Proceedings of 12th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2022 Month: October Volume: 12

Fantasy Premier League - Performance Prediction

Pratik Pokharel [a], Arun Timalsina [b], Sanjeeb Panday [c], Bikram Acharya [d]

[a, b, c, d] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal*

[a] 076msdsa008.pratik@pcampus.edu.np , [b] t.arun@pcampus.edu.np,
[c] sanjeeb@ioe.edu.np, [d] aacharya.bikram@gmail.com

Abstract

Fantasy Premier League (FPL) participants often use personal bias, favoritism and recency logic, when picking the squad or making the transfers. Their decision is often based on which club they support and on the so called ”star image” of a player. But if these factors are put aside and decisions are made focusing on possible return on investment(ROI), it could be a wiser option. This paper presents a rational approach to the player selection, team drafting and transfer- by predicting return on investments- using xgboost regression. In addition, the effects of fixture congestion on FPL points is also assessed by using the mid-week cup fixture data. On evaluation using FPL global ranking- using the initial drafted team throughout the season without transfers performed better. The transfer algorithm had shortcomings due to its dependency on accuracy of regression model. The mean rmse score for all players was 2.048. The effect of cup fixture congestion was found to be insignificant as far as FPL points is concerned.

Keywords

FPL, xgboost regression, rmse, ROI

1. Introduction

Fantasy Premier League- often abbreviated as FPL is the most popular fantasy sport in the world. It is the official fantasy football game for the English Premier League- top tier of English League football system- and is organized and managed every year by the official website of the Premier League. The unpredictable nature of football matches makes FPL extremely engaging and the risks of the game gives participants a unique adrenaline rush similar to that of an adventure sport.

A participant is provided with a virtual budget of 100 million pounds to select a fantasy football squad of 15 players. The squad should be composed of: two goalkeepers, five defenders, five midfielders and three forwards. The restriction while selecting players is that one can't have over three players from a single Premier League club in their team. Of the 15 players team created, every week 11 players among that has to be selected for the staring XI. The selected 11 players amass points based on their performance in the actual matches played in the premier league over the game-week. A captain is allowed to be nominated from the starting XI, whose total points is doubled and added to the participant's total score. The rules of FPL points scoring is defined by the official website of FPL. Apart from the team selection, each week, participants are allowed to make a transfer. Using transfer feature, participants can remove a player from the squad, and bring on a new player for the same position, keeping in mind the budget restrictions and maximum players limit. If more than one transfer is made for a game-week, then for each extra transfer made, the participant is penalized with four points.

The uncertainty around team selection makes it hard for FPL managers to make what is the most important choices of the season. Game-week 1 is all where it starts and it is one of the most crucial stages in FPL. The foundation of the season is laid with the team a participant drafts for game-week 1. If it is managed somehow, to identify consistent performers, predict and include the big hitters early, half of the battle is won [1]. A general intuition while drafting a team is- certain players should be selected such that they can be kept in the team for the entire season, and team should be built around those players. Some cheap players should be selected based on their average fixture difficulty rating for the upcoming 3 or 4 game-weeks.

