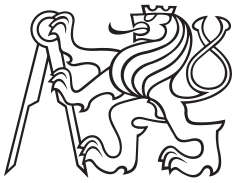
Bachelor Project



Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Competing in fantasy sports using machine learning

David Mlčoch

Supervisor: Ing. Ondřej Hubáček
Field of study: Cybernetics and Robotics
May 2021

BACHELOR'S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Mlčoch David** Personal ID number: **474542**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Control Engineering**

Study program: **Cybernetics and Robotics**

II. Bachelor's thesis details

Bachelor's thesis title in English:

Competing in fantasy sports using machine learning

Bachelor's thesis title in Czech:

Soutěžení ve fantasy sportech s pomocí strojového učení

Guidelines:

The recent rise of popularity of daily fantasy sports presents new challenges and opportunities for ML models. In a fantasy sport competition, a participant selects sportsmen for his virtual lineup(s) following set of constraints given by the organizer. After the sportsmen compete in a real-life, their performance is scored according to rules given by the organizer and fantasy points assigned to the lineups. Finally, the participants split the bank according to their performance in the fantasy competition.
The participant therefore has to take into account not only the expected performance of the sportsmen but also the profitability of his/her whole set of lineups as well as decisions of other participants. The goal of this thesis is to focus on the problem of maximizing profit during daily fantasy sports competitions.
1) Introduce the daily fantasy sports
2) Research state of the art in modeling fantasy sports.
3) Select a suitable sport domain and collect relevant historical data.
4) Analyze the specifics of the selected domain and its competitions.
5) Compare different approaches to modeling the value of player/lineup.
6) Develop a system for lineup(s) selection.
7) Evaluate the overall performance of your system from different perspectives.

Bibliography / sources:

[1] Hunter, D.S., Vielma, J.P. and Zaman, T., 2016. Picking Winners in Daily Fantasy Sports Using Integer Programming. arXiv preprint arXiv:1604.01455.
[2] Pantuso, G., 2017. The football team composition problem: a stochastic programming approach. Journal of Quantitative Analysis in Sports, 13(3), pp.113-129.
[3] Bonomo, F., Durán, G. and Marenco, J., 2014. Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game. International Transactions in Operational Research, 21(3), pp.399-414.
[4] Becker, A. and Sun, X.A., 2016. An analytical approach for fantasy football draft and lineup management. Journal of Quantitative Analysis in Sports, 12(1), pp.17-30.

Name and workplace of bachelor's thesis supervisor:

Ing. Ondřej Hubáček, Intelligent Data Analysis, FEE

Name and workplace of second bachelor's thesis supervisor or consultant:

Date of bachelor's thesis assignment: **25.05.2020** Deadline for bachelor thesis submission: _____

Assignment valid until:
by the end of winter semester 2021/2022

Ing. Ondřej Hubáček	prof. Ing. Michael Šebek, DrSc.	prof. Mgr. Petr Páta, Ph.D.
Supervisor's signature	Head of department's signature	Dean's signature

III. Assignment receipt

The student acknowledges that the bachelor's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor's thesis, the author must state the names of consultants and include a list of references.

. _____
Date of assignment receipt

_____
Student's signature

Acknowledgements

I would like to thank my supervisor Ing. Ondřej Hubáček, for his most welcomed advice, willingness, and guidance during our consultations, and also to my family, Stack Overflow and friends, who supported me throughout my studies.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses. In Prague, 20. May 2021

Abstract

The goal of my bachelor thesis was to design a model for the creation of fantasy football lineup portfolios in Daily fantasy tournaments with top-heavy payoff structures. In these tournaments, most of the winnings go only to the top participants. Therefore, we not only aimed to maximize players expected performance but also their variance and covariance. Our objective was also to minimize the correlation between the portfolio's lineups. By sampling probabilistic models of individual fantasy player statistics, we estimated fantasy point distribution for all players. We used players fantasy points mean and covariance prediction in our subsequent mixed-integer quadratic program (MIQP). We tested the created model on real fantasy data crawled from the fantasy sports provider. The results are very promising. The model finished in profit at the end of the season.

Keywords: Daily fantasy sports, Sports analytics, Portfolio optimization, Mixed-integer quadratic programming, Probabilistic models

Supervisor: Ing. Ondřej Hubáček

Abstrakt

Cílem mojí bakalářské práce bylo navrhnout model na tvorbu portfolia fotbalových fantasy soupisek pro Daily fantasy turnaje s top-heavy výplatní strukturou. V těchto turnajích připadne většina zisku pouze nejlepším hráčům. Naším cílem tedy nebylo pouze maximalizovat očekáváný hráčský výkon, ale také varianci a kovarianci hráčů. Dalším cílem bylo minimalizovat korelaci mezi soupiskami portfolia. Vzorkováním pravděpodobnostních modelů jednotlivých fantasy statistik hráčů jsme odhadli distribuci fantasy bodů všech hráčů. Predikovaný průměr a kovarianci fantasy bodů hráčů jsme využili v následné námi navržené úloze smíšeného celočíselného kvadratického programování (MIQP). Vytvořený model jsme otestovali na reálných datech nacrawlovaných od poskytovatele fantasy sportů. Výsledky jsou velice slibné. Model skončil na konci sezóny v zisku.

Klíčová slova: Denní fantasy sporty, Sportovní analýza, Optimalizace portfolia, Smíšené celočíselné kvadratické programování, Pravděpodobností modely

Překlad názvu: Soutěžení ve fantasy sportech s pomocí strojového učení

Contents

Figures

Tables

Chapter 1

Introduction

Daily fantasy sports have recently become significantly more popular. The possibility of winning high prices with low buy-in attracts a growing number of participants. Fantasy sports classification as not a game of chance makes them ideal for statistical modelling. Our goal is to create a model for constructing portfolios of fantasy lineups. We'll focus on the football tournaments with a top-heavy pay scale. We'll train and test the model on the real data crawled from the fantasy provider. Our aim is to finish the season with our model in profit.

Chapter 2 introduces daily fantasy sports in more detail. The thesis follows with Chapter 3, where we put our approach in context with the previous research. Chapter 4 shows the players statistics we have crawled and used for modelling. We explain our model in Chapter 5. We divide our model into two parts. First, we simulate the football matches with probabilistic models to obtain players fantasy points predictions and covariances. These statistics are then used in the subsequent mathematical optimization model. We formulate the lineup portfolio formation as a mixed-integer quadratic program. We maximize the expected fantasy points of a lineup, together with players variance and their covariance. Our objective is also to minimize the correlation between the portfolio's lineups. Our parameterized objective function can be tuned for different strategies. We show and discuss our results in Chapter 6.

Chapter 2

Daily Fantasy Sports

Fantasy sport is a game where participants act as sports managers. They assemble virtual teams made out of real players. Every player has a price based on his skills. Better player = higher cost. Prior to the actual matches, participants create virtual lineups limited by constraints (e.g. limited budget, only one goalkeeper etc.). Players then compete in series of matches in the real world. Their performance is rated by the scoring rules designed by the fantasy game provider. In football, for example, a player gets 5 points when he scores a goal, loses 1 point for a yellow card, etc. These fantasy points are then added up over all players in the lineup and credited to the participant. Before the entry to the tournament, participants need to pay an entrance fee. Fantasy sports provider takes some portion of the collected money and divides the rest between some percentage of top-ranking participants.

The length of the competitions can vary significantly. Seasonal fantasy games last for an entire season. Participants create a team at the beginning of the season and then throughout the season, they are allowed to make some changes to the team. They are paid out at the end of the season. Because in seasonal fantasy games, players need to wait a long time for the results, daily fantasy sports have emerged. They last for a shorter period, usually a maximum of one week. Player's performances from only one match are ranked in contrast to all the matches in the season-long game. This thesis will focus only on the daily fantasy sports.

2.1 Terminology

- **Participant** = Participant of the fantasy game
- **Player** = Real football player
- **Fantasy points (fp)** = Points awarded to players for their performance in the matches.
- **Lineup** = Participant's selection of eleven players, who will compete in the fantasy competition.

- **Prize pool** = Amount of money distributed between winners of the fantasy competition.

2.2 Types of Competitions

This section describes the most common types of competitions and their payout structures.

2.2.1 Guaranteed Prize Pools

In this type of game, the prize pool distribution is known in advance, no matter how many people enter the contest. Provider guarantees prices for particular rankings. Guaranteed Prize Pools are often associated with a top-heavy pay scale. Only 10-25% of participants with the highest number of points win, and the prices are not distributed evenly. A handful of top players win most of the money. These games are attractive for their possibility to win high prices with a small entry fee. However, it is also much harder to win at least some price as there is more competition and less winners. Many providers allow participants to enter multiple lineups in one game to increase their chances. [LegalSportsReport, 2020] We will focus only on the Guaranteed Prize Pools in this thesis.

