hand. On the downside, models, like *Support Vector Machines* or *Decision Trees*, that promised success as well did not offer accurate predictions.

In the experiment on the test dataset, the Random Forest model won a prize twice in nine rounds and thus achieved the desired goal, winning a total prize of €33.86. On average, the model misjudged each player by about 100 points, which is quite a lot for an average score of 200 per player. Nevertheless, even this accuracy seems to be enough to win in the competition *SPITCH* regularly. Another aspect this work proved in this way is that football is a challenging sport to predict, and therefore the entire competition offered by *SPITCH* resembles many characteristics of gambling. In addition, the calculated *Cohen's d* of 0.313 between the treatment and baseline models indicates that the betting odds indeed have a small but positive effect on the predictions, even though the calculated *p-value* of 0.5212 implies an inconclusive sample. Finally, the central research question can be answered as follows: Individual player performances cannot be predicted to the exact point. There are far too many factors to consider, which are not or only very difficult to realise in a model. Nevertheless, machine learning models can help to assemble a line-up that is occasionally better than 75% of the competitors.

In the future, to further improve the models, more features can be added. Examples would be the official transfer market value or characteristics about the playing style of different players or teams. In order to further validate the impact of betting odds and generally to increase the models' accuracies, more data can be collected and used. In addition, other forecast values can be examined, such as expected goals or spread bets.

List of Figures

| | | |
|----|--|----|
| 1 | Top 10 Ranking FREE Pitch, 8th matchday 2021/22 | 25 |
| 2 | Line-Ups of Top 3 Managers from 8th matchday 2021/22 | 26 |
| 3 | Offline Phase Process | 28 |
| 4 | Online Phase Process | 29 |
| 5 | Effects between Dependent and Independent Variables | 36 |
| 6 | Distribution of Events per Player | 38 |
| 7 | Number of Players per Position | 39 |
| 8 | Number of Players per Team | 40 |
| 9 | Distribution of Betting Odds | 41 |
| 10 | Percentage Distribution of Occurences | 42 |
| 11 | Autocorrelation of Score | 42 |
| 12 | Distribution of Average Score per Player | 43 |
| 13 | Top 10 Players with highest Average Score | 44 |
| 14 | Distribution of Total Score per Player | 45 |
| 15 | Top 10 Players with highest Total Score | 45 |
| 16 | Average Score per Position | 46 |
| 17 | Spearman Correlation Matrix between Odds and Goal Difference | 47 |
| 18 | Percentage Distribution of Match Results for Home Teams | 48 |
| 19 | Percentage of Prize per Rank for the €30 Pitch | 50 |
| 20 | Microservice-Architecture of the Application | 57 |
| 21 | Screenshot of the Applications Messages | 58 |

List of Tables

| | | |
|---|---|----|
| 1 | SPITCH Glossary | 9 |
| 2 | Concept Matrix | 16 |
| 3 | Overview of Features and their Datatype | 37 |
| 4 | Dummy Models Results | 51 |
| 5 | Cross-Validation Results | 52 |
| 6 | Models' Results on the Test Dataset | 54 |
| 7 | Two-Sided t-Test Results | 56 |

Bibliography

- About ASA. (n.d.). Retrieved July 20, 2021, from <https://www.amstat.org/ASA/about/home.aspx?hkey=6a706b5c-e60b-496b-b0c6-195c953ffdbc>
- Anik, A. I., Yeaser, S., Hossain, A. G. M. I., & Chakrabarty, A. (2018). Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms, In *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEICT)*, Dhaka, Bangladesh, IEEE. <https://doi.org/10.1109/CEEICT.2018.8628118>
- Becker, A., & Sun, X. A. (2016). An analytical approach for fantasy football draft and lineup management [Publisher: De Gruyter Section: Journal of Quantitative Analysis in Sports]. *Journal of Quantitative Analysis in Sports*, 12(1), 17–30. <https://doi.org/10.1515/jqas-2013-0009>
- Beliën, J., Goossens, D., & Reeth, D. V. (2017). Optimization modeling for analyzing fantasy sport games, 16.
- Bonomo, F., Durán, G., & Marenco, J. (2014). Mathematical programming as a tool for virtual soccer coaches: A case study of a fantasy sport game. *International Transactions in Operational Research*, 21(3), 399–414. <https://doi.org/10.1111/itor.12068>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Retrieved June 18, 2021, from </paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72>
- Demediuk, S., Kokkinakis, A., Patra, M. S., Robertson, J., Kirman, B., Coates, A., Chitayat, A., Hook, J., Nolle, I., Slawson, D., Ursu, M., Block, F., & Drachen, A. (2021). Performance Index: A New Way To Compare Players, 16.

