Team Recommendation System for Fantasy Premier League

Nesma Talaat Mahmoud, Mahmoud Kamel Shoush, Mohamed Maher Abdelrahman
University of Tartu, Institute of Computer Science

Project Repository

Introduction

The English Premier League is the most-watched sports league in the world, it has billions of fans.

Fantasy Premier League (fantasy.premierleague.com) is also the most popular fantasy soccer game, with more than **5 million users**, in the soccer world. Get your fantasy season started by drafting players to build a solid team and using coaching tips to keep your team going strong.

Every week, Fantasy Premier League users are competing to choose their team of players based on different criteria. The success of your fantasy football team hinges on your 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 chose the most useful team at the start of any league to get a high score.

The aim of this project is to:
- **Extract features** that can be used to predict player performance each game week from available datasets on Kaggle, and Fantasy API.
- Use different Machine Learning and Deep Learning algorithms to **build a model** that can predict player points each week.
- **Recommend** a team line-up based on the results of the model that can achieves total score exceeding fantasy users scores consistently.
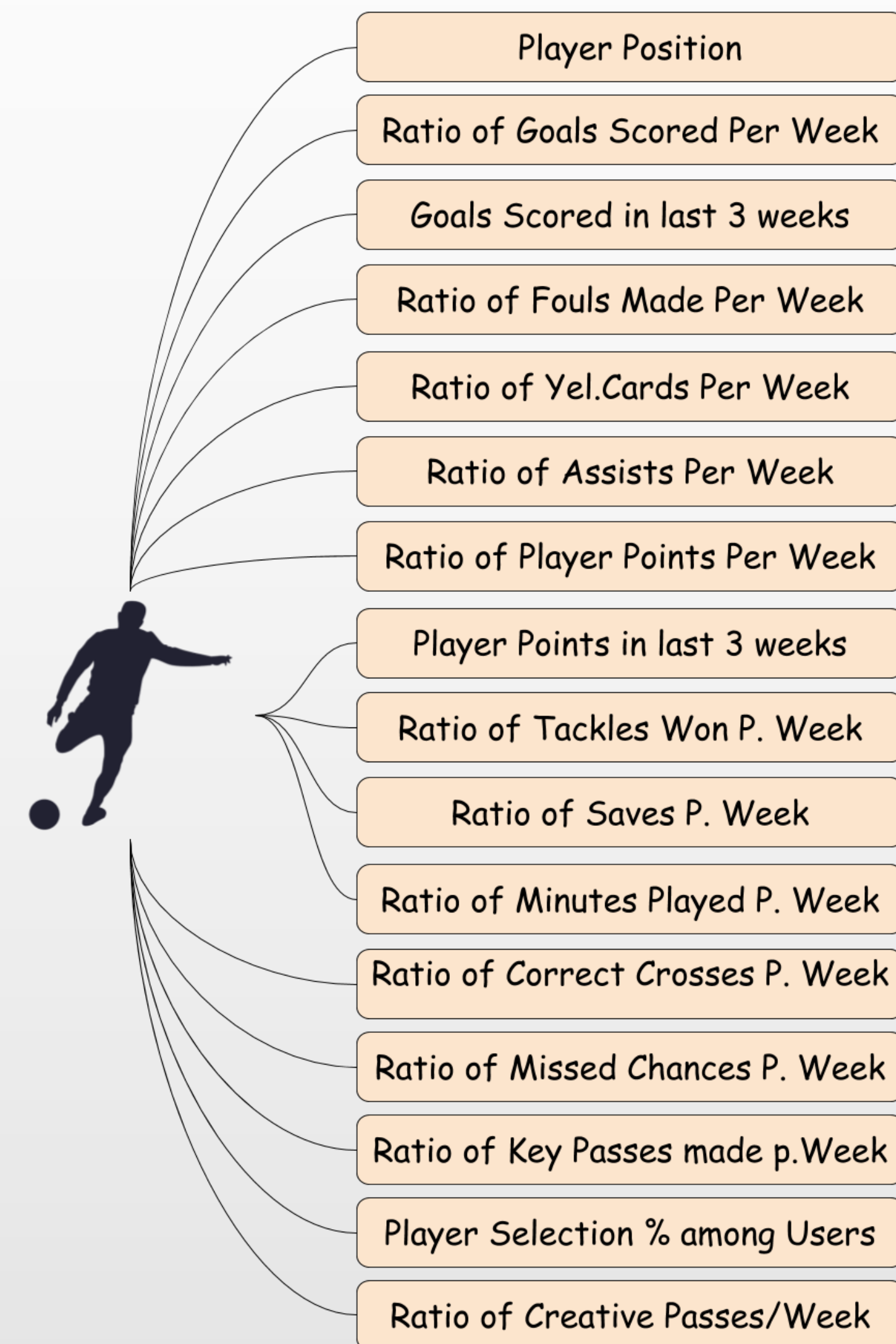
Data Description

