Drafting a Fantasy Football Team

Machine Learning Project (CS/DS 864)

Rishabh Manoj IMT2013035
Nigel Fernandez IMT2013027
Anirudh Ravi IMT2013005

IIIT-Bangalore
December 22, 2016

Contents

List of Tables

List of Figures

Abstract

This report details our attempt to apply machine learning models to draft the optimal fantasy football team in the English Premier League. Three models namely Support Vector Machines,Random Forests and Naive Bayes were used to predict the points scored by a player. A constraint satisfaction problem was formulated with required fantasy football rules to draft the team. An advantage of 28.5 *points or* 61.59% *was observed over teams drafted by the average user. By drafting a large volume of teams, our probability of winning will increase. Further improvement (assumed) can be made to our algorithm by incorporating career historical records and including a riskiness measure.*

1 Introduction

Fantasy sports have grown exponentially in popularity in the last 10 years with as estimated market size of $70 billion per year [1] for the National Football League in the United States. We have chosen the fantasy sport of the English Premier League (EPL) referred to as the Fantasy Premier League (FPL) [2] as our sport of choice. Our algorithm can be used for other fantasy sports due to their similar structure and objectives.

The FPL attracts over 4 million competitors or managers. The structure of the FPL includes 380 fixtures or football matches split across 38 textsl-gameweeks. Each gameweek consists of 10 fixtures involving each of the 20 EPL teams once. Each football player is appraised with a price to include in a manager's team and is assigned a playing position. Before each gameweek, a manager drafts a fantasy team by selecting 15 players satisfying certain constraints including cost [3]. A manager earns points depending on the real match performance of the players during the gameweek included in his fantasy team and aims to maximize the points earned. In standard fantasy leagues, a manager wins an amount of money proportional to the percentile standing of the points earned by him. In FPL, there is officially no monetary rewards but managers are rewarded with equivalent prizes.

We chose to work on the fantasy team problem due to the inherent predictive nature of the problem aptly suited for machine learning models, the availability of large amounts of data and the novelty of the problem including the fun aspect.

2 Problem Framework

As stated above, before each gameweek, a manager drafts a fantasy team by selecting 15 players from over 600 available. Each player can cost on average between €2 to €12 million euros. The total cost or budget of the team should not exceed €100 million. In addition, the team must consist of exactly 2 goalkeepers, 5 defenders, 5 midfielders and 3 forwards. There cannot be more than 3 players selected from the same football club. A manager earns points depending on the performance of his included players where actions like scoring a goal, performing a save and other events are associated with point earnings. Table 1 provides a partial example of 7 events and their associated point earnings.

Event	**Points**
For playing 60 minutes or more	2
For each goal scored by a forward	4
For each goal assist	3
For a clean sheet by a midfielder	1
For every 3 shot saves by a goalkeeper	1
For each penalty save	5
For each red card	-3

Table 1: Scoring System of FPL

We split our problem into two parts namely (1) predicting the points scored by a player in his next fixture and (2) drafting an optimum fantasy team which maximizes expected points. The first part of our problem was solved using machine learning models namely Support Vector Machines, Random Forests and Naive Bayes which were trained on historical data to predict the points expected to be scored by a player. The second part of our problem was modelled as a linear optimization problem. The constraints of the LP included the budget and position limits. The maximization function was the expected points earned by the selected team. We did not include the third constraint of a limit of 3 players selected from the same football club. We believed this constraint would be violated with very low probability [4].

3 Choosing our Feature Set

3.1 Data Collection and Challenges

Data collection proved to be quite a challenge. Publicly available data was either stale or did not have the relevant features we were looking for. FPL's statistics repository [5] has current data updated after every gameweek with over 70 relevant attributes of performance and match day statistics. Since they do not provide an official API, we extensively explored their website for API endpoints and used Python's `urllib` library to scrape the data. The legal guidelines are a grey area but point towards permitting scraping of data for academic or personal use.