If the total number of combinations of teams that can

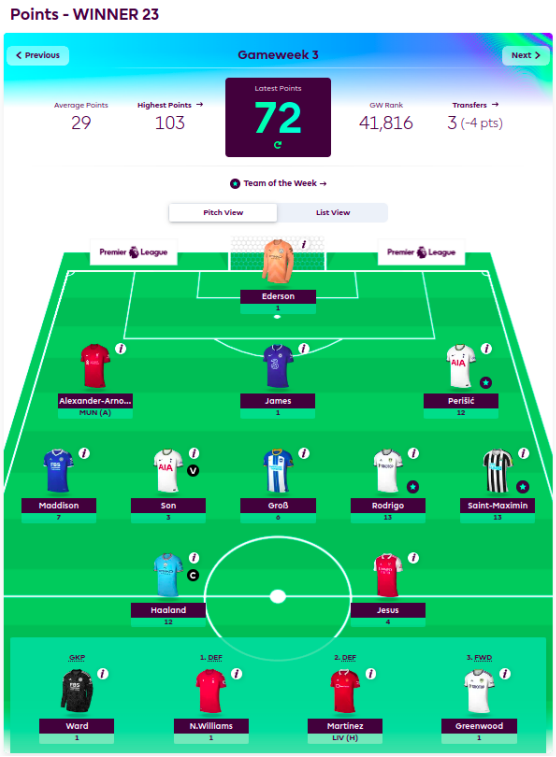

Figure 1: FPL Console

be formed from available players and restrictions is calculated, then there is a possibility of over 50 octillion combinations of teams. Even if Exploratory Data Analysis is performed on posteriori stats and options are concentrated to the high performers only, there would still be a significantly big selection dilemma.

The researchers of FPL analytics often focus on finding the best transfers and high performers using historical statistical data. The issue with this approach is that external factors such as: statistics of other tournaments, mid-week fixture fatigue effects and manager's approach to squad rotation can never be incorporated into the final predictions. Certain teams play in the mid-week in European cup fixtures as well as domestic cup tournaments. Managers of these teams often rotate their squad, and leave some of their best players on the bench in the league matches, if either the opponents of cup matches are of higher difficulty or the cup fixture is of higher importance.

2. Related Literature

The general inclination observed in previous works has been to use historical statistical data in combination with machine learning methods to predict future scores [2]. Using the statistics of previous game-weeks together with Gaussian Naive Bayes algorithm, Thapaliya predicted future performances with a reported accuracy of 86 percent. He classified the data labels into two classes- the ones that have amassed 6 or more points, and the ones with less than 6 [3]. His model was one of the very early FPL predictor models, but had shortcomings. His model dealt with a class imbalance problem, with the class having less than 6 points as the majority class. As far as FPL is concerned, we are interested in the minority class- the players with 6 or more points. So, 86 percent accuracy was not a true reflector of the performance of his model. Apart from that, injuries were also not considered into the final model.

Bonomo *et al.* developed a mathematical optimization model using integer linear programming to predict ideal line ups every game-week in Argentinian Football League. They used historical data combined with information from manager's press conference before matches were played. They used their model on posteriori stats to determine factors that could possibly help in building predictive models [4].

Matthews presented a more sophisticated fantasy football predictor that consisted of belief-state Markov Decision Process algorithms combined with Bayesian Q-learning to train on the past five years of football data. He combined expert knowledge along with statistical player data and was able to achieve a rank of 113,921 out of 8 million participants [5].

Bonello incorporated human feeedbacks into predictions by combining standard statistical measures along with betting market data, posts from social media, web articles, twitter, reddit, etc. While testing on 2018/19 season, his team generated a rank of 20,000 out of 7.5 million participants that season [6].

Godin used historical data with tweets to predict football match results for beating bookmaker accuracies, showing that combination of multi-stream data was useful to predict match outcomes. He considered various factors including sentiment analysis and produced 62 percent correct prediction of results of upcoming matches by using data from previous 5 games [7].

3. Methodology

The block diagram of the system is depicted in the figure 2.

