Abstract

Contents

1 Introduction **1**
1.1 Motivation 1
1.2 Aims 1
1.3 Methodology 1

2 Foundations **2**
2.1 Fantasy Leagues 2
2.2 SPITCH 5

3 Literature Review **9**

4 Implementation **19**
4.1 Business Context 19
4.2 Data 19
4.2.1 Procurement 19
4.2.2 Preparation 19
4.2.3 Exploration 19
4.3 Models 19
4.4 Application 19

5 Conclusion **20**

List of Figures **A**

List of Tables **B**

Source Code **C**

Bibliography **F**
A.1 Diagrams . . . . K
A.2 Tables . . . . K
A.3 Screenshots . . . . K
A.4 Graphs . . . . K

Decleration of Authenticity **L**

Chapter 1

Introduction

1.1 Motivation

1.2 Aims

1.3 Methodology

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 football, 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 themself. 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 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 prominent example of differences 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 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, $\{\, i \mid i \in \{1, 2, ..., 711\} \,\}$ has a score $S_{P,R}$ for each of the 34 rounds R_j, $\{\, j \mid j \in \{1, 2, ..., 34\} \,\}$. 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} :

$$S_{\mathrm{L}} = \sum_{i=1}^{11} S_{\mathrm{P_i}} \tag{1}$$

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, $\{\, k \mid k \in \{1, 2, ..., 33\} \,\}$ 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:

$$S_{\mathrm{P_i}} = \sum_{k=1}^{33} O_{\mathrm{k}} * p_{\mathrm{k}} \tag{2}$$

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

$$S_{\mathrm{L}} = \sum_{i=1}^{11} \sum_{k=1}^{33} O_{\mathrm{ik}} * p_{\mathrm{ik}} \tag{3}$$

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:

$$S_{\mathrm{M}} = \frac{\Delta_{\mathrm{Budget}} \cdot 0.8}{100{,}000} = \Delta_{\mathrm{Budget}} \cdot 0.8 \cdot 10^{-5} \tag{4}$$

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 $-€10\mathrm{m} \cdot 0.8 \cdot 10^{-5} = -80$ manager score. Consequently, manager points do not increase or decrease exponentially 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 $\mathrm{S}_{\mathrm{P_i}}$, equation (2) can be supplemented by equation (4):

$$S_{\mathrm{P_i}} = -V_{\mathrm{i}} \cdot 0.8 \cdot 10^{-5} + \sum_{k=1}^{33} O_{\mathrm{k}} * p_{\mathrm{k}} \tag{5}$$

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

$$S_{\mathrm{LM}} = S_{\mathrm{L}} + S_{\mathrm{M}} = \sum_{i=1}^{11} \sum_{k=1}^{33} O_{\mathrm{ik}} * p_{\mathrm{ik}} + \Delta_{\mathrm{Budget}} \cdot 0.8 \cdot 10^{-5} \tag{6}$$

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 *fields*, such as the *€2 field* or the *€30 field*. Only the participants of the individual fields compete against each other. In the *free field*, the top 10 managers in the ranking, i.e., the ten managers with the highest adjusted final line-up scores $\mathrm{S_{LM}}$, win. In all other fields, 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 in the coming chapters.

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 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 and Watson, 2002, p. 14). First, as recommended by 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.

According to 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 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* 2021). 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).

In the process mentioned above, 22 papers were examined and compared in a concept matrix (see Figure 2 on page 12) 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 analysed. 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 optimisation 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 optimised 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 analysed. Due to the always different data, a generalised 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 recognised 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

	Published In					Solution Approach				Task		Solution Objective			Key Features		
Paper	**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	X	X
Wheatcroft (2020)	X			JQAS	soccer		X			X	X		X	X	X	X	X
Shah et al. (2021)		X		n.R.	soccer		X			X	X			X	X	X	X
Spann and Skiera (2009)	X			n.R.	soccer					X	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		
Egidi 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)		X		n.R.	soccer				X	X	X	X			X		
Skinner and Guy (2015)	X			n.R.	basketball			X			X	X			X		
Pappalardo et al. (2019)	X			B	soccer		X	X		X		X			X		
Yurko et al. (2019)	X			JQAS	football		X		X	X		X			X		
Demediuk et al. (2021)		X		n.R.	e-sports		X			X		X			X		
Edwards (2018)		X		n.R.	football	X				X		X	X		X		
Karthik et al. (2021)		X		C	cricket		X	X	X		X		X		X		
Beliën et al. (2017)	X			n.R.	cycling	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	
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 Perforamance, Team = Team Performance