We obtained three broad tables of data related to (1) player statistics, (2) team statistics and (3) gameweek fantasy play statistics. The player statistics table consisted of 616 players with over 60 attributes quantifying their performance in the last 18 gameweeks of the $2016-17$ season. We did not include career historical records from previous seasons which are available from the $2007-08$ season due to data collection issues. This could be a possible improvement in the next iteration of our project. The team statistics table consisted of the rating across 7 attributes of the 20 EPL teams based on their performance. The gameweek fantasy play statistics table provided us with the average and highest points earned by managers in various gameweeks. These attributes will help us measure the accuracy of our algorithm.

We also anticipated to face challenges in our algorithm's accuracy due to the unpredictability of real life football match outcomes and player performances. A case in point in when 14th place Leicester City in the $2014-15$ premier league season scripted a historical upset akin to a fairytale story to win the premier league season of $2015-16$.

3.2 Feature Selection

Choosing the right feature set for your machine learning model is arguably the most important step. Through a mix of intuition, literature review and informativeness measures done through descriptive analysis experiments we narrowed down on our feature set. We decided to include features grouped in three categories namely (1) `current player performance` (2) `player form` and (3) `fixture difficulty rating (FDR)`. A partial overview of our

feature set in this perspective is provided in table 2. Note that player form has the same attributes as current player performance except the measurements are running averages over the past 5 games. The number of features in our feature set were above 60.

Player Form	**FDR**
minutes	strength overall home
yellow cards	strength overall away
red cards	strength attack home
clean sheets	strength attack away
own goals	strength defence away
goals scored	strength defence away

Table 2: Subset of Feature Set: Perspective 1

Current player performance contains attributes measuring the performance of a player in the past fixture only. Player form is a collection of attributes which incorporates the *streakiness* of performance which is the observation that a good run of performances is often continued over to future matches. The natural question therefore is the definition of the length of the streak, i.e., the choice of the number of games over which to average the attribute measures to calculate the running player form. After experimenting with different choices of number of matches over which to calculate a player's form or streak, we found that 5 matches is an optimal number to improve the accuracy of our machine learning models applied later. This fact is visualized in figure [link].

The choice of including fixture difficulty rating in our feature set was derived from the results of an experiment in our descriptive analysis in which we found out that teams and players perform better when they play at their home stadium over playing at away location. Home advantage is an important factor and is visualized in figure [link]. An average player has a better chance of scoring points when playing at home versus playing away. The collection of attributes included in the FDR measured performance statistics of a team's capability in attack and defence in home and away locations.

The second aspect of our feature selection was grouping players depending on the player type or position namely (1) goalkeeper, (2) defender, (3) midfielder or (4) forward and then choosing relevant attributes independently in each group for each of the three feature categories namely (1) current

player performance (2) player form and (3) FDR [6]. Goalkeeper and defenders had a nearly similar feature set while midfielders and forwards had a nearly similar feature set. A partial overview of our feature set in this perspective is provided in table 3. This was done as different player types score points based on different performance events. A partial snapshot of the scoring system in the FPL was earlier presented in table 1. Therefore the attribute `penalties saved` is relevant to the performance of a goalkeeper and the attribute `goal assists` is relevant to the performance of a midfielder or forward. There were some attributes which were common across playing positions like `yellow cards`.

Goalkeepers / Defenders	**Midfielders / Forwards**	**Common**
saves	big chances missed	minutes
clean sheets	offsides	yellow cards
goals conceded	assists	red cards
penalties saved	completed passes	own goals
tackles	target missed	goals scored

Table 3: Subset of Feature Set: Perspective 2

We also tried to incorporate *riskiness* in a limited way in our feature set. In our literature review [4], we found that riskiness is a relevant feature set decider. Riskiness is defined as the ratio of a players' standard deviation to their mean of points scored. This leads to the formation of a *unique* team differentiated from the average team with a small probability of a risky player scoring a high amount of points and placing the fantasy team in the top percentile of the the competition. The relation to a player's risk to the points scored by him is visualized in figure 1. To include riskiness and obtain less frequently picked players from our model we left out the `purchase price` attribute from our feature set. We found a correlation between the purchase price and points scored by a player with costlier players having scored higher points in the past. This is visualized in figure 2. By leaving out the purchase price we intended to avoid biasing our model to pick popular high scoring players and instead predict lesser known players to score high points. Further improvement and quantification of riskiness can be done in the future iteration of our algorithm.