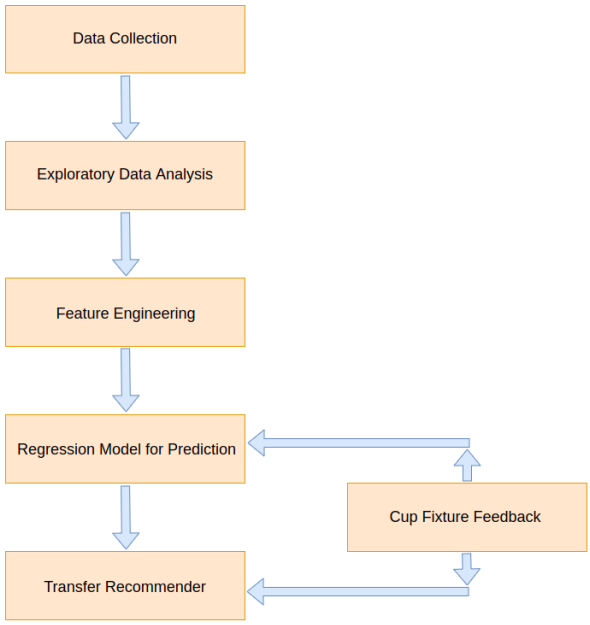

Figure 2: System Block Diagram

3.1 Dataset Description

A publicly available dataset of FPL statistics from season 2016/17 to 2021/22 was used for training and testing of the proposed model. The dataset consists of statistics related to over 1000 players [8]. For feedback data of other tournaments, data for the detailed fixture lists of each of the 20 clubs playing in the premier league in the season 2021-22, which includes UEFA competitions, FA cup and English League cup was collected. The collected data ranges from the start of the 2018-19 season to March 2022.

3.2 Exploratory Data Analysis

Some preliminary EDA was performed on 2019-20 season data to test out some of the hypothesis. One of it was whether players belonging to the teams in the top half of the table had more points than from players belonging to the bottom half teams. The hypothesis came out to be true on a higher level, yet some outliers were found which were too good to be ignored. These outliers are the hidden gems as far as FPL is concerned. They can be visualized from the figure 3.

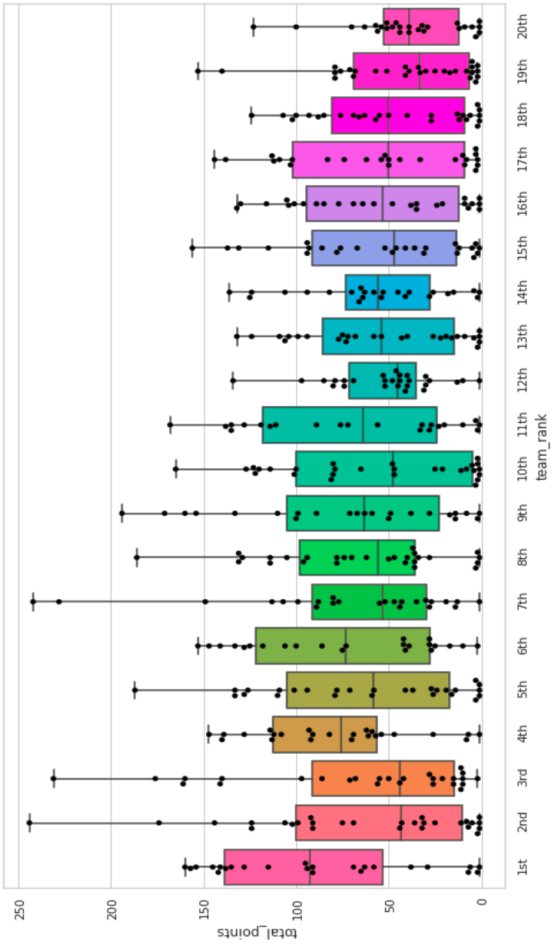

Figure 3: Total Points vs Team Rank of players- season 2019-20

Furthermore, how clubs ranked in terms of FPL points earned was checked. Interesting to see that Man City, Leicester City, Sheffield United and West Ham United outperformed teams higher up in the league standings up to some extend. This can be seen from figure 4.

