



universität
wien

MASTERARBEIT / MASTER THESIS

Titel der Masterarbeit / Title of the Master Thesis

„Daily Fantasy Sports:
Ecology of a P2P Zero-Sum Game“

verfasst von / submitted by
David Georg Selig, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Diplom-Ingenieur (Dipl.-Ing.)

Wien, 2018 / Vienna, 2018

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet:

A 066 935

Studienrichtung lt. Studienblatt /
degree programme as it appears on
the student record sheet:

Masterstudium Medieninformatik

Betreut von / Supervisor:

Univ.-Prof. Dipl.-Ing. Dr. Helmut Hlavacs

Contents

1	Introduction	1
2	Literature Survey	5
2.1	General Data Analysis in Football	5
2.2	Traditional Sports Betting	7
2.3	Fantasy Games	9
3	Background	13
3.1	History	13
3.2	Market	14
3.3	Operators	15
3.4	Data Sources	16
3.5	3rd Party Tools	17
3.6	Challenges	19
3.7	Game Mechanics	21
3.7.1	Rules	22
3.7.2	Scoring	23
3.7.3	Formats	23
3.7.4	Multi Entries	24
3.7.5	Guaranteed Prize Pool (GPP)	24
3.7.6	Rake	24
3.7.7	Parimutuel vs. Fixed Odds	26
3.7.8	Bookmakers (Soft vs. Sharp)	26
3.8	DFS: A Peer-to-Peer Zero-Sum Game	27
3.8.1	Peer-to-Peer (P2P)	27
3.8.2	Zero-Sum Games	27
3.8.3	Strategies (and their EV)	27
3.9	Random Forest	28
3.10	Knapsack: A Linear Integer Program	29
4	Methodology	31
4.1	Study Design	31
4.2	Limitations	33
4.3	Datasets	33

4.4	Machine Learning	35
5	Practical Part I: Preparation & Machine Learning	37
5.1	Scraper	37
5.2	Preprocessing	40
5.3	Fantasy Points	41
5.3.1	Limitations	41
5.3.2	Distribution of Fantasy Points	42
5.4	Fantasy Market Values	43
5.4.1	Limitations	44
5.4.2	Distribution of Fantasy Market Values	44
5.5	Machine Learning: Random Forest	45
5.6	Machine Learning: AutoML (H2o)	51
6	Practical Part II: Simulation & Betting-Strategies	53
6.1	Betting Strategies	53
6.2	Lineup Creation	54
6.3	Simulation	56
7	Results	59
7.1	Plots	59
7.2	Tests	66
7.2.1	Projected Points	66
7.2.2	Backtested Points	66
8	Discussion and Conclusions	69
	Bibliography	73
	Appendices	81
	Abstract	91
	Abstract (Deutsch)	93
	Acknowledgments	95

List of Figures

3.1	Average performances of bettors (Picture taken from McKinsey & Company [68])	20
5.1	Fantasy Points: Histogram	42
5.2	Market Prices: Histogram	46
5.3	Random Forest: QQPlot of residuals (DRF)	46
5.4	Random Forest: Variable Importance (DRF)	48
5.5	Random forest: Variable Importance (XRT)	48
5.6	Random Forest: Scoring History (DRF) - RMSE	49
5.7	Random Forest: Scoring History (XRT) - RMSE	49
7.1	Projected Points for each Strategy and Matchday	60
7.2	Simulation Results on ground truth data for each Strategy and Matchday	60
7.3	Projected Points for each Strategy	61
7.4	Simulation Results on ground truth data for each Strategy	61
7.5	Cumulative Projected Points (Captain Bonus)	62
7.6	Cumulative Simulation Results on ground truth data (Captain Bonus)	62
7.7	Projected Points (deducted Fantasy Scores for Missing Players) .	63
7.8	Simulation Results on ground truth data (deducted Fantasy Scores for Missing Players)	63
7.9	Simulation Results on ground truth data (Total Buyins Won/Lost per Tournament-Type)	64
7.10	Simulation Results on ground truth data (Total Buyins Won/Lost per Strategy and Tournament-Type)	64
7.11	Simulation Results on ground truth data (Total Buyins Won/Lost per Strategy)	65

List of Tables

3.1	Scoring Points (Fanteam)	23
3.2	Example: Payout Structure	24
3.3	Fantasy Tournaments: General Information	25
3.4	Example: Heads-Up Game	25
3.5	ROI vs. Winrate	26
4.1	Technical Setup	35
5.1	Fantasy Points: Summary	43
5.2	Market Prices (1 st version): Summary	45
5.3	Market Prices (2 nd version): Summary	45
5.4	Results for DRF and XRT	47
5.5	Variations of Input Parameters	51
5.6	AutoML Results	52
6.1	Strategies for picking Fantasy Players	53
6.2	Fantasy Lineups: Missing Players	54
7.1	Tests for Normality (Shapiro-Wilk)	66
7.2	Comparison Matrix: Projected Points	67
7.3	Comparison Matrix: Backtested Points	67
1	Game Specific Details	83
2	Player Specific Data: FIFA	84
3	Player Specific Data: PES	86
4	Weather	87
5	Historical Betting Odds	87
6	Additional Features for Machine Learning	88

Listings

5.1	Example: R / Request (JSON)	37
5.2	Example: R / Request (rvest)	38
5.3	Example: R / Request (R-Package)	38
5.4	Example: R / Parsing	39
5.5	Example: Python / Scraper (Webdriver)	39
5.6	Example: R / Market Prices	43
5.7	Example: R / Market Prices (rescale)	44
6.1	Example: R / LP-Solver	54
6.2	Example: R / Simulation for Tournaments	56

Abbreviations

AIC Aikake Information Criterion

AUC Area Under Curve

DFS Daily Fantasy Sports

DRF Distributed Random Forest

EV Expected Value

ExpG Expected Goals

GBM Gradient Boosting Machine

GLM Generalized Linear Model

GPP Guaranteed Prize Pool

ITM In The Money

MAE Mean Absolute Error

MSE Mean Squared Error

P2P Peer-to-Peer

RF Random Forest

RMSE Root Mean Squared Error

ROI Return On Investment

TSR Total Shots Ratio

XRT Extremely Randomized Trees

Chapter 1

Introduction

Daily Fantasy Sports (DFS) is a sports betting product originating in the USA during the 1960s. The last few years have seen a tremendous up-rise in popularity due to a great variety of reasons, ranging from an increased availability of personal computers, mobile phones and internet access to ever-changing legal aspects as well as fine adjustments of the game's principles (e.g. daily contests instead of season-wide competitions).

Today, an estimated number of more than 56 million people (in the United States and Canada alone) are taking part in the game on a regular basis, transforming the ecology of DFS to an approximately \$7.22 billion market, while professional stakeholders within the industry try to gain additional momentum in Europe and emerging markets [1].

The games' principles are relatively simple: in contrast to traditional sports betting, people are not wagering their bets on a single team, but instead on their own fantasy lineup consisting of multiple individual players, which they have to pick following certain guidelines. An example for such a constraint would be to select a total of eleven players – with varying prices depending on their past performances – within a total budget of \$100M virtual dollars. Subsequently, punters can join a great variety of tournaments with their created fantasy lineups and compete against each other for money. On the following matchday, all events of corresponding games are assessed with a scoring system (each operator provides their own rules) and each previously selected fantasy player is awarded accordingly with fantasy points at the end of the competition. Typically, the owner of the lineup with the biggest total sum of fantasy points wins the tournament. The challenge of the game is therefore to predict the outcome for hundreds of real players on a given matchday as accurately as possible; as well as to maximize expected fantasy points of a – ideally – unique lineup (created by the punter).

While predicting results of football games is difficult due to the variance of results (even more so on player level), there exist a lot of well-known approaches, data sources and strategies for building fantasy teams. In consequence of the professionalization in sport over the last decades – accompanied by the intro-

duction of universally available statistics and betting odds – today’s games can be expected to take place on a fairly high competitive level. As a matter of fact, myriads of platforms evolved throughout the last couple of years, dedicated to helping their member base to share latest (injury) news, discuss strategies and even provide each other with advanced statistical knowledge or ready-to-use lineups – and the community itself plays a large role in customer retention and monetarization.

Given a diverse player pool consisting of regular, semi-professional sports data analysts with up-to-date information using predictive models and automatic lineup creation – as well as hobby punters composing their teams with pen and paper a few days well in advance before the start of a game (on top of that often picking fantasy players based on their personal preferences) –, it makes sense to expect very different outcomes for competing bettors if one takes a closer look at the underlying game principles. In fact, while the average Joe is reported to lose roughly about \$650 each year, so-called *heavy users* may well exceed those numbers relatively easily [1]. This also reflects experiences with other betting products in the grey area between gambling and skill; e.g. online poker, which hit a massive setback on Black Friday 2011 and since has been replaced by fantasy sports as a socially accepted game played for money [2].

Within the frame of my master thesis, I investigate the principles and mechanics of the game by collecting statistics on team and player level (completed by datasets including player ratings, weather information, historic betting odds and FIFA/Pro Evolution Soccer player data) and consecutively applying various machine learning methods to create predictions for fantasy scores.

In contrast to existing research in hindsight to sports betting (and more specifically Daily Fantasy Sports) – which typically focus on the predictive modeling part and rarely on meaningful metrics in the scope of fantasy games – a core concept of this work is to realistically imitate lineup creation based on punters’ beliefs and eventually compare resulting performances by backtesting all strategies in real-life scenarios (for various tournament types).

As *‘stacking blocks’* (selecting multiple players from the same team in hopes of a multiplication effect, e.g. when banking a clean sheet) is said to be one of the most reasonable approaches, multiple versions will be implemented and tested. Results are expected to show significant differences in mean fantasy points and won/lost buyins between strategies.

Existing academic research in the scope of sports betting will be reviewed in Chapter 2, mainly revolving around prediction models and further concepts in the scope of skill/gambling products.

In Chapter 3, an extensive overview about fantasy sports’ current state is given: starting with its history, information about the market and its stakeholders (e.g. operators, data providers, developers of 3rd party tools); providing details of the game (rules, formats); completed by some key basic concepts (bookmakers/odds, Random Forest/Knapsack algorithms) which are vital for understanding the context of this thesis.

Chapter 4 briefly lays down the steps which have been taken (and the respective methodology which has been followed), and further describes the technical setup, structure and reasoning behind my approach.

Eventually, Practical Part I (see Chapter 5) and Practical Part II (see Chapter 6) document the implementation step-by-step; from scraping, preprocessing and wrangling data to the creation of various models, leading to the simulations and tests for all betting strategies.

Final results are presented in Chapter 7 and subsequently discussed in Chapter 8 (extended by conclusions), which marks the end of my work.

Chapter 2

Literature Survey

Academic research within the scope of soccer can be roughly broken up into three main categories: a) basic concepts in regards to data analysis of football itself, b) a more specific approach in hindsight to traditional sports betting and c) fantasy game models and predictions. However, there are many points of intersections between these areas, as we will see in just a moment.

2.1 General Data Analysis in Football

Football is a mass sport played by professionals, amateurs and hobby enthusiasts all around the globe. There exist national leagues as well as international competitions, qualifiers, cups, friendly matches and many more. They are either organized on a state level (e.g. ÖFB¹ in Austria) or, on a higher level, for different regions (AFC², CAF³, CONCACAF⁴, CONMEBOL⁵, UEFA⁶, OFC⁷). Their umbrella association is the well-known FIFA⁸. Soccer organizations provide their members with access to a huge network of sponsors and broadcasters. A reported number of 1,013 billion people have watched the last world cup final (2014, Brazil) for at least a minute (695 million more than 20 minutes) [3].

Today, soccer is omnipresent due to its extensive media coverage: television, radio, news magazines, online blogs, social media – results, statistics and strategies are discussed by experts as well as respective fan bases. Whereas traditional perspective on the game (based on aggregated stats like possession, number of corners/offsides, pass success) still has its merits and is the quasi gold-standard for the general public; the modern age of ubiquitous internet ac-

¹<https://www.oefb.at/>

²<https://www.the-afc.com/>

³<https://www.cafonline.com/>

⁴<https://www.concacaf.com/>

⁵<https://www.conmebol.com/>

⁶<https://www.uefa.com/>

⁷<https://www.oceaniafootball.com/>

⁸<http://www.fifa.com/>

cess and electronic devices would already allow for an even deeper and more detailed approach – even for amateurs apart from a professional environment. It therefore comes as no surprise that many tactic blogs arose during the last couple of years (e.g. ^{9, 10, 11, 12, 13}). In contrast to typical discussions in sports media coverage, their authors focus on different reasoning – based on tactical variants, formations and player instructions – as opposed to the traditional method of analyzing single events of a game, supported only by sparse data about a specific player. New phrases, methods and concepts found their way into the discourse; and while there is still a lot to do to catch up on American Football or Baseball strategy and statistics, there is no denying that systematic approaches to soccer steadily gain in importance.

One such an example is the story of FC Midtjylland, a danish club which almost went bankrupt in 2012 and was bought by former hedgefund manager and successful sports bettor Matthew Benham. He instructed everyone of the club to put a lot more emphasis on statistical analyses of players, recurring game situations and even table rankings (*'model rating always, in any case, overrules the league position in evaluating our performance [...] The table is lying'*) [4].

This scientific approach has a lot of benefits as well as implications on existing processes within a club's operations – for example, scouts are now asked to focus on a player's character instead of his ability, as his skill has already been determined by analyzing his data.

While many clubs nowadays try to engage a so-called *concept trainer* in the hopes of having him implement a systematic framework for scouting, hiring, training and all fundamental game principles; a top-down approach like FC Midtjylland still is the exception. On the other hand, even if not every club and its environment allows for such a drastic *data revolution* to happen, more and more stakeholders recognize the advantages of statistical analyses – ranging from traditional, goal-oriented training control based on athletic performances of individual players (e.g. how many sprints at which speed per game have been measured or can be expected) to early detection of diseases or injuries (and many more). Academics and professionals of various fields work interdisciplinary together, combining recent findings in sports, medicine and computer science.

On a higher level, there exist statistical analyses for arising questions of any kind, debunking popular myths and misconceptions. It has been shown that the replacement of a coach within a given season has no measurable impact on a teams performance [5] – still, every single season, professional coaches lose their job due to the hope of an existing *'trainer effect'*.

Other insights from data analyses revolve around the value of scoring first or receiving a red card [6], goal scoring patterns [7, 8] or passing sequences in relation to shots and goals [9]. Most of these studies classify teams depending on their results and further investigate their similarities and differences on ob-

⁹<http://ballverliebt.eu/>

¹⁰<https://www.90minuten.at/>

¹¹<https://www.fourfourtwo.com/performance/tactics>

¹²<http://www.soccerpilot.com/tactic/tactic-selection.html>

¹³<https://spielverlagerung.com/2015/05/08/tactical-theory-compactness/>

served events (e.g. detecting situations as counterattacks, analyze success rates depending on player positions, pace, number of passes) [10].

Some key takeaways would be: 1) most statistics have no real value, except if they are combined with other information (e.g. interceptions are logged as a single event by most data providers – but without further knowledge about the specific situation, it is hard to draw meaningful conclusions) [11]; 2) as a consequence, new statistics and key performance indices have to be found (e.g. *'packing stats'* – a value which tells how many opponents were outplayed by a given action¹⁴); and 3) new tracking devices, algorithms and visualization methods will make their way into the community, further enabling players and teams to improve and track their performances (e.g. player heatmaps, coaching tools, training routines, ...) ¹⁵.

2.2 Traditional Sports Betting

While the general public still relies on outdated concepts, partly negating the explanatory power of data analyses, academic research of traditional sports betting focuses almost exclusively on statistical methods. One of the most important finding came from Maher [12] in 1982. He presented a model for goal scoring, based on the average (shot and conceded) goals of a given team, its opponent and their location (home or away). The proposed implementation of two independent Poisson distributions therefore has four parameters for each team: attacking strength (home and away) and defense strength (home and away). Though the Poisson distribution (as well as negative binomial distribution) has been seen as a good approximation for soccer results for decades [13], it has a few drawbacks – most of all, it underestimates low-scoring draws (which unfortunately play a major role in every league). A lot of improvements have been suggested by various authors, e.g. by bringing in additional parameters (based on the average number of goals in the very same competition) [14, 15] and correcting for the underestimation [16].

Apart from the underlying model of the goals' distribution, another very important factor in creating models for prediction of soccer results are other additional available data sources reflecting the performance of teams within a game. Unfortunately, most of public available statistics (such as received cards, fouls committed or conceded) show no correlation and significant influence on future outcomes [17].

Although many punters believe exclusively in the value of publicly available odds and therefore base their estimations about a game's expected result solely on them, it has been shown that models could be further improved relatively easily by additional factors (e.g. significance of a match, *'big team effect'* and more) [18]. As most of the bookmakers are offering fixed odds several days before a match, it is quite common that more recent information (e.g. a player's injury) will have no impact on the odds (at least most of the times). Even if a

¹⁴<https://www.impect.com/de/>

¹⁵<https://www.pro-soccerdrills.com/soccer-drills/soccer-tactics>

punter uses information of a sharp bookmaker – where odds will be reflecting the estimation of a match’s outcome of the whole playerpool, leading to the very best public available estimation – he is just guaranteed to work with the very same information as everyone else, has to consider whether the operator spread the obligatory rake evenly across the various possible game results and still possibly works with inaccurate numbers because of the underlying used distributions [19] in his model. One strategy or addition to an existing model though could be to track odds movement and react accordingly to spikes shortly before a game starts, as this could be an indicator for new information not taken into account yet.

As neither odds nor existing statistics provided enough explanatory power, new variations of expressing a team’s attacking or defending strength have been tested. A well-known way of adding more significant information to a model would be the total shots ratio (TSR) – simple relating goals shot to the total number of shots [20]. Depending on the level of detail, shots can be classified according to the position of a player on the field and his distance to the goal, the amount of opponents between him and the net, whether he shot with his preferred foot (or head) and many more variables.

This lead to the introduction of another very important concept within the area of soccer data analysis: expected goals (ExpG). Situations within a game are classified by their location on the field and inherent nature (open play shots, open play headers, penalties, in/direct freekicks, corners, throw-ins, rebounds from a goalkeeper save) and eventually awarded with their respective probability of leading to a goal. A penalty, for example, has a chance of roughly about 76% to be successful, thus leading to a very high score compared to an indirect freekick from midrange [21]. This approach has shown to be a superior predictor for future outcomes, as it combines quantitative data points with a structured interpretation of their specific contexts. Other useful information can be gained by looking at (percentage, number and distance of completed) passes in the final third or deep zone or total shots [22].

Research about a decay factor for performances (to express form of a team or account for changes in a gamesquad) has lead to inconsistent results. While incorporating weights to adjust model parameters to reflect a team’s current ability has been used by several authors to further improve the predictive ability of models [23, 24] many other authors refrain from using temporal information for their models.

In general, a lot of different algorithms and methods have been used in combination with a broad variety of underlying statistics – ranging from simple linear models [25] to hierarchical bayesian poisson models [26, 27] or iterative simulations (Markov Chain Monte Carlo, as in [15]). A comparison of different methods has been done by [28] (linear from stochastic gradient descent, naive bayes, hidden markov model, support vector machine and random forest) and [29] (linear regression, logistic regression, decision tree, bayesian networks).

Besides the great variety of models based on different variants of mathematical formulas and distributions, diverging data sources and computed statistics, there are also models which are simply based on other existing models. The au-

thors of [30] used ELO ratings within an ordered logit regression model, therefore defining expected match results as a function of rating differences.

Another use of publicly available data has been shown by the authors of [31]. By collecting data from Twitter and counting the occurrences of mentioned team and player names, they were able to rank them by their frequency, using the result as predictions for match outcomes (*'wisdom of ignorant crowds'*). While this simple approach lead to a consistent performance above chance (similar to the result of predictions based on official rankings), other methods – e.g. based on betting odds – have been proven to be superior.

Additionally, existing models are often extended by incorporating factors such as total transfer spending, league points, number of injured players, managerial changes, classification of teams (last year winner, additional international competitions, promoted team), team stress and fatigue (days of recovery between games) or squad instability (because of rotation or leaving players) [32].

Besides laying the theoretical framework of creating predictive models for football, there have also been numerous tests to implement value betting systems. In [33], the authors tested various money management systems, ranging from fixed bets or returns to the Kelly ratio or Markowitz portfolio management. Another interesting approach was presented by [34] – by incorporating individual player attribute data from EA SPORTS and KONAMI (FIFA, Pro Evolution Soccer) as well as the EA SPORTS Player Performance Indicator (PPI) ratings¹⁶, a model was created and tested using a Kelly [35] betting strategy.

However, no matter which models and betting systems are used [36, 37], they all share the same disadvantage for aspiring punters: the number of identified valuebets per season is typically too low, as is the expected return on investment.

2.3 Fantasy Games

Most of the literature in regards to fantasy games revolves around the creation of meaningful models for an individual player's performance – similar to traditional sports betting. Due to the harder task of predicting not only a single match outcome (or number of goals for both teams), but instead the performance of twenty-two different players (depending on an arbitrary scoring system), research is still in its early stages, especially for football.

Nevertheless, research about traditional sports betting still has its merits, since teams usually consist of players with a similar skill level and fantasy scoring systems value team-events (such as clean sheets) as well as individual contributions. Therefore, it is still important to predict a team's overall performance, which is then ideally combined with new approaches to account for a player's actions.

