Abstract

Contents

Chapter 1

Introduction

1.1 Motivation

1.2 Aims

1.3 Methodology

Chapter 2

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 2.1 on page 6) 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 (2019), even if it investigates football instead of soccer.

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 (2019) 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. 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 important 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. Finally, missing data can be dealt with in three ways: from setting it to zero, giving it a mean value from similar players, and estimating it accurately. The latter is promising the most success.

Figure 2.1: 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

Placeholder: explain chosen gap / research question

Chapter 3

SPITCH

3.1 General

3.2 Game Rules

3.3 Scoring System

3.4 Competitors

Chapter 4

Data

4.1 Used Data

4.2 Descriptive Analytics

4.3 Data Preparation

Chapter 5

Modelling

5.1 Optimization

5.2 Prediction

Chapter 6

Machine Learning

6.1 Feature Selection

6.2 Model Selection

6.3 Model Training

Chapter 7

Evaluation

7.1 Combinatorial Model

7.2 Machine Learning Model

7.3 Performance Comparison

Chapter 8

Conclusion

List of Figures

List of Tables

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] 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).

[4] 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).

[5] 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).

[6] 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).

[7] 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 (June 2019), 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).

[8] 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).

[9] 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).

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

[11] 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.

[12] 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