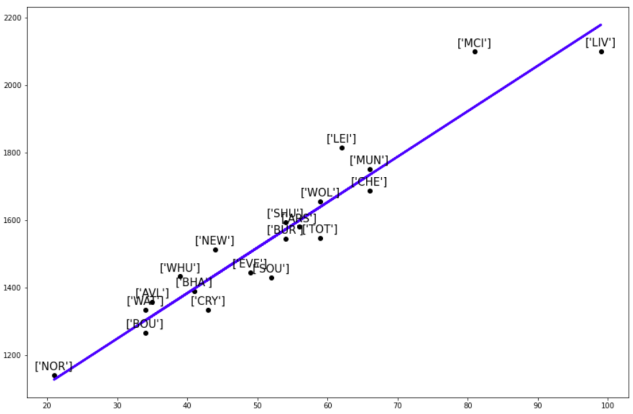

Figure 4: Total Fantasy Points vs Total League Points- season 2019-20

Another hypothesis was whether a club's overall FPL points is the true reflection of their performance on every parts of the pitch. The hypothesis turned out to be false as shown in the figure 5. Man City, Despite ranking on top in terms of total points earned amongst the 20 PL teams, had the worst defensive record with only 31.1 % defensive contribution to the total points earned. A certain club named Sheffield United had a 61.48 % defensive contribution. In terms of midfielders, Man City, Spurs and West Ham were the best buy and for strikers- Arsenal and Southampton.

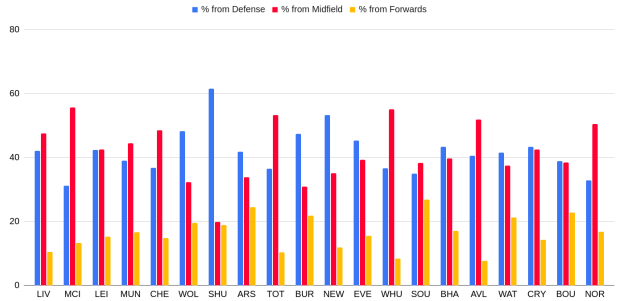

Figure 5: Points contribution percentage from each playing positions for all PL clubs- season 2019-20

It was checked whether high scoring players from the overall list gave good return on investment(ROI) or not. ROI is determined by calculating total points per million cost for every 90 minutes of matches played. The main logic behind using ROI is that if the return on investment is maximized within the budget constraint and the full budget is used, then points would be maximized in the long run. For instance, consider two teams A and B with identical budget constraints of 100 million. If ROI is calculated for both teams such that:

$$ROI_A \geq ROI_B,$$

$$then, TotalPoints_A/Budget_A \geq TotalPoints_B/Budget_B$$

$$since, Budget_A = Budget_B$$

$$finally, TotalPoints_A \geq TotalPoints_B$$

This shows that if the use of provided budget is maximized, the team with the highest ROI will be the team that produces the highest points. Thus a participant's goal should be to increase Return on investment from those players who play more often. This research uses the same metric on players playing frequently to predict their performance.

