

Team Recommendation System for Fantasy Premier League

Team Members:

Mahmoud Kamel Shoush

Mohamed Maher Abdelrahman

Nesma Talaat Mahmoud

1. Business understanding

1.1. Background

Fantasy football is a fun, competitive, and addictive hobby. Get your fantasy season started by drafting players to build a solid team and using coaching tips to keep your team going strong. Fantasy Premier League is competition for Soccer fans to choose their team of players based on different criteria. The success of your fantasy football team hinges on your draft picks, and there is a different strategy to pick your draft. The main target is to minimize risk, maximize gain, and make the tough decisions when it comes to fantasy football draft day to choose the most useful team at the start of any league to get a high score above the average score of all fantasy users, we are concentrated on premier league. The English Premier League is the most-watched sports league in the world, it has billions of fans. Our targeted segment is the soccer fans which is a wide segment because they are billions.

1.2. Business goals

- Ensure the quality of our predicted team each week with the allowed team budget and constraints by fantasy EPL game.
- Helping Fantasy Players to achieve higher scores and make correct team selections.
- The growth of our project to participate in the game and get a higher score than the average score.

1.3. Business success criteria

- Helping fantasy fans in making their line-up selections to improve their scores.

- Helping football managers to predict the performance of either their own players or the opponent team players.

1.4. Inventory of resources

People: Three Data Science Course Students.

Dataset: ~232MB dataset collected from different sources

Hardware: 3 personal computers with 16GB Ram and 8 cores for processing.

1.5. Requirements, assumptions, and constraints

Requirements:

The effort from each resource in the project to work at least 24hr. Also, We need to achieve a team score that exceeds fantasy premier league average users scores each week consistently.

Assumptions:

The collected dataset is fair enough to predict player performance upon it and to predict the best team in the upcoming weeks.

Constraints:

We need to construct a team from all players in the premier league teams where this team can achieve a score above the average of all fantasy users consistently and avoid forming a team with (1. At least 3 players of the same team (Example: we can't form a team of 4 players from the Liverpool team) 2. Having a team budget of less than 100 units according to fantasy premier league goals.

1.6. Risks and contingencies

People – key people unable to contribute to the project because of, for example, illness or career change.

Technical – The proposed algorithms didn't result in a good accuracy

1.7. Terminology

EPL: English Premier League

1.8. Costs and benefits

Costs:

-- Not applicable--

Benefits:

- Being able to predict player performance each week can help managers to make their choices and form up their team or take care of some players of the opponent teams.
- Increasing chance of being able to have a high score in fantasy premier league as top-ranked users worldwide can **win valuable prizes**.

1.9. Data-mining goals

- Dataset Collection
- Dataset Filtration
- Feature selection
- Develop a model to predict the best team for the upcoming week
- Model Evaluation
- Visualization

1.10. Data-mining success criteria

- Being able to form a team line-up that achieves a high score consistently based on the players predicted scores each week. We aim to achieve a score higher than the average score of fantasy EPL users.

2. Data understanding

2.1. Gathering data

2.1.1. Outline data requirements:

We need to use the available data resources from datasets available on Kaggle and Fantasy Premier League API to get intuitive features about player performance in the past weeks and that enables us to predict the player score in the coming week.

2.1.2. Verify data availability:

Multiple datasets are already available on Kaggle [\[1\]](#), [\[2\]](#), [\[3\]](#) that can be used to extract the features needed to build the dataset. These datasets contain real data and statistics about teams and players from the past premier league and current seasons. In case of missing some data for a specific season, the official fantasy premier league website has an [API](#) that can offer all the history of data and it can be used to complete all missing features.

2.1.3. Define selection criteria:

We have chosen some features to be calculated for each player in order to be able to predict his performance. These features can be divided into three categories:

2.1.3.1 Player History:

Player Position (Forward, Midfielder, Defender, Goalkeeper)

Ratio of Goals Scored/ week in season for player

Player Goals scored last 3 weeks

Player Yellow Cards last 3 weeks

Ratio of fouls/week made by the player

Ratio of Goals Scored/ week in season for player

Ratio of Assists/ week in season for player

Ratio of player points/week in last 3 weeks

Ratio of player points/week in last season

Ratio of minutes played/week in current season

Ratio of saves/week for the player in the current season (GK)