As has been stated before, *wisdom of ignorant crowds* can be applied to prediction tasks [31]. In hindsight to fantasy games, a lot of meaningful information could be gained from publicly available ratings assessed by fans and

¹⁶<http://www.salford.ac.uk/research/case-studies/case-studies/business-and-economy/a-player-performance-index-for-english-football>

experts. There exist data sources for each single match and player of a team¹⁷ (rated by public community) as well as professional player databases (e.g. ratings from computer games like FIFA¹⁸ or Pro Evolution Soccer¹⁹ – these values are updated on a regular basis, estimated by scouts and represent overall abilities as well as position-specific statistics). Their applications are manifold – FIFA video game ratings already have been used by multiple authors and for different reasons: in [38], ratings have been incorporated into a model and used to predict match results; while the authors of [39] studied whether FIFA’s individual player statistics reflect well-known strategic playing-styles (e.g. *tiki-taka*) by doing cluster analyses; and in [40], these ratings have been fed into a neural network to predict the marketvalue of a player.

Very interesting findings have been presented by Sarah Newell [41]. To the best of my knowledge, it is the only study so far focusing explicitly on *Daily Fantasy Sports* (contrary to researching season-wide challenges as in [42]). In her thesis about DFS for American Football, she describes her approach of using basic integer programming as well as stochastic integer programming for creating fantasy lineups. Instead of simply feeding the algorithm with discrete expected points, it is possible to define a probability distribution (assuming normality) for each individual player, thus allowing for optimizing a lineup for an overall high expected mean and variance. As fantasy tournaments can attract tens of thousand of participants, finding a unique set of players with a high ceiling is key for increasing expected payout long term – at least for tournaments with a top-heavy payout structure, contrary to e.g. Double-Up or Heads-Up formats. With the information of historic averages of fantasy points needed for placing in various payout tiers, a threshold can be defined within the program to guarantee that a created lineup meets certain minimum criteria, further allowing for calculating expected payouts. On top of that, by entering 35 MLB double up contests with randomly selected teams (and not finishing once in the money!), she was able to prove the very unlikeliness of DFS being a game of luck rather than a game of skill (as the probability of losing all 35 tournaments is 1 in 312,681,518) [41].

However, in hindsight to my work, a few limitations have been identified: as main topic of Newell has been modeling of the optimization of lineup-creation, expected points for players were predicted by simply taking the mean of the most recent season performances and a normal distribution was assumed for each and every player. Both of these issues were addressed by herself, also suggesting improving predictions and fitting individual distributions for each player. Furthermore, team variance within the stochastic integer programming was calculated by summing up all individual players’ variances (and taking the standard deviation). This does not account for players possibly performing better depending on their colleagues or for considering all possible resulting match scenarios depending on pre-match probabilities, which could turn earning predefined fantasy points thresholds for various payout tiers a little bit less

¹⁷<https://www.whoscored.com/>

¹⁸<https://www.easports.com/fifa?>

¹⁹<https://www.konami.com/wepes/2018/eu/de/>

meaningful, as I will discuss in more detail in Chapter 8. Newell states that taking covariances into the equation could contribute to solving that issue. Last but not least, fantasy rules of American Football differ vastly compared to traditional Football – particularly clean-sheet bonus for the whole defense, number of players per team/position and, of course, scoring rules within respective game mechanics.

Another major contribution to optimizing lineups in DFS contests has been provided by the authors of [43], accounting for their opponents’ typical favourite picks (popular fantasy players which get selected disproportionately often) and incorporating them into their version of an integer linear program. As a further consequence, they were able to investigate the gain in edges if, for example, an insider had access to the lineups of his opponent – or, alternatively, is able to predict his rivals’ picking choices by studying past behaviour of the playerpool.

An overview over recent research has been given – the overwhelmingly majority of authors created (a broad variety of) models by incorporating general team statistics as well as individual player ratings and special features (e.g. team fatigue, squad stability). They generally share in common that subsequent model evaluation is based on basic metrics such as root mean squared error (RMSE), area under curve (AUC) or similar – which makes a reasonable interpretation and comparison of a model’s quality for fantasy games hard, as true results depend as well on (unique) combinations of lineups, performance of opponents or arbitrary scoring rules provided by different operators.

This stands in contrast to research of traditional sports betting, which oftentimes includes testing predictions by backtesting historic data using specific betting strategies. By adding this component, one can investigate impact of different lineup-building strategies and further make assumptions about longterm profitability of certain approaches.

Precisely because there are so many publicly available existing models, ratings and player recommendations – while at the same time lacking reasonable evaluation of resulting predictions –, main focus of this thesis lies on the additional steps of building lineup combinations according to defined strategies and eventually running them against each other in typical competitive scenarios. In Chapter 6, a detailed step-by-step documentation will be given.

Chapter 3

Background

There are several concepts which have been covered in the following sections to further explain the mechanisms behind Daily Fantasy Sports tournaments. Besides specific rules and formats, which build the foundation for later analyses and simulations, a closer look is taken at bookmakers, betting odds, general game mechanics and algorithms used for predictions and creation of lineups. Beginning with its origin and evolution, an extensive overview of the ecology of DFS is given, including its stakeholders, 3rd Party Tools and current challenges.

3.1 History

Fantasy sports have gone a long way since Wilfred Winkenbach, an Oakland businessman and sports enthusiast, invented the first version of today's well-known drafting game in the 1950s. Back then, it was just meant to be a playful competition among friends – picking a team of professional golf players, simulating results by using historical data and comparing scores afterwards [44, 45]. A few years later, new rules and scoring systems for American Football and Baseball followed. Still, at that time, fantasy sports have been mostly organized and played by academics, who proceeded to pick up the idea and spread it within their circles [46, 47].

It took another decade until the game's principles were adapted to reflect real and recent results instead of just running realistic simulations. Ever since magazine writer and editor Daniel Okrent started playing with his friends and introduced fantasy sports to other journalists, things got rolling. In his essay *'The Year George Foster Wasn't Worth \$36'* in Inside Sports [48], Okrent explained the rules of the game for the first time to a broader audience. From that point on, more and more people started to take interest in playing. Guide books, first statistical tools and analyses followed [49]. A few years later, thousands of leagues were organized in the USA. Fantasy sports had become a widespread hobby by the end of the 1980s [50].

Around the turn of the millennium, ubiquitous availability of internet access

and computers enabled even more users to participate in fantasy leagues. Media publishers, sports companies and gambling operators tried to create profitable business models to monetize fans’ engagement [51]. One of the first gambling companies to introduce DFS to their existing portfolio was the Montana Lottery in 2008 [52]. Many others followed – but most notably, market entrance from two rivaling companies, DraftKings and FanDuel, marked a significant shift in the growth process. With the help of massive investments, extensive advertising and sponsoring of official sport leagues, both of them managed to acquire an overwhelming chunk of market share until 2014 [53, 54, 55].

Nowadays, fantasy sports tournaments are much faster paced compared to their traditional season-wide versions – each and every real matchday can be used to offer new challenges to fans. Due to the massive amounts of simultaneously available tournaments, its large user base and the up-rise of statistical tools and strategy discussions, DFS can be seen as an emerging alternative to online poker [53, 56]. However, this also includes all issues and complex problems which may arise in regards to legality in different states. Right now, each and every state in the USA as well as in Europe make their own decisions based on their respective laws. Most important factor is the declaration of DFS being primarily a game of chance or skill [57, 58].

Up until now, fantasy products in Europe have not quite taken off yet. There are many explanations for that, ranging from criticism about the product itself (e.g. *‘too stats-based’*) to more profound reasons – like missing legal requirements, a lack of political willpower or simply completely different market conditions (such as widespread existence of legal live sports betting in Europe, which is not allowed in most states of the USA) [59]. Only time will tell whether Daily Fantasy Sports can gain new momentum and attract even more fans all around the globe.

3.2 Market

A lot of things have changed since Winkenbach invited his colleagues to a fantasy golf game. In 2003, almost 50 years after its invention, more than 15 million punters were already playing fantasy football, spending an average of more than \$150 per year for their new favourite hobby. DFS had become a billion dollar industry and growth still proceeded [60].

In 2017, about 60 million punters were betting on fantasy games (USA and Canada only!) – compared to 500,000 back in the 80s. The typical DFS player is male (71%), 32 years old, has a college degree or more (50%), is full-time employed (67%), has an household income of \$75k+ (53%) and spends \$653 per year (2012: \$80, 2015: \$465, 2016: \$556) [61].

Unsurprisingly, more and more companies try to take advantage of the market’s potential, as it has been shown that offering fantasy sports can also lead to increased engagement. Fans participating in fantasy leagues watch significant more games on television, which results in greater demand and therefore profitable advertising opportunities [62, 63]. A survey of the NFL in 2002 revealed

that fantasy players were consuming 8.4 hours of NFL per week, compared to the average customer with 6.6 hours [64].

On top of that, acquisition costs are said to be lower (1%) compared to more traditional sports betting or casino games, as there is less competition within this new sector. Oulala¹, one of the first european fantasy football providers, claims to attain a cost per click of 0.84 € (casino: 88.59 €). Furthermore, punters tend to become loyal enthusiasts (playing fantasy sports for an average of 9.51 years) [65].

3.3 Operators

It is relatively easy to acquire a betting licence – at least if sports betting is defined as a skill game in a country’s law. Commercial operators offer games for real cash (e.g. USD, €, Bitcoins, ...) as well as free versions of their product:

- <https://www.draftkings.com/>
- <https://www.fanduel.com/>
- <https://www.fantasybet.com/>
- <https://www.fantasydraft.com/>
- <https://www.fanteam.com/>
- <https://www.oulala.com/>
- <https://www.playon.co/>

Besides commercial operators, there also exist alternatives completely free of charge:

- <https://www.dreamteamfc.com/>
- <https://www.fantasy.premierleague.com/>
- <https://www.sports.yahoo.com/dailyfantasy/>

In central europe (Germany/Austria), similar products already exist or have been developed, though they have been adapted for local markets:

- <https://www.comunio.com/>
- <https://www.foosio.com/>
- <https://www.kicker.de/games/classic/startseite/>

On top of that, there are many B2B-solutions available:

¹www.oulala.com

- <https://www.betconstruct.com/>
- <https://www.gdfsi.com/>
- <https://www.oulala.com/>
- <https://www.scoutgaminggroup.com/>

3.4 Data Sources

There are many commercial data providers which offer subscription services depending on their customer's needs – most of them are targeting bookmakers or other professional companies in the scope of sports betting, analyses or games. Though they differ in regards to their pricing models (according to which sport, leagues or parameters are requested), they all make use of REST-APIs or similar technical implementations:

- <https://www.nowgoal.com/>
- <https://www.optasports.com/>
- <https://www.spapi.pw/>
- <https://www.statscore.com/>

Apart from that, there exist a couple of communities and initiatives to collect and share freely available sports data. All too often, their data is not up-to-date and there are issues in regards to data quality. Datasets are usually stored in git-repositories or as *.csv-files:

- <https://www.footytube.com/>
- <https://www.github.com/jokecamp/FootballData/>
- <https://www.github.com/openfootball/>
- <https://www.openligadb.de/>

Somewhere in between are projects revolving around linked data [66] as well as other (partly commercial) providers. They rarely offer convenient interfaces for exporting data, instead focusing on publishing information visually appealing for fans:

- <https://www.espnfc.com/>
- <https://www.eplindex.com/>
- <https://www.fourfourtwo.com/>
- <https://www.fussballdaten.de/>

- <https://www.fussballstatistik.net/>
- <https://www.fussball-statistiken.com/>
- <https://www.kicker.de/>
- <https://www.soccervista.com/>
- <https://www.squawka.com/>
- <https://www.uefa.com/>
- <https://www.whoscored.com/>

An extensive list of available football data sources can be found at:

- <https://www.jokecamp.com/blog/guide-to-football-and-soccer-data-and-apis/>

A more detailed look into which statistics were collected (match-events, timelines, aggregated stats, lineups, player ratings, historic tables, ...) can be found in Section 4.3.

3.5 3rd Party Tools

Alongside the various platforms to play and compete against other users, there are more and more additional features, apps and tools provided by other stakeholders within the domain of fantasy sports. To some extent, their creators and owners also try to build a community around their product. For example, there are solutions for (private) league owners to create and organize leagues or even to collect dues from participating users:

- <https://www.clickydraft.com/draftapp/>
- <https://www.draftnight.com/>
- <https://www.fandraft.com/>
- <https://www.leaguesafe.com/>

Sites like Rotowire² provide punters with recent news, lineup updates, injuries – and on top of that, strategic advice, tournament tracker for various DFS-betting sites and a large, active community. Similar to that, Fantasy-FootballNerd³ offers weather forecasts, player rankings and projections, picking suggestions or even more advanced tools (lineup analyzer). Other sites within this range would be:

²<https://rotogrinders.com/>

³<https://www.fantasyfootballnerd.com/>

- <https://www.euroclubindex.com/>
- <https://www.bettingexpert.com/>
- <https://www.fantasyoverlord.com/FPL/TeamPointForecasts/>
- <https://www.matchodds.org/>
- <https://www.predictz.com/>
- <https://www.11heroes.com/>

Most of these communities do not provide their users with pre-built and ready-to-use lineups, instead urging them to take part in discussions and to make up their own mind about the following game week. Nevertheless, a lot of subscription services for automatic lineups can be found very easily, though they usually focus on single fantasy platforms, rarely run an official business and practically never give insight into their lineup-building process, therefore appealing shady or at least nontransparent.

More reliable and useful information (e.g. official lineups, last-minute injuries) is published about an hour before the start of each match, updated on a regular basis and can be found at Rotowire⁴, Twitter⁵ or similar services. Of course there are also well-known providers of traditional sports bets, whose APIs can be used to acquire knowledge about a team's form, results, statistics or odds:

- <https://www.developer.betfair.com/exchange-api/>
- <https://www.developer.sportradar.us/page> (free 90 day trial)
- <https://www.developer.yahoo.com/fantasysports/>
- <https://www.pinnaclesports.com/en/api/> (free for bettors and affiliates)

Interestingly, as the game evolves, more and more similarities to the up-rise of online poker can be seen: just like Sharkscope⁶ has provided dedicated online poker players with detailed statistics of their opponents, fantasy tournament results are scraped and made available on sites like DFSstats⁷. By simply typing in the nickname of a player, historical results are shown for Fanteam⁸, Fantasybet⁹, Drafton¹⁰ and Playon¹¹, allowing for selecting opponents based on their past performances.

⁴<https://www.rotowire.com/soccer/soccer-lineups.htm>

⁵<https://www.twitter.com/>

⁶<https://de.sharkscope.com>

⁷<http://dfstats.net/>

⁸<https://www.fanteam.com/>

⁹<https://www.fantasybet.com/>

¹⁰<https://drafton.com/>

¹¹<https://playon.co/>

Besides all mentioned sources, football fans around the globe do their research by studying tables, recent performances of teams and players, news or blogs committed to football strategy:

- <https://www.90minuten.at/>
- <https://www.abseits.at/>
- <https://www.ballverliebt.eu/>

As has been shown, there exist a broad variety of publicly available information sources, ranging from raw quantitative data about a specific match, to complete lineups or even detailed tactical interpretations.

3.6 Challenges

Daily Fantasy Sports certainly face a lot of challenges given their nature of being a game in the grey area between chance and skill. Besides the variety and complexity of already existing laws and their enforcement in any given country, global access to DFS games enables operators based in Malta, Gibraltar or the Isle of Man to offer their products cross-border. Whereas licensing of sports betting can be challenging or even impossible in other states with a regulated gambling industry – due to the efforts in regards to responsible gaming and protection of players –, these countries support companies by offering them a safe haven for their operations.

Not only are there legal issues, there are also recurring debates about game mechanics. One such an example was Woofbet’s (now: Fanteam¹²) model for fantasy player market prices: the more points a player collected in the course of a season, the more expensive he became (because all of his scored points up to that point were simply accumulated to calculate a respective market value). This led to ever increasing prices for fantasy players, leaving customers unsatisfied because building lineups towards the end of the season became almost impossible within the provided virtual budget of \$100M. A typical solution for this problem could be to include a player’s ownership in his market-price calculation – the more he got selected by punters in the preceding game week, the more expensive he should become in the following week, leading to a new dynamic by decreasing his price-value-ratio in regards to his expected performance. This can be summed up as a problem of inefficient pricing – salaries could be assigned more efficiently by using *wisdom of the crowd* (user ratings).

Tweaking game mechanics is also key in respect to user churn rates and retention, or the so-called ‘*ecology*’ in general. As people get more and more advanced in using statistical analyses and tools to build their lineups, it becomes harder for new and recreational players to win their fair share of the prize money. With the introduction of multi-entries at tournaments, *sharks* (winning players) benefited even more. A sample consisting of Baseball fantasy competitions in

¹²<https://www.fanteam.com/>

the first half of 2015 has shown that only a very small percentage of players (1.3%) won 91% of all profits [67] - while only contributing for around 40% of entry fees [68] (see Figure 3.1).

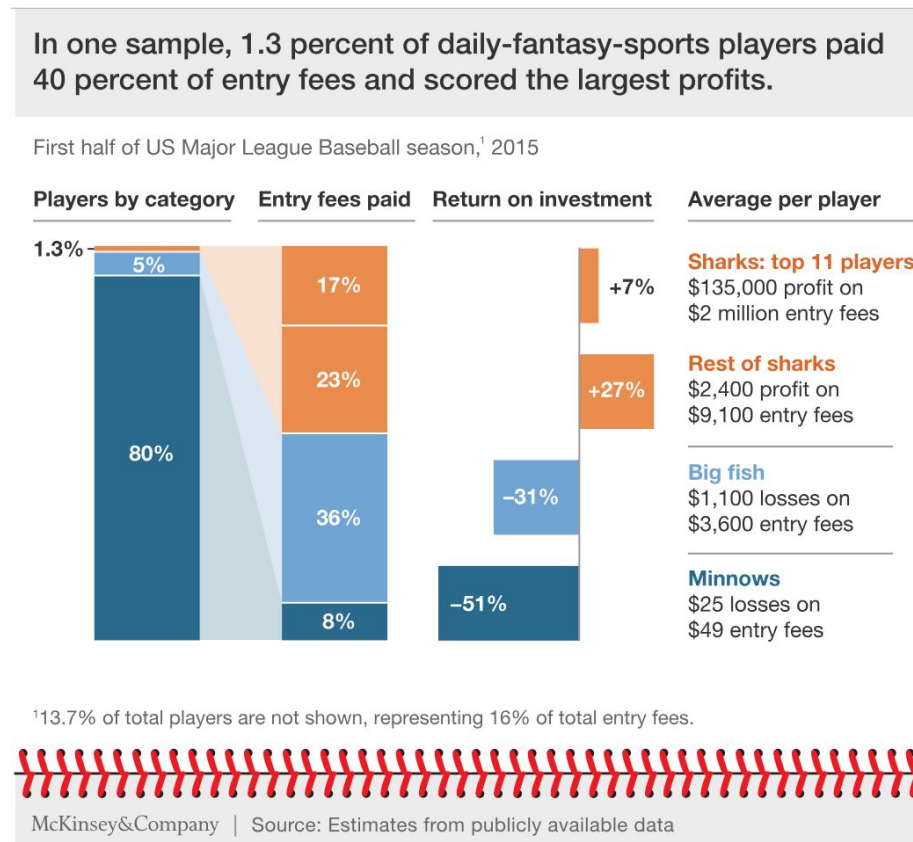


Figure 3.1: Average performances of bettors (Picture taken from McKinsey & Company [68])

In detail, the top 11 players spent an average of \$2 million in entry fees, netting an average profit of \$135,000 for a return on investment of 7%. Contrary to that, the bottom 5% of players ended with a loss rate of 31%, accumulating to astonishing 75% of all losses (more than \$4,000 per year and person).

Whereas the big majority of players in between loses money in the long run, too, variance of the game itself may disguise this fact for most of the punters (leaving them at least entertained or lucky once in a while in their best case scenario). The current state of DFS therefore can be seen as a reflection of other games in the grey area of chance and skill, following the universal principle of competitive semi-gambling products: *heavy users* bringing in the money, *regulars* raising the traffic and average Joes looking for opportunities

and entertainment on a non-regular basis (also called *recreationals*) [67].

It is hard to argue for Daily Fantasy Sports being either a game of chance or skill exclusively, as the answer depends on one's respective perspective. On the one hand, there obviously exist a number of consistent winners across multiple platforms, sport types and game formats. Advanced statistical knowledge, higher quality of information and expertise could give punters an edge over their competitors (even with rake being as high as 10%). On the other hand, laws often specify games of chances differently, defining them e.g. to be '*less than 50% skill-based*' or dependant on '*external, non-quantifiable factors*' [69]. Alternatively, DFS punters – similar to poker players – generally argue with sample-size, relative amount of rake and possible edges – contrary to that, courts may focus on a single poker hand or fantasy tournament to define the impact of chance [70].

While this discussion is led mostly by industry experts, lawyers and business partners, individual players face other, more practical challenges: be it winning too much too early (as companies try to '*make them early winners*' [71] to decrease user retention and increase their lifetime customer value), losing too fast (due to increased edges between players because of statistical tools, result tracking sites and advanced strategies) or simply losing too much (because of risky gambling behaviour [72]).

But even more problems may (again) arise on the horizon. New business models like freemium games may cover up expenses and losses of punters, making it even harder for them to understand how much they are losing. On top of that, unlicensed operators and anonymous payment methods like Bitcoin could provide unlimited and unrestricted access to gambling products for underaged – an example for another hot topic in regards to betting of minors has been unregulated skin-betting in the realms of esports [73]. Even football itself - just like any popular sport - also has its well-known scandals and problems, ranging from corruption to money laundering or fraud.