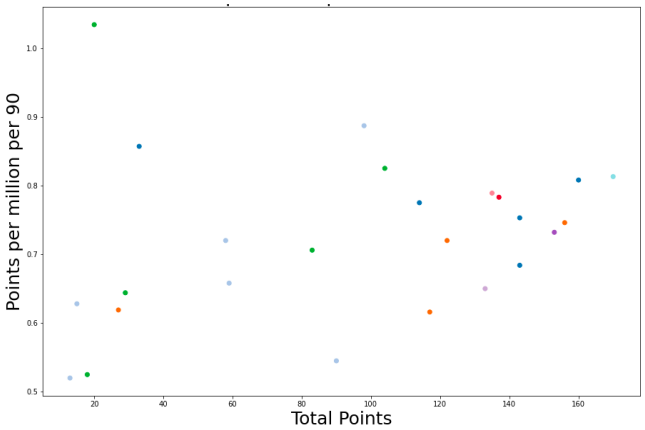

Figure 6: ROI- Goalkeepers

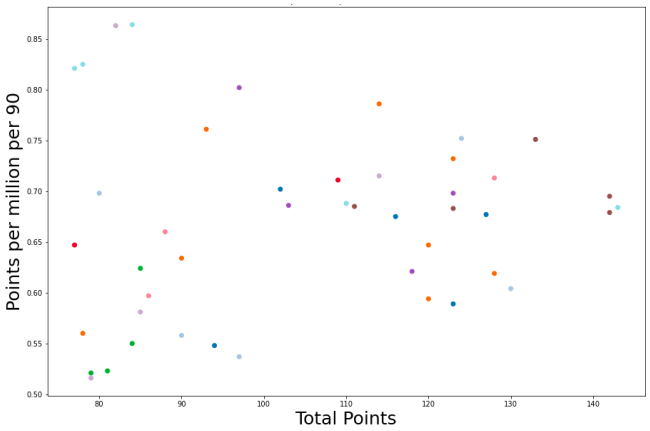

Figure 7: ROI- Defenders

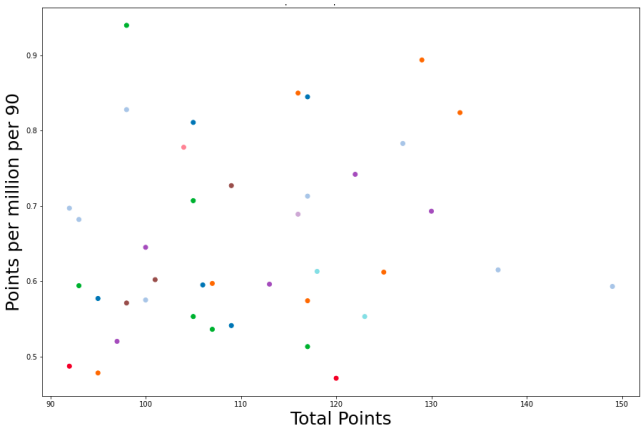

Figure 8: ROI- Midfielders

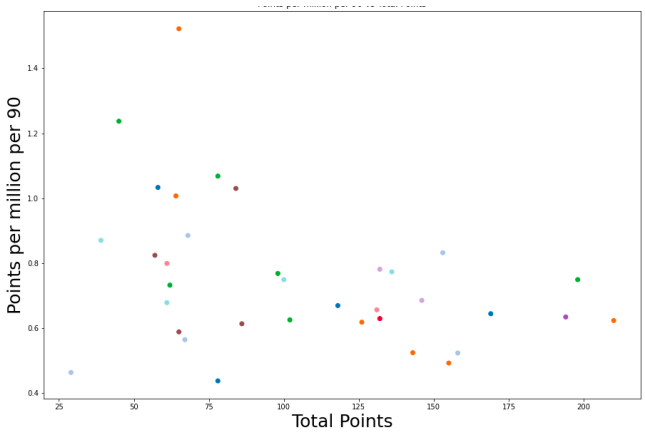

Figure 9: ROI- Forwards

The ROI plots for filtered players on each positions as shown in figures 6,7,8 and 9- based on overall points gives a comprehensive information about who not to miss and who not to select in a team. Based on the plots it is evident that the players that lie closer to the top right are good investments.

3.3 Feature Engineering

After performing EDA and visualizing EDA plot , a threshold is set for players having a history of good performances in each position. The options are now narrowed, which will help in making the team drafting step faster. Before using regression model, feature engineering was performed. The techniques used were: feature removal(using information from correlation matrix), feature scaling, one hot encoding, rolling average of non-deterministic features, etc.

3.4 Points Prediction

One of the most crucial part of this project is performance prediction or points prediction. The prediction is done using regression model, using xgboost regressor. The labeled data was first scaled to Return on Investment(ROI). Four different models were created for the four playing position categories- goalkeepers, defenders, midfielders and forwards. 17-19 features were selected for model training. The number of features varied for different playing positions. Of the features used, around 13 features had skewed distribution. The skewed dataset was treated using log transformation, square and cube root transformation. Other features showing normal distributions were treated with standard and min-max scaling techniques. The normalized features were then used to build the model. For tuning the hyperparameters, a randomized search was performed. Randomized search makes a combination of each of the parameter values- learning rate, max depth, min child weight, gamma, colsample by tree- and finds out for which combination it gives highest accuracy [9, 10]. One shortcoming of this approach was that the time and resources was significantly spent on tuning the hyperparameters. But since our problem is to track the random nature of football and xgboost has known to outperform most other models, it was worth a shot. It was necessary to be wary of overfitting since the nature of data can change dramatially across seasons(eg: due to player transfers mid-season, improvement or decline of form of players and clubs, etc.)[11, 12, 13].