Ratio of clean sheets for player own team in the last 10 matches

Ratio of fouls/week made by player

Ratio of tackles won/ week made by player

2.1.3.2 Own Team History:

Player Team Rank

Ratio of Goals scored by team/week in season for own team

Ratio of Goals conceded by team / week in season for own team

Ratio of wins/losses for player own team in the last 10 matches

2.1.3.3 Opponent Team History:

Opponent Team Rank

Ratio of Goals scored by team/week in season for the opponent team

Ratio of Goals conceded by team / week in season for the opponent team

Ratio of clean sheets for the opponent team in the last 10 matches

Ratio of wins/losses for the opponent team in the last 10 matches

2.2. Describing data:

Statistical features need to be collected about the player performance himself in the past weeks (Ratio of his goals/ assists/ clean sheets/etc.). in addition to some features about his team (win/loss rate/e.t.c) and the opponent team.

The number of features is ~30 of a numerical data type. The dataset has around 43323 instances which are the number of all players in all fantasy teams of the EPL multiplied by the number of weeks in the league in the last three seasons

2.3. Exploring data:

All the available datasets contain real data about players performance in each week. For example, we can get how many goals, assists, tackles won, yellow card, red cards, e.t.c made by Player X in week N. In addition, history of records of matches between teams are also available that can help in extracting features about team records and ratio of wins/losses of the team. Most of the columns are pretty self-explanatory. we have some historical data for the player performance in previous seasons, as well as future match fixtures. Also, since a lot of this data is used to drive the UI of a Web Application, there are a lot of redundancies, all of which could do with being cleaned up.

2.4. Verifying data quality:

All the collected data is considered as correct data and have no missing or wrong information as it is collected from reputable sources for the official premier league recording system and they reflect each player score based on his performance in reality during the past matches.

3. Setting up and planning your project

3.1. Github Repository: [URL](#)

3.2. Slide URL:

https://docs.google.com/presentation/d/1RHDUPsJVvtwVfPp8-WxsK8udEpYOmf4Ki9NjtbgpLDU/edit#slide=id.g496346b8fd_38_5

3.3. Detailed Plan

- **Dataset Collection:** trying to collect more labeled data since the more labeled we have, the better our model performs. Here I'm going to show ways to get labelled data through scrapping from the web.

- **Dataset Filtration and cleaning:** trying to detect and correct or remove corrupt or unuseful either feature or instances from our data set. After cleansing, a data set should be consistent with other similar data sets in the system for our model.
- **Feature Extraction and Selection:** we are going to compute some statistical features from the available data like the average of goals scored by the player in past weeks from the true values.
- **Develop a model:** Trying different machine learning algorithms to predict the score of each player in a certain week. In our case, we need to try different regression algorithms and trying to select the model that can achieve the lowest error between the predicted player score and the actual score.
- **Team Formation:** Based on the score predicted and the constraints we mentioned above (team budget and number of player allowance from each team), we are going to construct the best available team line-up for different formations like (3-5-2, 3-4-4, 4-4-2, e.t.c)
- **Model Evaluation and assessment:** We are going to evaluate the model based on the average score obtained by fantasy premier league users. We need to consistently achieve a higher score than this average score.
- **Visualization:** We are going to visualize our results by having a bar chart that shows score achieved by our model in different weeks and the average fantasy users score to be able to recognize the difference and if there is an improvement made by the model.

4. Timeline plan:

4.1. Week 1 -- Part 1

Nesma Mahmoud → Dataset Collection → 4 hrs

Mahmoud Kamel → Data set Filtration → 5 hrs

Mohamed Maher → Feature Extraction and Selection → 8 hrs

4.2. Week 1 -- Part 2

Nesma Mahmoud → Model Selection & Model Building → 12 hrs

Mahmoud Kamel → Model Selection & Model Building→ 10 hrs

Mohamed Maher → Model Selection & Model Building→ 10 hrs

4.3. Week 2 -- Part 1

Nesma Mahmoud → Model Evaluation and assessment → 5 hrs

Mahmoud Kamel → Model Evaluation and assessment→ 5 hrs

Mohamed Maher → Team Formation→ 4 hrs

4.4. Week 2 -- part 2

Nesma Mahmoud → Visualization 4 hrs

Mahmoud Kamel → Visualization 4 hrs

Mohamed Maher → Visualization 2 hrs