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. Lutz (2015), Egidi and Gabry (2018), and Yurko, Ventura, and Horowitz (2019) focus in their papers on what to do if this data is not available. Yurko, Ventura, and Horowitz (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 guess 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) and 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. Lutz (2015), Demediuk et al. (2021), and Egidi and Gabry (2018) 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 and Zhong, 2020, p. 4). That confirms the discovery of Rein and Memmert, who claim that at the current state of research, *'most [Machine Learning] soccer analyses are performed by computer scientist research group with little apparent involvement by sports scientists.'* (Rein and 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.

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 optimisation and prediction tasks. Only two different methodologies were chosen for the optimisation task: either brute force optimisation (Landers and Duperrouzel, 2017) or mixed-integer programming (Becker and Sun, 2016; Edwards, 2018; Beliën, Goossens, and Reeth, 2017; Bonomo, Durán, and Marenco, 2014; Matthews, Ramchurn, and Chalkiadakis, 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)* (Landers and Duperrouzel, 2017; Deng and Zhong, 2020), *Random Forest* (Deng and Zhong, 2020; Shah, Hyman, and Samangy, 2021; Demediuk et al., 2021; Karthik et al., 2021) and *Deep Neural Networks (DNN)* (Karthik et al., 2021; Skinner and Guy, 2015; Deng and Zhong, 2020; Lutz, 2015; Landers and Duperrouzel, 2017). 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 and Zhong, 2020, p. 4) They note that while Decision Tree models compute faster and require less 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, 5 activation layers, 2 dropout layers and one batch normalisation layer.'* (Deng and 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 emphasise 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 normalise or scale the features. (cf. Landers and Duperrouzel, 2017, p. 6)

In the researches from Shah, Hyman, and Samangy (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, Hyman, and Samangy, 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 optimised 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 analysed 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.

The final concept discussed in this chapter 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, Hyman, and Samangy (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, Durán, and Marenco, 2014; Landers and Duperrouzel, 2017; Shah, Hyman, and Samangy, 2021; Deng and Zhong, 2020) 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 bettings 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, McAfee, and Suri (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.

In the following concluding paragraph 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 and 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, McAfee, and Suri (2014) into account.
The central research question is first divided into two sub-questions. These questions are *'How accurately can individual soccer player performances be predicted using historical data?'* and *'How accurately can individual soccer player performances be predicted using betting odds?'* After these two questions are answered respectively; the central research question can be answered, which reads as follows:

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

Chapter 4

Implementation

4.1 Business Context

4.2 Data

4.2.1 Procurement

4.2.2 Preparation

4.2.3 Exploration

4.3 Models

4.4 Application

Chapter 5

Conclusion

List of Figures

List of Tables

1 SPITCH Glossary 5
2 Concept Matrix 12

Source Code

Bibliography

[1] *About ASA*. URL: https://www.amstat.org/ASA/about/home.aspx?hkey=6a706b5c-e60b-496b-b0c6-195c953ffdbc (visited on 07/20/2021).

[2] Aminul Islam Anik et al. "Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms". en. In: *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*. Dhaka, Bangladesh: IEEE, Sept. 2018, pp. 500–505. ISBN: 978-1-5386-8279-1. DOI: 10.1109/CEEICT.2018.8628118. URL: https://ieeexplore.ieee.org/document/8628118/ (visited on 07/01/2021).

[3] Adrian Becker and Xu Andy Sun. "An analytical approach for fantasy football draft and lineup management". en. In: *Journal of Quantitative Analysis in Sports* 12.1 (Mar. 2016). Publisher: De Gruyter Section: Journal of Quantitative Analysis in Sports, pp. 17–30. ISSN: 1559-0410. DOI: 10.1515/jqas-2013-0009. URL: https://www.degruyter.com/document/doi/10.1515/jqas-2013-0009/html (visited on 06/02/2021).

[4] J Beliën, D Goossens, and D Van Reeth. "Optimization modeling for analyzing fantasy sport games". en. In: (2017), p. 16.