2.2.2 Head-to-Head

In a Head-to-Head game, the participant's lineup competes with only one other participant's lineup. The lineup with the higher score wins and gets the opponent's entry fee minus a rake (provider's commission fee).

2.2.3 Double-up or 50/50

Double-up and 50/50 are the same names for a game where those participants who finish in the top half split the cash pool evenly, with every participant nearly doubling their entry fee. The payout is a bit less than half due to the rake.

2.2.4 Multipliers

Multipliers are similar to Double-Ups. In Double-Up, it is possible to win 2x of the money. In multipliers, it is possible to win other amounts as well - 2x, 3x, 4x, or 5x of the buy-in. A smaller amount of people can win, but the price is proportionally higher. [Chase, 2018]

2.3 Fantasy Sports Providers

2.3.1 DraftKings and FanDuel

DraftKings and FanDuel are the biggest fantasy sports providers in the US. Together they account for more than 90 percent of the market share. [LegalSportsReport, 2020] With FanDuel founded earlier than DraftKings in 2009, DraftKings now has a more extensive user base. As of July 2017, DraftKings had eight million users. [Tepper, 2017] Both companies offer large varieties of fantasy sports. DraftKings and FanDuel are operating under US law and are unavailable in most of the European countries. We mention them here to make a comparison between DraftKings, FanDuel and Fanteam football scoring rules.

2.3.2 FanTeam

FanTeam is the biggest daily fantasy site in Europe [FanTeam, 2020]. In this thesis, we will use data from Fanteam's English Premier League daily fantasy tournaments. The following subsections will explain FanTeam soccer scoring rules, lineup requirements, payout structures and compare them with the most prominent US providers, DraftKings and FandDuel.

Lineup Requirements and Rules

Lineups for the FanTeam tournaments we will model need to follow these rules [FanTeam, 2018]:

- Sum of player's prices must be less than $100.0M.
- A fantasy team consists of 11 players. Each lineup has a goalkeeper and a combination of other players. Allowed formations are: 5(defenders)-4(midfielders)-1(attacker), 5-3-2, 4-3-3, 4-4-2, 4-5-1, 3-5-2, 3-4-3, 5-2-3
- Lineup can not include more than 3 players from same team.
- Multiple fantasy teams per participant are allowed.
- Participant must choose a captain for the fantasy team, and that player will receive 2x points.
- Participant must choose a vice captain for the fantasy team, and that player will receive 2x points if the captain does not play.
- If two teams are tied in a game, the team with the most remaining funds wins.
- Safety-net is enabled. All non-starting players will be replaced at the beginning of their match, with a player from the same team, same position, and equal or cheaper price. Closest price chosen first.

Scoring Rules

After players finish their matches, their statistics are recorded and transferred into fantasy points according to the provider's scoring rules. A more successful player receives more points. Every provider has a different set of scoring rules, which leads to the necessity of adapting the lineup formation strategy to each provider.

We compare FanTeam, DraftKings, and Fanduel football (soccer) scoring rules in Tables 2.1 and 2.2. There are multiple similarities, but also significant differences. DraftKings have more detailed rules (e.g., accurate pass, passes intercepted, and crosses). Measuring more specific events can lead to better accuracy in the evaluation of the player's real impact. Fanteam is mainly concerned with goals, assists, clean sheets, time spend on the field, cards and penalties. Both FanTeam and DraftKings have 22 scoring rules, but FanTeam has more negative scoring rules. FanDuel has only 13 rules, and they are less dependent on the position. FanDuel also has more detailed events than FanTeam (tackles, clearances, interceptions, etc.) but does not consider cards. A significant difference is that only FanTeam ranks goal differently depending on the position (forward, midfielder, defender, goalkeeper). This needs to be taken into account during the lineup formation. Also, only FanTeam rates time spend on the field and winning/losing during the period player is on the field. [FanTeam, 2018], [DraftKings, 2018], [FanDuel, 2018]

FanTeam Football Scoring Rules	
Midfielder or attacker plays the full match	**1**
Forward scores a goal	**4**
Midfielder scores a goal	**5**
Defender scores a goal	**6**
Goalkeeper scores a goal	**8**
Assist or Fantasy Assist	**3**
Midfielder keeps a clean sheet	**1**
Defender keeps a clean sheet	**4**
Goalkeeper keeps a clean sheet	**4**
Goalkeeper or defender concede 2 goals	**-1**
Score an own goal	**-2**
Penalty miss	**-2**
Goalkeeper made a save	**0.5**
Goalkeeper saves penalty	**5**
Playing time up to 60 minutes	**1**
Playing time over 60 minutes	**1**
Yellow card	**-1**
Red card	**-3**
Scoring freekick caused	**-2**
Caused a penalty	**-2**
Impact = Team "wins" in the period player is on the pitch	**1**
Impact = Team "loses" in the period player is on the pitch	**-1**

Table 2.1: FanTeam football scoring rules

DraftKings Soccer Scoring Rules		FanDuel Soccer Scoring Rules	
Goal	**10**	Goal	**15**
Assist	**6**	Assist	**7**
Shot	**1**	Shot on goal	**5**
Shot on goal	**1**	Chances created	**3**
Crosses	**0.7**	Tackles	**1.3**
Assisted shot	**1**	Clearances	**1.3**
Accurate pass	**0.02**	Interceptions	**1.3**
Fouls Drawn	**1**	Blocked Shots	**1.3**
Fouls conceded	**0.5**	Defender Clean Sheet	**5**
Tackle won	**1**	Goalkeeper Clean Sheet	**10**
Passes intercepted (D,M,F)	**0.5**	Goalkeeper concedes goal	**-2.5**
Yellow card	**-1.5**	Goalkeeper save	**3**
Red Card	**-3**	Goalkeeper win bonus	**7**
Clean Sheet (D)	**3**		
Shootout goal	**1.5**		
Shootout miss	**-1**		
Goalkeeper save	**2**		
Goalkeeper concedes goal	**-2**		
Goalkeeper has clean sheet	**5**		
Goalkeeper if team wins	**5**		
Goalkeeper saves penalty	**3**		
Goalkeeper saves shootout	**1.5**		

Table 2.2: DraftKings and Fanduel soccer scoring rules

Payout Structure of Guaranteed Prize Pool

Each fantasy sports provider use different price distribution with Guaranteed Prize Pools games. We will focus on the FanTeam's Guaranteed Prize Pool payout distribution which we can see in the Figure 2.1. Graph shows the prize distribution of one of the FanTeam tournaments. Distribution is decided by a ratio of 15 standard score calculation. [FanTeam, 2020] To attract more players, FanTeam promotes a massive price for the first place. Winnings from the next places then decrease rapidly. In this tournament, 20% of participants won. Out of 1555 participants, 311 participants won at least some price. However, we can see that most of the money was accumulated only by the top players.

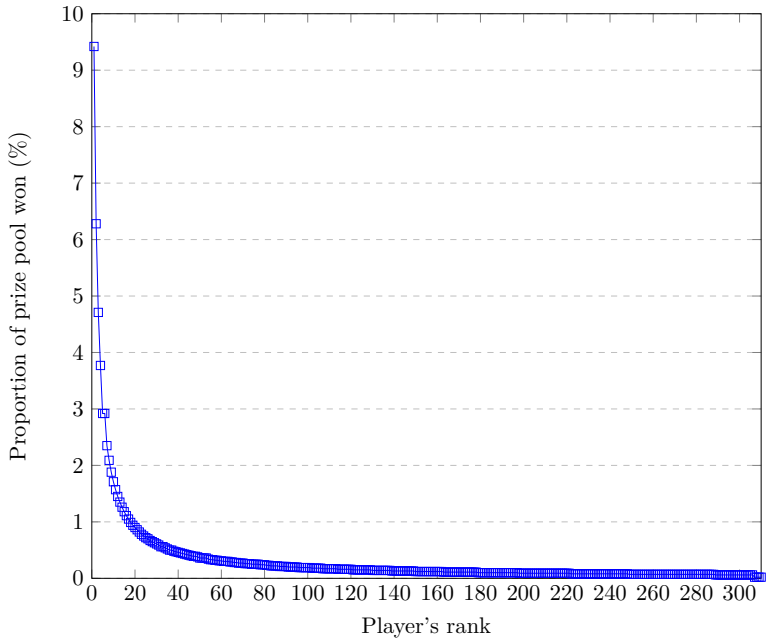


Figure 2.1: FanTeam top-heavy prize distribution for tournament with 1555 participants

2.4 Game of Skill

There has been an ongoing discussion whether to classify fantasy sports as a form of gambling or not. In case fantasy sports were classified as gambling, they would be considered illegal in the US.