Still, the biggest setback for fantasy sports has been self-made: in 2016, a DraftKing employee participated in tournaments on a competing platform. By using internal information about player's ownership, he was able to create unique lineups and won \$350,000. These events brought up questions regarding insider trading, data ownership and responsibility [74, 75]).

3.7 Game Mechanics

Due to their ubiquitous availability and rising popularity amongst punters in the course of the last decade, Daily Fantasy Sports have been further developed to match user needs by successively adding new formats and sports to their portfolio. On top of that, the well-known traditional season-long games have been divided into shorter cycles (e.g. just betting on a single matchday of a season on any given weekend), hence leading to a faster-paced version of the original game. An overview about the current state of game mechanics is covered below.

3.7.1 Rules

At the beginning of a round, a couple of real matches are selected by the operator (e.g. matchday 32, Germany's 1st division, kick-off 28.04.2018 15:30):

- FC Bayern Munich (vs.) Eintracht Frankfurt
- Hertha BSC (vs.) FC Augsburg
- VfL Wolfsburg (vs.) Hamburger SV
- SC Freiburg (vs.) 1. FC Koeln
- FC Schalke (vs.) VS Borussia Moenchengladbach

Every punter is only allowed to select players from those teams previously specified and shown in a tournament's lobby. As soon as he enters a fantasy tournament by paying the participation fee, he receives a certain amount of virtual money (typically a hundred virtual millions) in exchange.

From that moment on, a user's goal would be to select eleven players (plus one substitute) out of all available teams which he expects to perform well in real life. Events which are rewarded with fantasy points are defined in the corresponding game rules of a platform (e.g. Fanteam's rules¹³ – which have been used for assessing performances of betting strategies throughout this thesis).

While the price of a certain fantasy player depends mainly on his past performances, it may also be influenced by the predicted skill of his real-life opponents or by his ownership across all lineups within a competition. As a consequence, star players at Fanteam can cost up to around \$15M – compared to non-playing substitute squad members, which usually are priced around \$2.5M (even if they have not played a single game yet). Besides managing budget restrictions, punters have to meet the following preconditions:

- create a valid formation (GK=goalkeeper – DEF=defender – MID=midfielder FWD=forward), e.g. 1-5-4-1, 1-5-3-2, 1-4-3-3, 1-4-4-2, 1-4-5-1, 1-3-5-2, 1-3-4-3, 1-5-2-3
- define a SUB=substitute player (if a selected player misses a game, he will be automatically replaced)
- a maximum of three players from each team is allowed
- pick a captain/vice-captain (fantasy points of this player will be doubled)
- lineup can be altered as long as none of the matches has started yet
- at the end of a round, all fantasy points gained by the players of a chosen lineup are summed up and teams are ranked according to their totals
- depending on the format, a given number of participants receive their payout

¹³<https://www.fanteam.com/support-center/football-rules>

3.7.2 Scoring

Selected fantasy players earn points for various events (see Table 3.1, taken from Fanteam). Scoring rules as well as constraints for creation of lineups vary across different fantasy football sites, leading to unique game characteristics depending on number and type of statistics used.

Table 3.1: Scoring Points (Fanteam)

Event	Fantasy Points
Playing time up to 60 minutes	1
Playing time over 60 minutes	2
Playing full match (no substitution)	3
Goal (defender or goalkeeper)	6
Goal (midfielder)	5
Goal (attacker)	4
Assist	3
Clean sheet (defender or goalkeeper)	4
Clean sheet (midfielder)	1
Yellow card	-1
Red card	-3
Keeper saves (for every 3 saved shots on goal)	1
Causing a freekick (which leads to direct goal)	-2
Causing a penalty (goal or not)	-2
Achieving a penalty (if a teammate scores)	3
Penalty kick save by keeper	5
Penalty missed by shooter	-2
For every 2 goals conceded (defender and goalkeeper)	-1
Own goal	-2
Team 'wins' in the periode player is on the pitch	1
Team 'draws' in the periode player is on the pitch	0
Team 'loses' in the periode player is on the pitch	-1
Captain	double

3.7.3 Formats

There exist at least four different formats for a typical daily fantasy tournament:

- **Heads-Up (1 : 1)**: Two punters challenge each other with their chosen lineup, the winner receives the money of both bettors minus rake (except if there is a tie – in that case, the money is spread accordingly between both users).
- **Double-Up (50 : 50)**: If there is no limit for participants, any number of punters can join – but only the upper half of them wins (about twice their initial bet – minus rake).

- **One-vs-All** (1 : *ALL*): Similar to the double-up-format, multiple users can join this tournament, but typically only the best 15% of lineups will be rewarded. The payout-structure is top-heavy and similar to poker tournaments.
- **Winner Takes It All** (1 : *ALL*): Similar to 1:ALL, except that only the first rank will be awarded (100% - minus rake).

Table 3.2 shows an example payout structure for a typical top-heavy One-vs-All (1 : *ALL*) tournament, while Table 3.3 lists number of participants and places paid as well as points needed for a tournament win or a position in the money (ITM) on Fanteam between January and May 2018.

Table 3.2: Example: Payout Structure

Position	Payout in %
1	12.73
2	8.49
3	6.36
4	5.09
5	4.24
6	3.46
7	3.18
8	2.83
9	2.55
10	2.31
11-76	between 0.33 and 2.12
77-506	0.00

3.7.4 Multi Entries

These are tournaments where multiple lineups per punter are allowed. By varying formations, fantasy players or captain selection, variance of results can be decreased.

3.7.5 Guaranteed Prize Pool (GPP)

Tournaments with a GPP are provided by operators and offer a valuable betting option for punters, if there are too few teams participating in the contest. Existing overlay is often called '*dead money*'.

3.7.6 Rake

This is the fee charged by the operator – typically amounting to 10% of the buy-in (e.g. 10+1 €). It has direct implications to game theory/strategies,

Table 3.3: Fantasy Tournaments: General Information

ID	Players	Paid	Points (ITM)	Points (1st)	Payout % (1st)
96520	1010	154	89	126	9.86
95348	668	102	56	89	11.88
94535	704	110	82	106	11.75
93758	681	104	76	95	11.83
92891	713	108	69	96	11.73
92090	721	109	53	75	11.68
90765	709	108	88	115	11.73
90043	466	72	87	106	2.24
89230	782	119	82	113	11.47
88334	790	113	80	100	11.44
86529	778	115	68	87	11.49
84669	785	118	66	89	11.47
83695	737	110	70	97	11.63
81351	903	137	94	130	11.11
Total	746.21	112.79	75.71	101.71	10.81

expected win/loss rate, return on investment or risk of ruin and therefore plays a central role. Similar to the poker ecology, there exist possibilities to decrease the relative rake by making use of a first deposit bonus, free tournament tickets or rakeback (which is less common nowadays).

Let's take a look at an example: a user plays 1:1 Heads-Up fantasy sports tournaments against an opponent. The buy-in for both participants accounts to 10+1 € (each), therefore a total of 20 € is added to the prize pool and 2 € are deducted by the operator for its services, infrastructure and staff.

Let's say both punters play a hundred games against each other and they both win exactly 50% of the time. Each one of them has to pay a hundred buy-ins ($2 \cdot 11 \cdot 100 = 2200$ € in total). 200 € will be deducted by the operator (rake), which leaves an average of $2000/2 = 1000$ € to win for both punters. Return on Investment (ROI) can be easily calculated (see Table 3.4): $((\text{Tournament Winnings} - \text{Tournament Buyins}) / \text{Tournament Buyins} * 100)$.

Table 3.4: Example: Heads-Up Game

Statistic	€
Paid Buy-Ins (per user)	1100
Rake (per user)	100
Total Prize Pool	2000
Prizes won (per user)	1000
Loss (per user)	-100
Return on Investment (ROI)	-9.09 %

Table 3.5: ROI vs. Winrate

Winrate	Calculation	ROI in %
0 %	$(0.00 \cdot 2000 / 1100 - 1) \cdot 100$	-100 %
40 %	$(0.40 \cdot 2000 / 1100 - 1) \cdot 100$	-27.27 %
50 %	$(0.55 \cdot 2000 / 1100 - 1) \cdot 100$	-9.09 %
55 %	$(0.55 \cdot 2000 / 1100 - 1) \cdot 100$	0.00 %
56 %	$(0.56 \cdot 2000 / 1100 - 1) \cdot 100$	1.82 %
57 %	$(0.57 \cdot 2000 / 1100 - 1) \cdot 100$	3.64 %
58 %	$(0.58 \cdot 2000 / 1100 - 1) \cdot 100$	5.45 %
59 %	$(0.59 \cdot 2000 / 1100 - 1) \cdot 100$	7.27 %
60 %	$(0.60 \cdot 2000 / 1100 - 1) \cdot 100$	9.09 %

It is easy to see that both of them are expected to lose money long-term because of the rake (given their strategies deliver the very same results). To be able to play at least break-even, a user would have to win a minimum of $1100/2000 \cdot 100 = 55\%$ of the games. Table 3.5 lists examples for winrates and respective expected ROIs (calculated by $(Winrate * Total Prizepool / Buyins - 1) \cdot 100 = ROI \text{ in } \%$). It is evident that even seemingly small changes in winrate have a large effect on the long-term ROI for a punter.

3.7.7 Parimutuel vs. Fixed Odds

In parimutuel betting systems, an individual's outcome depends on the amount of all other wagers within the very same betting round. Only when all bets have been closed, odds for the various game results are calculated and approved. Contrary to that, fixed odds are provided within a given timeframe before a match. Though odds can always be altered by bookmakers with regard to new information, a user usually keeps his initial wage and odds (if he does not decide to cancel his first bet and place another one instead) [76].

3.7.8 Bookmakers (Soft vs. Sharp)

Bookmakers offer a broad variety of different sports and betting markets to punters, ranging from ordinary game results to (first) goal scorers, clean sheets, over/under, handicap and many more. Usually odds are provided in the background by B2B companies like Betradar¹⁴ and then accordingly adjusted for a bookie's specific customer segment to increase expected margin and minimize payout risk.

A soft bookmaker meets these requirements by considering the well-known '*favourite-longshot bias*' [77]), which results in decreasing the odds for a favourite's win in advance, as most of the bettors tend to prefer to pick a '*safe*' bet. Although odds can still be lowered even further as more and more users are placing

¹⁴<https://www.betradar.com/>

their bets, changes are made just occasionally and oftentimes manually. In contrast to that, sharp bookmakers simply set the initial odds for a game and open the betting market for users. Depending on the amount of money placed on a specific result, odds change accordingly, reflecting all the information about a game's result tendency (*wisdom of the crowd*). An example for such a (sharp) bookmaker would be Pinnacle¹⁵.

It is important to understand the difference between these two approaches: odds provided by soft bookmakers do not reflect the probabilities of a game, they are calculated to allow operators to safely earn their margin and are therefore at least a bit skewed against the favourite. Though they often do not differ by much (if various rake levels are considered), odds from sharp bookmakers are a better indicator of people's true beliefs about the outcome of a match.

3.8 DFS: A Peer-to-Peer Zero-Sum Game

3.8.1 Peer-to-Peer (P2P)

Due to the ubiquitous availability of personal computers, mobile devices, access to internet, (football) sports data and betting exchanges, punters have been enabled to betting online against each other without the need for a bookmaker to set fixed odds for them [34]. Therefore, offering sports betting products with lower risk can be accomplished by licensed providers, supplying customers with Guaranteed Price Pools, forms of rake-back, deposit bonus or similar. At the same time, it allows bettors to face other people instead of an anonymous playerpool and as a result, to select their opponents depending on their past performances and estimated skill level.

3.8.2 Zero-Sum Games

It has been shown that products in the grey area of gambling and skill can be interpreted as zero-sum games [78]. This is also expected to be true for Daily Fantasy Sports, as many studies already successfully have investigated the influence of skill being predominant for results (e.g. [41]), despite not all of the collected fees being attributed to the prize money as operators take out a certain percentage for their own benefit – the so-called rake. For all contestants combined, this means the game has a negative expected value (EV) – nevertheless, depending on the impact of skill (e.g. predictions of fantasy points, optimizing of teams, uniqueness of lineup, collecting last-minute information), some participants still are able to accomplish a positive EV in the long run.

3.8.3 Strategies (and their EV)

Due to the nature of the game, finding a theoretical proof for a dominant strategy which overrules all others would make up for quite a challenging - or rather

¹⁵<https://www.pinnacle.com/>

unsolvable - task. Up to this day, prediction models for football results are simply not remotely accurate enough, which naturally leads to large variances in their outcome. Nevertheless, even if all scenarios on a given matchday would have the very same probability, there could be possibilities for individual punters to at least improve their (probably negative) EV by choosing unique lineups instead of going for their favourite teams or similar irrational decisions. A study about parimutuel soccer betting (*Toto*, a totalisator game from Austrian Lotteries with the goal of correctly predicting tendencies for twelve soccer games per round), has shown that the well-known favourite longshot bias in betting leads to an overuse of certain betting-combinations. In that case, if punters win the jackpot having selected typical favourites, there is a great chance of them having to split the prize with many, many others [79].

3.9 Random Forest

Before H2o's auto-machine learning is used to create a Stacked Ensemble of multiple various models during practical work (see Chapter 5), a pre-analysis is done by making use of (Distributed) Random Forest (DRF) (see Chapter 5.5).

Random Forest (RF) is a well-known machine learning algorithm which has widespread application due to its simplicity, speed and reasonable performance for most tasks. It can be used for classification as well as regression problems. Furthermore, it allows for valuable insights into how the resulting model processes data input by calculating feature importance [80]. While RF is typically used as a supervised algorithm, some of its variants can also be implemented as unsupervised machine learning [81].

Introduced by Ho in 1995 [82], RF showed to deliver superior results compared to traditional decision trees, which are typically prone to overfitting due to their sensitiveness towards noise in data sets. This problem can be attacked by building several individual decision trees on randomized training data (*'bootstrapping'*), followed by a process of combining and averaging their results - which usually leads to decreased variance as well as higher accuracy. However, a fundamental condition which has to be fulfilled for this approach to succeed, is that all fitted trees are not correlated with each other (as this would again increase generalization error). This can happen if only a small number of features influence predictions heavily, as they will show up in almost any single tree. For this reason, *'feature bagging'* (also known as *'subspace method'*) has been introduced, which results in iterative training on randomized selections of features instead of just different training samples [83, 84, 85].

Another additional form of randomization is implemented by Extremely Randomized Trees (XRT). On top of sampling training input data as well as using random feature selection, even the splitting rules of trees - which are usually chosen by identifying the most discriminative thresholds - are randomized [86]. That way, variance can be typically expected to decrease, though at the same time bias tends to increase¹⁶.

¹⁶<https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drfs.html>

In regards to the topic of my thesis, RF have been used widely by multiple authors (see [28, 29, 87, 88, 89]), who are using the algorithm’s versatility for classification as well as regression tasks.

3.10 Knapsack: A Linear Integer Program

In Chapter 6, a variant of a linear integer program will be implemented for the task of fantasy lineup creation - including constraints for budget, maximum players per team and their respective expected value.

This problem can be solved by the ‘*knapsack*’ algorithm, whose origins date back in 1897 [90]. Due to its capability of finding proper combinations of items with different weights and values - eventually maximizing their total sum -, it has widespread application in a great variety of fields, ranging from portfolio management [91] to cryptosystems [92].

There exist multiple versions of knapsack linear programming implementations, which differ primarily in their restrictions in hindsight to input or output data. For example, a problem could be formulated as *bounded* (without the restriction of selecting only one of each individual items, but instead a newly defined maximum number of occurrences per item) or *unbounded* (generally removing the restriction for a maximum number of occurrences of an individual item). For the purpose of creating lineups (in hindsight to all constraints like a player’s market price, his expected fantasy points, his position and team), a multi-dimensional knapsack problem perfectly meets our requirements. Each item’s weight i is defined by a D-dimensional vector $\vec{w}_i = (w_{i1}, \dots, w_{iD})$ (consisting of a player’s attributes), which will be allocated to a D-dimensional capacity vector $(W_1, \dots, W_D)$ representing our squad slots (as has been shown in [93]). General problem formulation is taken from [91] (see Equation 3.1).

$$\begin{aligned}
& \text{maximize} && \sum_{i=1}^n v_i x_i, \\
& \text{subject to} && \sum_{i=1}^n w_{ij} x_i \leq c_j, && j = 1, \dots, m, \\
& && x_i \in \{0, 1\}, && i = 1, \dots, n.
\end{aligned} \tag{3.1}$$

Variants of knapsack algorithm for creation of fantasy lineups have been studied intensively in academic literature ([41, 43, 94, 95]) and are widely used by punters, as they are easily adaptable and available in a lot of commonly used tools (e.g. Microsoft Excel¹⁷).

¹⁷<https://products.office.com/de-at/excel>

Chapter 4

Methodology

As has been shown in Chapter 2, most of the academic studies in the scope of fantasy games revolve mainly around the task of fetching data and building predictive models for players. Unfortunately, there still is an absence of evaluations of backtested simulations on ground truth data, which actually has been a typical study object within the field of sports and betting analyses for decades (e.g. [33, 34, 35, 36, 37]).

Especially due to the peer-to-peer nature of fantasy games, which produces a lot of big losers as well as a few big winners (see Figure 3.1), a deeper look at outcomes in real-life scenarios could contribute to understanding effects of various strategies.

On top of that, to the best of my knowledge, there exists no similar approach up to this date to imitate typical lineup building decisions by punters.

Furthermore, as fantasy sports emerged from the USA, a lot of literature can be found about American Football, Baseball, Ice Hockey or even Cricket, while analyses about fantasy football (*'soccer'*) tends to be neglected at best.

Last but not least, all well-known tournament types have been included because they are expected to have a great impact on results of strategies, which rarely is taken into account by other studies.

For all of these reasons, a new approach had to be found to investigate real-life scenarios, various tournament types and picking strategies for fantasy football; though many of the insights presented in Chapter 2 can be partly integrated and reused.

4.1 Study Design

For the most part, study design follows typical data analyses which have been intensively studied and described throughout this thesis.

Three common key stages have been identified to conduct an comprehensive analysis of betting models and strategies for Daily Fantasy Sports, marking the path of my work:

1. Data

- **Research:** an extensive overview of existing literature has already been given, insights were taken into account
- **Scraping:** various ways of collecting data for machine learning tasks will be investigated and presented
- **Quality:** as multiple sources are used, validity of data can be studied
- **Preprocessing:** due to the different formats and structures, reliable transforming/cleaning of data will be a key factor
- **Wrangling:** data from various sources has to be matched (primarily by string comparison)

2. Modeling

- **implementing Fanteam's scoring rules:** allowing for quick adaptations if a new scoring system is to be studied
- **calculating realistic fantasy market prices:** as there are no historical values available, reasonable prices will have to be calculated on my own
- **using (Distributed) Random Forest to create a first model:** this allows for insights in regards to variable importance and functions as a baseline model
- **using H2o auto-machine learning to create final model:** by combining multiple different machine learning algorithms, a Stacked Ensemble model will be created
- **predicting fantasy points on ground truth data:** by using the previously built Stacked Ensemble model

3. Evaluation

- **defining strategies commonly used by punters:** multiple strategies will be defined, mainly focusing on '*stacking blocks*' of players (e.g. maximum number of defenders allowed for a given team)
- **building lineups by using knapsack algorithm and expected fantasy points:** results will be used for a real-life scenario backtest afterwards
- **running simulations for various tournament types:** by calculating a strategy's outcome (fantasy points) for any matchday on ground truth data
- **running statistical tests for significant differences in outcomes (mean fantasy points):** allowing for conclusions in hindsight to impact of strategy selection

4.2 Limitations

Due to the massive effort needed to collect and melt data, only matches from Germany's 1st division ('1. Deutsche Bundesliga'), ranging from season 2012/2013 to 2016/2017 were taken into account. This dataset could be extended for additional leagues as well as international competitions, as this could help fill sparse data points (due to a player transferring from another country at the beginning of - or even within - a new season).

This also holds true for Fanteam's scoring rules (one could implement Draft-King's or any other operator's constraints), and particularly in regards to historical fantasy market prices. As scraping these values was not possible (because they simply did not exist on Fanteam at the point of writing), a simple and reproducible solution to this problem had to be found. However, there are a lot of competing variants of calculating market prices, each one having its own advantages and drawbacks.

While Random Forest algorithm is widely used throughout academic literature in regards to sports betting and fantasy sports and helps identifying most important input attributes, other machine learning methods could lead to superior results in the long run. Regardless, by using H2o's auto machine learning, finding a competitive model (in hindsight to a typical punter's ability to predict performances of fantasy players) is expected to be an achievable task.

Picking strategies of punters are imitated by focusing on approaches around '*stacking blocks*' of multiple players of a single team. However, other variants could be tested in future to find out about superior lineup building decisions, as my implementation allows only for a narrow range of possible lineups.

Furthermore, power of final statistical tests for significant differences in outcomes (in particular mean fantasy points) could be improved by larger sample sizes or by choosing other tests.

4.3 Datasets

An extensive research has been done to identify and scrape available datasets of matches, teams, players, historic odds and weather data (detailed steps for collecting and preprocessing can be found in Section 5.2).