The training was done using data from 2017-18 season to 2020-21 season, and the prediction was done on the 2021-22 season data. For prediction, only those players were subjected whose chances of playing according to FPL dataset was 75 percent or more. This helped filter suspended players, injured players, and players that were unavailable to play for various reasons.

3.5 Transfer Recommender

A transfer recommender is an algorithm that suggests which player to kick out of the team and which to replace him with. The player generating worst ROI in recent game-weeks in the team and highest average fixture difficulty ratio in the upcoming 3 game-weeks is replaced with a player with good ROI and comparitively easier fixture difficulty ratio in the upcoming game-weeks.

3.6 Cup Fixture Feedback

The historical data available had no information about mid-week cup fixtures that the teams play during the course of the season. On a season there are 38 matches each teams have to play, and on top of that teams can play upto 25 extra competitive matches in cup competitions. These fixtures create congestion and may create fatigue in players thus affecting the squad selection policy of manager and the player performance too. These information can be crucial in predicting FPL points of players. Keeping that in mind, the cup fixture information for the premier league teams playing in 21-22 season was collected from the start of 18-19 season until March 2022. This feedback was done to compare results and validate if fixture congestion had a bearing on FPL points.

Home	Away	Date
Young Boys	Man Utd	2018.09.20
Man Utd	Derby County	2018.09.26
Man Utd	Valencia	2018.10.03
Man Utd	Juventus	2018.10.24
Juventus	Man Utd	2018.11.08
Man Utd	Young Boys	2018.11.28
Valencia	Man Utd	2018.12.13
Man Utd	Reading	2019.01.05
Arsenal	Man Utd	2019.01.26
Man Utd	PSG	2019.02.13
Chelsea	Man Utd	2019.02.19
PSG	Man Utd	2019.03.07
Wolves	Man Utd	2019.03.17

Table 1: Sample cup fixtures data

3.7 Evaluation Metrics

For the evaluation of the model, rmse score was used. The prediction was done on Return on Investment, which is in fact a relative measure. Using predicted ROI measure, absolute measure of predicted points was generated and compared to actual points amassed using rmse error value. Apart from rmse, the global FPL ranking was also used to evaluate the overall performance of the team generated by the model.

4. Results and Analysis

The results of regression model were intended for initial team drafting as well as transfer algorithm. The results obtained were interesting. The star performers often generate very high FPL points so they are present towards the tail of skewed histogram. Average and subpar players are concentrated towards the peak. Because of this reason, the predictions of average and subpar players was found better than that of star performers.

