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

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

I. Project Overview

FANTASY PREMIER LEAGUE (FPL) is an online fantasy game wherein managers create teams with the goal of scoring highest number of points. The basis of scores is on the real life match results and performances of the players. Here, a manager refers to the person participating in FPL to win the the league 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 at least 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 Game Week (GW). 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 throughout 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 bench boost 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.

II. Current Progress

In the proposal report, we presented a model of team formation with the aim to maximize the points under certain FPL constraints and achieve a higher rank. To start with, we proposed a baseline model to create a team of 15 players and devise swapping strategies for imminent maximum score further extended for the complete season. Significant factors observed that affected the scores includes position of players in respective teams, cost of teams, importance of player transfers and the dynamics of the team. Finally, we discussed the directions that could optimize the score predictions over the baseline model. This optimized model would incorporate weighted ranking techniques, linear optimization for the team selection and time series analysis to optimize the score prediction by taking the past GWs scores and future fixtures into consideration. The model generated scores would be validated against the real life league scores and leader-board points aggregation.

The progress on the project till now has been on creating and evaluating our baseline model as planned in the proposal report. With more analysis, we could recollect more interesting findings and challenges ahead in the project. A better understanding of an optimized model of the solution approach has been stated in this report.

A. Implementation of Baseline Model

As stated in the proposal report, we formulated and implemented the baseline model to construct a benchmark for comparison of the model performance. This also allowed us to understand how naive approach for team selection and formation performs over the entire league. The data used for the model creation is for the seasons from 2015-2020. The model evaluation was performed against actual data from the season 2020-2021.

To restate, the factors used in the indexing of the players were as follows. 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)*.

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

The historical data of the players from the aforementioned period is quite extensive. Since the merged data includes players' performances in terms of their scores, the games played by each player and the matches between teams.

For the analysis of past performance of the players, we used the the scores per games played rather than using the raw score data. This was done to standardize the judgment scale of how good or bad a player performs. In other words, if a player plays large number of games but was able to perform extremely well in only a handful of them, could be perceived as high-performing player. In contrast, a consistent player who performed well in average over the games played, could be a better choice in team selection.

B. Ranking and grouping of players

Based on the above index and average scores of the players, a grouping was performed on the positions of players in the team. After grouping, the players were ranked in top down order to select the best players in each group.

C. Team Selection using Linear Programming

Once the rankings are calculated for the players, the next step was to form a team of 15 players that can be formed for the budget limit of £100M and defined constraints on positions of players & team restrictions. This is a variation of classic Knapsack Optimization problem. The aim was to form the team by using the limit of budget allowed and best possible ranked player selection.

Since there is a linear relationship between the constraints for team formation, this becomes a linear optimization problem and hence linear programming was applied to optimize the team selection.

Linear Programming consists of an objective function (consisting of a number of variables) that is maximized or minimized under the given set of constraints [1]. Here, the constraints were on the cost of the player than can be selected in the team against the available budget & the position of the players i.e. the limits on the number of players from each position that can be included to form the team.

The top rewarding player out of the selected 15 players is made the captain, since the scores of the captain are doubled. The least rewarding players are considered as substitutes depending upon the position constraints.

D. Swapping Strategy

For the current baseline model, we devised the following strategy for making swaps of players after completion of a GW.

The fixtures of next GW and the rankings of the teams is taken into consideration. The team with higher ranking would tend to dominate team with lower rank, and so the players from a higher ranked team would have advantage if he plays against such team in next GW. We assigned a scaled factor to each player's score based on above understanding and then decided which player should be transferred out.

Once decided, using the linear programming model again, we find the player that could be transferred in based on the available budget and free position.

E. Evaluation of Baseline Model

The above explained baseline model provides us a team with predicted high rewarding players for the GW 1. To evaluate this baseline model, we extrapolate the same team till the entire season, with swapping strategy in place and without.

Firstly, without adopting any swapping strategy, we compare the baseline model predicted points per game week for the season against the team suggested by the linear programming model using the actual scores of the players for subsequent GWs. The results and analysis have been explained in the next sections.