Following data sources have been used:

- **Whoscored.com**¹ (matchevents, playerstatistics, ratings)
- **Squawka.com**² (matchevents, playerstatistics, ratings)
- **Fourfourtwo.com**³ (matchevents, playerstatistics)
- **Football-data.uk**⁴ (matchevents, historic odds)

¹<https://www.whoscored.com/>

²<https://www.squawka.com/>

³<https://www.fourfourtwo.com/>

⁴<https://www.football-data.co.uk/data.php>

- **SOFIFA**⁵ (player attributes in FIFA/EAsports)
- **PESDB**⁶ (player attributes in PES/Konami)
- **R-library** (stationaRy⁷, API for meteorological stations)
- **Fanteam.com**⁸ (fantasy tournaments and scoring system)

Redundant information (e.g. match results, fixtures, ...) have been compared and cross-checked – except for minor details, correlation between all available datasets is extremely high (≥ 0.99). The only neglectable differences occur when it comes to total shots (some sites do not count shots hitting the post, some are).

Eventually, all statistics are wrangled to fit into one dataset which then can be used for machine learning. Each row represents an individual player’s statistics for a specific date and consists of several categories of data, which will be presented up next.

Game Specific Details

Data has been collected for Germany’s 1st division (‘1. Deutsche Bundesliga’), ranging from season 2012/2013 to 2016/2017. This includes all general information as well as individual match events, e.g. *overall* (country, league, season, team-info, players, lineup), *fixtures* (location, date/time and referee), *events* (goal shot/conceded, assist, headers won/lost, minutes played ...) and *ratings* (by community via polls, by operator via statistical models, by scoring-system for real events ...). Events have been parsed for every match on team and player level. A list of attributes classified as Game Specific Detail can be found in Appendix 1. Furthermore, all numeric game specific details of a specific match-day are used to calculate aggregated statistics (team/player statistics for last 6 games) and get filtered out for prediction tasks. Statistics include maximum, minimum, mean, median, variance, standard deviation and interquartile range for both teams (hero, villain).

Player Specific Data

Additional information (extended player details) is provided by well-known game developers – publishers like EA Sports work together with thousands of people to rate players and determine their attribute values for a more realistic gameplay. Sites like SOFIFA and PESDB publish this information with regular updates on a monthly basis. Each player can be mapped with his actual rating of his abilities at a certain point of time. A list of attributes fetched for both games can be found in Appendix 2 and Appendix 3.

⁵<https://sofifa.com/>

⁶<https://pesdb.net/pes2018/>

⁷<https://github.com/rich-iannone/stationaRy>

⁸<https://www.fanteam.com/>

Weather

To further investigate a possible impact of weather on match results, data for 2011-2017 has been fetched. By calculating distances between meteorological stations and all stadiums (using R-library *geosphere*⁹), appropriate data points have been matched accordingly and can be inspected in Appendix 4.

Historical Betting Odds

An extensive dataset for odds from a great variety of bookmakers has been used, including traditional betting markets (tendency, over/under, min/max/mean odds) as well as general information about matches, which made melting data-points with all other collected data relatively simple. Attributes are listed in Appendix 5.

Additional Features

Besides collecting data, several features have been calculated in hindsight to the process of machine learning. For all statistics, values have been collected, calculated (min, max, mean, median, sd, var and iqr) and melted for both teams (hero vs. villain). For a list of calculated attributes, see Appendix 6.

4.4 Machine Learning

Due to the relatively small dataset collected, requirements for machine learning are low and could be fulfilled by a typical laptop (see Table 4.1). Nevertheless, for final task of auto machine learning, a Digital Ocean droplet has been instantiated and used for computations.

Table 4.1: Technical Setup

Laptop	ASUS Zenbook (8 GB RAM, 2.5 GHz)
Tools	RStudio Desktop, RStudio Server, H2o
Languages	R, Python, Ruby
Server	Digital Ocean
... for scraping (PYT)	4 GB RAM, 2 vCPUs
... for machine learning (ML)	64 GB RAM, 32 vCPUs
Database	
... for staging	*.csv, *.json, *.yaml, *.rds
... for archiving	PostgreSQL / pgAdmin
... for analyses	loading scripts in R
REST-API	JHipster Web-GUI

For data exploration and first insights, (Distributed) Random Forest algorithm will be used, which can be applied to a multitude of various tasks due

⁹<https://cran.r-project.org/web/packages/geosphere>

to its robustness, performance and simplicity (see Section 3.9). After studying its results, H2o’s auto-machine learning feature will be tested to further improve modeling results, if possible. Eventually, the very best performing model will be used to predict quantitative outcomes (fantasy points) for every player. Subsequently, by defining a linear program for all given constraints (e.g. maximum number of players of the same team, budget restrictions) and feeding it with expected fantasy points, salary and team of all available players of a given matchday, valid lineups can be created easily and adjusted for various strategies. Ultimately, those resulting fantasy teams are run against each other and evaluated corresponding to different game formats (by using ground truth data of season 2016/2017 compared to training data consisting of matches between 2011-2016).

Following next, in Practical Part I data preparation steps are described (scraping, preprocessing), followed by machine learning using the library H2o¹⁰. Practical Part II: Simulation & Betting-Strategies revolves around defining reasonable strategies and testing them against each other.

¹⁰<https://www.h2o.ai/>

Chapter 5

Practical Part I: Preparation & Machine Learning

The following two chapters explain the process of my work step-by-step. Results can be found in Section 7.

5.1 Scraper

To acquire the needed data, I relied on free APIs or downloaded complete datasets whenever possible. Nevertheless, a huge amount of data from various (semi-)commercial providers still had to be scraped – I expected this to become a cumbersome procedure and, unfortunately, had been right. Following the guideline of *'being nice'* (not over-stressing the server) and not trying to hard to work around obviously intentional obstacles (website protection, e.g. IP bans, Captchas), a ban never occurred. Of course I also always looked out for alternative sources well in advance.

Some sites allowed for simple curl-calls (see Listing 5.1, *scrape_squawka.R*), returning requested data as JSON – simply tweaking the query parameters, which could be figured out previously by using Chrome's Web Inspector, worked very well.

Listing 5.1: Example: R / Request (JSON)

```
library(dplyr)
df_squawka <- list()
for (j in 1:length(vi_filter)) {
  url <- paste0("http://www.squawka.com/wp-content/
  themes/squawka_web/leaderboard_process-v2.php?
  type=Player%20Stats",
```

```

"&filter=", vi_filter[j],
"&league=", vi_league,
"&team=", vi_team,
"&played=All%20matches&position=All%20Player%20Positions",
"&agestart=16&ageend=99&noofmatch=0&seasonstart=",
format(vi_date, "%d/%m/%Y"),
"&seasonend=", format(vi_date, "%d/%m/%Y"),
"&by=season&timestart=0&timeend=90&showtype=total")

df_squawka[[j]] <- retry(call_JSON_to_DF(vi_url),
maxErrors=10, sleep=10)
if (!is.na(df_squawka[[j]])) {
  df_squawka[[j]] <- df_squawka[[j]] %>%
    filter(!is.na(info.full_name))
}
} saveRDS(df_squawka, file=vi_filename)

```

Other providers publish statistics and information in plain HTML-tables, which can easily be scraped by using R and its library *rvest*¹, allowing for fast prototyping and further processing steps (see Listing 5.2, *scrape_sofifa.R*).

Listing 5.2: Example: R / Request (rvest)

```

library(rvest)
library(dplyr)
df_res <- url_sofifa %>%
  read_html() %>%
  html_nodes(css='#pjax-container > table') %>%
  html_table(header=TRUE, fill=TRUE) %>%
  as.data.frame()

```

Other data sources even provided appropriate R-packages for requesting ready-to-use datasets (see Listing 5.3, *scrape_weather.R*).

Listing 5.3: Example: R / Request (R-Package)

```

library(dplyr)
library(stationary)
met_data <- get_isd_station_data(
  station_id = "007005-99999",
  startyear = 2012,
  endyear = 2017)

```

A much more tedious task has been parsing and data cleansing of information stored in strings, which oftentimes had to be further investigated and accordingly transformed (see Listing 5.4, *scrape_pes.R*).

¹<https://cran.r-project.org/web/packages/rvest/rvest.pdf>

Listing 5.4: Example: R / Parsing

```
library(dplyr)
# eg "Bayer 04 Leverkusen" "2011 ~ 2018"
vec <- (strsplit(df_res$Team,
split = "(?<=[a-zA-Z]|[0-9])(?=[0-9][0-9][0-9][0-9] ~)",
perl = T))
if( (length(strsplit(
df_res$Team[sapply(
lapply(vec, function(x) length(x)), "[", 1) == 1],
split = "[0-9]*,", perl = T ))>0 )
{
df_res$Team[sapply( lapply(vec, function(x)
length(x)), "[", 1) == 1] <-
paste0(
sub("\\s+$", "",
strsplit(
df_res$Team[sapply(
lapply(vec, function(x) length(x)),
"[", 1) == 1],
split = "[0-9]*,", perl = T
)[[1]][1]), "0000 ~ 0000")
}
df_res$Team <- sapply( (strsplit(df_res$Team,
split = "(?<=[a-zA-Z]|[0-9])(?=[0-9][0-9][0-9][0-9] ~)",
perl = T)) , "[", 1)
df_res$position <- sapply( (strsplit(df_res$Name,
split = "\r\n", fixed=TRUE)) , "[", 2)
df_res$Name <- sapply( (strsplit(df_res$Name,
split = "\r\n", fixed=TRUE)) , "[", 1)
df_res$LoanDate[df_res$LoanDate == "N/A"] <- NA
```

Some of these scraping-scripts have been rewritten in Python, extended with data validation, staging scripts for the PostgreSQL database and deployed to a Digital Ocean droplet, where they are started every twelve hours (using a job scheduler). If new matches have been added, they get scraped and stored.

Due to the massive use of JavaScript on some of the sites, I included Selenium WebDriver in R/Python for automation. Though this resulted in longer run-times, it also made scraping of some of the data possible in the first place – and as scraping all the information is not a time-sensitive process, stability outweighs speed. By adding varying breaks during the scraping-scripts, small up front checks to watch out for new data on the page and a cautious approach in general, problems with publishers were avoided. Following example written in python allows for parsing JavaScript codeblocks from Whoscored.com (see Listing 5.5, *scrape_whoscored.py*).

Listing 5.5: Example: Python / Scraper (Webdriver)

```
def scrape_matchlinks(link_match, WS_matchid):
    wd = webdriver.Chrome("chromedriver.exe")
    wd.get(link_match)
    time.sleep(randint(10,20))

    soup = BeautifulSoup(wd.page_source)
    dataobject = soup.select("script")
    with open('out.txt', 'w', encoding='utf-8') as file:
        file.write(str(soup))
    with open('out.txt', 'r', encoding='utf-8') as myfile:
        data=myfile.read().replace('\n', '')

    # cut string to parse JavaScript
    k = data[(data.find("var matchStats = ") + 17):]
    data = k[(k.find(";") + 1)]

    # parse variable from JavaScript to Python object
    match_json = demjson.decode(data)
    wd.quit()
```

Data has been stored in a PostgreSQL-database or in staging folders in different formats. In the end, I wrote some additional R-scripts to automate loading of all files directly from staging folders into R-objects because of simplicity and speed of that solution.

5.2 Preprocessing

Following steps have been taken to turn all information into one dataset suitable for future work:

- cleansing (NaN, NA, 0, empty strings)
- mapping values (teams, players, odds, weather, FIFA/PES values)
- calculating aggregated statistics (team, player, ratings) for last 6 games
- removing redundant information
- imputing (replacing missing variables)
- transforming variables (factors, numeric, character)
- creating own statistics (e.g. squad stability)

Besides the tedious task of trying different transformations and preprocessing steps for various machine learning algorithms, one of the most time consuming processes turned out to be the wrangling of all datasets. As there exists no

unique identifier which would be valid for such a variety of sources, I had to map all data sources by a player's name – which led to even more problems (inconsistent encodings, exotic names and/or symbols, full first names versus abbreviations or nicknames). Moreover, adding statistics about the weather to a certain game required a) fetching stadium location of each team, b) calculating distances to all available weather stations² and c) mapping them, if possible (a threshold of 10 kilometres has been chosen).

5.3 Fantasy Points

For each performance of a player in our dataset, fantasy points have been calculated by using the scoring rules from Fanteam³. Some adaptations were made to simplify this complex mechanism:

- achieving a penalty, if a teammate scores (3 points)
- Team 'wins' in the periode player is on the pitch (1 point)
- Team 'draw' in the periode player is on the pitch (0 points)
- Team 'loses' in the periode player is on the pitch (-1 point)
- Captain (double points)

5.3.1 Limitations

To identify whether a penalty was successfully taken after another player has been fouled, more tedious wrangling with match-event data would have been needed, furthermore adding another source. For simplification purposes, I decided to refrain from encoding these sparse events.

A similar demanding challenge would have been to correctly assign points for winning, drawing or losing a certain period within a game for every individual. As drawing does not change the total sum of fantasy points at all – and winning/losing is just valued with +1 or -1 points – I decided to just take the final result of the game for every player to allocate respective points (+1 for players of winning team, -1 for losing team).

Moreover, one player per team has to be selected as a captain, which means his total sum of fantasy points will be doubled for a given matchday. As this is dependant on the selection of punters, there was no need to include this rule at this point. Though not all rules have been implemented perfectly down to the last detail, resulting fantasy scores can be considered to be a very precise approximation.

²<https://cran.r-project.org/web/packages/geosphere/geosphere.pdf>

³<https://www.fanteam.com/support-center/football-rules>

5.3.2 Distribution of Fantasy Points

A summary of my own implementation of Fanteam’s scoring system can be seen in Table 5.1 – resulting in a mean of 3.942 fantasy points and individual performances ranging from -4 to 23 points. In Figure 5.1, a histogram is presented, showing the distribution of fantasy points assigned to real game data (1. Deutsche Bundesliga, Season 2011-2017).

Interestingly, two spikes can be identified: one roughly about 3 points, the other one around 7 points. This can be easily explained by Fanteam’s scoring system, as goalkeepers and defenders are awarded with 4 bonus points for a cleansheet (which means they did not concede any goal during a match) – and most lineups consist of 4 to 6 defensive players. Additionally, if an offensive player succeeds in netting a goal or accomplishes an assist, he will also earn a couple of points; shifting performances based on certain events either a few points up (e.g. cleansheet, goal/assist) or down (e.g. causing a penalty, red card). For most players this means either getting rewarded or penalized for their own actions as well as their team’s performance, splitting the dataset roughly into two parts. Result is a reasonable distribution of fantasy points which can be used for further experiments.

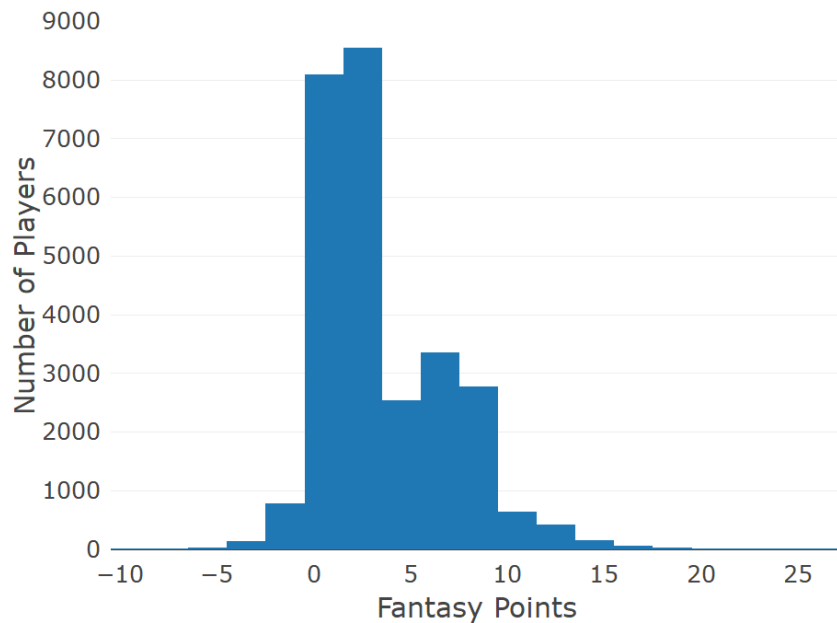


Figure 5.1: Fantasy Points: Histogram

Table 5.1: Fantasy Points: Summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-4	2	3	3.941	7	23

5.4 Fantasy Market Values

There is no information on Fanteam’s page regarding the calculation of transfer market values. Unfortunately, they also do not provide users with historic data (at least not for several months) – therefore I had to implement a self-designed system for assigning reasonable numbers. As mentioned in 3.6, there are manifold possible pitfalls when constructing market values (as accumulating market prices relative to total fantasy scores can lead to restricted possible lineup combinations at the end of a season). Due to the lack of appropriate data regarding the ownership of specific fantasy players (which would be a preferable model), I decided to approximate typical market values based on empirical data (min/max values on Fanteam) and the average performance of a player (last 6 games):

- calculate fantasy points average for last 6 games
- search for maximum value, assign \$15M as market price
- search for minimum value, assign \$2.5M as market price
- NA values: assign average market price of the same player (if possible)
- NA values: assign average market price of the same team
- (optional) scale data to have an average of \$100M/11=\$9.09M price per player
- round values to next 0.1

These steps have been implemented in R (see Listing 5.6, *calculate_marketprices.R*).

Listing 5.6: Example: R / Market Prices

```
library(dplyr)
library(rescale)
# search max and min for 6-day-moving-average
max(df_players$FP_last6_mean, na.rm=T) # 13.33
min(df_players$FP_last6_mean, na.rm=T) # -1.33

# if there are NA values, assign player his own average
df_players <- df_players %>%
  group_by(whoscored_name) %>%
  mutate(FP_last6_mean = ifelse(
```

```

        is.na(FP_last6_mean),
        mean(FP_last6_mean, na.rm = TRUE),
        FP_last6_mean))
# if there are still NA values, assign own team average
df_players <- df_players %>%
  group_by(whoscored_club) %>%
  mutate(FP_last6_mean = ifelse(
    is.na(FP_last6_mean),
    mean(FP_last6_mean, na.rm = TRUE),
    FP_last6_mean) )

# scale data accordingly to min/max values
df_players <- rescale(df_players$FP_last6_mean, c(2.5,15))
df_players <- round(df_players$FP_last6_mean, 1)

```

5.4.1 Limitations

Over the course of the construction of market prices it became evident that market prices were too low on average (see Table 5.2), leading to very similar fantasy lineups afterwards. Therefore, a mean market price of \$9.09M was chosen (as team budget consists of \$100M virtual millions for 11 players), while keeping the standard deviation of mean fantasy points per player (see Listing 5.7, *calculate_marketprices.R*).

Listing 5.7: Example: R / Market Prices (rescale)

```

library(psych)
mean(df_players$FP_last6_mean, na.rm=T) # res=3.496495

v_tmp <- rescale(df_players$FP_last6_mean,
mean=9.09, sd=sd(df_players$FP_last6_mean))

df_players$FP_last6_newMarketvalue <-
v_tmp$t.t.scale.x....sd...mean.

```

5.4.2 Distribution of Fantasy Market Values

Maximum price of a player increased to \$18.837M, which should still be in an acceptable range as it just affects a couple of individuals over the stretch of seven seasons (0.35% of all market prices were higher than my initial target of \$15M). Minimum price of a player increased to \$4.17M, which is also acceptable (filtered for players with a sample of more than 6 games altogether – injured or uncertain players are typically valued around \$2.5M, so it makes sense to value the worst of active players a bit higher than that). Advantages of this implementation for market values are predominantly its simplicity and its reasonable reflection

Table 5.2: Market Prices (1st version): Summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.5	5.6	6.5	6.618	7.5	15

Table 5.3: Market Prices (2nd version): Summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.170	7.837	8.837	9.000	10.004	18.837

of average fantasy points of players. Results can be found in Table 5.2 and Table 5.3 as well as Histogram depicted in Figure 5.2.

Preparations have been finished by collecting, cleaning and wrangling all required data sources; calculated fantasy points as well as market prices have been added and as a result, machine learning can be applied.

5.5 Machine Learning: Random Forest

In contrast to the exhausting process of fetching and wrangling the data, experimenting with different machine learning algorithms was relatively easy to conduct once the final dataset had been completed. Using H2o⁴ allowed for convenient tests, control and monitoring within a GUI in a browser window as well as an easy deployment on a faster server, whenever needed.

For data exploration and as a first baseline model (Figure 5.3 depicts a QQplot of residuals), Distributed Random Forest (DRF) as well as Extremely Randomized Trees (XRT) algorithms have been used as they have widespread possible applications and allow for inspecting feature importance, which can be seen in Figure 5.4 and Figure 5.5 – model details are presented in Table 5.4. Scoring History for both DRF and XRT was plotted for RMSE and can be found in Figure 5.6 and Figure 5.7.

Results will be discussed subsequently before moving on to auto machine learning with H2o.