Several datasets are available on Kaggle for different premier league seasons. These datasets include records for all matches results, and player performance after each game week. However, these datasets aren't consistent across all seasons. So, we used the Fantasy official API to fetch the same missing data of player records across all seasons.

In this project, we used data for seasons **2016\2017, 2017\2018**, and current season **2018\2019** till the 8th week. As the current data is only considered as records for each game week, we need to make use of this data to find suitable features that can help in predicting the player performance in following week.

Extracted Dataset

Each instance in the extracted dataset represents player information for a certain game week.
This information includes:
1) **Statistics of player performance** in the past game weeks of the same season.
2) **Team of the player** past results and performance during last game weeks of same season.
3) **Opponent Team** of the player past results and performance during last game weeks of same season

- Team Rank
- Ratio of Wins
- Ratio of Goals Scored
- Ratio of Goals Conceded
- Current Fixture Game Week No.
- Ratio of Clean Sheets
- Ratio of firstly Leading by Goal

- Player Position
- Ratio of Goals Scored Per Week
- Goals Scored in last 3 weeks
- Ratio of Fouls Made Per Week
- Ratio of Yel.Cards Per Week
- Ratio of Assists Per Week
- Ratio of Player Points Per Week
- Player Points in last 3 weeks
- Ratio of Tackles Won P. Week
- Ratio of Saves P. Week
- Ratio of Minutes Played P. Week
- Ratio of Correct Crosses P. Week
- Ratio of Missed Chances P. Week
- Ratio of Key Passes made p.Week
- Player Selection % among Users
- Ratio of Creative Passes/Week

Building a Model

The dataset has been spitted into 80% Training Set and 20% Testing Set. Then, Different Regression algorithms have been fitted over the training section.

Overall, The algorithms chosen are wide range of machine learning **linear regression algorithms**, **decision trees**, **ensemble algorithms**, and **KNN**. The hyper-parameters of all the algorithms have been tuned by following the Grid Search technique with 5-fold cross validation to get the best tuned model from each algorithm.

In addition, Deep Neural Networks have been trained over the dataset using recurrent network (**LSTM**) as data can be considered as time-series, and simple fully connected layers (**DNN**).

The following chart summarizes the performance (Mean Squared Error between predicted player points and actual points obtained) of all selected algorithms:

