Optimized selection-based technique of Score Prediction for Fantasy Premier League

Project Proposal Report - CSE-519: Data Science Fundamentals - Fall 2021 - Stony Brook University

***Abstract*—In Fantasy Premier League (FPL), managers create teams with the goal of scoring highest number of points. In this report, we propose a model of team creation with the aim to maximize the points under certain FPL constraints. To start with, we considered a baseline model to create a team of 15 players and devise swapping strategies for imminent maximum score further extended for the complete season. Certain factors observed that affected the scores include position of players in respective teams, cost of teams, and importance of player transfers. Finally, we discuss the directions that could optimize the score predictions over the baseline model. This optimized model incorporates weighted ranking techniques, regression algorithms for best feature selections and time series analysis to optimize the model extended over the entire season. The model generated scores are validated against the actual league scores and leader-board.**

I. Introduction

THERE is a growing interest in virtual games where people create teams and manage the teams to win points. A certain number of highest scorers then go on to win attractive cash prizes. They are usually modelled after real sporting events. This phenomenon is seen across many sports and is generating huge revenues. Fantasy Premier League or FPL is one such fantasy sports game which is played in sync with England's top most football league i.e the English Premier League. Participants of FPL create football teams based on certain constraints and compete with players from different parts of the world to win a cash prize of 100 million pounds.

In Fantasy Premier League, a manager must pick a team of 15 players at the start of the tournament. Here, a manager refers to the person participating in FPL to win the cash prize and a player refers to a real life football player who is playing in the English Premier League. While creating the team, the manager cannot choose more than 3 players from a club and each team must comprise of exactly 2 goalkeepers, 5 defenders, 5 midfielders and 3 forwards. A budget of 100 million pounds is provided at the start of the game. The manager has to decide on a playing XI which must comprise of atleast 1 goalkeeper, 3 defenders, 3 midfielders and 1 forward. Points are awarded based on the performance of players in a game week. A captain has to be chosen and the captains points are doubled in a gameweek. Also, top 3 players in each match are rewards bonus points between 1-3.

The manager is allowed free 1 swap per game week. A manager can choose to carry forwards the swaps but a manager can save upto 2 swaps only. Additional swaps are allowed over the free swaps but there is a penalty of 4 points per swap. If a team buys a player for 5 million and later the price of the player increases to 6 million and is swapped out, it leads to a increase of budget by 1 million. Gamechips are also available to the managers. These gamechips can be used only once throughtout the season except the wildcard and only 1 particular game chip can be used in a gameweek. Gamechips such as wildcard allows a manager to replace his entire team while a benchboost allows a manager to receive points for all 15 players in the team. Freehit chip is similar to wildcard with the exception that the team is restored back to original in the following week. The triple captain allows the points of the captain to be tripled.

Analysis on complete GWs can serve multiple purposes. Post GW analysis could give insights such as which players scored the best or which transfers boosted the team scores. It can help devise a strategy of selecting the best playing XI for the future GWs.

II. Problem Statement

The objective is to choose 15 players and select a playing XI amongst them such that the team is able to win as many points as possible. The points are won based on the performance of the players in a particular fixture in the game week. These points can be acquired based on certain factors such as the number of goals scored, the number of assists etc. Points can also be deducted for reasons such as own goals, red cards, yellow cards, missed penalty, and conceding goals. The team has to be created based on the criterion mentioned in the introduction. The task of predicting the teams entails selecting the best players. These best players have to be chosen based on past performance , pricing, budget constraint and several other features that affect their scores along with the fitness of the player. Another influential factor for player selection is the fixtures of matches over the period of the entire season. Players could be swapped based on the factors affecting the next game week. These factors include away games and fixtures of the next game week.

III. Data and Exploratory Analysis

We have fetched the data from FPL's official API and Github[1]. The dataset itself is very comprehensive and well self-contained. We have data for each Game Week(GW) containing information about each player's performance, and cost, minutes played, and match statistics. Some interesting fields are numbers of transfers in and out of a player in each game week, number of selections of a player in their team in

that game week, influence, creativity and threat of a player in a match, and how a particular FPL team has performed that Game Week. Some important correlations(Fig 1):

1) No. of selections of a player highly correlates with the points scored.(0.359).
2) More the cost of a player, more the chances of scoring big.(0.364).
3) Some teams perform better in home matches than in away matches because of the supporting crowd. But if we see player's average scores on home and away matches, it's not much correlated.
4) Defenders play more average number of minutes per match.