⁴<https://www.h2o.ai/>

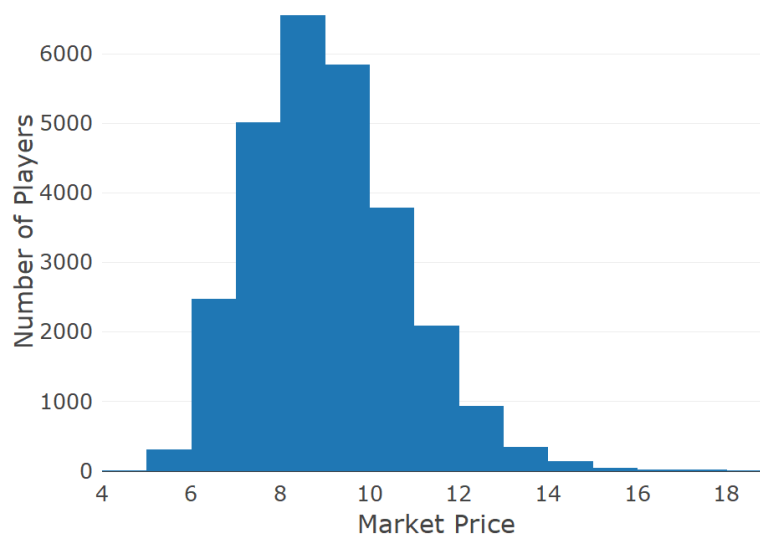


Figure 5.2: Market Prices: Histogram

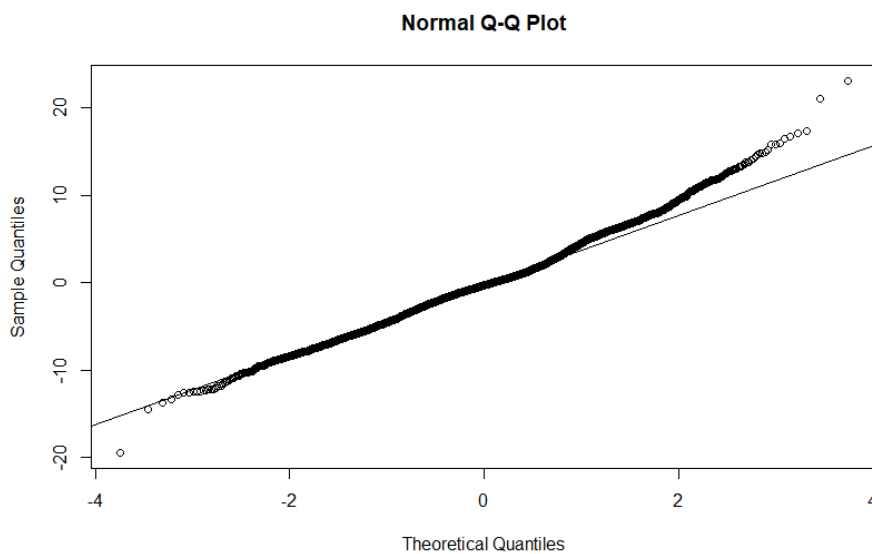


Figure 5.3: Random Forest: QQPlot of residuals (DRF)

Table 5.4: Results for DRF and XRT

Model	DRF	XRT
Number of Trees	22	5
Number of Internal Trees	22	5
Model Size in Bytes	865370	200345
Min Depth	20	20
Max Depth	20	20
Mean Depth	20	20
Min Leaves	2773	2879
Max Leaves	3593	3423
Mean Leaves	3120.182	3181.8
H2o Regression Metrics (validation data)		
Model-ID	DRF	XRT
MSE	5.121652	6.057723
RMSE	2.263107	2.461244
MAE	1.568336	1.721896
Mean Residual Deviance	5.121652	6.057723
H2o Regression Metrics (training data)		
Model-ID	DRF	XRT
MSE	5.684946	8.508333
RMSE	2.384313	2.916905
MAE	1.660962	1.996058
Mean Residual Deviance	5.684946	8.508333
Cross-Validation Metrics Summary (training data)		
Model-ID	DRF	XRT
MSE	5.282857	6.262031
RMSE	2.298447	2.502405
MAE	1.608554	1.766348
Mean Residual Deviance	5.282857	6.262031

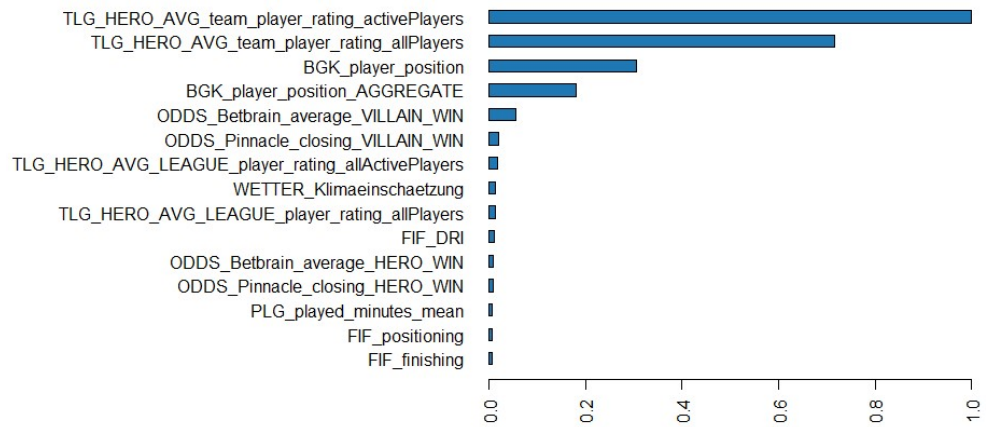


Figure 5.4: Random Forest: Variable Importance (DRF)

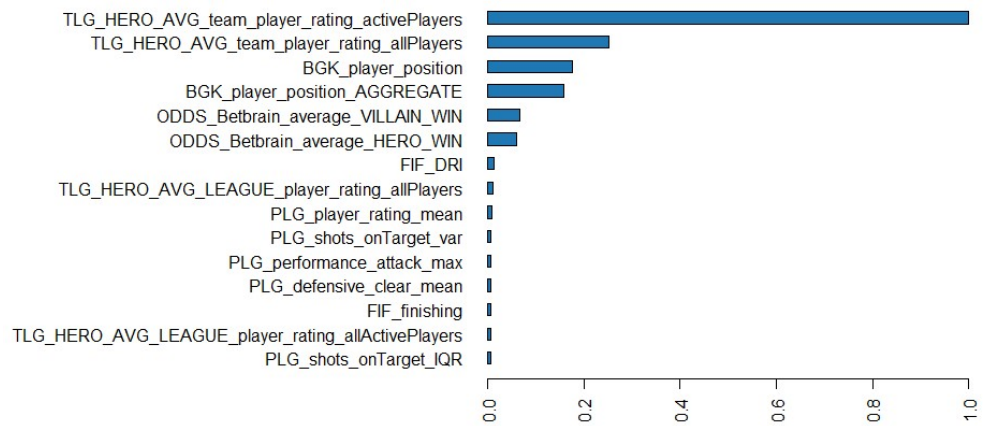


Figure 5.5: Random forest: Variable Importance (XRT)

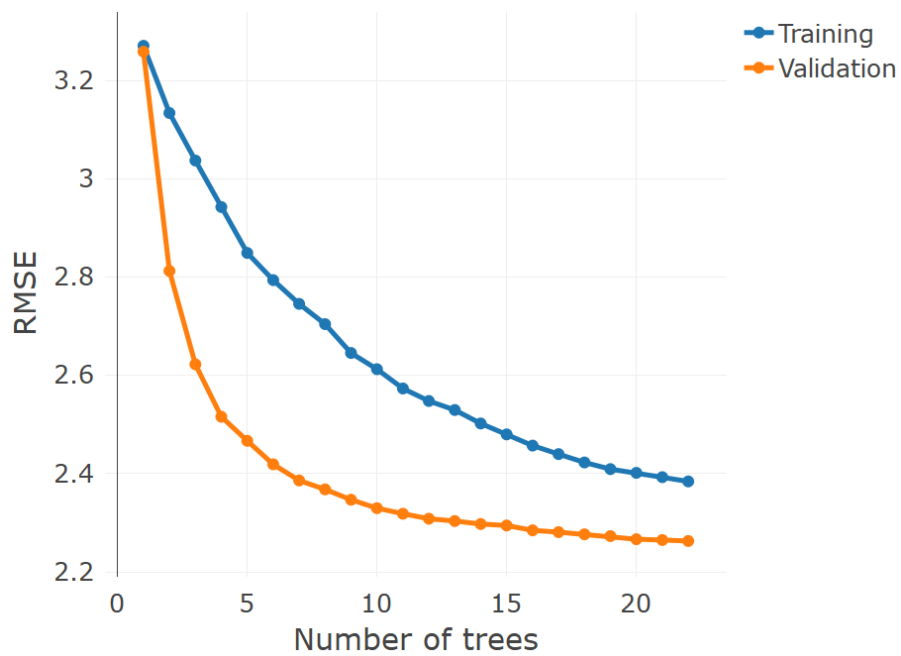


Figure 5.6: Random Forest: Scoring History (DRF) - RMSE

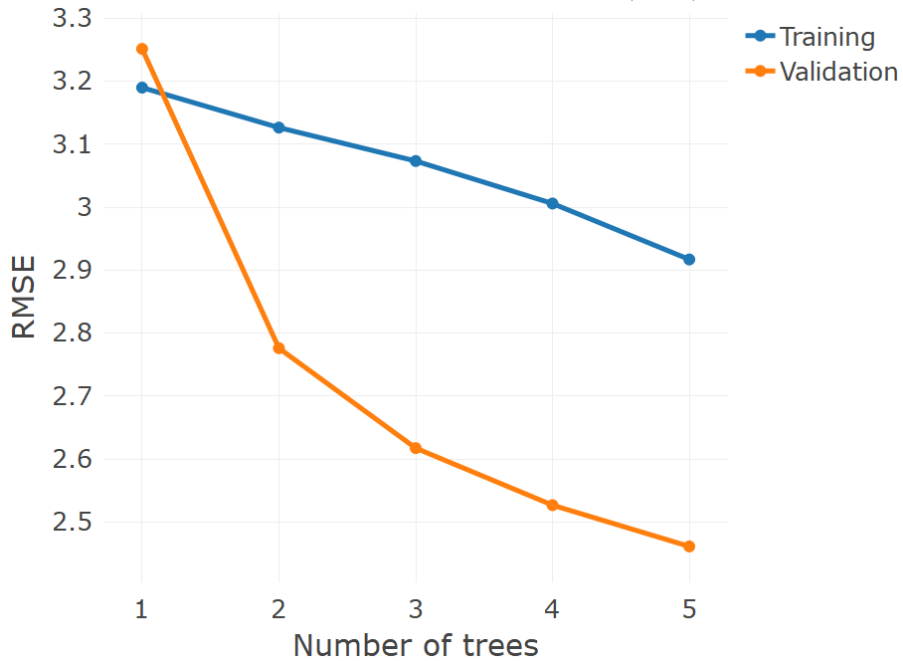


Figure 5.7: Random Forest: Scoring History (XRT) - RMSE

Both distributed random forest (DRF) as well as extremely randomized trees (XRT) have been implemented and tested. In general, DRF lead to slightly better results compared to XRT – mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE) and mean residual deviance were lower for validation data, training data and cross-validation on training data (see Table 5.4). Though results differ slightly in regards to variable importance, both share their top-five predictors in common:

- **TLG_HERO_AVG_team_player_rating_activePlayers** – average player-rating from WhoScored.com aggregated on team-level for all squad members who stood on field on a given matchday
- **TLG_HERO_AVG_team_player_rating_allPlayers** – average player-rating from WhoScored.com aggregated on team-level for all players of a team – including those on bench or injured
- **BGK_player_position** – detailed position of a player defined by FIFA or PES statistics – e.g. left centerback, right-offensive midfielder, second striker
- **BGK_player_position_AGGREGATE** – position of a player defined by Fanteam’s rules – e.g. defender, midfielder, forward
- **ODDS_Betbrain_average_VILLAIN_WIN** – average odds for winning, using all bookmakers

Intuitively, these findings seem to be plausible: first of all, average rating for all players of a team (whether they have played or not) should be expected to be a significant indicator for a team’s overall performance – in this case, averages have been calculated for the last six season games. On top of that, ratio between the average strength of the team on field versus all players of the very same team could help the algorithm to identify whether the ‘best’ possible lineup was currently available.

Secondly, player positions described by Fanteam are not as detailed as in the datasets of FIFA and PES, oftentimes even placing nominal forwards as midfielders (or vice versa – this also holds true for midfielders and defenders). Apparently, performances can be estimated more accurately by identifying players whose position is not stored properly.

A little bit surprisingly, variable importance of team win odds did not make it within the top-three – still, placing fifth using DRF as well as XRT.

All other variables did not show to have a great impact on predictions – average WhoScored.com rating of all players within a league, dribbling skill and finishing skill values from EASports’ FIFA database made it into the top-fifteen of both algorithms, completed by a wild mixture of weather information, average played minutes, other FIFA skill values (positioning), shots on target or successful defensive actions (clear).

Next up, a Stacked Ensemble model will be created by using H2o’s auto machine learning function, extended by testing various combinations of input attributes.

5.6 Machine Learning: AutoML (H2o)

Following pre-analysis with (Distributed) Random Forest, H2o's auto machine learning was used to create a Stacked Ensemble model. For this reason, 60% of collected data has been used for training, 20% for testing and another 20% for validation and backtesting (when using AutoML⁵, H2o tries to create multiple models, assesses them against each other and builds a final leaderframe). AutoML applies and uses several algorithms for training models – including Random Forest (RF), Extremely Randomized Trees (XRT), a combination of both (DRF), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM) and DeepLearning. Maximum runtime was set to 30 minutes – using a server from DigitalOcean with 64 GB RAM and 32 vCPUs.

Following different variants of training inputs have been tested by dropping columns:

- *ML_1*: without odds and whoscored-player-ratings
- *ML_2*: without whoscored-player-ratings
- *ML_3*: without odds
- *ML_4*: all available columns

As expected, leader model turned out to be a Stacked Ensemble model created by using all available columns (*ML_4*), which is further used for all following steps – though dropping odds (*ML_3*) delivered very similar results. Table 5.5 lists results for all four variants. *ML_1* as well as *ML_2* lead to slightly worse results compared to initial tests with RF (see previous Section 5.5).

Table 5.5: Variations of Input Parameters

AutoML	MSE	RMSE	R2	MAE	AIC	Mean res dev
ML_1	6,078	2,465	0,475	1,763	25498,413	6,078
ML_2	5,818	2,412	0,498	1,706	25270,694	5,818
ML_3	4,950	2,225	0,572	1,546	24379,639	4,950
ML_4	4,931	2,221	0,574	1,539	24369,532	4,932

Eventually, the leading model (see Table 5.6) was used to predict fantasy points on our ground truth data – which directly leads us to Part II of the practical part of my work.

⁵<https://www.docs.h2o.ai/latest-stable/h2o-docs/automl.html>

Table 5.6: AutoML Results

	model_id	Mean res	RMSE	MAE
1	StackedEnsemble_AllModels_0.084123	4,93	2,22	1,54
2	StackedEnsemble_AllModels_0.081108	4,95	2,22	1,55
3	GBM_grid_0.084123_model_4	5,10	2,26	1,60
4	GBM_grid_0.081108_model_4	5,11	2,26	1,61
5	GBM_grid_0.084123_model_6	5,11	2,26	1,60
6	XRT_0.084123	5,14	2,27	1,58
7	DRF_0.084123	5,15	2,27	1,58
8	GBM_grid_0.084123_model_2	5,16	2,27	1,61
9	DRF_0.081108	5,17	2,27	1,59
10	XRT_0.081108	5,18	2,27	1,59
11	GBM_grid_0.084123_model_3	5,19	2,28	1,59
12	GBM_grid_0.081108_model_3	5,21	2,28	1,59
13	GBM_grid_0.081108_model_2	5,24	2,29	1,63
14	GBM_grid_0.084123_model_1	5,25	2,29	1,63
15	GBM_grid_0.081108_model_1	5,27	2,30	1,63
16	GBM_grid_0.084123_model_0	5,32	2,30	1,66
17	GBM_grid_0.081108_model_0	5,39	2,32	1,67
18	StackedEnsemble_AllModels_0.074101	5,82	2,41	1,71
19	StackedEnsemble_AllModels_0.071035	6,08	2,47	1,76
20	GBM_grid_0.074101_model_3	6,16	2,48	1,79
21	StackedEnsemble_AllModels_0.063904	6,17	2,48	1,79
22	GLM_grid_0.084123_model_0	6,20	2,49	1,90
23	GLM_grid_0.081108_model_0	6,23	2,50	1,90
24	GBM_grid_0.074101_model_2	6,26	2,50	1,81
25	GBM_grid_0.074101_model_1	6,36	2,52	1,83
26	StackedEnsemble_BestOffFamily_0.063904	6,36	2,52	1,82
27	GBM_grid_0.074101_model_0	6,40	2,53	1,85
28	GBM_grid_0.063904_model_2	6,43	2,54	1,84
29	GBM_grid_0.074101_model_4	6,47	2,54	1,86
30	DeepLearning_0.081108	6,48	2,55	1,81
31	GBM_grid_0.071035_model_2	6,52	2,55	1,85
32	GBM_grid_0.063904_model_1	6,52	2,55	1,85
33	GBM_grid_0.071035_model_1	6,53	2,56	1,87
34	GBM_grid_0.071035_model_3	6,55	2,56	1,85
35	XRT_0.074101	6,56	2,56	1,88
	...			
50	GBM_grid_0.084123_model_7	11,34	3,37	2,73

Chapter 6

Practical Part II: Simulation & Betting-Strategies

6.1 Betting Strategies

Following betting strategies (see Table 6.1) were defined and deployed by an implementation of a linear program (to optimize combination of players depending on their expected fantasy points and their market price).

Table 6.1: Strategies for picking Fantasy Players

	Id	Name	Description
1	S_1	knapsack	depends on expected fantasy points
2	S_2	random	random selection of players
3	S_3	2def_best	defensive blocks consist of 2 teams (best available)
4	S_4	2def_worst	defensive blocks consist of 2 teams (worst available)
5	S_5	2def_random	defensive blocks consist of 2 teams (random)
6	S_6	2off_random	offensive blocks consist of 2 teams (random)
7	S_7	2off_best	offensive blocks consist of 2 teams (best available)
8	S_8	combo	defensive and offensive consist of 2 teams each
9	S_9	optimum	optimum of a given matchday

Furthermore, additional strategies need to consider lineups with missing players (due to misinformation or last-minute injuries) as well as accounting for a captain's bonus (double points for an individual player). Both of these versions can be calculated by using our base-strategies.

For missing players, a small dataset of Fanteam tournaments has been analyzed. These insights allow for adding additional strategies later on. Table 6.2

Table 6.2: Fantasy Lineups: Missing Players

Tournament	Teams	Missing players	AVG missing players
T1	787	186	0.236
T2	816	59	0.072
T3	755	155	0.205
T4	824	702	0.852
T5	727	323	0.444
T6	737	215	0.292
T7	727	108	0.149
Total	5373	1748	0.325

presents averages from 8 tournaments with 600 to 800 different participating lineups each.

Having more than 3+ missing players in a starting formation occurs very rarely and can be interpreted as outliers caused by turbulent game-weeks (e.g. a lot of international competitions), injuries, misinformation in media coverage or simply a too long time-window between the start of a fantasy tournament and the real match itself (making it harder to predict a manager's choice).

6.2 Lineup Creation

Stacking a team of players together while accounting for their expected points and market prices can be solved by a variant of the knapsack problem, which can be formulated as an integer linear programming program and has been well-studied in multiple contributions to optimizing fantasy lineups (e.g. [43, 94, 95]). By adapting an implementation for NFL teams on DraftKings¹ additional constraints have been included, matching Fanteam's ruleset for fantasy lineups:

- team budget (not more than \$100M)
- valid formations
- maximum of 3 players from each team

A code snippet in R (see Listing 6.1, *create_lineups.R*) shows the implementation of the function for lineup creation, including mandatory constraints.

Listing 6.1: Example: R / LP-Solver

```
library(lpSolveAPI)

create_lineup <- function(player_DF) {
  lpmodel<-make.lp(0,nrow(player_DF))
  set.objfn(lpmodel, player_DF$expected_points)
```

¹<https://github.com/geekman1/fantasylineup>

```

set.type(lpmodel, seq(1, nrow(player_DF), by=1),
type = c("binary"))

c_gk <- ifelse(player_DF$type == "GK", 1, 0)
c_def <- ifelse(player_DF$type == "DEF", 1, 0)
c_mid <- ifelse(player_DF$type == "MID", 1, 0)
c_fwd <- ifelse(player_DF$type == "FWD", 1, 0)

add.constraint(lpmodel, c_gk, "=", 1)
add.constraint(lpmodel, c_def, ">=", 3)
add.constraint(lpmodel, c_def, "<=", 5)
add.constraint(lpmodel, c_mid, ">=", 2)
add.constraint(lpmodel, c_mid, "<=", 5)
add.constraint(lpmodel, c_fwd, ">=", 1)
add.constraint(lpmodel, c_fwd, "<=", 3)

# total of eleven players
players_total <- c_gk+c_def+c_mid+c_fwd
add.constraint(lpmodel, players_total, "=", 11)

# maximum of three players from each team
team_names <- levels(factor(player_DF$team))
for(i in 1:length(team_names)) {
  no_three <- ifelse(player_DF$team==team_names[i],1,0)
  add.constraint(lpmodel, no_three, "<=", 3)
}

# total team budget below $100M
add.constraint(lpmodel, player_DF$market_price, "<=", 100)

lp.control(lpmodel, sense='max')
solve(lpmodel)
get.objective(lpmodel)
write.lp(lpmodel, 'model.lp', type='lp')

team <- data.frame(matrix(0, 0, ncol(player_DF)))
colnames(team) <- c(colnames(player_DF))
team_select <- subset(data.frame(player_DF,
get.variables(lpmodel)), get.variables(lpmodel)==1)
team_select$get.variables.lpmodel. <- NULL
team_select$market_price <- team_select$market_price
team_select$expected_points <- team_select$expected_points
team <- rbind(team, team_select)
team <- team[order(team$type),]
return(team)
}

```

By filtering the source data frame *player_DF* appropriately, fulfilling of constraints – e.g. only including goalkeepers and defenders from two different teams – can be guaranteed. Result is a fast and valid approach for building lineups following previously defined strategies.