Most fantasy sports providers argue that fantasy sports are games of skill. In determining whether the game depends on the chance or the skill, most US states employ the Predominance Test - *"The test inquires whether the outcome of the activity at bar is determined more by a participant's skill or by uncontrollable chance. The level of chance only becomes significant in the analysis when it can be shown to be the predominant element in the outcome."* [Meehan, 2015] Therefore, if the player's skill determines at least 51 % of the game's outcome, then by the Predominance Test, that game is a game of skill. In [Meehan, 2015], it has been demonstrated that under this and other tests, fantasy sports are games of skill similarly to poker.

One of the essential pieces of evidence is the ability of a small percentage of players to win most of the prices. DraftKings presented aggregate data at the 2014 Sloan Sports Analytics Conference. It showed that only 10% of its players were profitable in 2013. However, 80% of those profits were made by 5% of the profitable players on DraftKings. [Robins, 2018] Such a big difference can hardly be attributed to a chance, but rather to the player's skill. A crucial skill of an experienced fantasy player is the manner they manage their bankroll. Managing a bankroll is a form of risk assessment similar to risk management of a stock portfolio. Players need to decide on many variables. How much money to put in, what contests to enter, when to enter etc. [Meehan, 2015] These skills are crucial. They separate professional daily fantasy players from recreational users.

Chapter 3

Related Work

In this chapter, we introduce related work. Most of the past fantasy sports studies have focused on seasonal contests, so it feels natural to analyze some of them first. Then we move to our main focus: Daily fantasy sports. Subsection names in this chapter correspond to lineup selection approaches cited authors have taken.

3.1 Seasonal Fantasy Sports

In a seasonal game, participants select starting lineup at the beginning of the season, and then they are allowed to make a few allowed changes after each round of matches.

3.1.1 Integer linear program

Article from [Bonomo et al., 2014] presents its integer linear program model. The model acts as a participant, selecting the best starting lineup and the changes after every round. Using integer linear programming, the model tries to maximize the lineup score at each round.

The model must first find the player's performance predictions based on the historical data. Player performance prediction is called player's index. Player's indexes are calculated as averages of the points from the rounds already played in the current tournament. If there are data about the player from the previous tournaments, they are also used. The model weights indexes by three factors - home/away, the league table position of his next rival and the current performance of the player. The predictive mathematical model then tries to maximize the global team index = sum of the individual player's indexes [Bonomo et al., 2014].

[Bonomo et al., 2014] tested their models on Argentinian fantasy soccer game Gran DT. Players came from the First Division of the Argentinian professional soccer league. The model was tested in 6 tournaments and scored well at all of the them. It positioned in the top 0.1% of the game participants

in 1 of the tournaments, in the top 4% in 4 of the tournaments, and in the top 10% in the remaining tournament. With all the results combined, the model came at 530th place out of 343 017 competitors who participated in all six tournaments. That positioned the model among the top 0.2% [Bonomo et al., 2014]

Authors suggest that even better results can be achieved. For example, the computed models can be used just as a complementary support tool for experts who can use their knowledge as well. They also created a model to find a perfect lineup after we know the tournament results. The ideal lineups created exceeded all the human competitors by 50% to 70%. Even the best human lineups are thus still very far from optimal. This could be mainly attributed to participants rarely relying on highly unpredictable one-time successful players. Another suggested improvement is the formation of high-risk teams composed of players with high points variance. These teams might lead to rare exceptional good results. The authors also mention a similarity between the selection of players portfolio and stock portfolio in finance. [Hunter et al., 2016] mentions similar idea.

3.1.2 Mixed-integer optimization program

Similar to [Bonomo et al., 2014], [Becker and Sun, 2016] developed a model to create and edit the best lineup possible based on the player's and team's performance during the season. Compared to other studies, the proposed mixed-integer optimization model used many diverse types of historical data. Used data included: individual player performances, fantasy draft rankings for each position from expert articles, summary reports of actual owner draft behavior for several public fantasy contests. More data were obtained for simulation purposes. The model was trained on the 2004-2006 NFL season data. It has not been tested against real players, but on created simulation engine while using data from 2007 and 2008 NFL seasons. The engine was supposed to simulate the draft, weekly play and playoff phase. The engine was using data available for the current and past season. With simulating over 300 trials on the 2007 NFL season, the model won 16.7% of the time and came at least the second-best at almost 27.3% of the time. In the 2008 season, the model finished first or second 20.7% of the time. These results look promising, but it is not clear whether the model would be that successful in future tournaments against real opponents. As an improvement, authors suggested modeling of opponent's behavior based on media predictions and opinions which may indicate the general strategy of all the participants.

3.2 Daily Fantasy Sports

3.2.1 Sequential integer program with greedy algorithm

The focus of [Hunter et al., 2016] was on the guaranteed prize pool tournaments with top-heavy payoff structures. In these contests, most of the winnings go just to a handful of top players. Thus, the optimization goal wasn't to have many winning lineups with average rewards but to maximize the probability of having a few exceptional lineups that will rank in the top. The authors call this *picking winners* problem. This framework was introduced in [Hunter et al., 2017] to help venture capital funds form the portfolio of start-up companies worth investing in.

[Hunter et al., 2016] used publicly available player points predictions from the websites Rotogrinders [Rotogrinders, 2015] and Daily Fantasy Nerd [Nerd, 2015]. Authors assert that these predictions are accurate enough. They focused on building a player's fantasy points predictive model while using these predictions.

A sequential integer program with a greedy algorithm is used to construct lineups subjected to constraints. The greedy algorithm solves integer programming problem at every step. This program employs some heuristic principles: *"First, the entries' scores should have a large expected value and variance. This increases the marginal probability of an entry winning. Second, the entries should also have low correlation with each other to make sure they cover a large number of possible outcomes."* [Hunter et al., 2016] The authors identify multiple types of lineup constraints:

- **Feasibility constraints** = constraints to create a valid lineup. These are maximum budget, positions limitations (only one goalkeeper, etc.), players from different teams, etc
- **Overlap constraints** = Sharing the same players between more lineups will decrease the variance. To adjust this overlapping, the authors use the maximum lineup overlap constraint. This parameter sets how many players can lineups share.
- **Stacking constraints** = These are some heuristically learned rules applied to the model. In this way, it's possible to use domain knowledge. For instance, in hockey, players in the same line are positively correlated.

With [Hunter et al., 2016] approach, it is possible to create a portfolio of lineups that maximizes lineups mean and variance. After testing in real daily fantasy sports competitions, their approach has proven to be hugely successful. Two hundred lineups were usually created and entered into the hockey or baseball competitions. One of the lineups ranked between the top-ten entries in hockey and baseball contests, with thousands of competing entries numerous times. [Hunter et al., 2016]

3.2.2 Modeling opponent's behavior

[Haugh and Singal, 2018] follows on [Hunter et al., 2016]'s work. Authors also maximize the expected reward subject to portfolio constraints using a greedy algorithm. This work's significant improvement is modelling other participant's lineups and incorporating opponents' behaviour into the lineup creation. It is the first Daily Fantasy Sports study that has attempted to do so. The possible value of information about the opponent's lineups before the contest is further examined. [Robins, 2018] argue in favour of this approach that *"The payoff thresholds are stochastic and depend on both the performances of the real-world players as well as the unknown team selections of their fellow fantasy sports competitors."* Therefore, we must take opponents' behaviour into account. For example, there are well-known player behaviours, such as preferring to choose more popular players with no regard for real statistics. [Haugh and Singal, 2018]

The focus of [Haugh and Singal, 2018] was not limited to top-heavy tournaments but also on double-up tournaments. Authors argue that [Hunter et al., 2016]'s approach is optimized not for top-heavy payoff structure but for the winner-takes-all payoff, which is an only approximation of the true payoff structure. Therefore in [Haugh and Singal, 2018] work, the payoff structure is properly reflected (double-up or top-heavy in this case). Similarly to [Hunter et al., 2016], maximum lineup overlap constraint is imposed on the entries to diversify the portfolio.