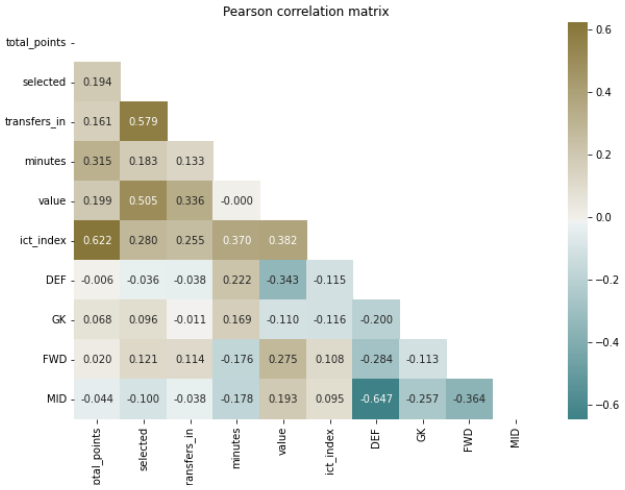

Fig. 1. Correlation matrix of features.

IV. Approach and Validation

The direction of the approach is to maximize the scores for a model that participates the entire season. The initial step is to create a model that can make selection for the initial squad of 15 players before the lock-in period begins for Game Week 1 (GW1). Once the squad is decided and locked in, the next phase is to select the playing 11 for the GW1. For the next Game Week, since the swapping window opens up, one player would be swapped based on a devised strategy.

The final phase is once after every GW completes, the scores are calculated for the selected players and the evaluation matrix will be updated. These evaluated data points would be used for Time Series Analysis to predict the scores for subsequent game weeks. This way optimum selection of players based on historical data points and the future fixtures is used to maximize the cumulative score of the season.

Baseline Model:

The baseline model has been formulated based on the exploratory analysis of the data and does not involve training on the data set. The steps for prediction on the main problem remains same, but the algorithm and selection strategy changes for baseline. This is to understand how the model efficacy can be increased by employing better techniques and algorithms.

(i) ***Ranking and selection of players***:

The player selection is based on two factors or features that are closely related to the overall objective of maximizing scores on constrained budget.

The attributes used are past *Performance (pf)* and the *Cost (ct)* of the Player. The aim is to process these factors to better analyze and form the ranking strategy of the players [3]. Using these factors, we calculate an index, called *Performance-Cost-Index (pci)*.

$$pci_i = \frac{pf_i}{ct_i} \qquad \forall i \in players$$

Based on the index calculated, the players will be ranked in top-bottom order.

$$ranks = [r_i, r_{i-1} \ldots] \qquad \forall i \in players$$

Once the rank is calculated for the players, the next step is to form a squad of 15 players that can be formed for the budget limit of £100M and defined constraints on positions of players.

This is a variation of classic Knapsack Optimization problem. The aim is to form the squad by using the limit of budget and best possible ranked player selection.

After the squad of 15 is formed, the selection of playing 11 is done solely based on the positions and better rankings of players in each positions.

(ii) ***Swapping of player at end of GW***:

For the baseline model, the player swapping strategy would be based on swapping the player with lowest index for actual score against predicted score i.e. depicting that the player did not perform well as expected. Let this player be called $player_{out}$.

The swapping would be performed based on the position of the player that is being removed from the squad. $player_{out}$ will be swapped against the player having next best pci in the pool of players in same position.

Once the swapping is complete, the selection strategy for deciding playing 11 remains same as discussed in (i).

(iii) ***Extrapolating baseline GW model for entire season***:

As explained in (i) and (ii), the formulated model would be calculated at the beginning of GW1 and extended for subsequent game weeks. The idea is to extend the same model and track the performance in terms of predicted scores for each game week against the actual scores of top players in FPL leaderboard.