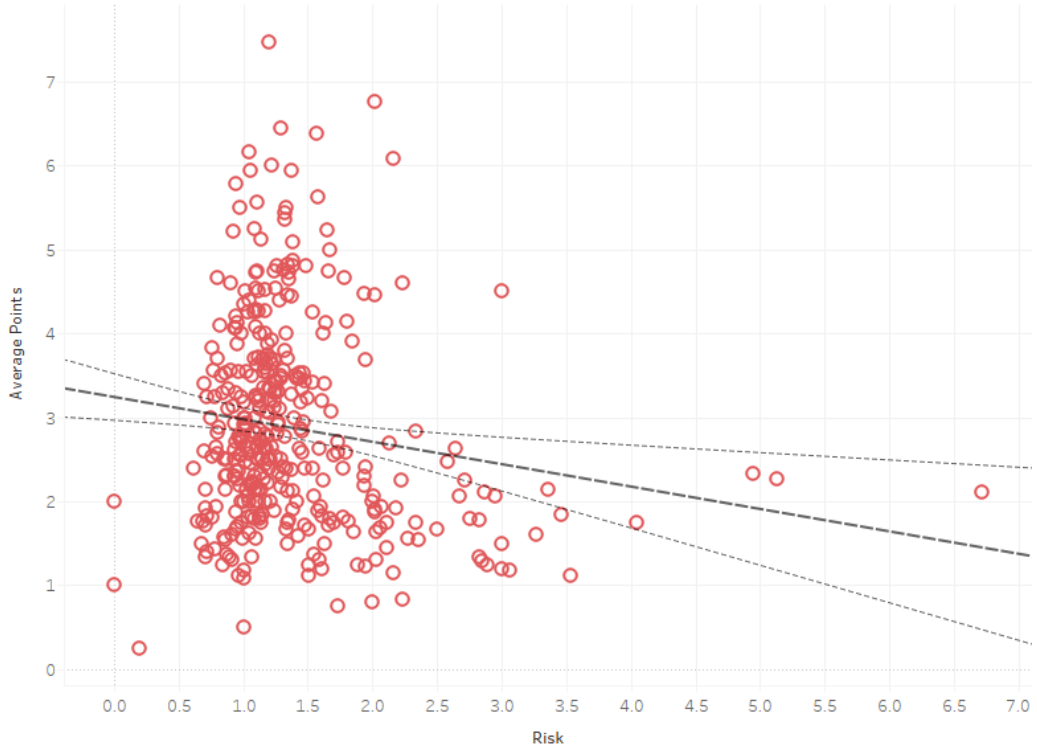

Figure 1: Points Scored vs Cost of Players

4 Score Prediction using Machine Learning

4.1 Support Vector Machines

After deciding on our feature set, we performed a z score standardization of the attribute measures to ensure that no one attribute biases our model. We chose a training to test data percentage of 70% - 30%. Once we had set up our feature space, we wrote an R script [7] to iteratively run an SVM model for each player. The output of the model was the expected points for a player. The accuracy of the SVM model is visualized in figure 3.

4.2 Naive Bayes

Similar to our SVM model, we wrote an R script to use Naive Bayes as our model. The accuracy of the Naive Bayes model is visualized in figure 4.