Comparison of Performance of difference Regression Algorithms
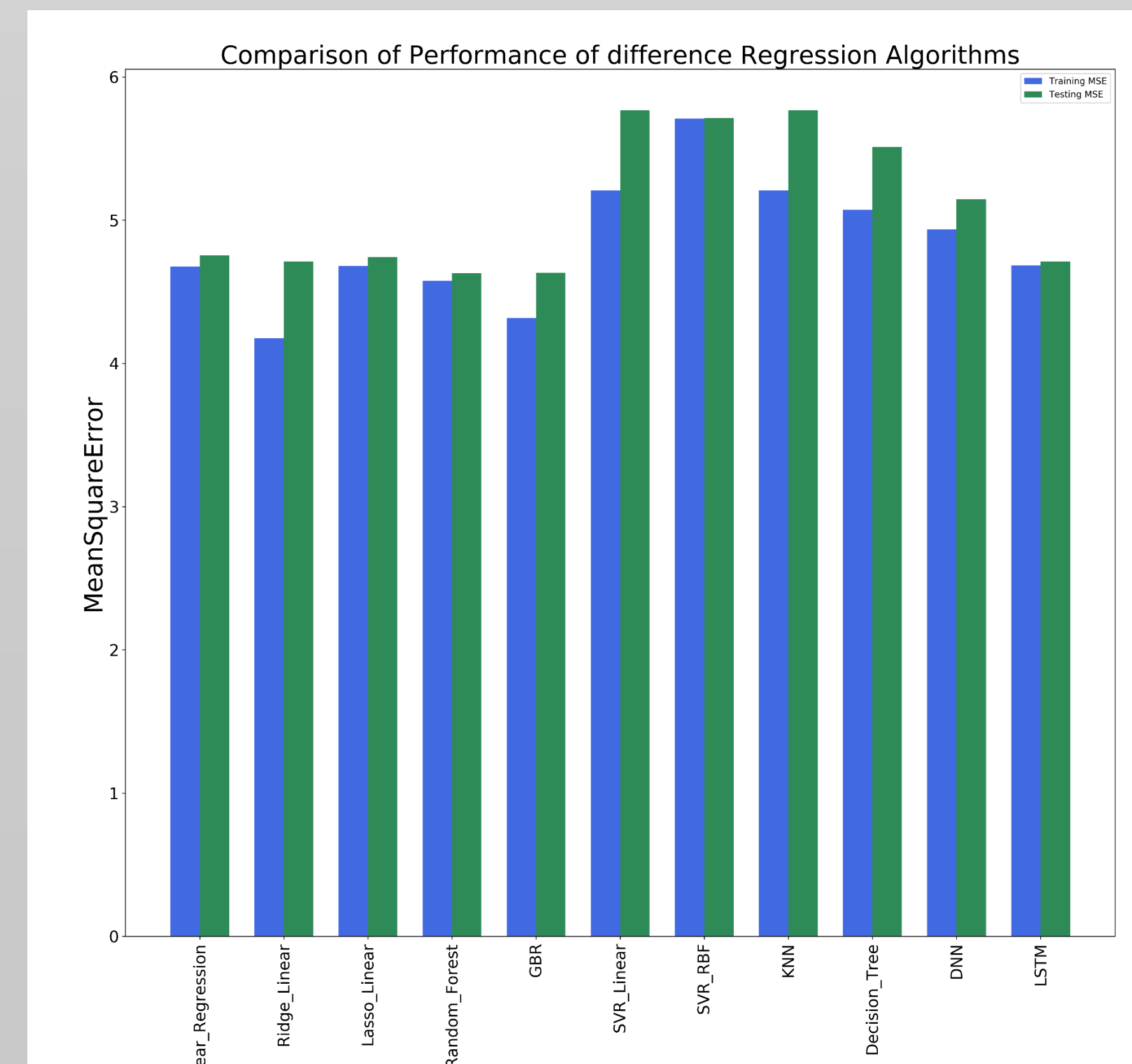
Based on the previous results, we have chosen to continue with the Random Forest Model. We have also used Model Interpretability Shapely technique to measure the feature importance upon which the model makes its decision.