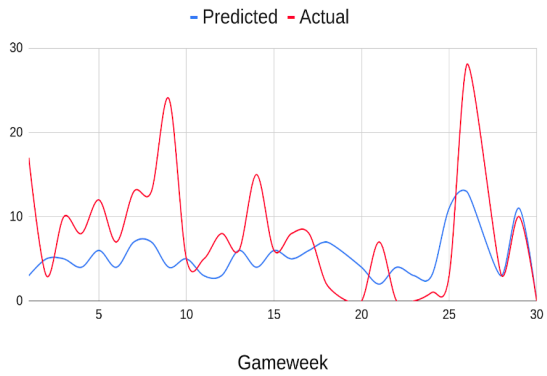

Figure 10: Predicted vs Actual Points- Weekly -Star Performer- Mo Salah

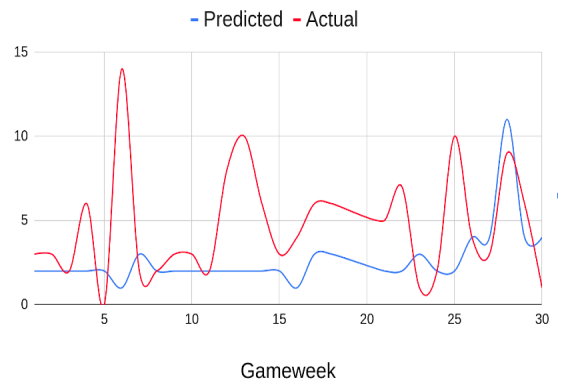

Figure 11: Predicted vs Actual Points- Weekly -Star Performer- Jose Sa

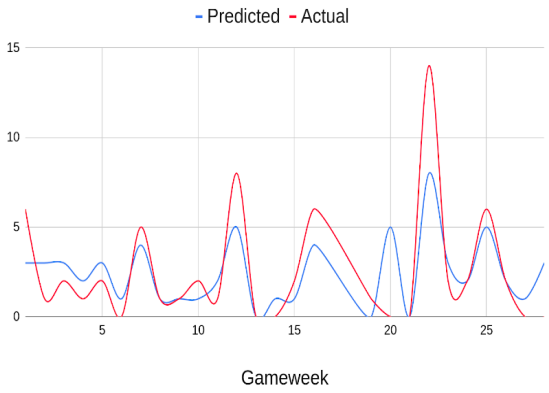

Figure 12: Predicted vs Actual Points- Average Performer- James Tarkowski

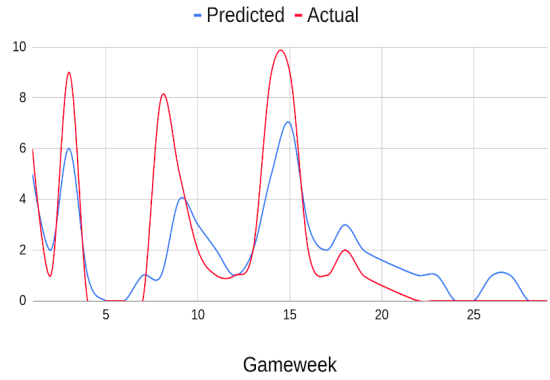

Figure 13: Predicted vs Actual Points- Weekly -Average Performer- Callum Wilson

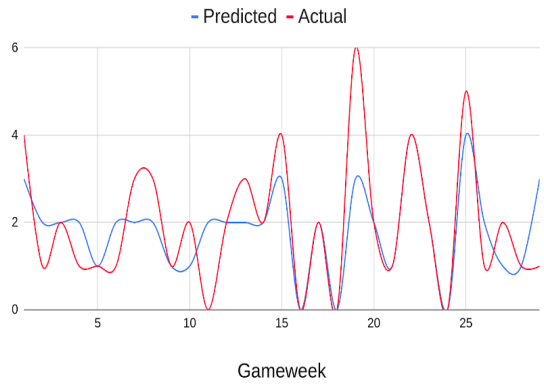

Figure 14: Predicted vs Actual Points- Weekly -Subpar Performer- Jakub Molder

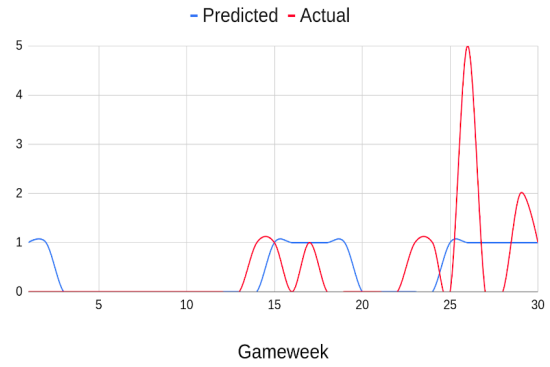

Figure 15: Predicted vs Actual Points- Weekly -Subpar Performer- Edouard Nketiah

Figures 10,11,12,13,14 and 15 show the comparison between predicted and actual points for six players of different calibre. The average rmse score for all players was calculated to be 2.048. Based on the total predicted scores of all 38 gameweeks, a team was drafted with 11 players within 84 million- Jose Sa (GKP), Trent-Alexander-Arnold (DEF), Joao Cancelo (DEF), Conor Coady (DEF), Virgil Van Dijk (DEF), Bukayo Saka (MID), Mohamad Salah (MID), Jared Bowen (MID), Heung Min Son (MID), Bernardo Silva (MID), Michail Antonio (FWD). The team had a formation of 4-5-1, and without making any transfers it generated a total score of 2344. This point gave a global rank of 340,000 out of 9.1 million FPL participants in the season. Over the same initial team draft, 1 transfer was made each game-week using the transfer algorithm. This gave a total point of 2080 with global rank of 2.25 million.

Additionally, we incorporated cup fixtures information into our model and generated point predictions again. The initial team drafted amassed 2351 points when no transfers were made- a negligeable improvement over previous result. Likewise, on making 1 transfer every game-week, based on transfer recommender, 2061 points were generated with a global rank of 2.6 million.

5. Conclusion

The error obtained with regression model was found very high considering the average points accumulated by players each game-week. The prediction on fringe players was more accurate than the star performers because of the skewed nature of the dataset. Since, the results of football are random, more factors need to be accounted for FPL analytics. Despite of using playing chance information from FPL data, often generated teams were shown to have injured and suspended players in the team. Prior notice to such information can be helpful to filter players to consider for a particular game-week. It was also seen that effects of mid-week fixtures is not significant as far as FPL points is concerned.