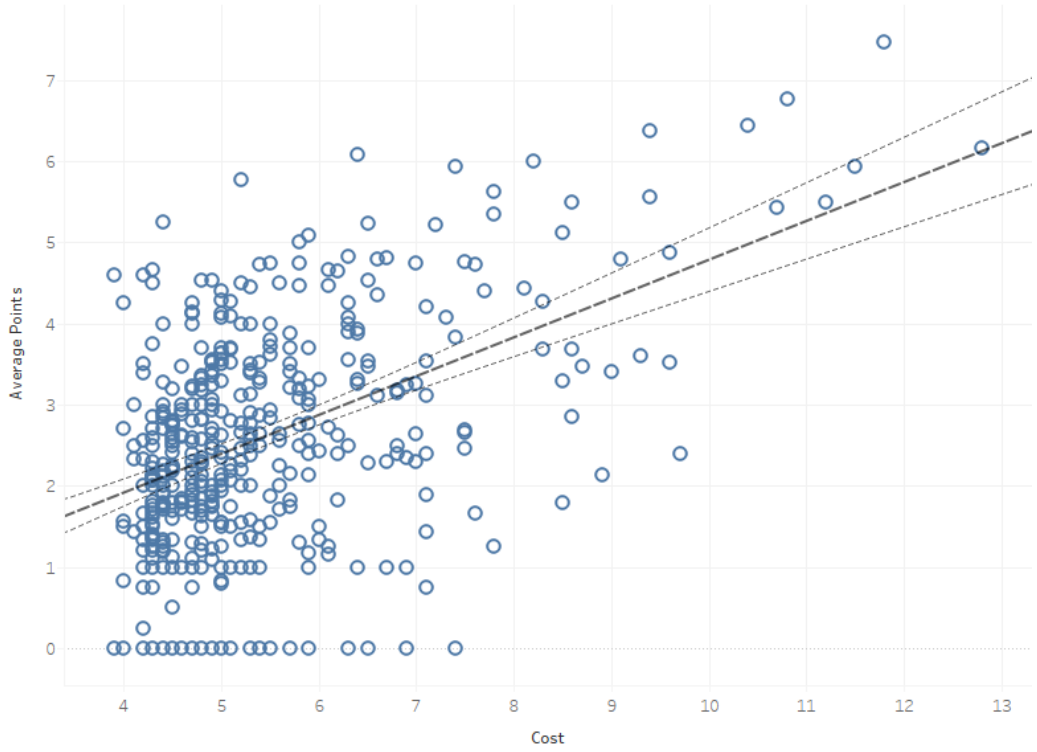

Figure 2: Points Scored vs Cost of Players

4.3 Random Forests

We also used an ensemble model in the form of random forests to prevent over fitting. The accuracy of the random forests model is visualized in figure 5.

5 Drafting our Fantasy Team

5.1 A Reasonability Check

We ran our machine learning model to predict the expected points scored for each player and cross checked whether our results were reasonable. Since the players predicted to score highly are likely to get drafted in the team, we concentrated on them. We found that among the players predicted to score