6.3 Simulation

For simulations, multiple tournament types have been chosen:

- Heads-Up (1 : 1)
- Double-Up (50 : 50)
- One-vs-all (1 : *ALL*)
- Winner Takes It All (1 : *ALL*)

Previously created lineups were loaded and run against each other for every matchday and tournament type, accounting for rake as well as respective payout structure. Result is a table with payout/total win, paid rake, number of played games and ROI (see Listing 6.2 for Double-Ups, *simulate_tournaments.R*).

Listing 6.2: Example: R / Simulation for Tournaments

```
library(dplyr)
df_5050 <- NA
for (i in 1:length(unique(df_strategies$matchday))) {
  # filter for matchday
  df_tmp <- df_strategies_copy %>% filter(matchday == i)
  # order by accumulated fantasy points
  df_tmp <- df_tmp[order(df_tmp$total_real_points,
    decreasing = TRUE), ]
  df_tmp$platzierung <- c(1:10)
  # upper half of participants doubles their buyin
  df_tmp$payout <- c(2,2,2,2,2,0,0,0,0,0)
  df_tmp$win <- df_tmp$payout - df_tmp$buyin
  df_5050 <- rbind(df_tmp, df_5050)
}
df_5050 <- df_5050[1:nrow(df_5050)-1,]
df_5050_AGG <- df_5050 %>%
  group_by(description) %>%
  summarise(total_buyin=sum(buyin),
    total_payout=sum(payout),
    total_rake=sum(rake), total_win=sum(win)) %>%
  as.data.frame()

df_5050_AGG$ROI <- round(df_5050_AGG$total_win /
  df_5050_AGG$total_buyin*100,2)
```


Other tournament results were simulated accordingly, further extended by three additional sub-strategies:

- 0.325 players missing in starting formation
- 1.000 players missing in starting formation
- Captain Bonus (add mean fantasy points)

For comparison, I also included a strategy with 1 instead of 0.325 missing players per lineup to study impact on ROI. Captain bonus has been determined by simply calculating the mean fantasy points of a team, which then get added to the total sum. Thus, real variance of results in real fantasy games is masked, as the choice of a player for a captain's bonus naturally has a large impact due to the large variance of fantasy points.

Up next, final results are presented (including outcomes of all strategies for projected points as well as for backtested real-life scenarios).

Chapter 7

Results

Strategies have been evaluated by their total *projected points* as well as their backtested results on ground truth data (*real points*). Contrary to all other tournament-types, 1 : 1 format requires a higher number of simulated tournaments (as opponents face each and every other player on a given matchday compared to a single tournament, where all punters compete at the very same time). Due to that reason, results have been scaled down to accommodate for increased sample-size and make comparisons more comprehensible.

7.1 Plots

Figure 7.1 depicts *projected* performance of all strategies by plotting expected total fantasy team points per matchday. Figure 7.2 visualizes results of *backtested* predictions on ground truth data, also adding a grey, dashed line for *optimal* results – this would have been the maximum possible fantasy team score on a given matchday. Figure 7.3 and Figure 7.4 show respective boxplots for *projected* and *backtested* fantasy points grouped by strategies.

Though differences between strategies are clearly visible (especially in regards to *projected points*), cumulative plots help understanding implications using a certain approach for multiple succeeding competitions – points for captain bonus have been added as well (see Figure 7.5 and Figure 7.6).

To visualize the effect of missing players (due to injuries, last-minute roster changes or similar), linecharts for strategies with deductions have been plotted accounting for 0.325, 1 or 2 absent squad members in Figure 7.7 (*projected points*) and Figure 7.8 (*backtested results*).

Figure 7.9 and Figure 7.10 depict total buyins won/lost grouped by tournament-type as well as per individual strategy within those competitions.

Eventually, overall results for all strategies (total buyins won or lost for all tournament types) are shown in Figure 7.11, followed by statistical tests.

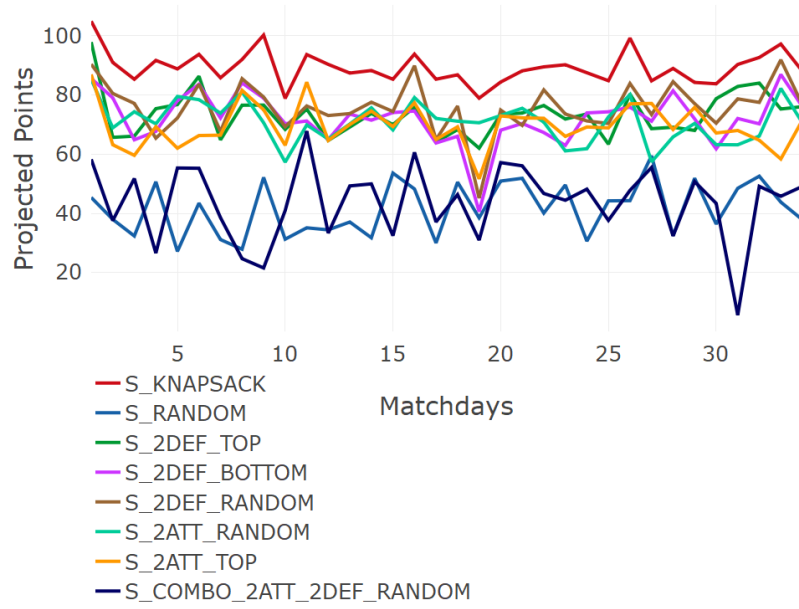


Figure 7.1: Projected Points for each Strategy and Matchday

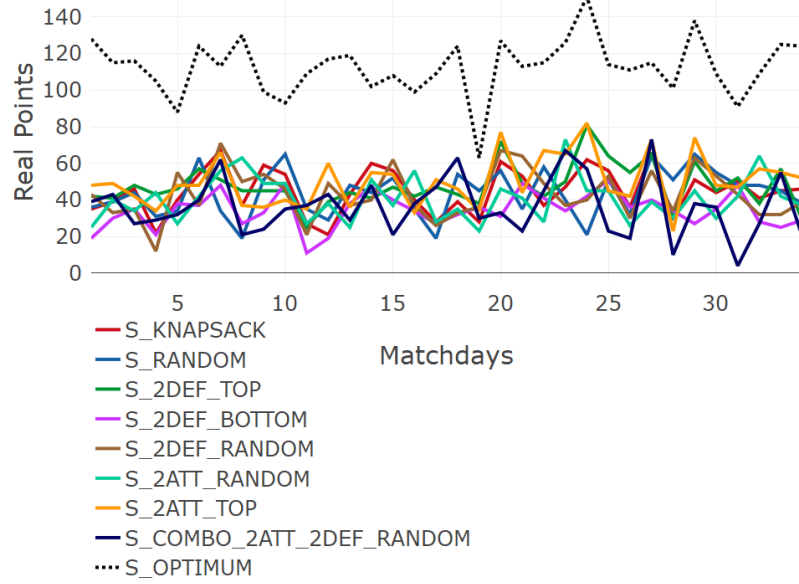


Figure 7.2: Simulation Results on ground truth data for each Strategy and Matchday

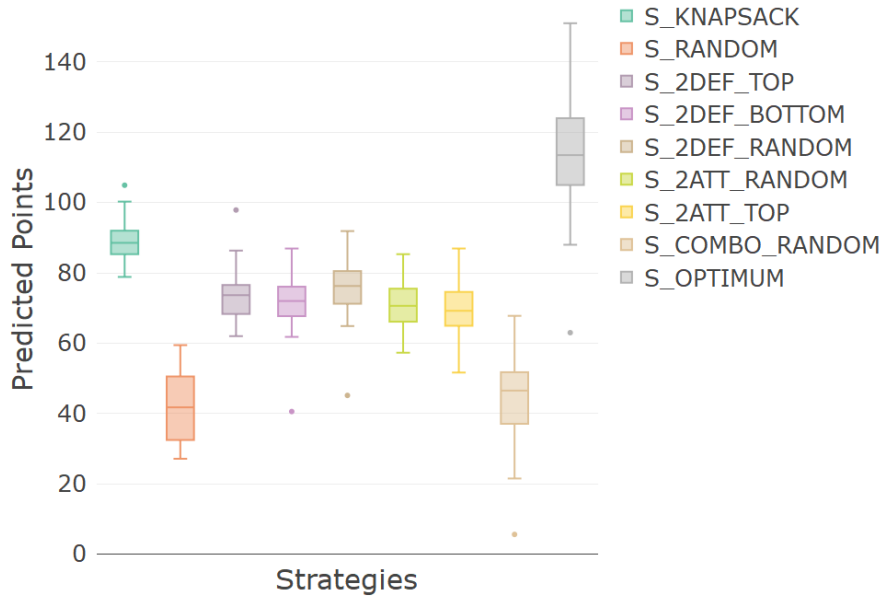


Figure 7.3: Projected Points for each Strategy

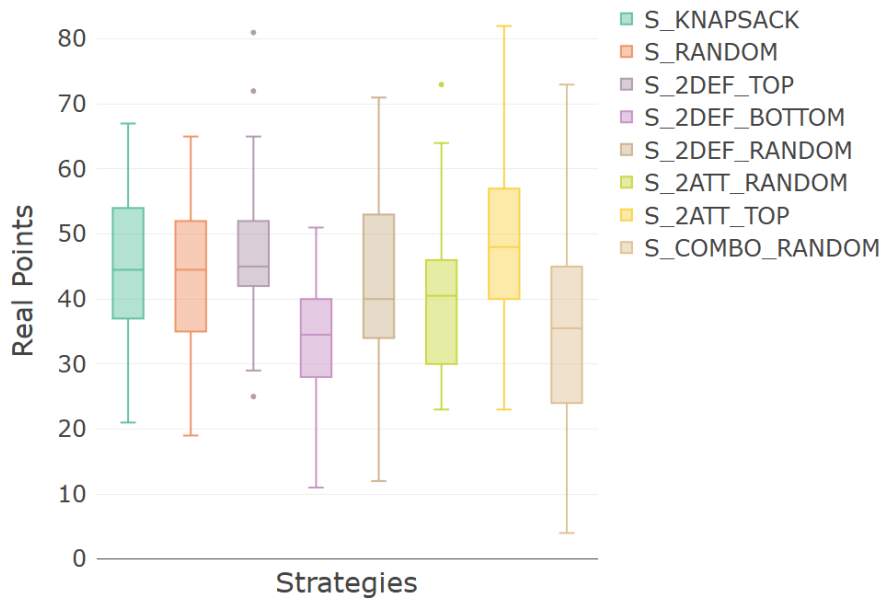


Figure 7.4: Simulation Results on ground truth data for each Strategy

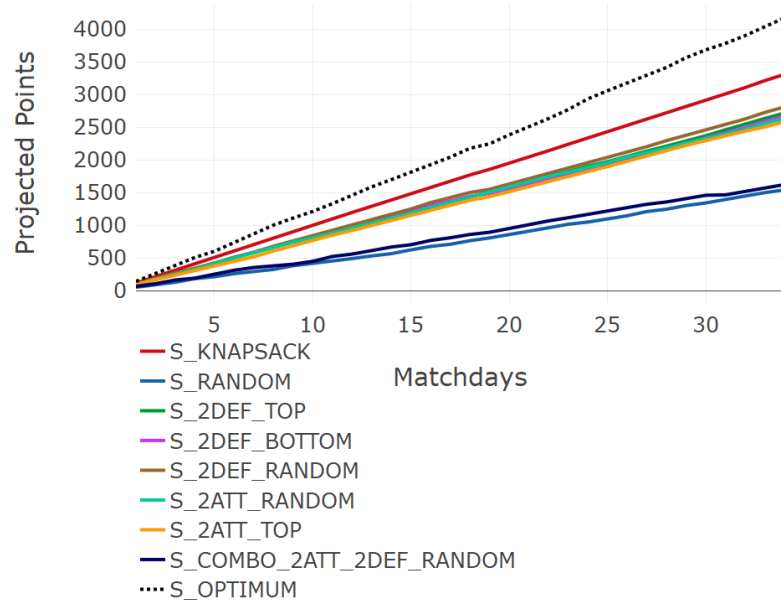


Figure 7.5: Cumulative Projected Points (Captain Bonus)

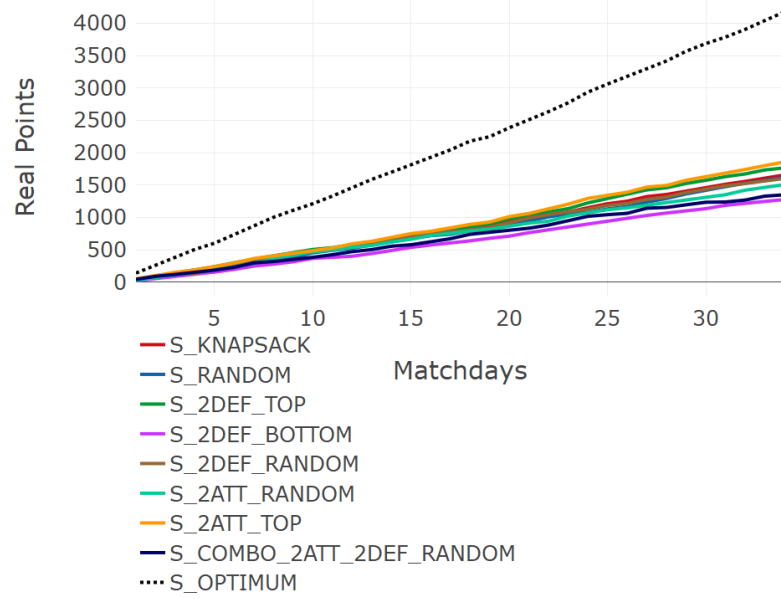


Figure 7.6: Cumulative Simulation Results on ground truth data (Captain Bonus)

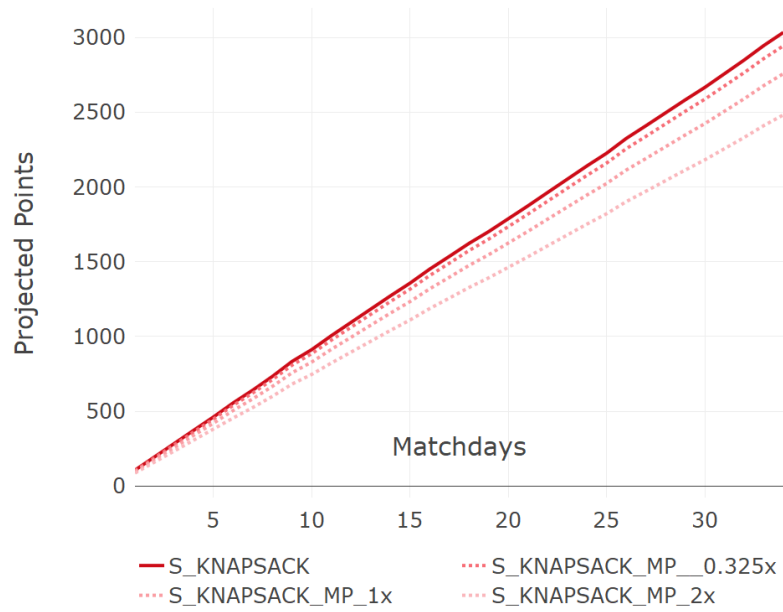


Figure 7.7: Projected Points (deducted Fantasy Scores for Missing Players)

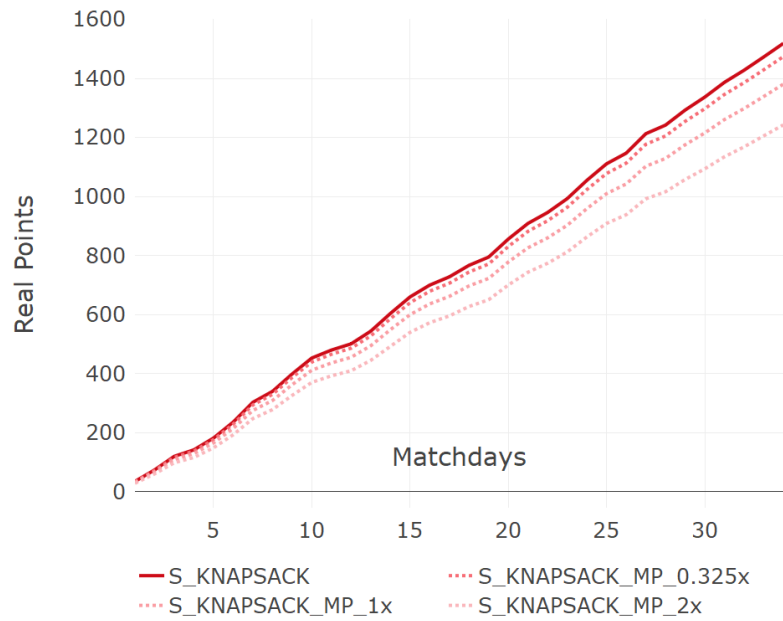


Figure 7.8: Simulation Results on ground truth data (deducted Fantasy Scores for Missing Players)

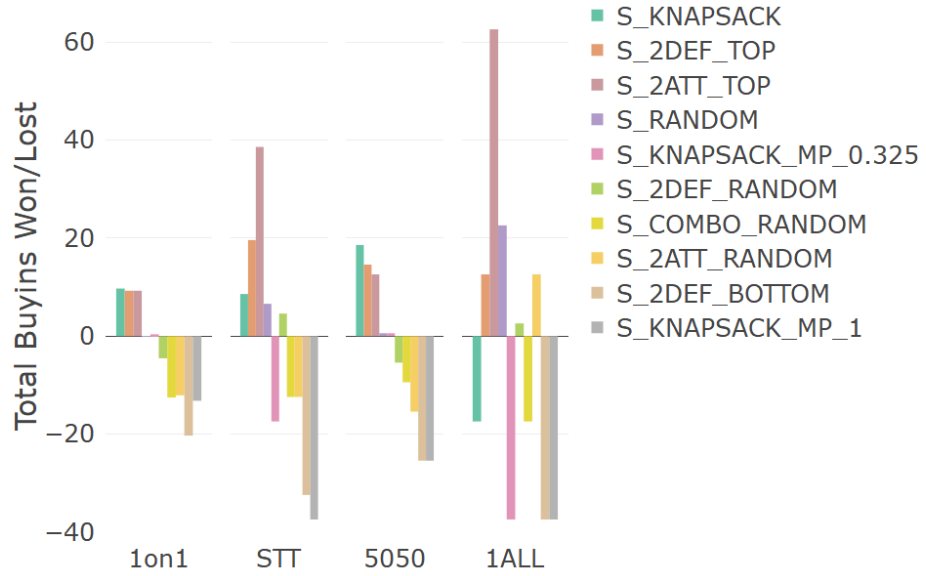


Figure 7.9: Simulation Results on ground truth data (Total Buyins Won/Lost per Tournament-Type)

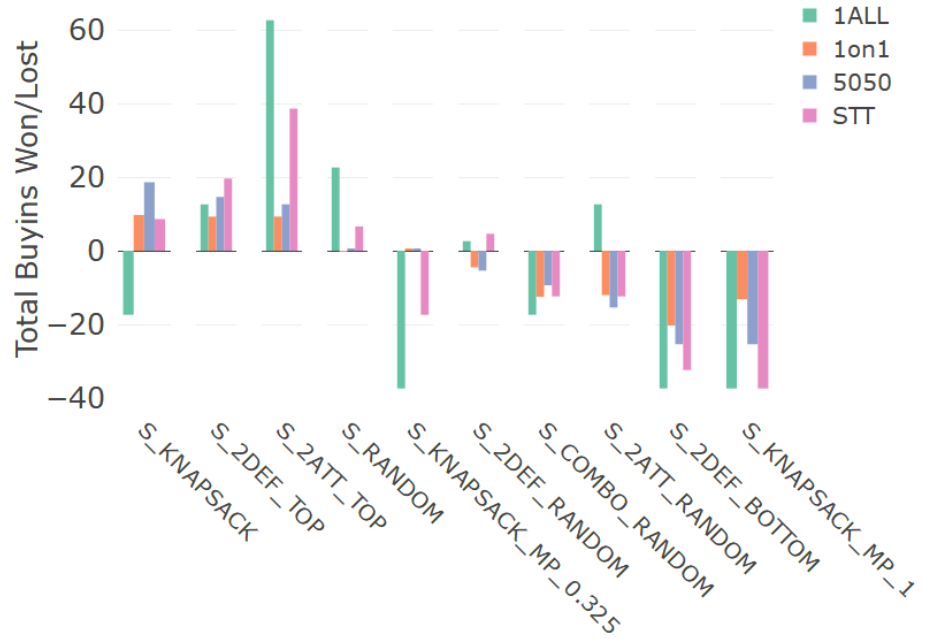


Figure 7.10: Simulation Results on ground truth data (Total Buyins Won/Lost per Strategy and Tournament-Type)

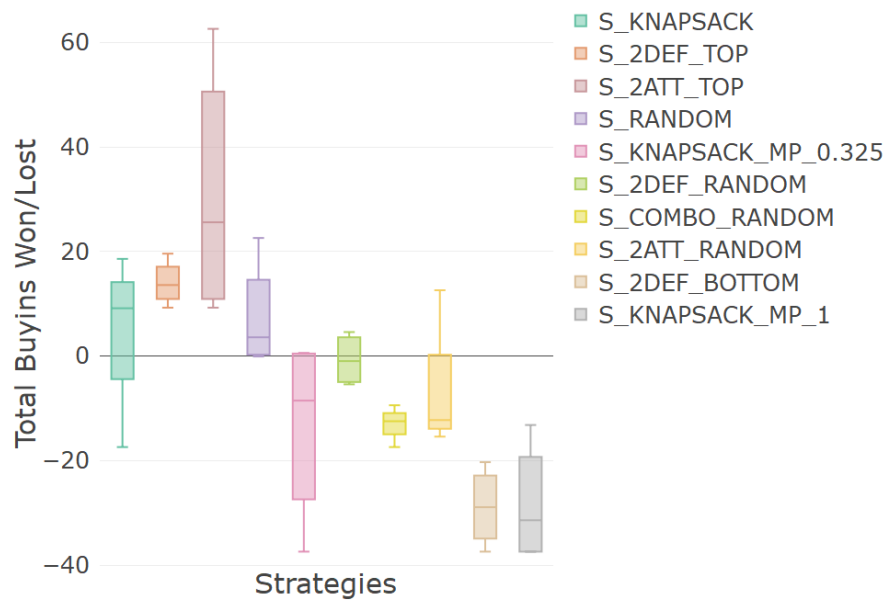


Figure 7.11: Simulation Results on ground truth data (Total Buyins Won/Lost per Strategy)

7.2 Tests

To determine whether resulting strategies differ significantly in their mean total points (*projected* and *backtested*), statistical tests have been conducted. For testing normality of all distributions of projected points, Shapiro-Wilk-Test was used at the 95% significance level. Null-Hypothesis ($p \geq \alpha$) was rejected for three of our final ten strategies ('*S_2def_bottom*', '*S_2def_random*', '*S_2def_top*') (see Table 7.1).

