Implementation of Machine Learning Techniques to Predict Player Performance using Underlying Statistics

MSc Research Project
Programme Name

Murtaza Saifi
Student ID: X18129463

School of Computing
National College of Ireland

Supervisor: Dr. Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Murtaza Saifi
Student ID:	X18129463
Programme:	Programme Name
Year:	2018
Module:	MSc Research Project
Supervisor:	Dr. Vladimir Milosavljevic
Submission Due Date:	20/12/2018
Project Title:	Implementation of Machine Learning Techniques to Predict Player Performance using Underlying Statistics
Word Count:	6939
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	25th January 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Implementation of Machine Learning Techniques to Predict Player Performance using Underlying Statistics

Murtaza Saifi
X18129463

Abstract

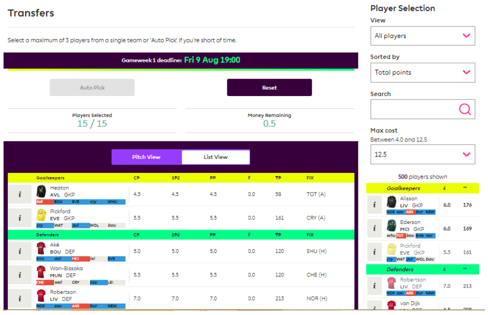
Fantasy sports have become a growing industry that is earning major revenue for the sports and media business. Player performances are being analysed through countless of data points to determine high quality players or match outcomes. This study focuses on analysing player performance in the English Premier League on the basis of the statistics present in their Fantasy tournament (Fantasy Premier League) with the addition of underlying statistics such as Expected Goals and Expected Assists and aims to determine how strong an attribute can underlying statistics be while predicting the player performance. Ensemble based classification modelling such as Random Forest and Extreme Gradient Boosting have been used with and without the presence of underlying statistics and a 0.19% increase in accuracy 0.2% increase in F1 score is observed with the presence of these data points. The models are validated using k-fold cross validation and a comparison analysis was conducted between 4 sampling techniques and evaluated with Accuracy, Specificity, sensitivity, Precision, recall and F1 score.

1 Introduction

The online sports industry has proved to become a successful avenue for sports consumers with a revenue generation of over 4 billion annually and the primary factor of participation in these events being the fact that participation stimulates players to keep themselves updated with latest information on teams and individual players which eventually leads to further consumption of live games and merchandise Lee et al. (2013). The thrill to beat friends and other participants, with the application of sports knowledge by keeping themselves constantly updated, also proves to be a strong source of motivation of involvement in such activities Kim and Ross (2016). This eventually led to the introduction of Fantasy Sports, primarily in the American continent, with options for baseball, basketball and American football and observing a total participation of roughly 30 million users in all sports in the last decade Drayer et al. (2010). Although the concept was introduced in the early 80s, it was limited to a local level as it worked on the newspaper and statistics back then were still being manually crunched. The Fantasy Sport industry was given a new breath of life in the era of internet towards the late 90s which allowed easier participation and access to statistical analysis leading it to venture into an international competition Farquhar and Meeds (2007). The notion of such a gaming event reached the European continent towards the early parts of the 21st century where each of the top European soccer nations introduced their Fantasy Leagues for their top-tier domestic leagues. One of these is the Fantasy Premier League (FPL) which is the fantasy tournament for the English Premier League (EPL) that boasts a participation of more than 7 million people across the globe. Such a high volume of participation has led to rapid commercialization of the tournament with the increase in written sports blogs and YouTube pages discussing various tips on improving FPL teams sharing insights in the form of statistics. Subscription to such individuals or their organizations has become a major revenue boost in these industries. With regards to sports, most algorithms have been focused on win-loss predictions or overall team selections for the entire tournament based on the match statistics. This research aims at creating a classification model for determining high player performance based on the fantasy statistics with the focus on the FPL league.

1.1 Background on the Competition

FPL is a yearlong tournament the begins in the month of August and goes on till May next year and works in parallel with its parent league (EPL). Any individual is allowed to participate as long as they have an email address. After registration with `https://fantasy.premierleague.com/`, a participant (known as "manager") is provided with 100 million as a budget and has to make a selection of 15 players in his team from the list of actual players who perform in the EPL. The value of the football players has no relation to their actual market and is decided based on past performances by the tournament organizers. A manager needs to ensure that the selected team of 15 players fulfils the minimum requirement of 2 goalkeepers, 5 defenders, 5 midfielders and 3 strikers with only a maximum of 3 players belonging to the same team. The manager is allowed to make multiple changes before the first week of the tournament to decide the "Starting 11" of their team while the remaining 4 stay on the bench. This "Starting XI" can be made up of multiple formations as long as it consists of 1 goalkeeper, 3 defenders, 3 midfielders and 1 striker. Figure 1 shows an example of player selection in the transfer section of the site.



Select a maximum of 3 players from a single team or 'Auto Pick' if you're short of time.