[Haugh and Singal, 2018] use Dirichlet multinomial data generating process to model opponents' lineups. The Dirichlet regression estimates the parameters of this model. The authors claim that we can reduce the optimization problem (with some simple assumptions and approximations) to a binary quadratic program. Expected fantasy points data were obtained through paid subscription from Fantasy Pros[FantasyPros, 2018]. An estimate of the correlation matrix was obtained from RotoViz [RotoViz, 2018]. The proposed framework has been compared to stochastic benchmarks and tested on double-up and top-heavy daily fantasy sports contests in the 2017 and 2018 NFL season. Results from top-heavy contests have been considerably better than from the double-up. The portfolio accounting with the opponent's behaviour has earned a cumulative profit of $384.74 in 12 weeks. It also outperformed the benchmark portfolio, which did not account for the opponent's lineups. Recommendations for future research mainly focused on improving estimations of opponent's portfolios. Also, as they have only analyzed football tournaments, they propose trying other sports with lower individual player variance, such as basketball, ice hockey or baseball. [Haugh and Singal, 2018]

3.2.3 Stochastic integer program

[Newell, 2017] has taken a slightly different approach and developed a stochastic integer program for optimizing the expected payout of a top-heavy tourna-

ment. Their program predicts each player's fantasy point distribution rather than predicting a single fantasy point value. Player's fantasy points are believed to be independent and normally distributed. Therefore, the team's fantasy points should be normal as well. Fantasy points earned by the team are the sum of the player's distributions. The final expected payout of the team is the sum of all payouts multiplied by the probability that the team achieves that payout level. [Newell, 2017]

The author argues that the presented algorithm produces an optimal lineup with maximized expected payout. The program is also computationally tractable as it takes less than 2 seconds to run on an average computer. When this program was tested for each week of the 2016-2017 NFL DraftKings® NFL Millionaire Maker contest, it did not show promising results. Most of the teams haven't scored enough fantasy points to reach even the lowest possible payout. The author suggests that a participant would have to wait a few seasons before winning a payout. He suggests possible improvements to the model. First, providing better estimates of the athletes' fantasy point distribution would increase the accuracy. Also, some players play better with other players on the field etc. These relationships could be reflected through covariance and added to the model. Another question to ask is whether the team's distribution is indeed normal. Incorporating other aspects relevant to the match (home/away, weather, injuries or others) could also show improved results. [Newell, 2017]

Chapter 4

Data

We have decided to test our approach on FanTeam's Premier League main tournaments. Premier League is the top league in the English football league system. The season starts in August and runs till May. It consists of 20 teams, each playing 38 matches - playing each team twice (home and away). Team usually plays once a week and most games are played on Saturday and Sunday afternoons. Daily fantasy sports last only a few days over which teams play. The Premier League is the most-watched sports league in the world. It is being broadcasted in 212 territories to 643 million homes and reaches a potential TV audience of 4.7 billion people. [Ebner, 2013], [John Dubber, 2015] Therefore the fantasy user base is very significant.

4.1 Fanteam tournaments

FanTeam main tournaments run every week of the season (38 in total), and they offer four different prize pools every week. Each tournament consists of 8 to 10 Premier league matches happening that week. All of the tournaments have a guaranteed prize pool money distribution explained in Chapter 2. There are four tournaments every week, each differing in prize pools and buy-ins. Number of participants also change as the season progress over time. We show average weekly prize pools, buy-ins, rakes and number of entrants (participants) for every tournament type in the Table 4.1. We can see that number of entrants depend on the buy in. Most notably, tournament type 4 with buy in over €100, usually has only around 26 entrants. Number of lineups we create will need to reflect the the tournament type. With increasing buy-in, number of lineups will probably have to be set smaller.

Tournament type	Buy in	Prizepool	Rake	Entrants
1	€1	€2224	€0.1	1866
2	€3	€4044	€0.3	1199
3	€10	€21351	€1.0	1654
4	€101.5	€2802	€8.5	26

Table 4.1: FanTeam weekly tournament types and their average statistics

4.2 Crawled data

As there was no available public dataset of FanTeam data, we had to put a considerable amount of effort into gathering and consolidating the data. We used Python with selenium and scrapy to crawl many parts of the FanTeam website and create a comprehensive database. We have assembled information about tournaments for season 2018 and information about players for seasons 2016, 2017 and 2018. It was not possible to crawl tournament information from older seasons as these data are not available anymore on the FanTeam website.

4.3 Prediction of match results

For the prediction of match results, we used the Double Poisson model, first introduced in [Maher, 1982]. This model provided us with λ_h and λ_a Poisson distribution parameters. With these parameters, we were able to create a Poisson distribution for home and away scored goals.

4.4 Database scheme

The Figure 4.1 displays our database scheme. The database includes crawled data from FanTeam and matches predictions described in the previous sections. From the player statistics, we've computed awarded fantasy points for each scoring rule described in Table 2.1. We refer to these statistics as fantasy points statistics. They're indicated in Player statistics table in Figure 4.1. These statistics include whether the player played a full match, scored a goal, get a yellow card etc. With these statistics we can evaluate players fantasy points performance in more detail than with a single fantasy point metric as for example [Hunter et al., 2016] did.

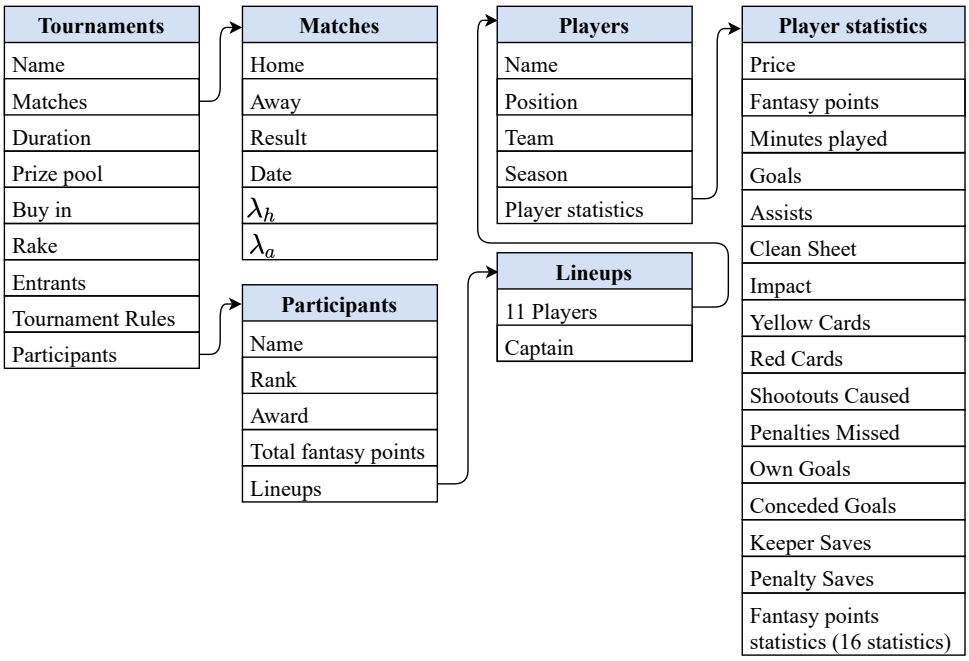


Figure 4.1: Database scheme

Chapter 5

Models

This chapter describes our model for creating a portfolio of lineups in daily fantasy sports competitions. We can divide our model into two subsequent parts:

1. Matches simulation

2. Mixed-integer quadratic program

5.1 Model pipeline

A high-level overview of our model is in Figure 5.1. On the left side of the figure, we have input parameters and data. On the right side of the pipeline is the model's output, which is the portfolio of lineups. Inside the model, we have two major parts: Matches simulation and Solver. First, the model simulates tournament matches to obtain predictions of player fantasy points, players variances and players covariance matrix. These statistics are then passed to an optimization solver, which solves our mixed-integer quadratic program and outputs a portfolio of lineups. This portfolio of lineups is the list of players that we should bet on in the tournament. We've implemented all parts of the model in Python 3 except for the mixed-integer quadratic program, for which we used Julia with the JuMP package [Dunning et al., 2017]. JuMP allows us to write the optimization problem in a solver independent way, in case we would like to try out different solvers. JuMP supports many open-source and commercial solvers for a variety of problem classes. We've chosen Gurobi, as it supports quadratic programs and is considered to be one of the most powerful mathematical solvers [Gurobi Optimization, 2021].