7.2.1 Projected Points

Table 7.1: Tests for Normality (Shapiro-Wilk)

Strategy	Description	Statistic	P.Value
S_1	knapsack	0.954	0.163
S_2	random	0.942	0.0690
S_3	2def_top	0.929	0.0294
S_4	2def_bottom	0.905	0.00635
S_5	2def_random	0.922	0.0184
S_6	2off_random	0.979	0.736
S_7	2off_top	0.983	0.872
S_8	combo	0.958	0.216
S_9	<i>optimum</i>	-	-
S_10	ks_missingPlayer_0.325	0.954	0.163
S_11	ks_missingPlayer_1	0.954	0.163

Given that constraints for normality were violated, nonparametric Kruskal-Wallis, a generalization of Mann-Whitney U test, has been chosen and resulted in rejection of Null-Hypothesis ($p - value < 2.2e - 16$). To find out about which specific strategies vary significantly, Post Hoc tests (pairwise Mann-Whitney-U as well as Conover-Iman Test of Multiple Comparisons Using Rank Sums¹) have been added, also applying the Bonferonni method (α/k) to match the need for correcting significance levels. Table 7.2 lists p-values ($\alpha = 0.005$) for all combinations of strategies.

7.2.2 Backtested Points

Kruskal-Wallis for *real points* resulted in rejecting Null-Hypothesis, too ($p - value < 1.909e - 05$). According p-values ($\alpha = 0.005$) for all combinations of strategies can be seen in Table 7.3.

¹<https://cran.r-project.org/web/packages/conover.test/conover.test.pdf>

Table 7.2: Comparison Matrix: Projected Points

	rand	2att rand	2att top	2def bttm	2def rand	2def top	combo	knap- sack	ks_mp 0.325
2att_rand	0.000*								
2att_top	0.000*	1.000							
2def_bottom	0.000*	1.000	1.000						
2def_rand	0.000*	0.015	0.000*	0.289					
2def_top	0.000*	1.000	1.000	1.000	0.852				
combo	1.000	0.000*	0.000*	0.000*	0.000*	0.000*			
knapsack	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*		
ks_mp_0.325	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	1.000	
ks_mp_1	0.000*	0.000*	0.000*	0.000*	0.006	0.000*	0.000*	0.000*	0.013

Table 7.3: Comparison Matrix: Backtested Points

	rand	2att rand	2att top	2def bttm	2def rand	2def top	combo	knap- sack	ks_mp 0.325
2att_rand	1.000								
2att_top	1.000	0.100							
2def_bottom	0.037	0.997	0.000*						
2def_rand	1.000	1.000	1.000	0.091					
2def_top	1.000	0.332	1.000	0.000*	1.000				
combo	0.401	1.000	0.001*	1.000	0.817	0.004*			
knapsack	1.000	1.000	1.000	0.010	1.000	1.000	0.140		
ks_mp_0.325	1.000	1.000	1.000	0.049	1.000	1.0000	0.496	1.000	
ks_mp_1	1.000	1.000	0.162	0.677	1.000	0.508	1.000	1.000	1.000

Chapter 8

Discussion and Conclusions

Findings of my work suggest that different approaches in building fantasy lineups can have significant impact on their performances in various competitions ranging from heads-up games to larger tournaments.

In regards to *projected* points, two strategies performed significantly worse than all other competing approaches: *S_random*, implementing a randomized lineup-creation, and *S_combo_2att_2def_random*, defined by always stacking defensive blocks as well as offensive blocks (for a total number of four different teams within a fantasy lineup). Apart from that, knapsack algorithm created astonishing results, beating all other strategies significantly.

In respect to *backtested* fantasy scores in real-life scenarios, results differ slightly: no strategy turned out to beat all others significantly – still, *S_combo_2att_2def_random* once again performed very poorly, only excelled by *S_2def_bottom*, which turned out to deliver even worse results. Both of these underperforming strategies vary significantly from strategy *S_2att_top* as well as *S_2def_top*.

Cumulative plots showed superior results for *S_knapsack* (*projected points*) and for strategies *S_2att_top* as well as *S_2def_top* for season-wide evaluation (*backtested*), followed by all other competitors.

Overall results suggest that picking a stacked combination of attacking – or defending – players from two top-teams is key for placing high in any kind of tournament and delivers significantly better results than simply combining players from four teams just for the sake of stacking more blocks. The same holds true for strategies which revolve around picking defending players from underdogs (either to create a unique lineup or to save money for midfielders and forwards).

For these reasons, punters probably should focus on building their lineups using blocks for attacking or defending players from top-teams, as clean sheet bonus or successful offensive events reward them with a bunch of points for each involved player (shifting their total accumulated points higher by a respectable amount). Unfortunately, combining both recommended approaches will not

work most of the time due to budget restrictions, which typically would make selecting too mediocre players to fill the rest of available squad slots inevitable. Eventually, picking the best possible strategy also depends partly on tournament type and even more on the competing playerpool’s average decisions.

In regards to my work, as expected, formats with a top-heavy payout-structure (such as *‘Winner takes it all’* and *‘One-vs-all’*) lead to a greater variance of results across different strategies than *‘Heads-Up’* or *‘Double-Up’* tournaments (see Figure 7.9).

More surprisingly, despite common beliefs of punters regarding different strategies being more or less applicable depending on tournament structure, this holds true only for a few candidates. Though strategies *‘knapsack’*, *‘2def_random’* and *‘2att_random’* lead to very different results depending on tournament type, all other strategies seem to be either clear winners or losers, no matter which competition they took part in (see Figure 7.10). While strategy *‘knapsack’* managed to create a positive return on investment in all competitions except for 1 : *ALL*, strategy *‘2att_random’* lead to the exact opposite result. Strategy *‘2def_random’* did well in bigger tournaments (1 : *ALL*, *STT*) in contrast to smaller tourney-types.

Interestingly, completely random selection of lineups did not even perform worst of all strategies: besides stacking players from underdogs, in particular missing players (0.325, but even more so for 1) turned out to lose even more (see Figure 7.11). This means that entering a contest early on, building a lineup well in advance and neglecting most recent news about starting squads and injuries seems to be by far the worst possible decision a punter can make.

It has been stated before in Section 2.3 that academic research in the scope of fantasy sports at this point is lacking appropriate methods for testing predictive models in real-life scenarios. Although there exist numerous studies which lay down in great detail how data has been collected, cleaned, wrangled, transformed and used for machine learning, the final step of building teams is often neglected.

Furthermore, even if predictions had been used to create lineups, their performances are generally compared to average scores of historic fantasy tournaments. While this still can tell us something about possible long-term profitability, it does not account for the various kinds of rules within a given tournament type, nor does it allow for studying ways of exploiting population tendencies, though both are key factors for real-life competitions outside the Lab.

In contrast to [41], more work has been put into creating reasonable predictions for player performances. Instead of simply taking the mean of the current season’s performances, a multitude of different attributes – melted with historic information about odds, weather and videogame player attributes – has been used to train a Stacked Ensemble model with H2o.

By taking a closer look into one of the base models created by using DRF algorithm, it has been shown that most important variables turned out to be a combination of additional player information (more detailed position in contrast to aggregated labels from Fanteam), average ratings for each active player (as

well as ratings relative to his team colleagues' most recent performances) and classic bookmaker-odds (for team win). Most of the other scraped data did not show to have any significant value for prediction tasks.

The resulting projections from the Stacked Ensemble model were put into a linear program, which successfully incorporates constraints of Fanteam's platform. Though stochastic linear programming could significantly add to the performance of lineup-creation, more adaptations would have been required to match the needs of traditional football: first of all, a very high mean (accompanied by a high variance) would not be necessary even for larger, top-heavy competitions as the number of competing participants generally is way lower (about 600 to 1000 per tournament compared to the tens of thousands in NFL contests). This could be prevented by reducing the threshold for required total expected points of a created lineup.

Another issue is the difference in scoring rules, which leads to goalkeepers and defenders in traditional fantasy football being (over)valued as a whole block, if their team manages to keep a clean sheet. This increases the effect of player performances being dependent on their colleagues, which has been mentioned before in Chapter 2.

For this reason, even if picking 6 defensive players (1 goalkeeper, 5 defenders) from 6 different teams still could turn out to lead to a high score, at the very same time a lot of hobby punters should be expected to place within a similar range of points with their teams: by simply picking 6 defensive players from only 2 different teams (out of all teams who banked a clean sheet), the very same result could be achieved – additionally allowing punters to pick cheap defensive and more expensive attacking players (if 6 teams turn out to concede zero goals on a given matchday, probability is high that at least a couple of them are underrated in their fantasy market price). Taking covariances into account and incorporating match scenarios with over/under, cleansheet and winnings odds into a stochastic linear program could add to further boost results and would be a massive contribution in regards to fantasy players selection.

Due to the setup and implementation of my work, several limitations have been identified: first of all, additional match-data on event-level could be used to add more accuracy to player's fantasy points calculation. Alternatively, scraping scores directly from fantasy sports sites would seem to be a valid option – unfortunately, one would have to track the data over the course of several years, as operators offer their statistics just for a limited period of time.

Furthermore, impact of a captain's bonus has not been inspected in detail, though this additional picking decision is well known for its huge influence on results.

On top of that, analyses and simulations could also be extended by incorporating information about a player's ownership within all participating teams. For example, it could be used to implement and test plenty variations of market value calculations for their impact on optimizing lineups. Pricing famous players higher than less popular ones could also turn out to account for more diverse lineups within the total player pool.

Moreover, different machine learning algorithms, predictions and strategies could be incorporated or extended.

Last but not least, by including various money management systems in the simulations, analyses about risk of ruin, return on investment and variance could be conducted.

In this thesis, I studied the principles of fantasy football by examining different strategies for picking a team of individual players. For this reason, match data was scraped from manifold sources, including statistics on team- as well as on player-level, historic betting odds, community ratings, attributes from video games and weather information.

Consecutively, fantasy scores and market prices were calculated, allowing for realistic picking decisions during simulations. By using R and H2o, a Stacked Ensemble model was trained and used for predictions on ground truth data.

Afterwards, multiple strategies for building fantasy lineups were defined and run against each other in different tournament formats.

Results suggest that there exist specific approaches which deliver superior numbers in comparison to others. It has also been shown that different tournament types have significant impact on the performance for some of the defined strategies.

Due to the limitations of my work – data sources and sample-size, calculation of fantasy points, market prices, chosen machine learning algorithms – it is possible, that some these findings could be proven wrong. Regardless, results point to the need for incorporating reasonable metrics for evaluation of predictive models for fantasy lineups.

Due to the codependency of outstanding performances within a single team – goalkeeper/defenders succeeding in banking a clean sheet, midfielder/forwards in netting a goal/assist – future researchers should consider investigating other approaches of building lineups.

Bibliography

- [1] FSTA, *Fantasy sports participation in North America rises to 57.4 million players*, FSTA, (2017), Retrieved from: <https://fsta.org/fantasy-sports-participation-in-north-america-rises-to-57-4-million-players/>.
- [2] Paul Oresteen, *Black Friday: Where Poker stands six years later*, Poker Central, (2017), Retrieved from: <https://www.pokercentral.com/articles/black-friday-where-poker-is-six-years-later/>.
- [3] FIFA.com, *2014 FIFA World Cup reached 3.2 billion viewers, one billion watched final*, Media Release, (Dec 16, 2015), Retrieved from <http://www.fifa.com/worldcup/news/2014-fifa-world-cuptm-reached-3-2-billion-viewers-one-billion-watched--2745519>.
- [4] Michiel De Hoog, *How data, not people, call the shots in Denmark*, De Correspondent, (Mar 24, 2015), Retrieved from <https://thecorrespondent.com/2607/how-data-not-people-call-the-shots-in-denmark/230219386155-d2948861>.
- [5] Andreas Heuer et al., *Usefulness of Dismissing and Changing the Coach in Professional Soccer*, (2012), DOI: <https://doi.org/10.1371/journal.pone.0017664>.
- [6] John Haigh, *On the ball*, Maths and Sport, (2015), Retrieved from <https://sport.maths.org/content/ball-0>.
- [7] Michalis Mitrotasios and Vasilis Armatas, *Analysis of Goal Scoring Patterns in the 2012 European Football Championship*, Contemporary Sports Issues, (2014).
- [8] A. Yiannakos and V. Armatas, *Evaluation of the goal scoring patterns in European Championship in Portugal 2004*, International Journal of Performance Analysis in Sport, 6:1, pp 178-188, (2017), DOI: 10.1080/24748668.2006.11868366.
- [9] Mike Hughes and Ian Franks, *Analysis of passing sequences, shots and goals in soccer*, Journal of Sports Sciences, 23:5, pp 509-514, (2007), DOI: 10.1080/02640410410001716779.

- [10] Joaquín Gonzalez-Rodenas, Ignacio Lopez-Bondia, Ferran Calabuig, Jose Antonio Pérez-Turpin and Rafael Aranda, *Association between playing tactics and creating scoring opportunities in counterattacks from United States Major League Soccer games*, International Journal of Performance Analysis in Sport, 16:2, pp 737-752, (2017), DOI: 10.1080/24748668.2016.11868920.
- [11] Mark Taylor, *Do Interceptions Matter to Teams ?*, The Power of Goals, (Jun 5, 2012), Retrieved from <http://thepowerofgoals.blogspot.com/2012/06/do-interceptions-matter-to-teams.html>.
- [12] M.J. Maher, *Modelling association football scores*, Statistica Neerlandica, 36:3, pp 109-118, (1982).
- [13] M.J. Moroney, *Facts from figures*, Penguin, (1956).
- [14] M.J. Dixon and S.G. Coles, *Modelling association football scores and inefficiencies in the football betting market*, Appl. Statist., 46:2, pp 265-280, (1997).
- [15] H. Rue and Ø. Salvesen, *Prediction and retrospective analysis of soccer matches in a league*, The Statistician, 49, pp 399-418, (2000).
- [16] D. Karlis and I. Ntzoufras, *Analysis of sports data by using bivariate poisson models*, The Statistician, 52:3, pp 381-393, (2003).
- [17] Mark Taylor, *Predicting and Explaining. How to use Statistics in Soccer*, The Power of Goals, (Oct 4, 2011), Retrieved from <http://thepowerofgoals.blogspot.com/2011/10/predicting-and-explaininghow-to-use.html>.
- [18] John Goddard, *Modelling football match results and the efficiency of fixed-odds betting*, (2018).
- [19] SBO.net, *Football Prediction Model: Poission Distribution*, (2018), Retrieved from <https://www.sbo.net/strategy/football-prediction-model-poisson-distribution/>.
- [20] Mark Taylor, *Total Shot Ratio & World Cup betting*, (May 16, 2014), Retrieved from <https://www.pinnacle.com/en/betting-articles/World-Cup/world-cup-total-shot-ratio/>.
- [21] 11tegen11, *Can Expected PDO teach us about luck in football?*, (Sep 22, 2015), Retrieved from <http://11tegen11.net/2015/09/22/can-expected-pdo-teach-us-about-luck-in-football/>.
- [22] 11tegen11, *Predicting league football using xG, and more*, (Sep 22, 2015), Retrieved from <http://11tegen11.net/2016/02/21/predicting-league-football-using-xg-and-more/>.

- [23] Opisthokonta.net, *A simple re-implementation of the Dixon-Coles model*, (May 06, 2018), Retrieved from <http://opisthokonta.net/?p=1685>.
- [24] Leonhard Knorr-Held, *Dynamic Rating of Sports Teams*, Collaborative Research Center 386, Discussion Paper 98, (1997).
- [25] He Yuan, *Predicting Market Value of Soccer Players Using Linear Modeling Techniques*, (2013), Retrieved from https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Yuan_He.pdf.
- [26] Rasmus Bååth, *Modeling Match Results in Soccer using a Hierarchical Bayesian Poisson Model*, UCS Minor, 18, (2015).
- [27] Gianluca Baio and Marta Blangiardo, *Bayesian hierarchical model for the prediction of football results*, Journal of Applied Statistics, 37:2, pp 253-264, (2010), DOI: 10.1080/02664760802684177.
- [28] Ben Ulmer and Matthew Fernandez, *Predicting Soccer Match Results in the English Premier League*, (2014).
- [29] Douwe Buursma, *Predicting sports events from past results Towards effective betting on football matches*, (2011).
- [30] Lars M. Hvattum and Halvard Arntzen, *Using ELO ratings for match result prediction in association football*, International Journal of Forecasting 26, pp 460-470, (2010).
- [31] Stefan M. Herzog and Ralph Hertwig, *The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition*, Judgment and Decision Making, 6:1, pp 58-72, (2011).
- [32] Anthony Constantinou, Norman Fenton, *Towards smart-data: Improving predictive accuracy in long-term football team performance*, Interfaces, 124, pp 93-104, (2017), DOI: <https://doi.org/10.1016/j.knosys.2017.03.005>.
- [33] Helge Langseth, *Beating the bookie: A look at statistical models for prediction of football matches*, Frontiers in Artificial Intelligence and Applications, 257, pp 165-174, (2013), DOI: 10.3233/978-1-61499-330-8-165.
- [34] Tarak Kharrat, *A Journey Across Football Modelling with Application to Algorithmic Trading*, (2016), DOI: 10.13140/RG.2.2.35651.60965.
- [35] John L. Kerry, *A new interpretation of information rate*, Bell System Technical Journal, 35(4):917-926, (1956).
- [36] Jeffrey A. L. Snyder, *What Actually Wins Soccer Matches: Prediction of the 2011-2012 Premier League for Fun and Profit*, (2013).

- [37] Thomas Peeters, *Testing the Wisdom of Crowds in the field: Transfer-market valuations and international soccer results*, International Journal of Forecasting, 34:1, pp 17-29, (2018), DOI: <https://doi.org/10.1016/j.ijforecast.2017.08.002>.
- [38] Xiaowei Liang, Zhuodi Liu and Rongqi Yan, *Result Prediction for European Football Games*, (2017), Retrieved from <https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a061.pdf>.
- [39] Leonardo Cotta et al. *Using FIFA Soccer video game data for soccer analytics*, (2016), Retrieved from http://www.large-scale-sports-analytics.org/Large-Scale-Sports-Analytics/Submissions_files/paperID13.pdf.
- [40] Sourya Dey, *Pricing Football Players using Neural Networks*, Final Project Report, (2017), Retrieved from <http://arxiv.org/abs/1711.05865>.
- [41] Sarah Newell, *Optimizing daily fantasy sports contests through stochastic integer programming*, (2017).
- [42] Xu Sun, *An Analytical Approach for Fantasy Football Draft and Lineup Management*, Journal of Quantitative Analysis in Sports, (2015), DOI: 12.10.1515/jqas-2013-0009.
- [43] Martin B. Haugh and Raghav Singal, *How to Play Fantasy Sports Strategically (and Win)*, (2018), Retrieved from http://www.columbia.edu/~mh2078/DFS_17April_2018.pdf.
- [44] Patrick Hruby, *The Founding Fathers of Fantasy*, SPORTS ON EARTH, (Dec. 2, 2013), Retrieved from <http://www.sportsonearth.com/article/64244480/>.
- [45] Corinne Green, *Wilfried ‘Bill’ Winkenbach invented Fantasy Football way back in 1962 with GOPPPL in Oakland*, Newsnetfive, (Sep. 11, 2014), Retrieved from <http://www.newsnet5.com/sports/wink-wilfred-bill-winkenbach-invented-fantasy-football-way-back-in-1962-with-gopppl-in-oakland/>.
- [46] Jon Saraceno, *As fantasy football fever continues to spread, it’s not all geeks to me*, USA Today, (Aug. 18, 2006), Retrieved from http://www.usatoday.com/sports/columnist/saraceno/2006-08-17-saraceno-fantasy-football_x.htm.
- [47] Ed Hilt, *Fantasy baseball league owners still bonding in their 32nd season*, Press of Atlantic City, (Jun. 26, 2007), Retrieved from <http://www.pressofatlanticcity.com/sports/local/story/7487701p-7382875c.html>.
- [48] Daniel Okrent, *The Year George Foster Wasn’t Worth \$36*, Inside Sports, (1981).