Secondly, we adopt a swapping strategy to swap a player in a GW under the allowed constraints and then evaluate how does this perform over the subsequent game weeks. Again, the analysis has been illustrated in the next section.

III. Intermediate Results and Analysis

The results and analysis from the baseline model performance are illustrated in this section.

The baseline model generated the scores of the players for the GW1 based on the past performances and the cost of each player. Once we had these rankings we selected 15 players based on the constraints provided on the position and the budget, by applying linear optimization. The selected players, their expected scores, costs and their positions have been listed as below.

```
Our Final playing XI as predicted by the model:
Name= Bruno Miguel Borges Fernandes, Expected Points = 8.357, Cost = 8.6071M$, Position = MID
Name= Dean Henderson, Expected Points = 4.21, Cost = 4.8657M$, Position = GK
Name= George Baldock, Expected Points = 3.736, Cost = 4.821M$, Position = DEF
Name= Jamie Vardy, Expected Points = 4.789, Cost = 9.2335M$, Position = FWD
Name= John Lundstram, Expected Points = 3.789, Cost = 4.7526M$, Position = MID
Name= Luke Thomas, Expected Points = 5.0, Cost = 4.0M$, Position = DEF
Name= Matt Doherty, Expected Points = 4.092, Cost = 5.5828M$, Position = DEF
Name= Mohamed Salah, Expected Points = 6.973, Cost = 11.8578M$, Position = MID
Name= Pierre-Emerick Aubameyang, Expected Points = 5.522, Cost = 10.9333M$, Position = MID
Name= Raheem Sterling, Expected Points = 5.333, Cost = 9.9065M$, Position = MID
Name= RaÃºl JimÃ©nez, Expected Points = 4.934, Cost = 6.9447M$, Position = FWD

Captain:
Name= Bruno Miguel Borges Fernandes, Expected Points = 8.357, Cost = 8.6071M$, Position = MID

List of Substitutes:
Name= Issa Diop, Expected Points = 2.144, Cost = 4.3657M$, Position = DEF
Name= John Egan, Expected Points = 3.5, Cost = 4.5289M$, Position = DEF
Name= Karlan Grant, Expected Points = 3.214, Cost = 5.0M$, Position = FWD
Name= Mathew Ryan, Expected Points = 3.377, Cost = 4.5701M$, Position = GK
```

Fig. 1. Team selected by the baseline model

Next we extrapolated the selected team of players over a period of 36 GWs for the FPL season 2020-21. The challenge in order to make the comparison with the model performance was unavailability of actual top leader-board teams. To tackle this, we used the actual scores of the players, selected top players and applied linear programming to select a team whose performance could be compared to our baseline model.

This team, could be called a "dream" team, as it could be possibly even better than the actual teams on leader-board as it's being formed using the actual scores and linearly optimizing over the constraints. The difference in the predicted and actual scores in the graph depict the same.

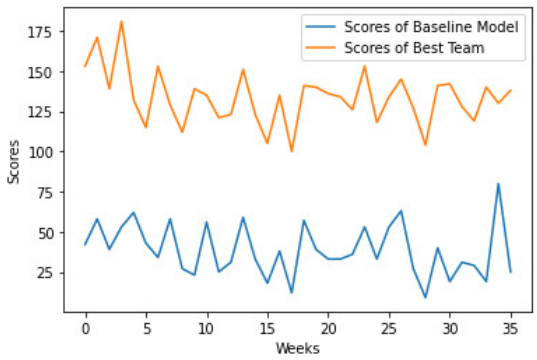

Fig. 2. Scores of Best Team Vs Scores of Team Selected by Baseline Model for Season 2020-21