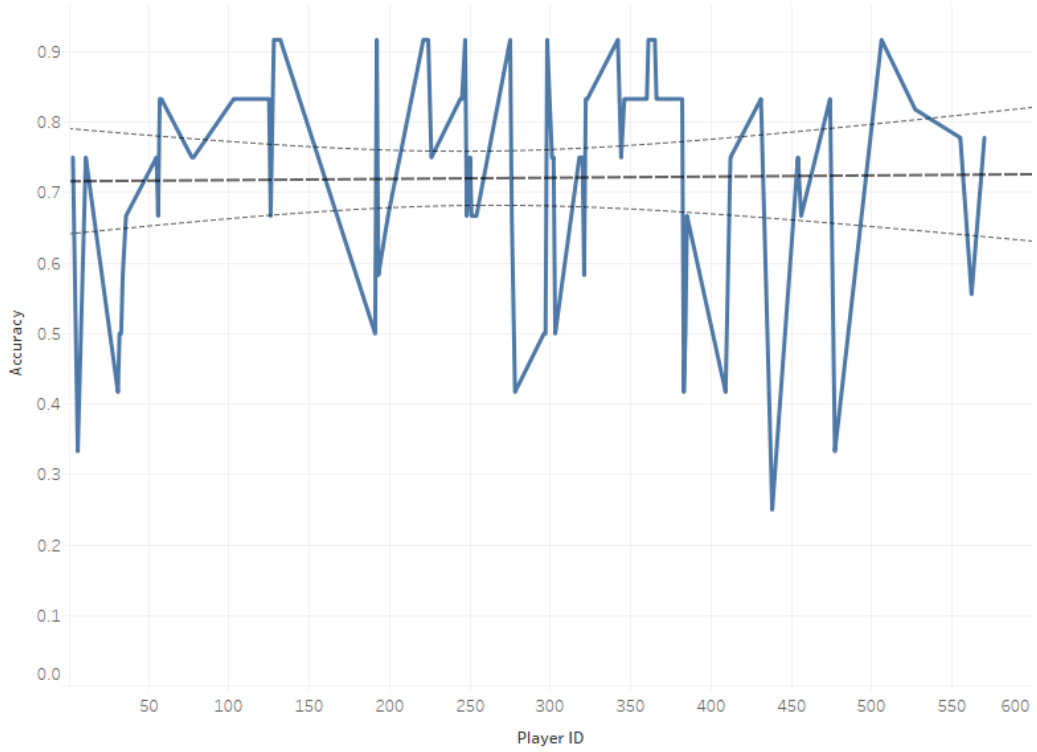

Figure 3: Points Scored vs Cost of Players

a high amount of points, 80% of players appear in the ranking charts of the FPL thereby assuring us of our model's accuracy.

For example Thibaut Courtois from Chelsea, César Azpilicueta from Chelsea, Alexis Sanchez from Arsenal and Diego Costa from Chelsea are ranked as top players by FPL based on previous performances and FPL's own analytics in the goalkeeper, defender, midfielder and forwards category respectively. All these players are predicted to score highly by our model. In addition our model also predicted less frequently chosen players who do not appear on the FPL charts as predicted to score highly. An example being Nathan Ake from Bournemouth.

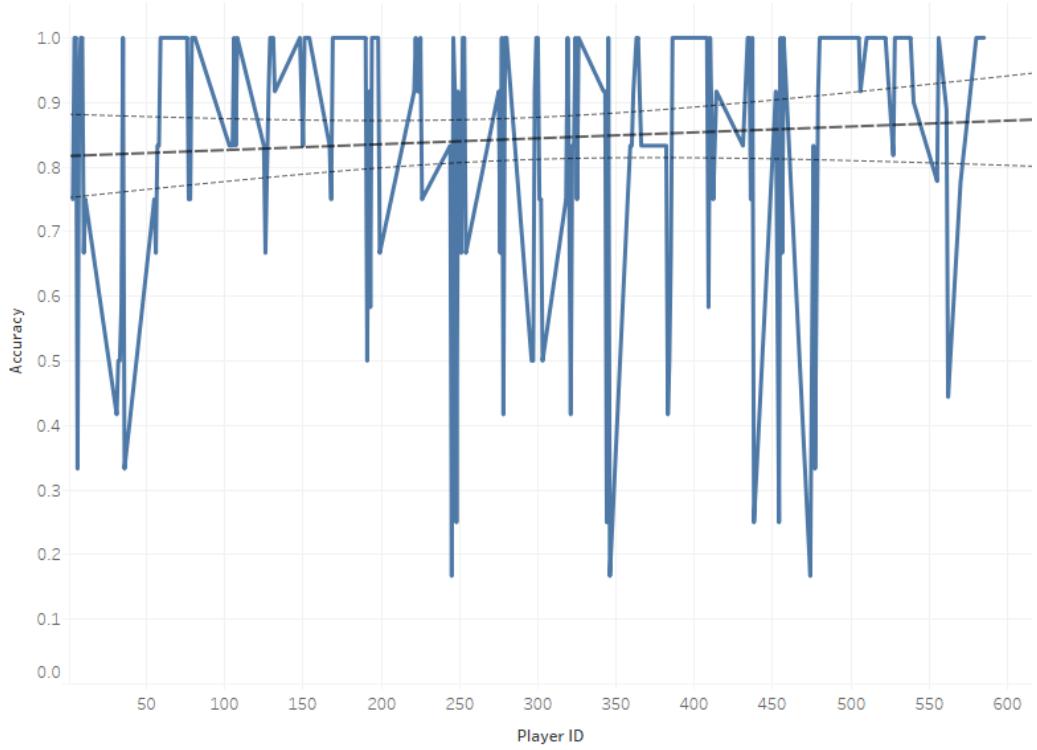

Figure 4: Points Scored vs Cost of Players

5.2 Linear Optimization Problem

Our next step was to take the list of players, their associated cost and predicted points to draft our fantasy football team. We needed to meet 2 main constraints namely (1) the total cost or budget of the team should not exceed €100 million and (2) the team must consist of exactly 2 goalkeepers, 5 defenders, 5 midfielders and 3 forwards. Notice that these constraints are linear in nature. Since we're trying to maximize the points scored by our team, we formulated our problem as a linear optimization problem [8]. The optimization problem is shown in figure [link].