Gameweek 1 deadline: Fri 9 Aug 19:00

Players Selected 15 / 15

Money Remaining 0.5

Goalkeepers	CP	SP2	PP	F	TP	FIX
Heaton AVL GKP tot BOU EVE cry WHU	4.5	4.5	4.5	0.0	58	TOT (A)
Pickford EVE GKP cry WAT avl WOL bou	5.5	5.5	5.5	0.0	161	CRY (A)
Defenders	**CP**	**SP2**	**PP**	**F**	**TP**	**FIX**
Aké BOU DEF SHU avl MCI lei EVE	5.0	5.0	5.0	0.0	120	SHU (H)
Wan-Bissaka MUN DEF CHE wol CRY sou LEI	5.5	5.5	5.5	0.0	120	CHE (H)
Robertson LIV DEF NOR sou ARS bur NEW	7.0	7.0	7.0	0.0	213	NOR (H)

View: All players

Sorted by: Total points

Max cost
Between 4.0 and 12.5
12.5

500 players shown

Goalkeepers	£	**
Alisson LIV GKP NOR sou ARS bur NEW	6.0	176
Ederson MCI GKP whu TOT bou BHA nor	6.0	169
Pickford EVE GKP cry WAT avl WOL bou	5.5	161
Defenders	**£**	**
Robertson LIV DEF NOR sou ARS bur NEW	7.0	213
van Dijk		

Figure 1: Screenshot of the Transfers section for selecting the Playing XI for the gameweek

The EPL consists of 20 teams which allows every team to play another team twice leading to a total of 38 games per team that are conducted in 38 gameweeks allowing every team to play one game every gameweek. Based on the live performance of the footballers, managers are given points on the selected players in their Playing XI and if a player does not play then, based on the substitution priority and formation, he is replaced from a player from the bench. A manager has the comfort of making one "Free Transfer" from their team of players with another from the open list. A deduction of 4 points takes place in the next week from a manager's score for every extra transfer made.[1]

1.2 Motivation

Various studies have been conducted on how fantasy sports are fueling the consumption of sports Beliën et al. (2017) and also highlights the fan motivation and the economic benefits the sports industry is achieving with its presence. In their study, the authors mention how modelling techniques for analyzing such sports tends to provide information and commercial value. Media outlets, whose fundamental trade relies on selling information, are the primary organizers of these online events and can display results of the analysis of these models on a regular basis for comparison. Also, it can prove to be critical in major tournaments to predict a players performance and consider selection of a team accordingly Iyer and Sharda (2009). This ensures that the best team members are sent for such tournaments and the possibility of winning a sport or, from an industrial standpoint, successful completion of a task increases substantially. Media outlets tend of obtain major commercial value by subscription to content discussing sports leagues and their respective fantasy leagues by the sole fact that fantasy sports lead to further consumption of that sport Beliën et al. (2017). Media outlets have recently started taking into consideration historic statistics as a valuable means of analysing the game and considering the possible outcome of future games but none of these have utilized underlying statistics in prediction algorithms. Also, team selection models have generally been developed with the consideration on the entire tournament Iyer and Sharda (2009) and only a few have worked upon the concept of dynamic team selection for each gameweek Matthews et al. (2012).

1.3 Research Question

How accurately can underlying match statistics be used to predict high scoring and undervalued players in the Fantasy Premier League team using statistical and machine learning techniques for the entire season??

The best application of the findings of this study would be for scouts of professional teams to help determine undervalued players that can prove to be a great addition to their respective teams. It would also be valuable to media outlets such as bloggers and sports talk show teams that can share accurate predictions of player performances which would further enhance the subscription of clients allowing further ingestion of their products and increased revenue. The information shared with the audience would tend to develop an interest in players of different teams that a participant would not usually follow leading to further consumption of information on those teams and possibly increased viewership of their games. Apart from the benefit to the sports industry, this study would be useful from an industrial perspective where multiple agents are required for the execution of a task and there are budget constraints involved.

[1] For further details on the rules and regulations on FPL visit: `https://fantasy.premierleague.com/help/rules`

2 Related Work

We initiate this section with discussions on why studies on Fantasy sports are essential and how are they impacting the sports and the media industry with it. This is followed by a quick discussion on prediction models with respect to player performances in different sports and we end with studies conducted in fantasy football.

2.1 Impact of Fantasy Sports to Sports and Media Industries

Upon their research on optimization modelling for fantasy sports, Beliën et al. (2017) mentioned how the audience is prone to wide participation by underrating the difficulty of the sport when models are publicly displayed and their results seem very much similar to the selection of regular participants. A promotional tactic of highlighting the huge gap in the optimal team and the winning team is also argued stating that such a difference tends to increase the confidence of a participant and serves as a fine motivation for continual participation.

Various industries can increase revenue by promoting such fantasy events as per multiple studies. Randle and Nyland (2008) suggested an increased use of traditional media thereby building customer loyalty by contesting in such tournaments after observing a strong correlation between participation in fantasy leagues and media consumption. In additional to agreeing with the above study, Drayer et al. (2010) adds that participants in these leagues tend to watch other team's performances along with their favorite team although this added interest in other matches does not result in a shift in allegiance towards their original favorite team. The author mentions only 1 out of 13 individuals would purchase merchandise of a team they do not support. Hence, we can conclude that a good result in the fantasy tournament may cover for the poor performance of the supported team.

Another revenue source has been observed by Nesbit and King (2010) pointing out a correlation in fantasy league participation and attendance in NFL games and this study was again replicated for the sport of baseball by Nesbit and King-Adzima (2012) using the same modelling technique. Although these studies counter the concept suggested by (Dwyer (2011), Fortunato (2011)) on the fact that television ratings are improved by participation in these tournaments, it can be understood that as multiple league games are conducted at the same time, people prefer checking results of multiple games at the same time to keep check on their scores by watching the games together or checking their scores online.

As the FPL is an open tournament which is free of cost, we have not covered gambling related studies as a motivation is our analysis.

2.2 Player performance prediction models in sports

One of the most prominent examples of predicting player performances would be by Lewis (2003), which describes the utilization of mathematical practices for supporting decisions in actual games, which was primarily showcased in baseball and the ideology was promoted so highly that it was even developed into a movie. Our attempt in this study focuses on machine learning techniques and hence sabermetrics would not come under consideration.

Riley (2017) conducted a recent study on the player performance in Ice Hockey using Structural equation modelling where the results observed that a highly offensive players tends to display poorer defensive performances which could possibly be considered a reason why attacking players in FPL are not given points on keeping a clean sheet (no goals conceded) in a match.