[5] F. Bonomo, G. Durán, and J. Marenco. "Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game". en. In: *International Transactions in Operational Research* 21.3 (May 2014), pp. 399–414. ISSN: 09696016. DOI: 10.1111/itor.12068. URL: http://doi.wiley.com/10.1111/itor.12068 (visited on 07/01/2021).

[6] Jan vom Brocke et al. "Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research". en. In: *Communications of the Association for Information Systems* 37 (2015). ISSN: 15293181. DOI:

10.17705/1CAIS.03709. URL: https://aisel.aisnet.org/cais/vol37/iss1/9/ (visited on 07/16/2021).

[7] Simon Demediuk et al. “Performance Index: A New Way To Compare Players”. en. In: (2021), p. 16.

[8] Wuhuan Deng and Eric Zhong. “Analysis and Prediction of Soccer Games: An Application to the Kaggle European Soccer Database”. en. In: *Insight - Statistics* 3.1 (Nov. 2020), p. 1. ISSN: 2661-3115. DOI: 10.18282/i-s.v3i1.332. URL: http://insight.piscomed.com/index.php/I-S/article/view/332 (visited on 07/01/2021).

[9] Steven J. Edwards. “Analyzing Fantasy Sport Competitions with Mixed Integer Programming”. en. In: *Data and Decision Sciences in Action 2*. Ed. by Andreas T. Ernst et al. Lecture Notes in Management and Industrial Engineering. Cham: Springer International Publishing, 2018, pp. 167–182. ISBN: 978-3-030-60135-5. DOI: 10.1007/978-3-030-60135-5_12.

[10] Leonardo Egidi and Jonah Gabry. “Bayesian hierarchical models for predicting individual performance in soccer”. en. In: *Journal of Quantitative Analysis in Sports* 14.3 (Sept. 2018). Publisher: De Gruyter Section: Journal of Quantitative Analysis in Sports, pp. 143–157. ISSN: 1559-0410. DOI: 10.1515/jqas-2017-0066. URL: https://www.degruyter.com/document/doi/10.1515/jqas-2017-0066/html (visited on 06/02/2021).

[11] Fabiano. *Fantasy football 101*. en-US. 2007. URL: https://www.nfl.com/news/fantasy-football-101-09000d5d80021ece (visited on 10/19/2021).

[12] Daniel G. Goldstein, Randolph Preston McAfee, and Siddharth Suri. “The wisdom of smaller, smarter crowds”. en. In: *Proceedings of the fifteenth ACM conference on Economics and computation*. Palo Alto California USA: ACM, June 2014, pp. 471–488. ISBN: 978-1-4503-2565-3. DOI: 10.1145/2600057.2602886. URL: https://dl.acm.org/doi/10.1145/2600057.2602886 (visited on 07/01/2021).

[13] Green. *’Wink’: Wilfred ’Bill’ Winkenbach invented Fantasy Football way back in 1962 with GOPPPL in Oakland - newsnet5.com Cleveland*. Sept. 2014. URL: 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 (visited on 10/19/2021).

[14] Michael Gusenbauer. “Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases”. en. In: *Scientometrics* 118.1 (Jan. 2019), pp. 177–214. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-018-2958-5. URL: http://link.springer.com/10.1007/s11192-018-2958-5 (visited on 07/16/2021).

[15] Hoffmann. *Millionen Daten pro Spiel: So werden Fußball-Statistiken erfasst.* de. Section: Sport. Nov. 2014. URL: https://www.hna.de/sport/fussball/millionen-daten-spiel-werden-fussball-statistiken-erfasst-4483893.html (visited on 10/19/2021).

[16] K. Karthik et al. “Analysis and Prediction of Fantasy Cricket Contest Winners Using Machine Learning Techniques”. en. In: *Evolution in Computational Intelligence*. Ed. by Vikrant Bhateja et al. Vol. 1176. Series Title: Advances in Intelligent Systems and Computing. Singapore: Springer Singapore, 2021, pp. 443–453. ISBN: 9789811557873 9789811557880. DOI: 10.1007/978-981-15-5788-0_43. URL: http://link.springer.com/10.1007/978-981-15-5788-0_43 (visited on 07/28/2021).

[17] Jonathan Robert Landers and Brian Duperrouzel. “Machine Learning Approaches to Competing in Fantasy Leagues for the NFL”. en. In: *IEEE Transactions on Games* 11.2 (2017), pp. 159–172. ISSN: 2475-1502, 2475-1510. DOI: 10.1109/TG.2018.2841057. URL: https://ieeexplore.ieee.org/document/8367900/ (visited on 07/01/2021).