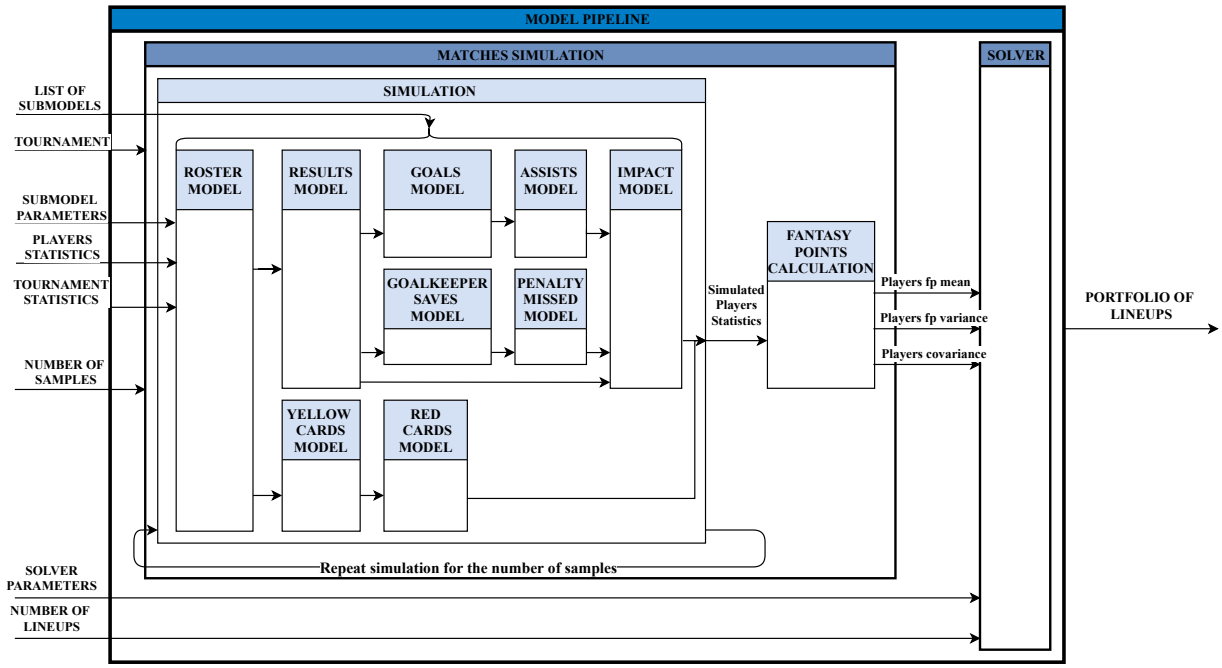


Figure 5.1: Model pipeline

5.2 Simulation

In this subsection, we focus on the matches simulation part of the Figure 5.1 Input to simulation are data described in Chapter 4 together with simulation parameters and a number of samples. The number of samples specifies how many times we sample the probability distributions of particular player statistics (goals, assists etc.).

Our simulation architecture is very modular. We model each player statistic (goal, assist, red card etc.) separately in simulation submodels. Submodels are the roster submodel, goals submodel, assist submodel etc., all depicted in the figure. These submodels are chained behind each other, or they can run in parallel. We depict these relationships with arrows that show the flow of the data. It is because some submodels require results from preceding submodels, but some rely solely on historical data. For example, the goals submodel need information about the match result (output of result model). After we know that the result was 2:1, we can assign these goals to particular players. But the yellow cards submodel does not necessarily need to know the match result to predict the number of yellow cards. Each of the submodels creates a probability distribution of the corresponding statistic during initialization. In most cases, we use Poisson distribution. This distribution is sampled in every iteration of the simulation.

We can easily replace each submodel with a different submodel as long it has the same input and output interface. Therefore, we can change only one submodel and investigate its effect on the model's output. It is also possible

to replace a model with ground truth and investigates its effects. In this way, we can find submodels whose improvement would be very beneficial for the whole model and, on the contrary, submodels with very little influence on the final output.

After we repeat the simulation for the specified number of times, we calculate players covariance matrix, means and variance. This happens in the fantasy points calculation block. These predictions are then passed into the optimization solver.

5.2.1 Goals model - Example of submodel

We created a different submodel for each player statistic. Submodels were simple prior models based on the previous weeks' statistics described in Chapter 4. We will explain the goals submodel in detail in this subsection. In the next subsection, we will briefly explain other submodels, which are very similar.

Input to the goals model is a roster prediction (11 players for each team that we predict to play), result prediction (home and away score) and players past goal performance. First, we calculate the probability that the goal was an own goal simply as a ratio between own goals so far and all goals received so far:

$$P(OwnGoal) = \frac{Own\ Goals}{Received\ goals} \tag{5.1}$$

Then (for the opponent team), we calculate a mean of scored goals, weighted by minutes played for each player. We set the means to be λ_i for the goals Poisson distribution.

$$\lambda_i = \frac{\boldsymbol{g_i^T m_i}}{\boldsymbol{1^T m_i}} \qquad i = 1, ..., N_p \tag{5.2}$$

where $\boldsymbol{g_i}$ is a vector of scored goals, $\boldsymbol{m_i}$ is a vector of minutes played and N_p is a number of players. From λ_i, we create a Poisson distribution for every player that assigns a probability of scoring goals in the interval $[0, Goals\ scored]$. $Goals\ scored$ are the output of the Result submodel. We know that the player couldn't score more goals than the score in the result.

Now we iterate in the interval $k = [1, Goals\ scored]$ and create a discrete probability distribution for each k. This distribution is made of players' probabilities of scoring a corresponding number of goals together with the likelihood of scoring an own goal. These probabilities are normalized to sum to 1. We can describe this distribution with the vector $\mathbf{p}$:

$$\mathbf{p}(k) = \frac{(P_1(k), ..., P_{N_p}(k), P(OwnGoal)))}{\sum_i^{N_p} P_i(k) + P(OwnGoal)} \tag{5.3}$$

where $P_i(k)$ is the probability of player i to score the k goal. This distribution is then sampled to get the player who scored the goal. If own goal was scored,

we assign the goal to the player with the most own goals scored so far. We update the probability $P_i(k+1)$ for the chosen player, so we reflect that this player already scored a goal. After we get all the players who scored, we update the players' fantasy points according to the FanTeam scoring rules. Each position receives a different amount of points (defender more than a forward etc.).

5.2.2 Other submodels

All other submodels are very similar to the goals submodel. We create Poisson probability distributions for the particular statistics for each player, normalize them together similarly to the 5.2 and sample them. These distributions are based on the historical means of the particular statistic weighted by the minutes played.

Roster submodel predicts 11 players that will play. We do not consider substitutes. The real lineup is usually known very well prior to the match and FanTeam replaces players who do not start with the most similar ones, therefore we choose 11 players that actually played.

Results submodel sample the Poisson distribution for home and away scores. λ_h and λ_a are inputs to the submodel from the data.

Yellow cards, red cards and penalty missed submodels are the same. We sample Bernoulli distribution with $p_i = \frac{\boldsymbol{c_i^T m_i}}{\boldsymbol{1^T m_i}}$, where $\boldsymbol{c_i}$ are received cards or penalties missed so far for the player i. In the assist submodel, we first also sample the Bernoulli distribution to know if there was an assist. Then we assign the assist to a particular player based on players assists averages. Of course, player who scored the goal can't get the assist.

In the goalkeeper saves submodel we have to take opponents' team statistics into account. We calculate average opponents shots on target and again create Poisson distribution. We sample the distribution and calculate final saves as $o_s - r$, where o_s are opponents shots on target, and r are received goals. Impact submodel creates Poisson distribution with three options {-1,0,1} (normalized with +1 to be positive) and λ, which is equal to the players' impact mean. We again sample the distribution to get the impact for every player.

5.3 Mixed-Integer quadratic program

In this section, we present our formulation of a mixed-integer quadratic program for selecting portfolio of lineups. First, we define our variables and then we formulate feasibility constraints with our objective function.

5.3.1 Variables and constants

L	Number of lineups
N_p	Number of players
N_t	Number of teams
$\mathbf{X} = (x_{ij}) \in \{0,1\}^{N_p \times L}$	Matrix indicating player i was selected into lineup j.
$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x_1} \\ \vdots \\ \boldsymbol{x_L} \end{bmatrix}$	Columns of matrix $\mathbf{X}$ ordered into one vector.
$\boldsymbol{x_i^G}, \boldsymbol{x_i^D}, \boldsymbol{x_i^M}, \boldsymbol{x_i^F}$	Subsets of $\boldsymbol{x_i}$ with only goalkeepers, defenders, midfielders or forwards included.
$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_{N_p} \end{bmatrix}$	Vector with players' fantasy points predictions.
$\boldsymbol{m} = \begin{bmatrix} \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{bmatrix}$	Mean predictions repeated L times in one vector.
$\boldsymbol{\Sigma}$	Players covariance matrix
$\boldsymbol{\Sigma}_0$ on the diagonal	Players Pearson covariance matrix with zeros
$\boldsymbol{c}$	Vector with players' prices
C	Budget limit
$\mathbf{T} = \begin{bmatrix} \boldsymbol{t_1} & \dots & \boldsymbol{t_{N_T}} \end{bmatrix} = (t_{ij}) \in \{0,1\}^{N_p \times N_t}$	Matrix indicating that player i is in the team j.
$\alpha \in \mathbb{R}, \beta \in \mathbb{R}, \gamma \in \mathbb{R}, \delta \in \mathbb{R}$	Optimization coefficients