Berrar et al. (2019) discussed the open challenge to analyze "to what extent was it possible to predict the outcome of a soccer match using commonly available match data?" which observed various studies the most prominent of which was by Hubáček et al. (2019) that used Gradient boosted tree algorithm as it's model and observed an accuracy of 52% but did not work on any fantasy related data. The author did argue that prediction challenges tend to be more testing as it can be conducted with the utilization of previous data rather in live updates. Hence, this gives us a strong reason to work with the summation of historical data in our work in every row for a player. Also gradient boosting happens to be a pretty popular choice when it comes to challenges Chen and Guestrin (2016) using ensemble modelling. Hence, XGBoost would be a model for consideration in the study.

There are often factors that are not known data points but can tend to be significant towards the prediction model. These would include ratings of a player where Magnus and Arntzen (2010) has used ELO coding to encode past data of the players and Lasek et al. (2013) using the ELO algorithm to analyze FIFA rankings. The unknown variables used by us are the player Influence, creativity, threat and a combined ICT Index which does not need encoding as these parameters are publicly available before the match.

Iyer and Sharda (2009) worked upon a Neural Network based model in the sport of cricket to forecast player performance for the World Cup tournament. A player's performance was classified into categories such as "Failure", "Moderate" and "Performer" and based on these definitions the data was trained and tested to evaluate the output with the World Cup. The multilayer perceptron was the profile that presented the highest accuracy of 84% for batsmen and 80% for bowlers. While it presented a reasonably strong result, not all positions (wicket-keeper batsmen, all-rounders, and fielders) were taken into consideration for this study.

Another study where all positions were not taken into consideration, and even acknowledged, was done by Mukherjee (2014) where player performance was calculated keeping into consideration the difficulty of the opponent faced. Here, players were scored and ranked as per their traits and the study managed to highlight that a model can successfully classify a player into a bowler or batsmen based on the quality of their attributes. This research was conducted using Social Network Analysis (SNA) using PageRank algorithms.

This drawback was, by a certain extent, overcome by discussing each position in intricate detail by Bin-Malek et al. (2018) where their research aimed at predicting an ideal cricket team based on the historical performances of the player's in the domestic cricket league of Bangladesh (Dhaka Premier League). A different model for performance prediction was run for different positions of player where SVM, conducted for the bowling department, showed the highest accuracy of 70% along with regression and decision trees while for batsmen and all-rounders there were k-means clustering, PageRank, linear regression and Naïve Bayes models run.

2.3 Data Analysis using Fantasy Data

Lutz (2015) conducts an analysis on the prediction of the fantasy score of a quarterback in the NFL using 5 years-worth of data using SVM and Neural Networks. Despite getting better results from neural networks in comparison, it was observed by the author that there were high errors in the models suggested. There could be a possibility that the high error rates are due to the fact that only one position is being taken into consideration for scoring.

The most prominent work of analysis using soccer fantasy data has been presented by Matthews et al. (2012) and Bonomo et al. (2014) where both have worked upon creating team selection models based on statistical analysis. Matthews et al. (2012) served as a primary benchmark towards sequential optimal team formation by formulating a belief-state Markov decision process for 3 years of FPL league data while abiding by the rules and budget constraints of the game and managed to observe the highest rate of success using Bayesian Q learning approach observing a rank as high as 20 thousand.

Another soccer fantasy analysis was later done by Bonomo et al. (2014) which was focused towards a fantasy league Gran DT that is based on the Argentinian soccer league. While the concept of Gran DT is the same as FPL, the rules tend to differentiate in various places. Here, the author ran 2 models of a priori and posteriori to determine an optimal line-up using an Integer linear programming model and a descriptive linear programming model respectively and managed to observe the results provided by the model would rank in the top 10% of the tournament.

To the best of our knowledge, we have not found any study utilizing fantasy data in football to determine individual player performance using machine learning techniques. Hence, with the novelty of added parameters in hand and understanding of above literature we intend to run a comparison analysis of multiple prediction models of player classification with and without the new parameters.

3 Methodology

This study utilizes the Knowledge Database Discovery (KDD) methodology. A basic understanding of the same can be seen in Figure 2.

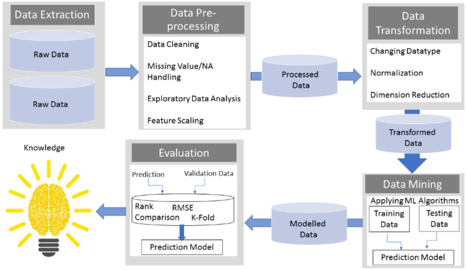


Figure 2: KKD Implementation Methodology

As we can observe from the above figure, the implementation of KDD methodology consists of the following phases:

- **Data Extraction**

It was initially planned to extract the data using the FPL API which was available on `https://fantasy.premierleague.com/api/bootstrap-static/` but post pulling data from the site it was observed that historical data is not held by the site after the start of the new season and the data present is a cumulation of values which does not provide with a week on week statistic. Hence, another approach was taken where the FPL dataset was taken from a Github repository that can be accessed from `https://github.com/vaastav/Fantasy-Premier-League` which consisted of weekly data for 3 seasons (i.e. from 2016/2017 – 2018/2019) along with a secondary data source which included underlying statistics that were obtained by request from `https://understat.com/` .

- **Data Pre-processing**

Once the data has been collected and stored at a local database, we had to work on adding position of the players and converting their team name from numerical to string as the same 20 teams do not play the next season. This was done using the raw player list present in the GitHub repository itself. Once the original values were restored, we were able to combine all the 3 datasets of individual seasons but in order to merge the data from understat we had to ensure the naming of all players matched. Post the handling of the improper names, we were able to merge the two datasets to obtain a file of 50 attributes.