[18] Roman Lutz. “Fantasy Football Prediction”. en. In: *arXiv:1505.06918 [cs]* (May 2015). arXiv: 1505.06918. URL: http://arxiv.org/abs/1505.06918 (visited on 06/02/2021).

[19] Tim Matthews, Sarvapali D Ramchurn, and Georgios Chalkiadakis. “Competing with Humans at Fantasy Football: Team Formation in Large Partially-Observable Domains”. en. In: (2012), p. 7.

[20] Alan M. Nevill and Roger L. Holder. “Home Advantage in Sport: An Overview of Studies on the Advantage of Playing at Home”. en. In: *Sports Medicine* 28.4 (1999), pp. 221–236. ISSN: 0112-1642. DOI: 10.2165/00007256-199928040-00001. URL: http://link.springer.com/10.2165/00007256-199928040-00001 (visited on 07/09/2021).

[21] Robert Rein and Daniel Memmert. “Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science”. en. In: *SpringerPlus* 5.1 (Dec. 2016), p. 1410. ISSN: 2193-1801. DOI: 10.1186/s40064-016-3108-2. URL: http://springerplus.springeropen.com/articles/10.1186/s40064-016-3108-2 (visited on 07/01/2021).

[22] Kushal Shah, James Hyman, and Dominic Samangy. “A Poisson Betting Model with a Kelly Criterion Element for European Soccer”. en. In: (2021), p. 16.

[23] Brian Skinner and Stephen J. Guy. “A Method for Using Player Tracking Data in Basketball to Learn Player Skills and Predict Team Performance”. en. In: *PLOS ONE* 10.9 (Sept. 2015). Ed. by Frank Emmert-Streib, e0136393. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0136393. URL: https://dx.plos.org/10.1371/journal.pone.0136393 (visited on 07/01/2021).

[24] Martin Spann and Bernd Skiera. “Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters”. en. In: *Journal of Forecasting* 28.1 (Jan. 2009), pp. 55–72. ISSN: 02776693, 1099131X. DOI: 10.1002/for.1091. URL: http://doi.wiley.com/10.1002/for.1091 (visited on 07/01/2021).

[25] SPITCH. *Points Catalogue*. URL: https://www.spitch.live/en/points-catalogue/ (visited on 10/21/2021).

[26] SPITCH. *Rules of Play and Participation:* URL: https://www.spitch.live/en/rules-of-play-and-participation/ (visited on 10/21/2021).

[27] SPITCH. *SPITCH | The Live Football Manager*. URL: https://www.spitch.live/en/ (visited on 10/21/2021).

[28] VHB e. V. *VHB-JOURQUAL3: Wirtschaftsinformatik*. 2015. URL: https://vhbonline.org/fileadmin/user_upload/JQ3_WI.pdf.

[29] Jane Webster and Richard T Watson. “Guest Editorial: Analyzing the Past to Prepare for the Future: Writing a literature Review”. en. In: (2002), p. 11.

[30] Edward Wheatcroft. “Profiting from overreaction in soccer betting odds”. en. In: *Journal of Quantitative Analysis in Sports* 16.3 (Sept. 2020). Publisher: De Gruyter Section: Journal of Quantitative Analysis in Sports, pp. 193–209. ISSN: 1559-0410. DOI: 10.1515/jqas-2019-0009. URL: https://www.degruyter.com/document/doi/10.1515/jqas-2019-0009/html (visited on 06/03/2021).

[31] Ronald Yurko, Samuel Ventura, and Maksim Horowitz. “nflWAR: a reproducible method for offensive player evaluation in football”. en. In: *Journal of Quantitative Analysis in Sports* 15.3 (Sept. 2019). Publisher: De Gruyter Section: Journal of Quantitative Analysis in Sports, pp. 163–183. ISSN: 1559-0410. DOI: 10.1515/jqas-2018-0010. URL: https://www.degruyter.com/document/doi/10.1515/jqas-2018-0010/html (visited on 06/02/2021).

Appendix A

A.1 Diagrams

A.2 Tables

A.3 Screenshots

A.4 Graphs

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.

Berlin, 13/09/2021

Jakob Heine