5.3.2 Feasibility constraints

Every lineup must comply with fantasy provider rules in order to be a valid lineup. We formulate all necessary FanTeam rules as linear constraints here:
Financial Constraint:

$$\boldsymbol{c^T x_i} \leq C \qquad i = 1, ..., L \tag{5.4}$$

11 players in one lineup:

$$\mathbf{1}^T \boldsymbol{x_i} = 11 \qquad i = 1, ..., L \tag{5.5}$$

1 goalkeeper in one lineup:

$$\mathbf{1}^T \boldsymbol{x}_i^G = 1 \qquad i = 1, ..., L \tag{5.6}$$

3-5 defenders in one lineup:

$$3 \leq \mathbf{1}^T \boldsymbol{x}_i^D \leq 5 \qquad i = 1, ..., L \tag{5.7}$$

2-5 midfielders in one lineup:

$$2 \leq \mathbf{1}^T \boldsymbol{x}_i^M \leq 5 \qquad i = 1, ..., L \tag{5.8}$$

1-3 forwards in one lineup:

$$1 \leq \mathbf{1}^T \boldsymbol{x}_i^F \leq 3 \qquad i = 1, ..., L \tag{5.9}$$

Maximum 3 players from a single team in one lineup:

$$\boldsymbol{t}_j^T \boldsymbol{x}_i \leq 3 \qquad j = 1, ..., N_T, \ i = 1, ..., L \tag{5.10}$$

5.3.3 Objective function

Due to the top-heavy payout structure, our goal is to maximise the probability of having at least one extraordinary lineup which will arrange most of the profit. Therefore, we will maximise not only fantasy points but also fantasy points variance. To cover a wider pool of options, we also don't want players to overlap between lineups. We will call this a lineup correlation. In each lineup, we want to have players that have positive influence with each other. Those are, for example, players from the same team, where if one score, some other player will probably receive an assist. On the contrary, we don't want to have players with negative influence on each other in our lineup (e.g. goalkeeper and forward from the opposing teams). We will call this players covariance.

These goals of our objective function are summarised here:

1. Maximize players fantasy points
2. Minimize lineup correlation
3. Maximize players variance
4. Maximize players covariance

We've created an objective function to reflect all of these goals. The goals are building on the approach taken by [Hunter et al., 2016]. The objective function is a sum of four terms, each reflecting one of the goals. Our objective function is in the Equation 5.11

$$\max_x(\boldsymbol{x^T m} - \mathbf{1^T X^T X 1} + \sum_{k=1}^{L} \boldsymbol{x_k^T diag(\Sigma)} + \sum_{k=1}^{L} \boldsymbol{x_k^T \Sigma_0 x_k}) \tag{5.11}$$

All of our constraints are linear and all of our variables are binary. The objective function is a sum of linear and quadratic forms. We can rewrite $\mathbf{1^T X^T X 1} = \sum_{i=1}^{L}\sum_{j=1}^{L} \boldsymbol{x_i^T x_j}$ to see this more clearly. Therefore, we see that our optimization problem is a mixed-integer quadratic program.

The first term $\boldsymbol{x^T m}$ maximizes fantasy points means. The second term $-\mathbf{1^T X^T X 1}$ minimizes correlation between lineups. We multiply all lineups together, so when a player is in both lineups, the multiplication result is one and the correlation increases, zero otherwise. Elements on the diagonal of $\boldsymbol{\Sigma}$ are players variances, so the third term $\sum_{k=1}^{L} \boldsymbol{x_k^T diag(\Sigma)}$ maximizes players variances for each lineup. The last term $\sum_{k=1}^{L} \boldsymbol{x_k^T \Sigma_0 x_k}$ is a sum of quadratic forms, where each form maximizes positive correlation of players inside the lineup. With this term, we can model the positive and negative influence of both teammates and opponents.

Even though our objective function now reflects our goals, it is not suitable for optimization directly in its present form. Each of the terms has different minimum and maximum boundaries, and so the range of values vary significantly. The terms are in different units. For example, the first term maximizes the fantasy points, but the second term counts how many lineups are correlated. Some of these terms can then have a much more significant effect on the objective value, which is not desirable. Therefore, we need to scale each of the terms to control its relative impact on the objective. We will scale each term with bijective mapping to the interval: $[0, 1]$. The formula for the bijective mapping is:

$$\frac{\boldsymbol{x} - min(\boldsymbol{x})}{max(\boldsymbol{x}) - min(\boldsymbol{x})} \tag{5.12}$$

where $\boldsymbol{x}$ is the scaled term. We see that we need to know the minimum and maximum of the terms. Fortunately, we always know this information prior to

the optimization. We summarize the boundaries of the terms in the following equations:

$$11 \cdot L \cdot \min(\boldsymbol{m}) \leq \boldsymbol{x^T m} \leq 11 \cdot L \cdot \max(\boldsymbol{m}) \tag{5.13}$$

$$11 \cdot L \leq \boldsymbol{1^T X^T X 1} \leq 11 \cdot L^2 \tag{5.14}$$

$$0 \leq \sum_{k=1}^{L} \boldsymbol{x_k^T diag(\Sigma)} \leq 11 \cdot L \cdot \max(\boldsymbol{diag(\Sigma)}) \tag{5.15}$$

$$-11^2 \cdot L \leq \sum_{k=1}^{L} \boldsymbol{x_k^T \Sigma_0 x_k} \leq 11^2 \cdot L \tag{5.16}$$

We don't need to calculate minimum and maximum for 5.16, as we know the limits for Pearson correlation coefficient are $[-1, 1]$. The final normalized objective function is in the equation 5.17:

$$\max_{\boldsymbol{x}}(\alpha \frac{\boldsymbol{x^T m} - 11L\min(\boldsymbol{m})}{11L\max(\boldsymbol{m}) - 11L\min(\boldsymbol{m})} - \beta \frac{\boldsymbol{1^T X^T X 1}}{11L^2 - 11L} + \gamma \frac{\sum_{k=1}^{L} \boldsymbol{x_k^T diag(\Sigma)}}{11L\max(\boldsymbol{diag(\Sigma)})} + \delta \frac{\sum_{k=1}^{L} \boldsymbol{x_k^T \Sigma_0 x_k} + 11L^2}{2 \cdot 11^2 L}) \tag{5.17}$$

To increase or decrease the effect of each term, we added four hyperparameters: α, β, γ and δ. We will refer to the terms of the objective functions as the alpha term, beta term, gamma term and delta term.

Solution of our mixed-integer quadratic program will be L most promising lineups. Unfortunately, the number of quadratic objective terms grow rapidly with the increasing number of players and lineups. Number of quadratic terms for the beta term is $\frac{L(L+1)}{2}N_p$. If we use the big O notation: $O(L^2 N_p)$. We see that quadratic terms grow quadratically in the number of lineups times the number of players. For the gamma term, the number of quadratic terms is equal to $\frac{N_p(N_p+1)}{2}L$, in the big O notation: $O(N_p^2 L)$. Alpha and delta terms are linear.

To limit the number of gamma quadratic terms, we introduce coefficient p_c. We set p_c value in the interval $[0, 1]$, and it's purpose is to null elements of matrix Σ_0, whose absolute value is smaller than p_c. By nulling small Σ_0 values, we can reduce the number of gamma quadratic terms considerably and thus simplify the problem.

The number of players (N_p) is fixed and will not change in any of our tests. On the other hand, the number of lineups (L) is a parameter that we set to the solver. Therefore, with setting higher L, we will need to make more effort to find optimal solver parameters. We also need to find optimal values for hyperparameters α, β, γ and δ. Setting all of them to one would mean that each term has the same weight on the objective function. But it might be that some terms are much more important than others. We will discuss the impacts of changing these optimization parameters in the next chapter.

Chapter 6

Results & Discussion

In this chapter we show and discuss our model results. We've divided our modelling in two parts: **validation** and **testing**.