- **Data Transformation**

This phase covered the handling of missing values that were observed along with the addition of the Form parameter which provided an average of the score of a player in the last 4 matches. After the addition of form parameter, we had to ensure that all the in-game statistics could not be used as a part of the model. Hence, we replaced all those attributes with the summation of their respective attributes till the previous game. This would act as historical data for every row. Also post analysis, we understood that scoring ability of a player differs based on their position and hence decided to set up a classifier that could determine if a player gave a high scoring or low scoring performance.

Every game tends to highlight multiple factors. As our dataset had roughly 50 attributes, we needed to run a Boruta algorithm for feature extraction that could determine which factors are important towards prediction of our required variable.

- **Data Mining**

As we have mentioned in our literature survey above, there has been very limited work done in fantasy football with respect to analyzing individual player performances. The only renowned work has been conducted using statistical techniques and mathematical programming. Hence, we plan on conducting a comparison analysis of multiple models with and without our novel parameters to give us an understanding of how they impact our model. The following models have been taken into consideration for the same:

RandomForest

It has been rightly pointed out the concept of the wisdom of crowds[2] which is the basic ideology behind an ensemble modelling technique such as random forest. In an ideal

[2] `https://towardsdatascience.com/understanding-random-forest-58381e0602d2`

scenario, the performance of a combination of multiple models will be better than that of just a single model. Hence, we can define Random Forest technique as multiple decision trees where the result of the majority trees is provided as the output of the model which can be seen in Figure 3.

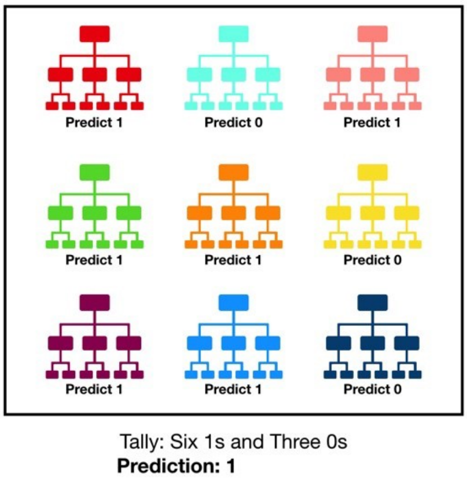


Figure 3: Random Forest Decision Making

XGBoost

After analyzing the study of Hubáček et al. (2019) and observing that it was providing the highest accuracy in the soccer prediction challenge, we decided to opt for the gradient boosting algorithm as one of our models. XGBoost works on a similar concept of gradient boosting but is considered to be one of the more efficient and resourceful technique of the two.(Chen and Guestrin (2016)).

4 Design Specification

We initiate with loading both datasets into the environment. Post cleaning and transformation, we merge the 2 datasets to initiate further processing to our model. The only drawback observed here is the secondary dataset holds the popularly known names of the players which is not the same as their registered names. Hence, there was a requirement to handle this naming convention as well. After merging the two datasets, we apply feature selection technique using Boruta algorithm to understand which attributes are important to the player performance. We then apply Random Forest and XGboost models with and without the presence of the xG and xA parameters under different techniques of handling class imbalance. We evaluate the results of the models using classification evaluation techniques. A pictorial representation of the design specification is shown in Figure 4.

5 Implementation

Upon extracting the data and bringing it into the environment, we observed that there was a requirement of merging weekly datasets into seasonal datasets and combine individual season sets to a final combined dataset only after which could the merge of the primary and secondary dataset take place. As team names had been converted into factors, they