IV. Transfer Strategy in Baseline Model

In out transfer strategy, we are considering both the cases, with and without penalty. Without penalty, we are considering only one transfer and we are considering penalty in such cases where we definitely need to include a player who is in red hot form and is a easy home match in the next game week. As per the expectations scores, to swap a player, we are giving a penalty to the all players in our Playing XI who have a tougher away opposition match and a reward to players who have an easy home match.

```
Swapping for GameWeek1:
Transferred in: Name= CÃ©sar Azpilicueta, Expected Points = 5.164, Cost = 6.3407M$, Position = DEF
Transferred out: Name= Issa Diop, Expected Points = 2.144, Cost = 4.3657M$, Position = DEF
Transferred out: Name= John Egan, Expected Points = 3.5, Cost = 4.5289M$, Position = DEF
Transferred in: Name= Marcos Alonso, Expected Points = 5.046, Cost = 6.5892M$, Position = DEF
Transferred in: Name= Martin Dubravka, Expected Points = 3.674, Cost = 4.937M$, Position = GK
Transferred in: Name= Mason Mount, Expected Points = 4.605, Cost = 6.35M$, Position = MID
Transferred out: Name= Mathew Ryan, Expected Points = 3.377, Cost = 4.5701M$, Position = GK
Transferred out: Name= Pierre-Emerick Aubameyang, Expected Points = 5.522, Cost = 10.9333M$, Position = MID
```

Fig. 3. Swaps for a particular GW where Chelsea has an easy fixture

```
Final team after swapping for GameWeek1:

Playing XI:
Name= Bruno Miguel Borges Fernandes, Expected Points = 8.357, Cost = 8.6071M$, Position = MID *
Name= Dean Henderson, Expected Points = 4.21, Cost = 4.8657M$, Position = GK
Name= George Baldock, Expected Points = 3.736, Cost = 4.821M$, Position = DEF
Name= Jamie Vardy, Expected Points = 5.289, Cost = 9.2335M$, Position = FWD
Name= John Lundstram, Expected Points = 3.789, Cost = 4.7526M$, Position = MID
Name= Luke Thomas, Expected Points = 5.5, Cost = 4.0M$, Position = DEF
Name= Matt Doherty, Expected Points = 4.092, Cost = 5.5828M$, Position = DEF
Name= Mohamed Salah, Expected Points = 6.973, Cost = 11.8578M$, Position = MID
Name= Pierre-Emerick Aubameyang, Expected Points = 5.522, Cost = 10.9333M$, Position = MID
Name= Raheem Sterling, Expected Points = 5.333, Cost = 9.9065M$, Position = MID
Name= RaÃºl JimÃ©nez, Expected Points = 4.934, Cost = 6.9447M$, Position = FWD

Substitutes:
Name= Issa Diop, Expected Points = 2.144, Cost = 4.3657M$, Position = DEF
Name= John Egan, Expected Points = 3.5, Cost = 4.5289M$, Position = DEF
Name= Karlan Grant, Expected Points = 3.214, Cost = 5.0M$, Position = FWD
Name= Mathew Ryan, Expected Points = 3.377, Cost = 4.5701M$, Position = GK

Budget Left: 0.03

Total expected score = 68.527
```

Fig. 4. Final Playing XI after transfer and expected points after penalty

In the above example, we have seen that Chelsea had the easy fixture against Fulham and Chelsea's defenfer are in red hot form and their record against Fulham is amazing. Our model has tried to give a bigger reward to Chelsea's player and tried to include all such players. Even after the penalty of each extra transfer(which is 2, Our model has predicted a score of 68.5 for this team. (76.5-4*2))

```
Transferred in: Name= Mohamed Salah, Expected Points = 6.973684210526316, Cost = 11.8578M$, Position = MID
Transferred out: Name= Harry Kane, Expected Points = 4.993421052631579, Cost = 11.8302M$, Position = FWD
```