- Deng, W., & Zhong, E. (2020). Analysis and Prediction of Soccer Games: An Application to the Kaggle European Soccer Database. *Insight - Statistics*, 3(1), 1. <https://doi.org/10.18282/i-s.v3i1.332>
- Edwards, S. J. (2018). Analyzing Fantasy Sport Competitions with Mixed Integer Programming (A. T. Ernst, S. Dunstall, R. García-Flores, M. Grobler, & D. Marlow, Eds.). In A. T. Ernst, S. Dunstall, R. García-Flores, M. Grobler, & D. Marlow (Eds.), *Data and Decision Sciences in Action 2*, Cham, Springer International Publishing. https://doi.org/10.1007/978-3-030-60135-5_12
- Egidi, L., & Gabry, J. (2018). Bayesian hierarchical models for predicting individual performance in soccer [Publisher: De Gruyter Section: Journal of Quantitative Analysis in Sports]. *Journal of Quantitative Analysis in Sports*, 14(3), 143–157. <https://doi.org/10.1515/jqas-2017-0066>
- e. V., V. (2015). VHB-JOURQUAL3: Wirtschaftsinformatik. https://vhbonline.org/fileadmin/user_upload/JQ3_WI.pdf
- Fabiano. (2007). Fantasy football 101. Retrieved October 19, 2021, from <https://www.nfl.com/news/fantasy-football-101-09000d5d80021ece>
- Football-Data. (n.d.). Germany Football Results and Betting Odds. Retrieved November 2, 2021, from <https://www.football-data.co.uk/germanym.php>
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, UNITED STATES, O'Reilly Media, Incorporated. Retrieved October 25, 2021, from <http://ebookcentral.proquest.com/lib/htw-berlin/detail.action?docID=5892320>
- Goldstein, D. G., McAfee, R. P., & Suri, S. (2014). The wisdom of smaller, smarter crowds, In *Proceedings of the fifteenth ACM conference on Economics and computation*, Palo Alto California USA, ACM. <https://doi.org/10.1145/2600057.2602886>
- Green. (2014). 'Wink': Wilfred 'Bill' Winkenbach invented Fantasy Football way back in 1962 with GOPPPL in Oakland - newsnet5.com Cleveland. Retrieved October 19, 2021, from <https://web.archive.org/web/20150929163914/http://www.newsnet5.com/sports/wink-wilfred-bill-winkenbach-invented-fantasy-football-way-back-in-1962-with-gopppl-in-oakland>
- Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1), 177–214. <https://doi.org/10.1007/s11192-018-2958-5>