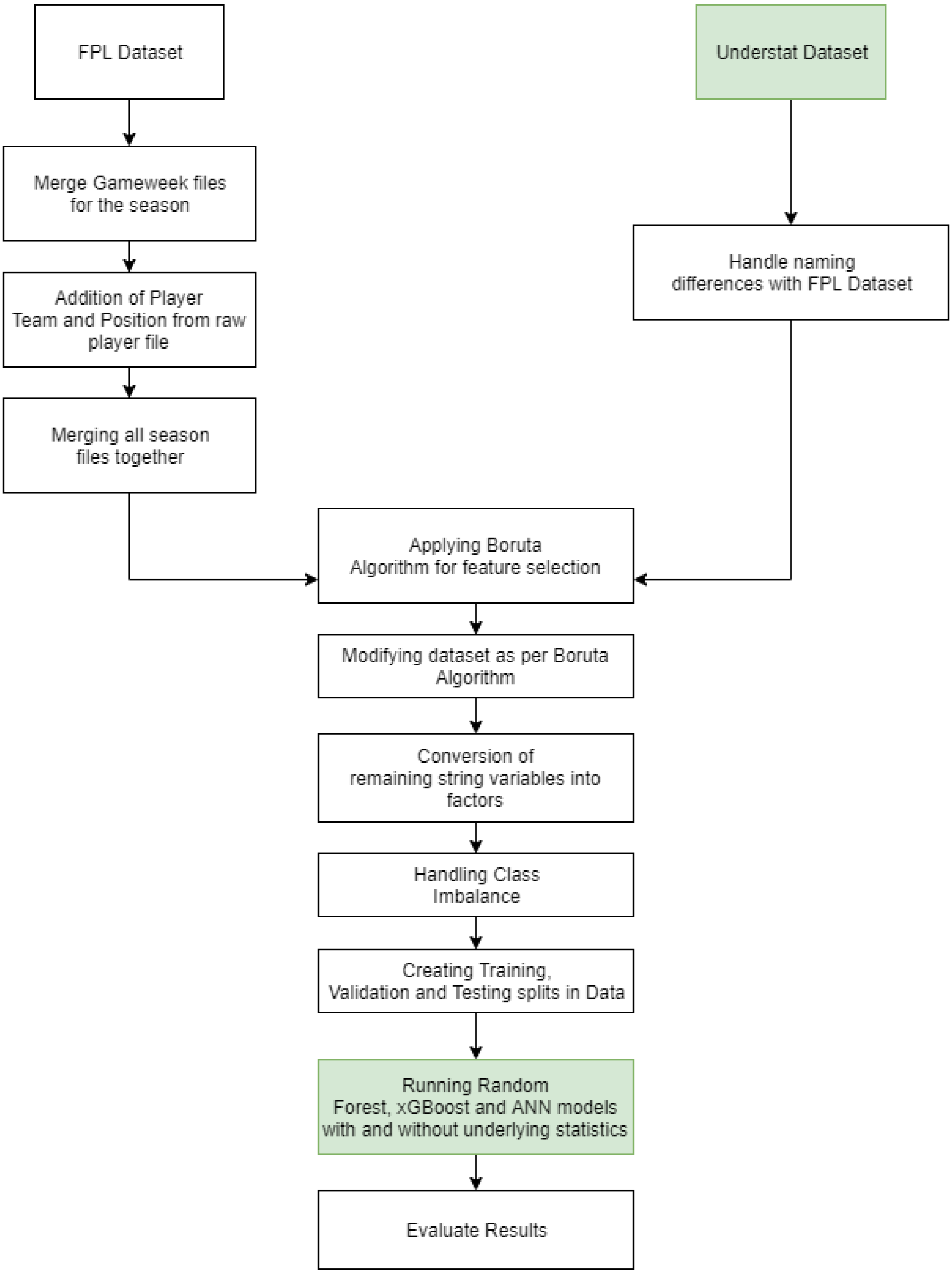

FPL Dataset
Understat Dataset
Merge Gameweek files for the season
Handle naming differences with FPL Dataset
Addition of Player Team and Position from raw player file
Merging all season files together
Applying Boruta Algorithm for feature selection
Modifying dataset as per Boruta Algorithm
Conversion of remaining string variables into factors
Handling Class Imbalance
Creating Training, Validation and Testing splits in Data
Running Random Forest, xGBoost and ANN models with and without underlying statistics
Evaluate Results

Figure 4: Initial Workflow Implementation

had to be renamed into their respective character values as the same 20 teams do not play the tournament in the following season. Also, player position column had been removed which had to be reinstated. These values were present in the raw player files present in the repository. Upon reinstating these values, the seasonal data was merged a final combined dataset was created.

Now in order to merge the two major data files we needed to ensure there were enough common parameters to identify the appropriate entry. The parameters “Player Name”, “Team”, “Opposition” and “Season” and “Month” were taken into consideration. While these parameters were sufficient to meet the requirement, we observed that the understat dataset had popular names of football players rather than the officially registered names. Hence, we had to work upon handling the incorrect names and observed around 83 players to have either a spelling difference or a nickname. This change was addressed post which we managed to combined the 2 datasets for our final working dataset. This combination included roughly 60 attributes out of which 12 pre-match attributes are shown below in Table 1. The last 4 parameters Influence, Creativity, Threat and ICT Index are underlying statistical parameters not observed in any other studies as it was introduced in 2016 much after the only other known study on FPL Matthews et al. (2012)(Matthews, Ramchurn and Chalkiadakis, 2012). Apart from this, the other underlying statistics presented from the secondary dataset are Expected Goals(xG) and Expected Assists(xA).

Sr.No	**Attribute Name**	**Type**	**Details**
1	Player Name	String	Name of the player
2	Team	String	Team the player plays for
3	Position	String	Playing Position i.e Goalkeeper, Defender, Midfielder or Striker
4	SeasonNo	int	Highlighting the season of the entry
5	Round	int	Highlighting the round in which match is played
6	value	int	Player value with the last digit being the first decimal
7	Opp	String	Opposition Team
8	was_home	Boolean	Is match in consideration being played at home?
9	Influence	int	The degree to which that player has made an impact on a single match or throughout the season
10	Creativity	int	It assesses player performance in terms of producing goal scoring opportunities for others
11	Threat	int	A value that examines a player’s threat on goal
12	ICT Index	int	Combination of Influence, Creativity and Threat

Table 1: Pre-match Attributes.

OPTA is the leading industry in maintaining sports statistics across the globe for multiple sporting events and have initiated these new parameters of analysis. After analyzing player data with a large amount of data points, they have managed to set these 2 new parameters where they can calculate the probability of a shot taken that could have been a goal as the expected goal and similarly the possibility of a key pass being an assist as an expected assist. These parameters can be used as an ideal understanding of how many goals or assist a player could have achieved in a match.

Hence, we can define Expected Goals (xG) in Eqn.1 as

$$ExpectedGoals = \sum_{i=0}^{n} P(Goal)i \tag{1}$$

where n = number of shots

And Expected Assists (xA) can be defined in Eqn.2 as

$$ExpectedGoals = \sum_{i=0}^{n} P(Assist)i \tag{2}$$

where n = number of key passes

While these underlying stats are calculated after the match, we are using this as a historical entry where this is being used as a data point of measure for the next game.

Upon merging the two, we observed these 2 attributed presenting a lot of missing values. Upon inspection we realized, that a majority of these belonged to people who had not played a single game. Hence, we could replace these values with a 0 and then decided to initiate imputing with the shots taken attribute. Here, a minimum value of 1 was given to players who had scored a goal. The remaining values were predicted using the ANOVA method through the RPart function in R Studio.