Fig. 5. One swap for next GW

```
Final team after swapping for GameWeek1:

Playing XI:
Name= Bruno Miguel Borges Fernandes, Expected Points = 8.357142857142858, Cost = 8.6071M$, Position = MID *
Name= Dean Henderson, Expected Points = 4.210526315789473S, Cost = 4.8657M$, Position = GK
Name= George Baldock, Expected Points = 3.736842105263158, Cost = 4.821M$, Position = DEF
Name= Jamie Vardy, Expected Points = 4.789473684210526S, Cost = 9.2335M$, Position = FWD
Name= John Lundstram, Expected Points = 3.789473684210526, Cost = 4.7526M$, Position = MID
Name= Luke Thomas, Expected Points = 5.0, Cost = 4.0M$, Position = DEF
Name= Matt Doherty, Expected Points = 4.092105263157895, Cost = 5.5828M$, Position = DEF
Name= Pierre-Emerick Aubameyang, Expected Points = 5.522222222222222, Cost = 10.9333M$, Position = MID
Name= Raheem Sterling, Expected Points = 5.333333333333333, Cost = 9.9065M$, Position = MID
Name= RaÃºl JimÃ©nez, Expected Points = 4.934210526315789, Cost = 6.9447M$, Position = FWD
Name= Harry Kane, Expected Points = 4.993421052631579, Cost = 11.8302M$, Position = FWD

Substitutes:
Name= Issa Diop, Expected Points = 2.1447368421052633, Cost = 4.3657M$, Position = DEF
Name= John Egan, Expected Points = 3.5, Cost = 4.5289M$, Position = DEF
Name= Karlan Grant, Expected Points = 3.2142857142857144, Cost = 5.0M$, Position = FWD
Name= Mathew Ryan, Expected Points = 3.3771929824561404, Cost = 4.5701M$, Position = GK

Budget Left: 0.06

Total expected score = 67.039
```

Fig. 6. Final team after a swap

V. Next Steps

As the performance of the baseline model can be analyzed, the prediction model clearly doesn't fare well against the actual points scored by managers and the scores of players. This is due to the reason that the baseline model works on a greedy approach and utilizes only the factors of past performance and cost of the players. The next step in the project is to optimize the model and use better co-related features to improve the accuracy.

- The optimized model would be considered using the features including the past scores of players, cost, performance of players in home vs away games, ict index, player's proneness to injuries, minutes played and the team dynamics. These features have been observed to have a strong correlation with the players' scores.
- The team selection model using linear optimization would remain the same. For added advantage, we can include the weighing of the players' scores, on deciding the playing 11 out of the selected squad of 15.
- Swapping strategy can be improved to better decide on the players transfers. This can be done by incorporating next GWs fixtures, players and team dynamics against opponents in upcoming GW, the home vs away fixtures and the current scores of the players in the season.
- Exploring and implementing Time Series Analysis [2] over the problem set using the feature set as stated above. We plan to explore and implement Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) for prediction of scores considering the moving window of time (GWs). We aim to benefit from the performance of LSTM-RNN and its advantage as a model with features having non-linear relationship.
- Evaluating the results and performance of our model against the actual scores and points of top managers in the FPL season of 2020-21 and 2021-22. This includes generating evaluation metric for the model, thereby giving better understanding of how the model actually performed against real life data.

VI. Challenges Ahead

There are certain challenges that lie ahead in the plan for this project affecting the model formation and points prediction.

- Since we are using the historical data of past couple of seasons, the performances of the players varies and it is understandable. However it is a challenge to accurately use the data when using for time series model, e.g. if the player performed very well in five seasons back but has not been able to do well since then, the model should ideally give lesser consideration while predicting scores.
- Another challenge is selecting and deciding on the threshold of budget to be consumed for substitute players, while forming a team. This is because if one allocates very low budget for substitutes, they may not select good players and if higher budget is allocated, then that may come at cost of better non-substitute players. Some order of "fudge" factor needs to be figured out to resolve this.
- The next challenge would be to formulate and analyze the patterns in values for the features such as ict index, transfers-in, transfers-out, changes in cost of players and team dynamics, to be used for time model. Since there is a non-linear relationship between the attributes, some form of preference or weight factor must be used so that the contribution of these factors is not skewed.

References

[1] Schulze, Mark. (2000). Linear Programming for Optimization
[2] 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