6.1 Validation

In the validation phase, we searched for optimal hyperparameters of our objective function. These hyperparameters are $\alpha, \beta, \gamma, \delta, L$ and p_c. As we had tournament data only for season 2018, we had to validate and test our model only on this one season. We used game weeks 20 to 28 for hyperparameters search (validation) and game weeks 29 to 38 for testing. We used 10 000 as a number of samples in the simulation part of the model. Explored grid search values are summarised in the Figure 6.1:

Hyperparameter	Possible values
α	1
β	0.5, 1, 1.5, 2, 3
γ	0.5, 1, 1.5, 2, 3
δ	0.5, 1, 1.5, 2, 3
p_c	0.1, 0.2, 0.3
L	1-40, 50

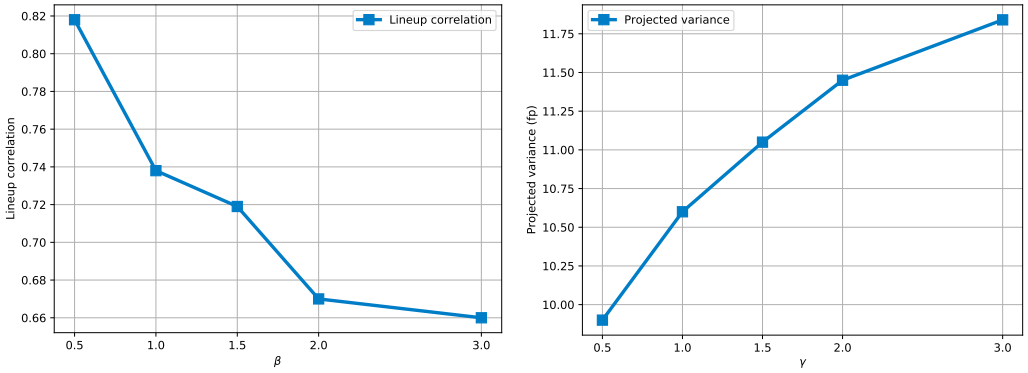
Table 6.1: Grid search values

We've fixed α to 1 and only varied other parameters. We've set the steps between our coefficient pretty wide. This is because the grid-search with tiny steps would require much more time. As an advantage, these bigger steps should prevent us from overfitting the model. For each set of parameters, our model outputted a portfolio of lineups. We've calculated fantasy points the lineup would win and compared the achieved fantasy points with actual fantasy points scored by other participants. We were able to position our lineup between real participants and find ranks and profits of our lineups.

Profits take buy-in into account, so the profit can be negative. We used multiple metrics to evaluate each portfolio from our model:

- **Portfolio profit** — Sum of all lineup profits. Most important metric, it tells us whether our model is profitable.
- **Total points** — Mean of fantasy points from all lineups. We want to increase total points. For tournaments with less participants, less points are necessary to win the tournament.
- **Lineup correlation** — Correlation of players between lineups. Zero means no correlation (No player is in multiple lineups)
- **Best rank** — Best ranked lineup in the competition. Due to the top-heavy distribution, our goal is not solely in decreasing ranks for all lineups, but rather trying to get one extraordinary lineup which will rank very well.
- **Projected variance** — Mean of projected variance from all lineups.

Each of our coefficients influences these metrics in some way. We show influence of β and γ in the Figures 6.1a and 6.1b, where we show means of the statistics over all tournaments. In the Figure 6.1a, we see that increasing β will result in decreasing lineup correlation. In the Figure 6.1b, we see that increasing γ will result in increasing projected variance. This proves that the terms in the objective function behave as expected. With parameters α, β, γ and δ, we can control our lineup formation strategy.



(a) : β vs. Lineup correlation mean

(b) : γ vs. Projected variance mean

Figure 6.1: Hyperparameters effect

6.1.1 Gurobi

Because solving the quadratic program for higher L was a more significant challenge for the Gurobi solver, we also had to find optimal configuration

parameters for the solver. The most important parameters we set were MIPFocus, which adjusts the high-level MIP solution strategy. To limit solving time, we set TimeLimit, which limits the total time expended. When the solver struggled to find an optimal solution, we slightly increased the MIPGap. MIPGap sets the relative MIP optimality gap. *"The MIP solver will terminate when the gap between the lower and upper objective bound is less than MIPGap times the absolute value of the incumbent objective value."* [Gurobi Optimization, 2021] To find appropriate hyperparameters, we used grid search. For the gurobi parameters, by default, we've used MIPFocus=1, TimeLimit=1500 and MIPGap=0.0005. In case Gurobi didn't find a solution, we increased the TimeLimit to 5000 and/or changed MIPFocus to 0 or 2. We also tried to increase the MIPGap, but maximum to 0.003.

6.2 Test

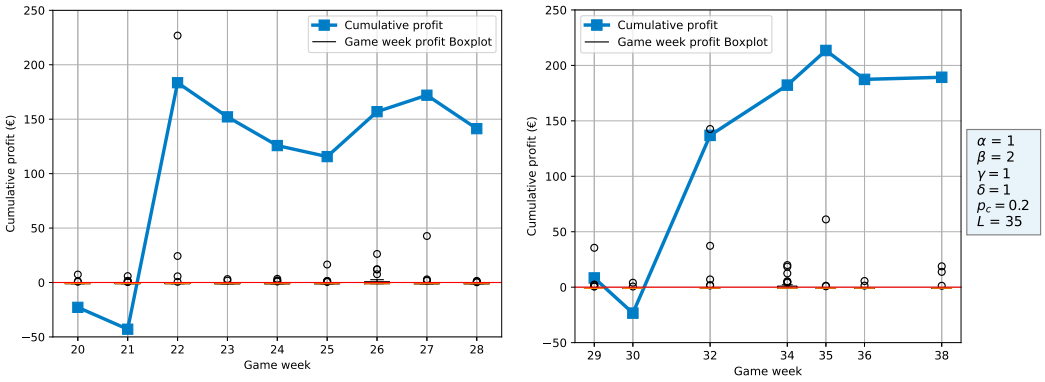
We tested the model on the game weeks 29 to 38. We chose set of hyperparameters that achieved the highest profit on the validation data for each tournament type. These hyperparameters we used for testing. Even though tournaments from the same weeks usually have the same real matches included, each tournament type has different buy-in, prize pool and consequently the number of participants. These statistics are summarised in Table 4.1. Our competition from other participants varies with the tournament type. Therefore, we need to consider each tournament type separately. The most important parameter that we need to change with the tournament type is L. For example, setting $L = 30$ for tournament type 4 wouldn't make any sense since the average number of participants for tournament type 4 is below 30. Testing hyperparameters are summarised in the Table 6.2.

Tournament type	α	β	γ	δ	p_c	L
1	1	2	1	1	0.2	35
2	1	2	1	1	0.2	35
3	1	1	1.5	1.5	0.1	50
4	1	1	0.5	1	0.1	3

Table 6.2: Testing hyperparameters

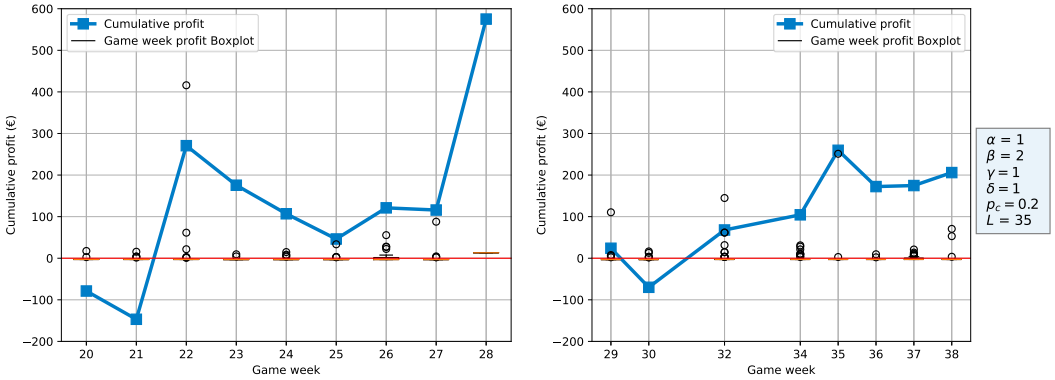
In the following Figures 6.2 to 6.5, we show the performance of our model for validation and testing data. Validation plots are on the left and testing on the right. We plot accumulated profit over time. We start with zero profit, and every week we add or deduct the profit achieved in that game week. In addition to the accumulated profit, we plot all portfolio profits as a boxplot for each week. We can see the average profit and, more importantly, whether there were some significant outliers. No tournaments were held during the game weeks 31 and 33. Tournament type 4 wasn't played in the game weeks 25 and 26. We're missing game weeks 29 and 30 for tournament

type 3, as the solver wasn't able to find the optimal solution within our limits (MIPGap=0.003 and TimeLimit=5000). The high number of lineups probably caused the difficulty.