The combination of the 2 major datasets provided us with 68 attributed, 12 of which were pre-match related details that are being used while the remaining contain various data points which are accounted for during the end of the match. Hence, in order to ensure the appropriate functioning of the model, those parameters were converted into a summation of all the previous entries to present itself as historical data. Also, an existing attribute in FPL called player form has been introduced.

The Boruta algorithm was run for feature selection on almost 50 variables out of which only 1 was rejected. Given below is the plot of the Boruta output in Figure 5.

As none of the variables were even being considered as tentative, we had to take a call of removing the 10 least important variables from the list.

Upon analyzing the dataset, it was observed that players with low base value also tend to get high scores. But due to inconsistencies or possibly injuries they cannot maintain a long run of high scores (as seen in Figure 6).

Hence, we draw a comparison analysis of the highest scorers for each position for every season and draw an average points per week average analysis as can be observed in Table 2.

Position	GK			DEF			MID			FWD		
Season	1	2	3	1	2	3	1	2	3	1	2	3
Highest Score	149	172	176	178	175	213	264	303	259	224	217	205
Points per game	3.92	4.53	4.63	4.68	4.61	5.61	6.95	7.97	6.82	5.89	5.71	5.39
Classification Score	4			5			6			4		

Table 2: Classification score table

This allows us to determine a fixed score for every position that can be used to identify a “above average” and “below average” performance status. This status is the dependent variable which is to be predicted. As the dependent variable has a binary value, the classification modelling was adapted to predict the output. The output predictions were

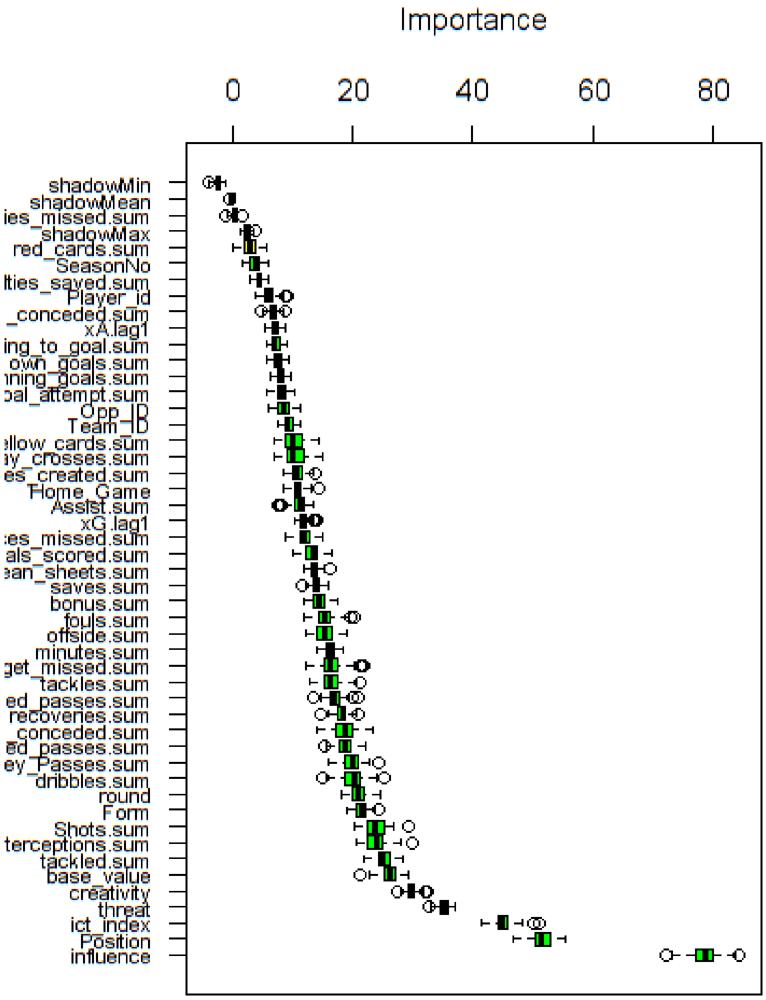


Figure 5: Boruta plot for feature selection highlighting parameters based on level of importance

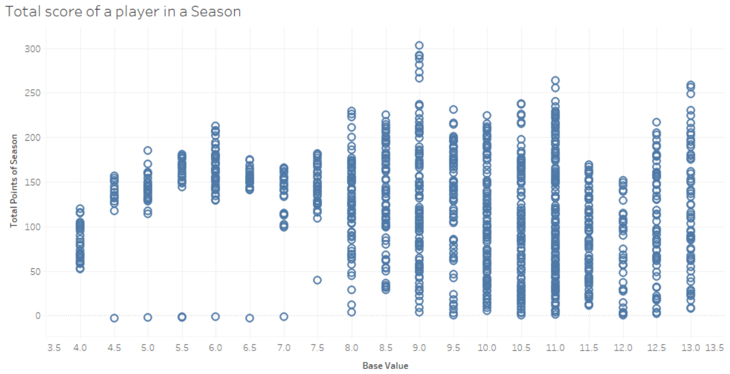


Figure 6: Player Value vs Total Points scored in a season

evaluated on the basis of the general classification evaluation matrices such as Accuracy, Precision, Recall and F1 score.

6 Evaluation

The initial intention of the study was to observe if underlying statistics play a major role in the prediction of player performance and hence each model was intended to run twice with experiment A being with the parameters and B being without the parameters.

During the splitting of the dataset into training and testing sets, class imbalance was observed between the count of the binary output. Hence, in order to account for the class imbalance, we conducted 4 different sampling methods: Over sampling, under-sampling, a combination of oversampling and under-sampling and Random Over Sampling which is a part of the ROSE package in RStudio. Hence, experiment A resulted in 4 models and B resulted in 4 models of the same algorithm. These models where then trained and their predictions were then evaluated using Confusion Matrix which provides us with the required evaluation parameters. Given below is a brief analysis of the same.