In figure [link], the first line states our objective function which is to maximize the points scored by the team. Each player is given a variable name corresponding to his `player ID` attribute. Constraint 1 is our budget constraint stating that the team picked must cost less than €100 million. Constraint 2 states that the number of goalkeepers picked in the fantasy

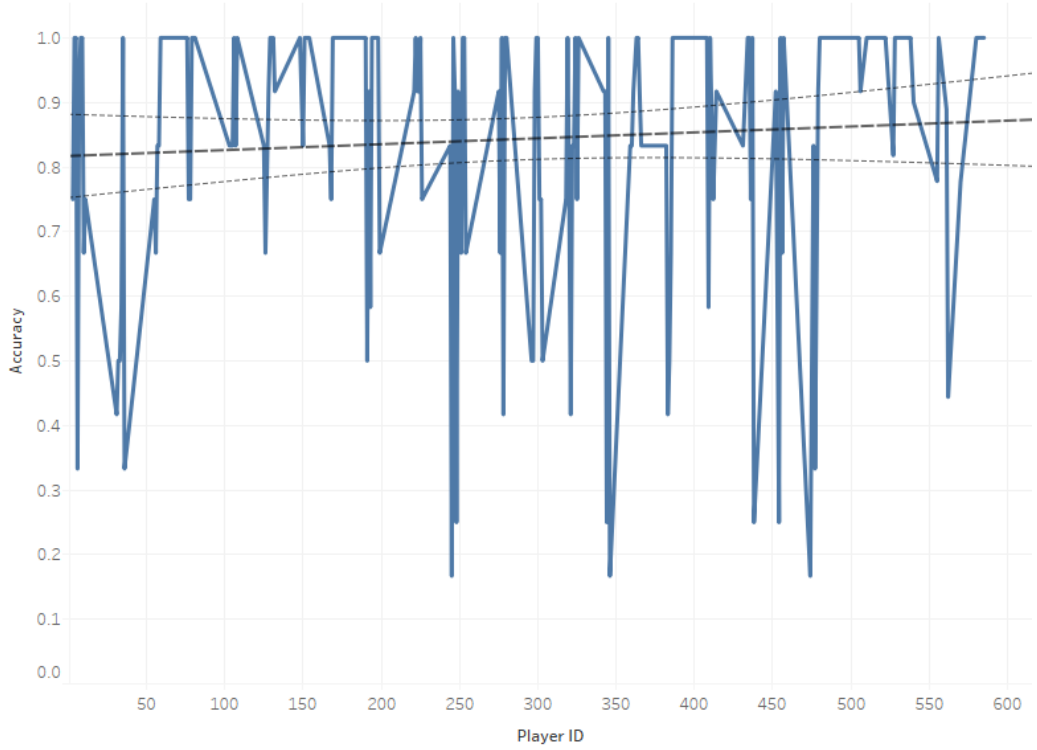

Figure 5: Points Scored vs Cost of Players

team must be exactly 2. Similarly constraints 3 to 5 limit the number of defenders, midfielders and forwards picked. Statement 6 indicated that all variables are binary in nature.

To solve our optimization problem, we used the open source `lp_solve` program [9]. lp_solve is a Mixed Integer Linear Programming (MILP) solver. We wrote a Python script to take in as input the predicted points for each player by our machine learning model, individual player files and output a constraint file. This constraint file was passed to lp_solve which provided a list of players and a binary indicator of 1 or 0 depending if the player was picked in our fantasy team or not.

We used the output of our SVM model to predict player points since it had the highest accuracy among the models. The resultant fantasy team drafted is shown in table 4 and figure 6. The substitutes drafted are shown in figure 7. This team was rated optimal for gameweek 18 of the EPL starting

on December 26, 2016. Among the 15 players picked, 13 players are ranked as top picks by FPL providing some kind of verifiability of the accuracy of our algorithm. We feel, the inclusion of the remaining 2 players who do not appear on the FPL charts is a result of our attempt to add some *riskiness* into our team. The 2 players are defender Nathan Ake from Bournemouth and defender Erik Pieters from Stoke City. This would add some uniqueness to our fantasy team differentiating it from the average manager's team. We assume the average manager would pick based on the rankings by the FPL. The average manager scores 46.411 points per gameweek while the average highest points per gameweek is 121.235 points. All 15 players picked are predicted to score level `H` points translating to a worst case of atleast 5 points per player. This would result in a total of 75 points which is 61.59% higher than an average manager and 61.718% lower than the gameweek's top manager.