- [49] Sam Walker, *Fantasyland*, Viking, (2006), Retrieved from <http://www.fantasyland.com/>.
- [50] Rachel Shushter, *The game is fantasy but the addiction is real*, USA Today, (May 12, 1988).
- [51] Yahoo!, *Yahoo! Sports Hits Home Run With Free Fantasy Baseball*, Yahoo!, (February 23, 1999).
- [52] Montana Lottery, *How it all started*, Montana Lottery, Retrieved from <https://www.montanalottery.com/en/view/about/>.
- [53] Adam Kilgore, *Daily fantasy sports Web sites find riches in Internet gaming law loophole*, The Washington Post, (Mar 27, 2015).
- [54] Yuliya Chernova, *FanDuel Raises \$70 Million as Daily Fantasy Sports Battle Heats Up*, The Wall Street Journal, (Sep 2, 2014).
- [55] Lora Kolodny, *Fantasy Sports Create Billion-Dollar Startups*, The Wall Street Journal, (Jul 14, 2015).
- [56] Lora Kolodny, *Legality of Daily Fantasy Sports Under Nevada Law*, The Wall Street Journal, (Jul 14, 2015).
- [57] Lora Kolodny, *How Will the Government Change the Game for Daily Fantasy Sports?*, Bloomberg News, (Oct 15, 2015).
- [58] Ira Brouday, Joshua Brustein, *Daily fantasy sports state-by-state tracker*, ESPN, (Jul 26, 2017).
- [59] Yousif Al-Dujaili, *Is DFS stalling in Europe?*, Gambling Insider, (Sep 26, 2017).
- [60] Business Wire, *Fantasy Sports Industry Now over 15 Million*, Business Wire, (Aug 14, 2003).
- [61] FSTA, *Fantasy Sports Demographic Information*, FSTA, (2017), Retrieved from: <https://fsta.org/fantasy-sports-participation-in-north-america-rises-to-57-4-million-players/>.
- [62] Todd M. Nesbit and Kerry A. King, *The Impact of Fantasy Sports on Television Viewership*, Journal of Media Economics, 23:1, (2010).
- [63] John A. Fortunato, *The Relationship of Fantasy Football Participation with NFL Television Ratings*, Journal of Sport Administration & Supervision, 3:1, (2011).
- [64] Terry Lefton, *Survey results push nfl.com to increase fantasy content*, Sports Business Journal, (Sep 09, 2009).

- [65] Valery Bollier, *Daily Fantasy Sports (Presentation)*, CEEGC, (2016). Retrieved from <https://ceegc.eu/static/PP-Presentations-2016/Valery-Bollier/Oulala.pptx>.
- [66] Tanja Bergmann et al., *Linked Soccer Data*, Proceedings of the I-SEMANTICS, (2013)
- [67] Ed Miller, *For daily fantasy-sports operators, the curse of too much skill*, Sports Business Journal, (Jul 2015).
- [68] Ed Miller and Daniel Singer, *For daily fantasy-sports operators, the curse of too much skill*, McKinsey & Company, (2015), Retrieved from: <https://www.mckinsey.com/industries/media-and-entertainment/our-insights/for-daily-fantasy-sports-operators-the-curse-of-too-much-skill>.
- [69] Peter Hammon, *Analyzing FanDuel's Statistical Arguments On Skill Vs. Chance At The New York Hearing*, Legal Sports Report, (Dec 1, 2015).
- [70] Seth Young, *I Believe Daily Fantasy Sports Is a Game Of Skill, And Here's The Proof*, Legal Sports Report, (Aug 6, 2015).
- [71] Gambling Insider, *Don't Leave Me This Way*, Gambling Insider, pp 16-17, (Aug 6, 2015).
- [72] Tony Leino, *The Relationship Between Structural Game Characteristics and Gambling Behavior: A Population-Level Study*, Journal of Gambling Studies, 31:4, (Dec 2015).
- [73] John T. Holden and Sam C. Ehrlich, *Esports, Skins Betting, and Wire Fraud Vulnerability*, gaming Law Review, 21:8, (Oct 1, 2017).
- [74] Chris Rabalais, *AllSportsMarket (ASM) Breaks \$1 Billion Ahead of U.S. Supreme Court Betting Case*, The New Sports Economy Institute, (Nov 29, 2017).
- [75] Chris Grove, *DraftKings Lineup Leak Rockes Daily Fantasy Sports Industry: Questions And Answers*, Legal Sports Report, (Oct 4, 2015).
- [76] Marco Ottaviani and Peter Sørensen, *Parimutuel versus Fixed-Odds Markets*, Preliminary Draft, (Apr 2015).
- [77] Michael Cain, David Law, David Peel, *The Favourite-Longshot Bias, Book-maker Margins and Insider Trading in a Variety of Betting Markets*, Bulletin of Economic Research, 55:3, pp 263-273, (Jun 27, 2003).
- [78] Stephen M. Weiss, Robert M. Demski and George J. Backen, *Fantasy baseball: A new way to gamble or just another game?*, Journal of Gambling Issues, 26:126-145, (2011), Retrieved from: <http://jgi.camh.net/index.php/jgi/article/view/3854/3911>.

- [79] Ursula Hauser, Ulrich König and Elizaveta Krylova, *Gamblers' Rationality in Parimutuel Soccer Betting: Strategisches Spielverhalten bei Totalisatorspielen mit nicht-gleichwahrscheinlichen Spielausgängen wie Toto und Torwette*, Reihe Ökonomie Economics Series, 87, (2000).
- [80] Leo Breiman, *Random Forests*, Machine Learning, 45:5-32, (2001), DOI: <https://doi.org/10.1023/A:1010933404324>.
- [81] Nicolas Goix, Nicolas Drougard, Romain Brault, Mael Chiapino, *One Class Splitting Criteria for Random Forests*, (2016), Retrieved from: <https://arxiv.org/abs/1611.01971>.
- [82] Tin Kam Ho, *Random Decision Forests*, Proceedings of the 3rd International Conference on Document Analysis and Recognition, 1:278-282, (1995), DOI: 10.1109/ICDAR.1995.598994.
- [83] Tin Kam Ho, *A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors*, Pattern Analysis and Applications, 5:102-112, (1995), DOI: <https://doi.org/10.1007/s100440200009>.
- [84] Tin Kam Ho, *The Random Subspace Method for Constructing Decision Forests*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20:832-844, (1995), DOI: 10.1109/34.709601.
- [85] Robert Bryll, *Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets*, Pattern Recognition, 36:1291-1302, (2002), DOI: [https://doi.org/10.1016/S0031-3203\(02\)00121-8](https://doi.org/10.1016/S0031-3203(02)00121-8).
- [86] Pierre Geurts, Damien Ernst, Louis Wehenkel, *Extremely randomized trees*, Machine Learning, 63:3-42, (2002), DOI: 10.1007/s10994-006-6226-1.
- [87] Josip Hucaljuk and Alen Rakipovic, *Predicting football scores using machine learning techniques*, Proceedings of the 34th International Convention MIPRO, (2011), Retrieved from <https://ieeexplore.ieee.org/document/5967321/>.
- [88] Gunjan Kumar, *Machine Learning for Soccer Analytics*, Thesis submitted for the degree of Master of Science in Artificial Intelligence, (2013), Retrieved from https://www.researchgate.net/profile/Gunjan_Kumar2/publication/257048220_Machine_Learning_for_Soccer_Analytics/links/0c96052441dfabfc87000000.pdf.
- [89] Ben Ulmer and Matthew Fernandez, *Predicting Soccer Match Results in the English Premier League*, (2014), Retrieved from <https://www.semanticscholar.org/paper/Predicting-Soccer-Match-Results-in-the-English-Ulmer-Fernandez/d4dfbb3d1454416799b72004d3d88bae6bbd660a>.

- [90] George B. Mathews *On the partition of numbers*, Proceedings of the London Mathematical Society, 28:486–490, (1897), DOI: 10.1112/plms/s1-28.1.486.
- [91] Hans Kellerer, Ulrich Pferschy and David Pisinger, *Knapsack Problems*, (2004), Retrieved from: <https://www.springer.com/de/book/9783540402862>.
- [92] Andrew M. Odlyzko, *The Rise and Fall of Knapsack Cryptosystems*, Proceedings of Symposia in Applied Mathematics, (1990), DOI: 10.1090/psapm/042/1095552.
- [93] Tim Matthews, Sarvapali D. Ramchurn and Georgios Chalkiadakis, *Competing with Humans at Fantasy Football: Team Formation in Large Partially-Observable Domains*, Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, (2012), Retrieved from: <http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5088>.
- [94] Debarghya Das, *Moneyballer: An Integer Optimization Framework for Fantasy Cricket League Selection and Substitution*, (2014), Retrieved from <http://debarghyadas.com/files/IPLpaper.pdf>.
- [95] David S. Hunter, Juan P. Vielma and Tauhid Zaman, *Picking Winners Using Integer Programming*, (2016), Retrieved from <https://arxiv.org/abs/1604.01455>.

Appendices

Table 1: Game Specific Details

Attribute
Match ID
Date
Kickoff
Team Home/Away
Accurate Passes
Aerial Won/Lost
Attempt Penalty Goal
Attempt Penalty Miss
Attempt Penalty Post
Attempt Penalty Target
Blocked Scoring Attempts
Clearance Off Line
Contest Won/Lost
Corners
Defensive Blocked
Defensive Clear
Defensive Interception
Defensive Error Lead
Discipline Chat
Discipline Dive
Discipline Other
Discipline Red
Discipline Yellow
Duels Committed/Suffered
Duels Dribble Success/Fail
Duels Header Success/Fail
Duels Tackle Success/Fail
Error Lead To Goal
Fouls Committed/Suffered
Goal Assist
Goal Shot
Goals Head
Goals Inside Area
Goals Outside Area
Goals Right Foot
Goals Left Foot
Goalkeeper Saves
Good High Claim
Last Man Tackle
Man Of The Match
Own Goal
Passes Accuracy
Passes Assist

Passes Length
 Penalty Conceded
 Penalty Save
 Performance Attack (Squawka)
 Performance Defense (Squawka)
 Performance Possession (Squawka)
 Performance Total (Squawka)
 Played Minutes
 Player Rating (Whoscored)
 Possession
 Post Scoring Attempt
 Red Card
 Saves
 Second Yellow
 Shots all
 Shot off Target
 Shot on Target
 Six Yard Block
 Team: Goals Conceded
 Team: Goals Shot
 Throw In
 Total Offside
 Total Passes
 Total Scoring Attempts
 Total Tackles
 Touches
 Won Contest
 Yellow Card

Table 2: Player Specific Data: FIFA

Attribute

Name
 Height
 Weight
 Club
 Nationality
 Overall Skill
 Potential (for career mode)
 Transfer Market Value
 Wage
 Position(s)
 Preferred Foot
 Weak Foot
 Skill Moves

PAC: Pace
SHO: Shooting
PAS: Passing
DRI: Dribbling
DEF: Defending
PHY: Physical Strength
Attacking: Total
Attacking: Crossing
Attacking: Finishing
Attacking: Heading Accuracy
Attacking: Short Passing
Attacking: Volleys
Skill: Total
Skill: Dribbling
Skill: Curve
Skill: FK Accuracy
Skill: Long Passing
Skill: Ball Control
Movement: Total
Movement: Acceleration
Movement: Sprint Speed
Movement: Agility
Movement: Reactions
Movement: Balance
Power: Total
Power: Shot Power
Power: Jumping
Power: Stamina
Power: Strength
Power: Long Shots
Mentality: Total
Mentality: Aggression
Mentality: Interceptions
Mentality: Positioning
Mentality: Vision
Mentality: Penalties
Mentality: Composure
Defending: Total
Defending: Marking
Defending: Standing Tackle
Defending: Sliding Tackle
Goalkeeping: Total
Goalkeeping: GK Diving
Goalkeeping: GK Handling
Goalkeeping: GK Kicking
Goalkeeping: GK Positioning

Goalkeeping: GK Reflexes
Special: Preferred Position 1-4
Growth (compared to last update)
Attacking/Defensive Work Rate
Body Type
Traits
Specialities

Table 3: Player Specific Data: PES

Attribute
Name
Height
Weight
Club
Nationality
Position(s)
Stronger Foot
SPD: Speed
SHT: Shooting
PAS: Passing
DRI: Dribbling
DEF: Defending
PHY: Physical Strength
Playing Style(s)
Player Skill(s)
Attacking Prowess
Ball Control
Dribbling
Low Pass
Lofted Pass
Finishing
Set Piece Taking
Curve
Header
Defensive Prowess
Ball Winning
Kicking Power
Speed
Explosive Power
Body Control
Physical Contact
Jump
Goalkeeping
Catching

Clearing
 Reflexes
 Coverage
 Stamina
 NonDom Leg Usage
 NonDom Leg Precision
 Condition
 Injury Resistance

Table 4: Weather

Attribute
Region
City
Status
USAF
WBAN id
Station Name / Province
Country
Station Type
Date
Location
Elevation (ground, barometric)
Latitude
Longitude
Temperature (mean/max/min)
Rainfall
Wind (mean/max/min)
Climate Description

Table 5: Historical Betting Odds

Attribute
League Divison
Match Date
Home/Away Team
Half Time Team Goals
Full Time Team Goals
Full Time Result
Crowd Attendance
Referee
Team Shots
Team Shots on Target
Team Hit Woodwork

Team Corners
 Team Fouls Committed
 Team Free Kicks Conceded
 Team Offsides
 Team Yellow/Red Cards
 Odds: Bet365 (W/D/L)
 Odds: Blue Square (W/D/L)
 Odds: Bet&Win (W/D/L)
 Odds: Gameboookers (W/D/L)
 Odds: Interwetten (W/D/L)
 Odds: Ladbrokes (W/D/L)
 Odds: Pinnacle (W/D/L)
 Odds: Sporting Odds (W/D/L)
 Odds: Sportingbet (W/D/L)
 Odds: Stan James (W/D/L)
 Odds: Stanleybet (W/D/L)
 Odds: VC Bet (W/D/L)
 Odds: William Hill (W/D/L)
 Odds: Betbrain maximum (W/D/L)
 Odds: Betbrain average (W/D/L)
 Odds: Oddsportal maximum (W/D/L)
 Odds: Oddsportal average (W/D/L)
 Odds: Betbrain maximum (over/under 2.5 goals)
 Odds: Betbrain average (over/under 2.5 goals)
 Odds: Gamebookers (over/under 2.5 goals)
 Odds: Bet365 (over/under 2.5 goals)
 Odds: Betbrain maximum Asian (handicap home/away)
 Odds: Betbrain average Asian (handicap home/away)
 Odds: Gamebookers Asian (handicap home/away)
 Odds: Gamebookers (size of handicap home/away)
 Odds: Ladbrokes Asian (handicap home/away)
 Odds: Ladbrokes (size of handicap home/away)
 Odds: Bet365 Asian (handicap home/away)
 Odds: Bet365 (size of handicap home/away)
 Odds: Pinnacle closing odds (W/D/L)

Table 6: Additional Features for Machine Learning

Attribute	Source
Stadium Location	Wikipedia
Player Last Game	Whoscored
Played Last 3 Games	Whoscored
Played Last 6 Games	Whoscored
Gamesquad Changes (compared to previous match)	Whoscored
Days Since Last Match	Calculated

Attacking Strength	Calculated
Defensive Strength	Calculated
Expected Goals (Poisson Distribution)	Calculated
Player Ratings all (League)	Calculated
Player Ratings allActive (League)	Calculated
Player Ratings allInactive (League)	Calculated
Player Ratings all (Team)	Calculated
Player Ratings allActive (Team)	Calculated
Player Ratings allInactive (Team)	Calculated
FanDuel Fantasy Points (mean/max/min)	Calculated
Fanteam Fantasy Points (mean/max/min)	Calculated
Fanteam Market Value (mean/max/min)	Calculated

Abstract

Daily Fantasy Sports have been studied intensively by academics from various fields due to the game's nature within the area of skill and gambling.

Despite numerous approaches to investigate its principles and mechanisms by creating predictive models, there is a lack of extensive analyses in hindsight to lineup creation as well as evaluations for real-life scenarios.

By scraping and melting data from various sources including historic odds, weather information and player attributes from video-games, a Stacked Ensemble-model has been trained using R/H2o.

Subsequently, multiple strategies commonly used by punters have been defined and run in simulations against each other; allowing for assumptions whether fantasy football games could be beatable despite of the rake by following certain approaches.

In contrast to existing literature, results provide insights into performances of strategies in regards to various tournament types, showing significant differences in mean fantasy points.

In conclusion, new approaches for improvements are presented.

Abstract (Deutsch)

Daily Fantasy Sports sind aufgrund ihres Daseins im Graubereich inmitten von Geschicklichkeits- und Glücksspielen bereits Untersuchungsgegenstand zahlreicher akademischer Publikationen verschiedener Studienrichtungen.

Trotz deren Vielzahl an Zugängen, die zugrunde liegenden Prinzipien und Mechanismen des Spiels durch das Erstellen erklärender Modelle zu analysieren, existiert vor allem in Hinsicht auf die endgültige Zusammenstellung der Teams - sowie die Evaluierung ihrer Resultate in realen Szenarien - noch ein Mangel an geeigneten Vorgehensweisen.

Durch das Sammeln und Verknüpfen von Daten verschiedener Quellen, etwa historischer Buchmacher-Quoten, Wetterinformationen und Spieler-Attributen aus Videospielen, konnte ein geeignetes Supermodel mit Hilfe von R/H2o erstellt werden.

Im Anschluss wurden mehrere typische Strategien definiert, die üblicherweise von Spielern eingesetzt werden, um sie in weiterer Folge in Simulationen von realen Spieltagen gegeneinander antreten zu lassen.

Die Ergebnisse lassen Rückschlüsse auf die Schlagbarkeit von Fantasy-Spielen in Hinsicht auf die an den Anbieter zu zahlende Grundgebühr (Rake) zu. Im Gegensatz zur existierenden Literatur erlaubt dies einen besseren Einblick auf die Auswirkungen verschiedener Strategien in unterschiedlichen Turnier-Formaten.

Dabei konnten signifikante Abweichungen im Mittel der erreichten Punkte aufgezeigt werden.

Schlussendlich wurden außerdem weitere Möglichkeiten zur Verbesserung der bestehenden Zugänge vorgestellt.

Acknowledgments

First and foremost, I would like to thank my supervisor Univ.-Prof. Dipl.-Ing. Dr. Helmut Hlavacs for his constant support, his altruistic knowledge sharing and unmatched patience.

Thank you also to my colleagues Florian Schwarz and Werner Robitza, for myriads of discussions and alternative solutions, no matter which topic I could ever think of.

A big thank you to anyone still believing in me despite of my own self-doubts and disbeliefs.

Devoted to Georg “Juri” Selig, as he always inspired me to learn, grow and eat enough.

David Georg SELIG

PERSONAL DATA

PLACE AND DATE OF BIRTH: Linz, Austria | 05 Nov 1986
EMAIL: david.georg.selig@gmail.com

WORK EXPERIENCE

<i>Current</i> AUG 2016	Data Scientist at ÖSTERREICHISCHE LOTTERIEN GES.M.B.H., Vienna <i>Business Intelligence</i>
AUG 2016 FEB 2016	IT College at ÖSTERREICHISCHE LOTTERIEN GES.M.B.H., Vienna <i>Business Intelligence</i>
JAN 2015 MAR 2012	Poker-Coach at POKERSTRATEGY.COM, Vienna/Gibraltar <i>Video Production & Private Coaching</i>

EDUCATION

CURRENT 2013	Master of Science in MEDIA INFORMATICS, Univ. of Vienna <i>Thesis: "Daily Fantasy Sports: Ecology of a P2P zero-sum game"</i>
2013 2007	Bachelor of Science in MEDIA INFORMATICS, Univ. of Vienna <i>Thesis: "A Quality Model for YouTube"</i>
2005 1997	Matriculation at AKADEMISCHES GYMNASIUM, Linz <i>Secondary School</i>

ACADEMICS AND RESEARCH EXPERIENCE

FEB 2013 OCT 2012	Tutor at FUTURE COMMUNICATION, Univ. of Vienna <i>System Architectures for Future Networks</i>
MAR 2012 FEB 2011	Researcher at ENTERTAINMENT COMPUTING RESEARCH GROUP, Univ. of Vienna <i>Project "Content Aware Coding for Mobile TV"</i>
JUL 2010 FEB 2010	Tutor at FUTURE COMMUNICATION, Univ. of Vienna <i>Internet Technologies</i>
FEB 2010 SEP 2009	Tutor at FUTURE COMMUNICATION, Univ. of Vienna <i>Network Technologies and Multimedia Applications</i>