- Hetzner. (2021). About Us - Hetzner Online GmbH. Retrieved October 28, 2021, from <https://www.hetzner.com/unternehmen/ueber-uns>
- Hoffmann. (2014). Millionen Daten pro Spiel: So werden Fußball-Statistiken erfasst [Section: Sport]. Retrieved October 19, 2021, from <https://www.hna.de/sport/fussball/millionen-daten-spiel-werden-fussball-statistiken-erfasst-4483893.html>
- Karthik, K., Krishnan, G. S., Shetty, S., Bankapur, S. S., Kolkar, R. P., Ashwin, T. S., & Vanahalli, M. K. (2021). Analysis and Prediction of Fantasy Cricket Contest Winners Using Machine Learning Techniques [Series Title: Advances in Intelligent Systems and Computing]. In V. Bhateja, S.-L. Peng, S. C. Satapathy, & Y.-D. Zhang (Eds.), *Evolution in Computational Intelligence* (pp. 443–453). Series Title: Advances in Intelligent Systems and Computing. Singapore, Springer Singapore. https://doi.org/10.1007/978-981-15-5788-0_43
- Landers, J. R., & Duperrouzel, B. (2017). Machine Learning Approaches to Competing in Fantasy Leagues for the NFL. *IEEE Transactions on Games*, 11(2), 159–172. <https://doi.org/10.1109/TG.2018.2841057>
- Lutz, R. (2015). Fantasy Football Prediction [arXiv: 1505.06918]. *arXiv:1505.06918 [cs]*. Retrieved June 2, 2021, from <http://arxiv.org/abs/1505.06918>
- Magnusson, K. (2021). Interpreting Cohen’s d. Retrieved November 5, 2021, from <https://rpsychologist.com/cohend/>
- Matthews, T., Ramchurn, S. D., & Chalkiadakis, G. (2012). Competing with Humans at Fantasy Football: Team Formation in Large Partially-Observable Domains, 7.
- Nevill, A. M., & Holder, R. L. (1999). Home Advantage in Sport: An Overview of Studies on the Advantage of Playing at Home. *Sports Medicine*, 28(4), 221–236. <https://doi.org/10.2165/00007256-199928040-00001>
- Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus*, 5(1), 1410. <https://doi.org/10.1186/s40064-016-3108-2>
- Shah, K., Hyman, J., & Samangy, D. (2021). A Poisson Betting Model with a Kelly Criterion Element for European Soccer, 16.
- Skinner, B., & Guy, S. J. (2015). A Method for Using Player Tracking Data in Basketball to Learn Player Skills and Predict Team Performance (F. Emmert-Streib, Ed.). *PLOS ONE*, 10(9), e0136393. <https://doi.org/10.1371/journal.pone.0136393>

- Spann, M., & Skiera, B. (2009). Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28(1), 55–72. <https://doi.org/10.1002/for.1091>
- SPITCH. (2021a). Points Catalogue. Retrieved October 21, 2021, from <https://www.spitch.live/en/points-catalogue/>
- SPITCH. (2021b). Rules of Play and Participation: retrieved October 21, 2021, from <https://www.spitch.live/en/rules-of-play-and-participation/>
- SPITCH. (2021c). SPITCH | The Live Football Manager. Retrieved October 21, 2021, from <https://www.spitch.live/en/>
- vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the Association for Information Systems*, 37. <https://doi.org/10.17705/1CAIS.03709>
- Webster, J., & Watson, R. T. (2002). Guest Editorial: Analyzing the Past to Prepare for the Future: Writing a literature Review, 11.
- Wheatcroft, E. (2020). Profiting from overreaction in soccer betting odds [Publisher: De Gruyter Section: Journal of Quantitative Analysis in Sports]. *Journal of Quantitative Analysis in Sports*, 16(3), 193–209. <https://doi.org/10.1515/jqas-2019-0009>
- Yurko, R., Ventura, S., & Horowitz, M. (2019). nflWAR: A reproducible method for offensive player evaluation in football [Publisher: De Gruyter Section: Journal of Quantitative Analysis in Sports]. *Journal of Quantitative Analysis in Sports*, 15(3), 163–183. <https://doi.org/10.1515/jqas-2018-0010>

Decleration of Authenticity

I declare that I wrote this thesis on my own and did not use any unnamed sources or aid. Thus, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made by correct citation. This includes any thoughts taken over directly or indirectly from printed books and articles as well as all kinds of online material. It also includes my own translations from sources in a different language. The work contained in this thesis has not been previously submitted for examination. I also agree that the thesis may be tested for plagiarized content with the help of plagiarism software. I am aware that failure to comply with the rules of good scientific practice has grave consequences and may result in expulsion from the program.

Leipzig, 08.11.2021

Jakob Heine