Name	**Team**	**Cost**	**Predicted Points**
Forwards:			
Wayne Rooney	Man Utd	8.6	High: 5+
Jermain Defoe	Sunderland	7.8	High: 5+
Christian Benteke	Crystal Palace	7.7	High: 5+
Midfielders:			
Michail Antonio	West Ham	6.7	High: 5+
Wilfried Zaha	Crystal Palace	5.8	High: 5+
Matt Phillips	West Brom	5.5	High: 5+
Joe Allen	Stoke	5.2	High: 5+
James McArthur	Crystal Palace	5	High: 5+
Defenders:			
Ben Gibson	Middlesbrough	4.8	High: 5+
Simon Francis	Bournemouth	4.6	High: 5+
Antonio Barragan	Middlesbrough	4.6	High: 5+
Erik Pieters	Stoke	4.5	High: 5+
Nathan Ake	Bournemouth	4.4	High: 5+
Goalkeepers:			
Fraser Forster	Southampton	5.1	High: 5+
Darren Randolph	West Ham	4.4	High: 5+

Table 4: Predicted Fantasy Team for Gameweek 18

Figure 6: Predicted Fantasy Team for Gameweek 18

6 Conclusion and Future Work

As mentioned earlier, we could not add career historical records from previous seasons due to data collection issues. The addition of these attributes to our feature set would likely improve the accuracy of our model by ensuring historically strong players are not sidelined due to recent performance form dips. We can also take into account injury risk of players. In addition, there is a requirement to further quantify and use the novel *riskiness* factor in our model. By drafting a large number of resulting unique teams, the probability of one of them scoring high increases. We can also use the available data of

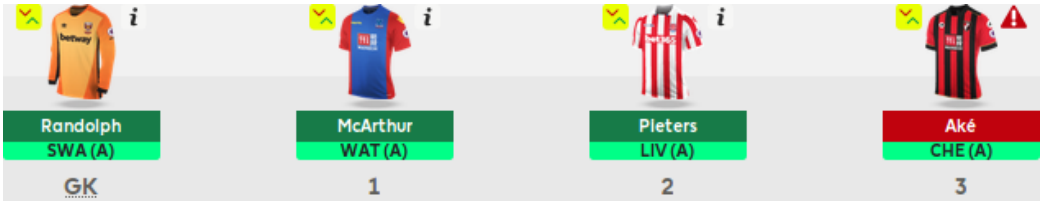

Figure 7: Predicted Substitutes for Gameweek 18

the percentage of managers who selected a player in their fantasy team to find an optimal balance between historical reliability and uniqueness.

Bibliography

[1] G. Brian. *The $70 Billion Fantasy Football Market*. URL: http://www.forbes.com/sites/briangoff/2013/08/20/the-70-billion-fantasy-football-market/#13b8ff1341b7.

[2] *Fantasy Premier League*. URL: https://fantasy.premierleague.com/.

[3] M. Tim, R. Sarvapali, and C. Georgios. "Competing with Humans at Fantasy Football: Team Formation in Large Partially-Observable Domains". In: *Association for the Advancement of Artificial Intelligence* (2012).

[4] H. Eric and N. Adebia. "Machine Learning Applications in Fantasy Basketball". In: *Stanford University* (2015).

[5] *Fantasy Premier League Statistics*. URL: https://www.premierleague.com/stats.

[6] S. Glenn and S. Travis. "Predicting Optimal Game Day Fantasy Football Teams". In: *Stanford University* (2015).

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. URL: https://www.R-project.org.

[8] M. Bill. *Choosing a Fantasy Football Team*. URL: https://llimllib.github.io/fantasypl/.

[9] B. Michel, E. Kjell, and N. Peter. *lp_solve: Open source (Mixed-Integer) Linear Programming system.* URL: http://lpsolve.sourceforge.net/.