The predicted scores $scores_{pred}$ can be calculated for the comparison.

$$scores_{pred} = [score_j, score_{j-1} \ldots]$$
$$\forall j \in predictedscores$$

The validation metric used can be Root Mean Squared Error (RMSE) to understand how the predicted scores have fared against the actual scores.

Further Enhanced Model:

The baseline model shows the idea of using seemingly strong contributing factors to make the selection and decision strategies. The overall approach is Naive and uses Greedy approach for the most parts.

However, further analysis of the data shows that there are many other features that have strong correlation and effect on the scores, depicting high probability of shift in model evaluation and results.

The enhanced model is targeted to incorporate multiple features and would use better ranking and optimization algorithms [2].

- Ranking of players and selection of 15 player squad[4]: Since there are multiple features having the potential of affecting the scores of players and in turn the overall score, we would be using regression models to generate the ranking index for players.

  The selection of 15 players out of the ranked players would follow the same Knapsack optimization approach.

- Selection of playing 11 out of 15 players squad: The selection of playing 11 will be different from the baseline model. The strategy here is to select the players based on the fixtures of the upcoming GW.

  The factors to consider here would be whether the fixture is home game or away game for the player and strength of player against the respective opponent based on past performance.

- Swapping strategy after completion of GW: There are two parts to this strategy. Firstly, to decide which player would be swapped out of the squad. For this, an index based on average scores per game and the cost of the player will be considered. The player whose index is low and probability of low performance based on next GW fixture, will be swapped out.

  The player to be swapped-in will be considered based on the same metric i.e. based on fixtures of next GW and also that fits the constraints of the budget available and the position to be replaced [6].

- Optimizing the model for subsequent GWs: The model predicting scores for a GW and next GW does not efficiently considers the performance of players in and out of the squad in the present season.

  To generalize the model, the key parameters that govern the selection of a player in the teams, including transfer in, transfer outs, increase/decrease in cost of players and the ICT indices for each players are considered as data points.

  These data points can be analyzed in terms of each GWs over the length of the season. Since, there is a development in these points over time, Time Series Modelling [7] could be used to optimize the points, resulting in improvement in score predictions against actual scores.

- Validation would be performed by comparing the scores predicted for the players and their actual scores over the completed GWs. The evaluation metric to be used for the validation is Root Mean Squared Error (RMSE).

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{predictedScore_i - actualScore_i}{\sigma_i}\right)^2}$$
$$\forall i \in Gameweeks$$

V. Visualizations and Initial Findings

We have done some visualizations to observe how the FPL teams perform wrt their players' cost. We can exploit such cases while building our team and give weight-age based on team's performance[5]. Findings(Fig 3):

- Spurs forwards are amazing and give best returns among all but costs a fortune as well. The Spurs Goalkeeper is cheap but is worth taking.
- Leicester's attack is cheap and the probability of scoring is more.
- The midfielders of Man City are league above the others but cost hefty as well.
- Aston villa's midfielders are way cheaper and provide good points.
- Don't go for Arsenal and Spurs defenders.
- Chelsea Defenders all the way. So Chelsea's GK would be considered for more points since there's higher probability of clean sheets.

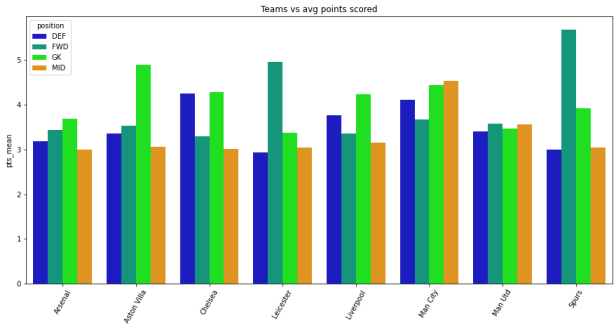

Fig. 2. Avg points scored by top 8 teams per GW at each position.