To improve predictions, use of information of injuries from news portals, manager press conferences, etc can be crucial. So, the use of Natural Language Processing to filter out unavailable players can help choosing players for consideration into FPL team for that game-week. Over the project we've never used the available bonus chips which are intended to improve total points. A proper algorithm to predict the best possible scenarios for using such bonus chips can be developed as enhancement to improve total points.

References

[1] Kaushal Shukla. Fantasy premier league, gameweek 1: Teams with best fixtures, players to buy and captaincy options, Aug 2021.

[2] Alun Owen. Dynamic bayesian forecasting models of football match outcomes. 01 2009.

[3] Roshan Thapaliya. Using machine learning to predict high- performing players in fantasy premier league, Mar 2017.

[4] F Bonomo, G Dur'an, and J Marenco. Mathematical programming as a tool for virtual soccer coaches: a case study of fantasy sport game. *International Transactions in Operational Research*, 21:399–414, Dec 2013.

[5] Tim Matthews, Sarvapali Ramchurn, and Georgios Chalkiadakis. Competing with humans at fantasy football: Team formation in large partially-observable domains. *Proceedings of the National Conference on Artificial Intelligence*, 2, 01 2012.

[6] Nicholas Bonello, Joeran Beel, Seamus Lawless, and Jeremy Debattista. Multi-stream data analytics for

enhanced performance prediction in fantasy football. 2019.

[7] Fréderic Godin, Jasper Zuallaert, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. Beating the bookmakers: Leveraging statistics and twitter microposts for predicting soccer results. 08 2014.

[8] Vaastav Anand. FPL Historical Dataset. Retrieved August 2022 from https://github.com/vaastav/Fantasy-Premier-League/, 2022.

[9] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 12 2013.

[10] Ramraj Santhanam, Nishant Uzir, Sunil Raman, and Shatadeep Banerjee. Experimenting xgboost algorithm for prediction and classification of different datasets. 03 2017.

[11] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 08 2016.

[12] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of xgboost, 11 2019.

[13] Tom Mitchell. *Machine Learning*. McGraw-Hill Education, 1997.