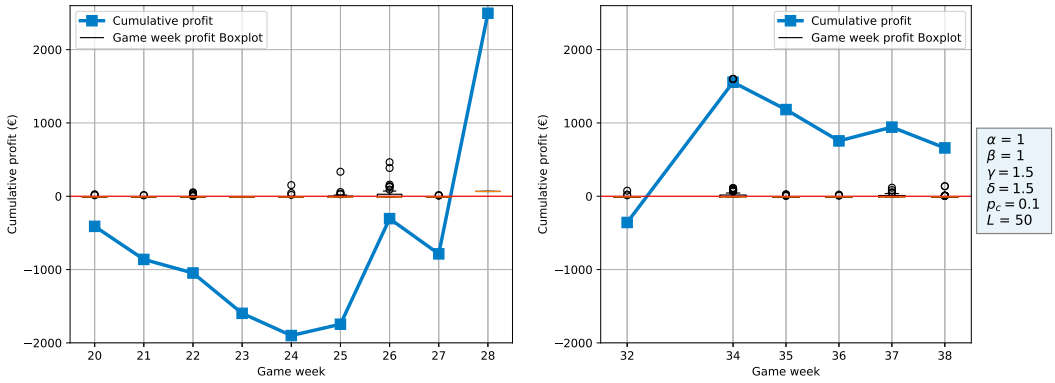
(a) : Validation **(b) :** Test

Figure 6.2: Cumulative profit for Tournament type 1



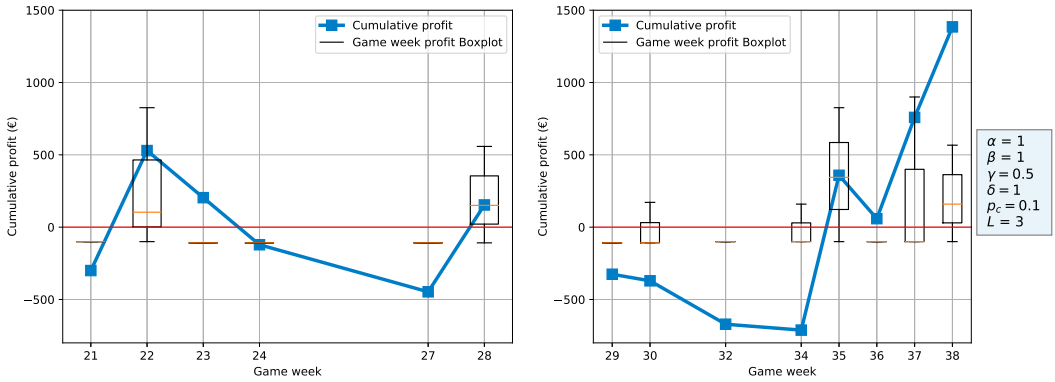
(a) : Validation **(b) :** Test

Figure 6.3: Cumulative profit for Tournament type 2



(a) : Validation **(b) :** Test

Figure 6.4: Cumulative profit for Tournament type 3



(a) : Validation **(b) :** Test

Figure 6.5: Cumulative profit for Tournament type 4

The most important result is that all tests ended in profit. As the tournaments are independent, we'd be able to bet in all of the tournaments without interfering with each other. If we'd used our model in the 2018 FanTeam tournaments, we'd achieve an accumulated profit of €2439.14 from the last eight tournaments of the season.

From the boxplots, we can see that the weekly profit average is always very close to zero. We can also see that a few outliers very often secured the profit. This is in line with our strategy to maximize the probability of having at least few exceptional lineups in the portfolio that will secure most of the profit.

We achieved the best results in all of our simulations with most of the objective function hyperparameters close to 1. This signifies that we did our normalization correctly, and all of the terms have a similar significance. For tournament type 1 and 2, the same hyperparameters proved to be the best on the validation data. It's not to a big surprise, as tournament types 1 and

2 are very similar, and therefore the same configuration should work for both types. For these tournament types, we used 35 as the number of lineups and 2 as β, which puts more weight on minimizing the lineup correlation.

For tournament type 3, we used the largest number of lineups (50). With the higher number of lineups, higher variance ($\gamma = 1.5$) and maximizing players correlation ($\delta = 1.5$) proved beneficial.

It is a small surprise that the highest profit (€1384.74) was achieved in tournament type 4. We expected that the approach of maximizing the variance and minimizing the lineup correlation would come into effect more significantly with more lineups in other tournament types. On the other hand, as there are fewer participants, the competition is smaller and fewer fantasy points are needed to win. Also, with lower γ, the model focused less on the variance and produced more moderate but stable results.

Chapter 7

Conclusion

We can conclude that our approach proved successful. If used in practice, our model would accumulate a profit of €2439.14 in the last eight weeks of the 2018 Premier League season. Our quadratic program is suitable for creating a portfolio of lineups for tournaments with top-heavy payout distribution. Thanks to the sampling of our probabilistic models, we obtain players fantasy points distributions that allow us to model the relationships between players. We maximize the covariance directly in the objective function together with players expected means and variance. We also minimize the lineup correlation in the objective function. We see our approach as a considerable improvement to previous integer linear programs. With comparison to most of previous work, our approach contains the full pipeline with both creating the fantasy points predictions and the subsequent optimization.

As we used quite basic prior models to model the fantasy players' statistics, the model can be improved with more complex probabilistic models. This is very easy, thanks to our modular architecture, where each fantasy rule can be modelled separately. Advantage of our generative model is that it doesn't place any assumptions on the distribution of the player's fantasy points. It could be very interesting to include data about other participants historical lineups in the model and use them to model the opponent's behaviour. Another possible improvement would be to explore more settings for the objective function hyperparameters and investigate their effect on the tournament rankings and profit.

Because we don't need to specify any heuristic rules for our quadratic program, our model is not domain-specific and can be used in other sports or in entirely different domains. A notable example is stock portfolio optimization.

Acknowledgment. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures and also. The access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics" is also gratefully acknowledged.

Bibliography

[Becker and Sun, 2016] Becker, A. and Sun, X. A. (2016). An analytical approach for fantasy football draft and lineup management. *Journal of Quantitative Analysis in Sports*, 12(1):17–30.

[Bonomo et al., 2014] Bonomo, F., Durán, G., and Marenco, J. (2014). Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game. *International Transactions in Operational Research*, 21(3):399–414.

[Chase, 2018] Chase, R. (2018). Types of games in daily fantasy sports. `https://www.dailyfantasycafe.com/academy/undergraduate/types-of-games`.

[DraftKings, 2018] DraftKings (2018). Soccer scoring rules. `https://www.draftkings.com/help/rules/soc`.

[Dunning et al., 2017] Dunning, I., Huchette, J., and Lubin, M. (2017). Jump: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320.

[Ebner, 2013] Ebner, S. (2013). History and time are key to power of football, says premier league chief. `https://www.thetimes.co.uk/article/history-and-time-are-key-to-power-of-football-says-premier-league-chief-3d3zf5kb35m`.

[FanDuel, 2018] FanDuel (2018). Rules and scoring. `https://www.fanduel.com/rules`.

[FantasyPros, 2018] FantasyPros (2018). `https://www.fantasypros.com/`.

[FanTeam, 2018] FanTeam (2018). Fantasy football rules. `https://www.fanteam.com/support-center/football-rules`.

[FanTeam, 2020] FanTeam (2020). Fanteam. `https://www.fanteam.com/`.

[Gurobi Optimization, 2021] Gurobi Optimization, L. (2021). Gurobi optimizer reference manual.

[Haugh and Singal, 2018] Haugh, M. B. and Singal, R. (2018). How to play strategically in fantasy sports (and win). In *Conference proceedings*.

[Hunter et al., 2017] Hunter, D. S., Saini, A., and Zaman, T. (2017). Picking winners: A data driven approach to evaluating the quality of startup companies. *arXiv preprint arXiv:1706.04229*.

[Hunter et al., 2016] Hunter, D. S., Vielma, J. P., and Zaman, T. (2016). Picking winners using integer programming. *arXiv preprint arXiv:1604.01455*.

[John Dubber, 2015] John Dubber, J. W. (2015). Playing the game: The soft power of sport. `https://www.britishcouncil.org/research-policy-insight/insight-articles/playing-game-soft-power-sport`.

[LegalSportsReport, 2020] LegalSportsReport (2020). Daily fantasy sports sites for real money - where to play dfs in 2020. `https://www.legalsportsreport.com/daily-fantasy-sports/`.

[Maher, 1982] Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118.

[Meehan, 2015] Meehan, J. C. (2015). The predominate goliath: Why pay-to-play daily fantasy sports are games of skill under the dominant factor test. *Marq. Sports L. Rev.*, 26:5.

[Nerd, 2015] Nerd, D. F. (2015). `https://dailyfantasynerd.com/`.

[Newell, 2017] Newell, S. (2017). *Optimizing daily fantasy sports contests through stochastic integer programming*. PhD thesis, Kansas State University.

[Robins, 2018] Robins, J. (2018). The analytics behind winning at daily fantasy sports - mit sloan analytics conference. `http://www.sloansportsconference.com/content/the-analytics-behind-winning-at-daily-fantasy-sports/`.

[Rotogrinders, 2015] Rotogrinders (2015). `https://rotogrinders.com/`.

[RotoViz, 2018] RotoViz (2018). `https://www.rotoviz.com/`.

[Tepper, 2017] Tepper, F. (2017). Draftkings and fanduel will no longer merge. `https://techcrunch.com/2017/07/13/draftkings-and-fanduel-will-no-longer-merge/`.