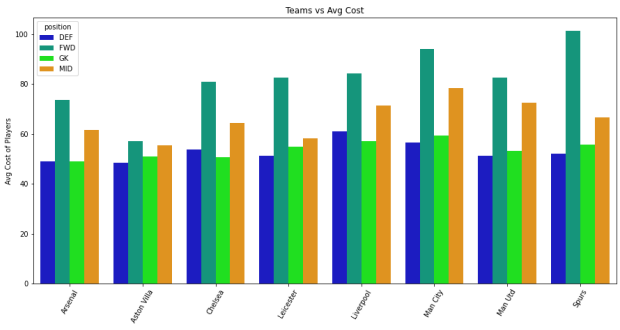

Fig. 3. Avg cost of players of top 8 teams per GW at each position.

Also, we have tried to visualize the impact of swapping strategy. Suppose, in the first week itself we create the best

possible team and hit the jackpot. If we don't change our team, let's see how this team fares every GW. Our findings as follows. Findings(Fig 4):

- Total number of points accumulated by this team keeps on decreasing and hits the average point per player of 2.66 next week itself from the massive average of 12 points per player.
- Thus, the total ICT index of these players starts dropping soon.
- Seeing the 1st week's performance, there had been multiple transfer-ins for such players which subsequently dropped every week as their avg points hadn't been good enough. Subsequently, a lot of people have started transferring them out from the team.

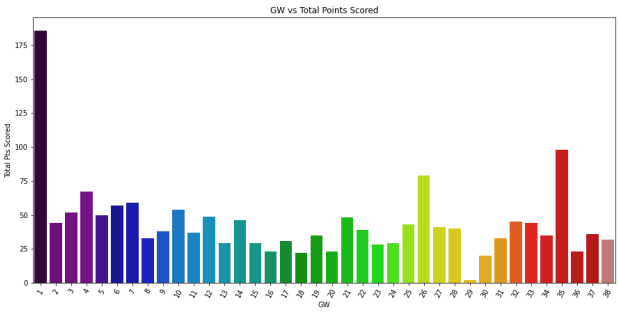

Fig. 4. Total points accumulated by the best players of GW1 over the season.

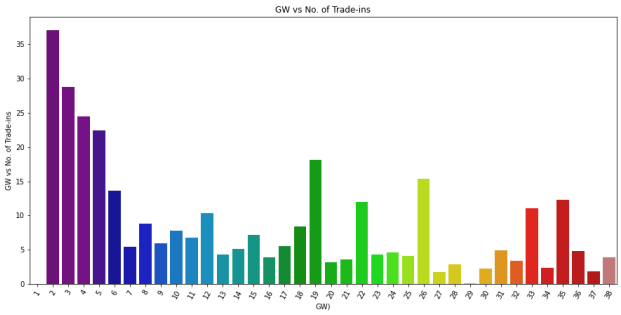

Fig. 5. Number of Trade-ins for the best players of GW1 over the season.

REFERENCES

[1] Datasets:fantasy.premierleague.com/api; https://github.com/vaastav/Fantasy-Premier-League

[2] William Eilertsen, Akash Gupta, Bjørn Kåre Kristiansen, "Developing a Forecast-Based Optimization Model for Fantasy Premier League", Industrial Economics and Technology Management, June 2018

[3] Matthews, T., Ramchurn, S.D., Chalkiadakis, G.: Competing with Humans at Fantasy Football: Team Formation in Large Partially—Cbservable Domains. AAAI (2012)

[4] Kunj Mehta, Fantasy Premier League x Data Analysis: Being Among the Top 2%, https://towardsdatascience.com/fantasy-premier-league-x-data-analysis-being-among-the-top-2-98a714a1d170, Feb 2021

[5] Dr Nick Barlow, Dr Angus Williams, A machine learning manager for Fantasy Premier League, https://www.turing.ac.uk/news/airsenal

[6] Nicholas Bonello, Joeran Beel, Seamus Lawless, Jeremy Debattista. "Multi-stream Data Analytics for Enhanced Performance Prediction in Fantasy Football". In 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science. 2019

[7] Gupta, Akhil. "Time Series Modeling for Dream Team in Fantasy Premier League". International Conference on Sports Engineering ICSE-2017, 23-25 October 2017, Jaipur, India