We have observed Evaluation parameters such as Accuracy, Sensitivity, Specificity, Precision, Recall and F1-score to determine quality of the models.

- **Experiment 1 Random Forest**

	Accuracy (%)	
Sampling	**A**	**B**
OverSampling	91.62	91.75
Under	83.34	83.59
Both	83.96	83.19
ROSE	91.06	92.30

Table 3: Accuracy comparison of Random Forest models

ROSE sampling provides the highest accuracy in case B while the lowest is observed in the combination of both cases. Accuracy is seen to increase in case B when attributes are removed.

	Sensitivity (%)	
Sampling	**A**	**B**
OverSampling	93.50	93.68
Under	98.93	98.96
Both	97.68	98.14
ROSE	92.19	92.40

Table 4: Sensitivity comparison of Random Forest models

Under sampling provides the highest sensitivity in case B while the lowest is observed in the ROSE sampling. Sensitivity is seen to increase in case B when attributes are removed.

Specificity is highest in the ROSE sampling in case B while the lowest in the combination of both. There is a significant drop in specificity observed in all cases and sampling methods.

	Specificity (%)	
Sampling	**A**	**B**
OverSampling	76.00	75.71
Under	45.39	45.36
Both	46.11	44.52
ROSE	79.07	79.23

Table 5: Specificity comparison of Random Forest models

- **Experiment 2 XGBoost**

	Accuracy (%)	
Sampling	**A**	**B**
OverSampling	95.37	95.47
Under	88.19	88.08
Both	87.60	86.50
ROSE	91.49	91.30

Table 6: Accuracy comparison of XGBoost models

Upon performing XGBoost modelling we observed the highest accuracy being observed in oversampling by 95.47% and the lowest in the combination of oversampling and under sampling. It can also be observed that apart from the over sampling method, accuracy has dropped in all sampling techniques when xG and xA are removed.

	Sensitivity (%)	
Sampling	**A**	**B**
OverSampling	91.35	91.50
Under	84.68	84.60
Both	85.00	84.04
ROSE	92.54	92.17

Table 7: Sensitivity comparison of XGBoost models

The ROSE sampling technique observed the highest sensitivity in experiment A while under sampling observed the lowest sensitivity in experiment B. Again, it can be seen that sensitivity drops in all cases barring oversampling.

Oversampling observed the highest specificity at 99.45% while the combination sampling observed the lowest specificity at 89.17% both in experiment B. Oversampling still follows the trend of going in the opposite direction as specificity increases for it in case B.

In case of precision, Oversampling provides the highest value in case B while the combination sampling shows the least also in case B. Precision is observed to decrease for all cases except oversampling.

Recall shows its highest value is ROSE sampling at 92.54% while its lowest is observed in the combination sampling at 84%.

F1-score is at peak value in Oversampling at 95.28% while its lowest is observed in the combination sampling at 86.62% in experiment B. Again F1-score shows a drop in experiment B for all sampling techniques barring oversampling.

	Specificity (%)	
Sampling	**A**	**B**
OverSampling	99.39	99.45
Under	91.70	91.56
Both	90.42	89.17
ROSE	90.43	90.42

Table 8: Specificity comparison of XGBoost models

	Precision (%)	
Sampling	**A**	**B**
OverSampling	99.34	99.40
Under	91.07	90.93
Both	90.57	89.37
ROSE	90.69	90.64

Table 9: Precision comparison of XGBoost models

- **Discussion**

From the above tables it is evidently clear that the underlying attributes provide a minimal impact towards the addition of the prediction model. Also, it can be seen that the accuracy takes a dip during the presence of the underlying parameters even if it were below 1%. As accuracy does not hold the strongest of value in a class imbalance scenario, we set our focus on sensitivity and specificity comparisons between the different modelling techniques. A quick look at Table 4 and Table 7, shows that while both models show high returns on sensitivity, Random Forest displays a relatively stronger output which highlights that Random Forest presents a high True Positive rate which shows a robust prediction of high performing player and less chances of incorrectly predicting the strong performers. But upon observing Table 5 and Table 8, we observe a drastic drop in specificity for Random Forest which falls below 50% for a couple of sampling techniques and shows a maximum of under 80% which when compared with Extreme Gradient Boosting (XGBoost), proves to be much poor. This drop observed in specificity show a weak attempt at predicting players with actual low performance which would result in a loss for clubs investing in players. Hence, XGBoost provides a more balanced approach towards player performance predictions. Also, as the precision and recall are both observed at high values and cannot provide a clear differentiation from Table 9 and Table 10, we observe the F1 score which is a harmonic mean of the 2 attributes to understand which sampling technique is more beneficial. Table 11 clearly shows Oversampling delivering the highest F1 score, but we need to also understand the oversampling may lead to over-fitting which can be observed with a 99% precision value. Therefore, we prefer holding onto the Random Oversampling Example technique (ROSE) as the ideal sampling technique for our imbalanced dataset as it ensures balanced samples using a smoothed bootstrap approach which is useful in classification techniques. It should also be noted that while the attributes xG and xA had low impact, the Accuracy and the F1 score of ROSE sampling in XGBoost drops by 0.19% and 0.2% respectively.

	Recall (%)	
Sampling	**A**	**B**
OverSampling	91.35	91.50
Under	84.68	84.60
Both	85.00	84.04
ROSE	92.54	92.17

Table 10: Recall comparison of XGBoost models

	F1-score (%)	
Sampling	**A**	**B**
OverSampling	95.18	95.28
Under	87.76	87.65
Both	87.70	86.62
ROSE	91.60	91.40

Table 11: F1-score comparison of XGBoost models

7 Conclusion and Future Work

The sports industry has looking at different avenues for raising revenue with introduction to gaming in different concepts. Fantasy tournaments such as FPL create the possibility of computing the performance of a player in a numerical score that can be calculated through a series of performance parameters. The presence of such data allows us to find potential replacements for well-established players most likely in the form of youth players. Scouting young talented players at an early age has always been a difficult task for various teams especially considering the resources at hand. Also, the introduction of unique data points from statistical organizations which have been under scrutiny over its reliability makes it difficult to analyse players. Hence, this study makes an attempt at determining if such underlying statistics can be used for the purpose of improving player performance predictions. The use of ensemble models has primarily been the area of focus as it provides as improved ability of prediction when compared to a standalone model and hence the use of Random Forest and XGBoost has taken place. The study highlighted how XGBoost provides a stronger result when both Sensitivity and Specificity and considered and that ROSE sampling technique is the most favourable option with the high F1 score. Also, we can also successfully conclude that presence of underlying statistics such as Expected goals and Assists as historical data points improves the accuracy levels by 0.2%.

Considering the fact that minimal approaches have been taken towards utilizing fantasy data in the past, there are various possibilities to conduct future research in the realms of team selection models on the basis of prediction models.

8 Acknowledgement

I would like to thank my parents for their constant support in allowing me to pursue my passion. I am extremely grateful to Dr. Vladimir Milosavljevic for his guidance in helping me implement this research and aiding me in thinking of the big picture while working on the research. I would also like to thank Vaastav Anand for allowing me to

use his Github repository and understat.com for sharing their dataset.

References

Beliën, J., Goossens, D. and Van Reeth, D. (2017). Optimization modelling for analyzing fantasy sport games, *Infor* **55**(4): 275–294.

Berrar, D., Lopes, P., Davis, J. and Dubitzky, W. (2019). Guest editorial: special issue on machine learning for soccer, *Machine Learning* **108**(1): 1–7.
URL: *https://doi.org/10.1007/s10994-018-5763-8*

Bin-Malek, M., Badhan, R. H., Shesir, M. I. and Fakir, N. H. (2018). Squad Selection For Cricket Team Using Machine Learning Algorithms, (October).

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.

Chen, T. and Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System.

Drayer, J., Shapiro, S. L., Dwyer, B., Morse, A. L. and White, J. (2010). The effects of fantasy football participation on NFL consumption: A qualitative analysis, *Sport Management Review* **13**(2): 129–141.
URL: *https://www.sciencedirect.com/science/article/abs/pii/S1441352309000217?via%3Dihub*

Dwyer, B. (2011). The Impact of Fantasy Football Involvement on Intentions to Watch National Football League Games on Television, *International Journal of Sport Communication* **4**(3): 375–396.

Farquhar, L. K. and Meeds, R. (2007). Types of fantasy sports users and their motivations, *Journal of Computer-Mediated Communication* **12**(4): 1208–1228.

Fortunato, J. A. (2011). The relationship of fantasy football participation with NFL television ratings, *Journal of Sport Administration & Supervision* **3**(1): 74–90.

Hubáček, O., Šourek, G. and Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees, *Machine Learning* **108**(1): 29–47.

Iyer, S. R. and Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection, *Expert Systems with Applications* **36**(3 PART 1): 5510–5522.
URL: *http://dx.doi.org/10.1016/j.eswa.2008.06.088*

Kim, Y. and Ross, S. D. (2016). An exploration of motives in sport video gaming, *International Journal of Sports Marketing and Sponsorship* **8**(1): 28–40.

Lasek, J., Szlávik, Z. and Bhulai, S. (2013). The predictive power of ranking systems in association football Zoltán Szlávik Sandjai Bhulai, **1**(1): 27–46.

Lee, S., Seo, W. J. and Green, B. C. (2013). Understanding why people play fantasy sport: development of the Fantasy Sport Motivation Inventory (FanSMI), *European Sport Management Quarterly* **13**(2): 166–199.

Lewis, M. (2003). Moneyball : The Art of Winning an Unfair Game.

Lutz, R. (2015). Fantasy Football Prediction, pp. 1–7.
URL: *http://arxiv.org/abs/1505.06918*

Magnus, L. and Arntzen, H. (2010). Using ELO ratings for match result prediction in association football, *International Journal of Forecasting* **26**(3): 460–470.
URL: *http://dx.doi.org/10.1016/j.ijforecast.2009.10.002*

Matthews, T., Ramchurn, S. and Chalkiadakis, G. (2012). Competing with humans at fantasy football: team formation in large partially-observable domains, pp. 1394–1400.
URL: *http://eprints.soton.ac.uk/340382/*

Mukherjee, S. (2014). Quantifying individual performance in Cricket - A network analysis of batsmen and bowlers, *Physica A: Statistical Mechanics and its Applications* **393**: 624–637.
URL: *http://dx.doi.org/10.1016/j.physa.2013.09.027*

Nesbit, T. M. and King-Adzima, K. A. (2012). Major League Baseball Attendance and the Role of Fantasy Baseball, *Journal of Sports Economics* **13**(5): 494–514.

Nesbit, T. M. and King, K. A. (2010). The impact of fantasy football participation on NFL attendance, *Atlantic Economic Journal* **38**(1): 95–108.

Randle, Q. and Nyland, R. (2008). Participation in internet fantasy sports leagues and mass media use, *Journal of Website Promotion* **3**(3-4): 143–152.

Riley, S. N. (2017). Investigating the multivariate nature of NHL player performance with structural equation modeling, *PLoS ONE* **12**(9): 1–29.