The results show that the player ratio of points obtained in last 3 game weeks, ratio of minutes played, opponent team rank, and player cost are the main factors that participate in changing the predictions value by the model.
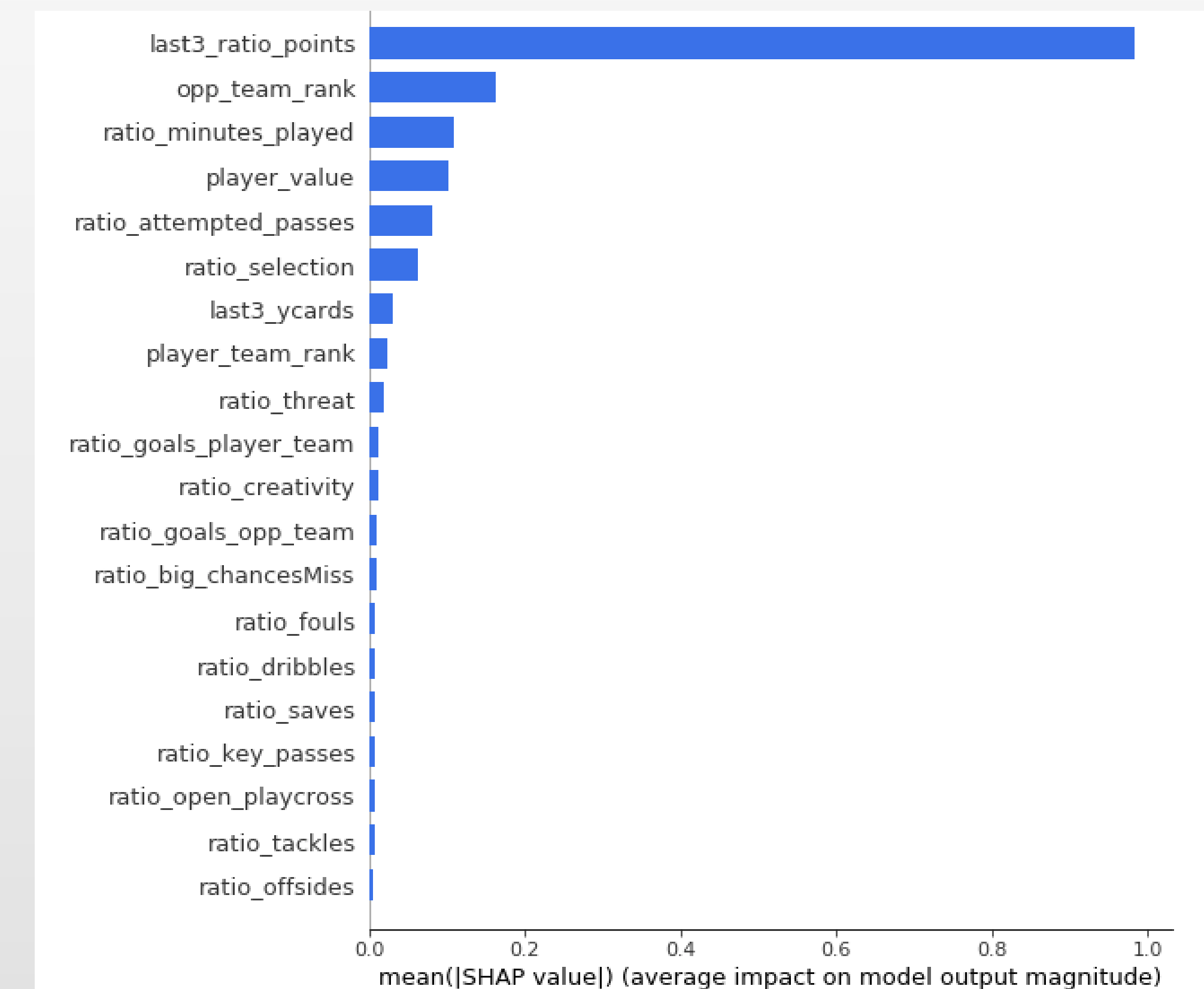

Team Line-Up Results

To Start forming a team-lineup based on the model predictions. We need to handle some constraints from Fantasy Game Rules:
1. Team **Budget** should be less than 100 units.
2. We can't choose more than **3 players** from the same team.
3. We should use one of the following **formation**s (3-4-3 / 3-5-2 / 4-4-2 / 4-3-3 / 4-5-1).
4. Score of Captain player selected is doubled.

Based on our analysis on dataset, average forwards players points are the highest, followed by midfielders, and defenders. So, we fixed 3-4-3 formation as our chosen formation for team-lineup selection.


Average Points of Each Line
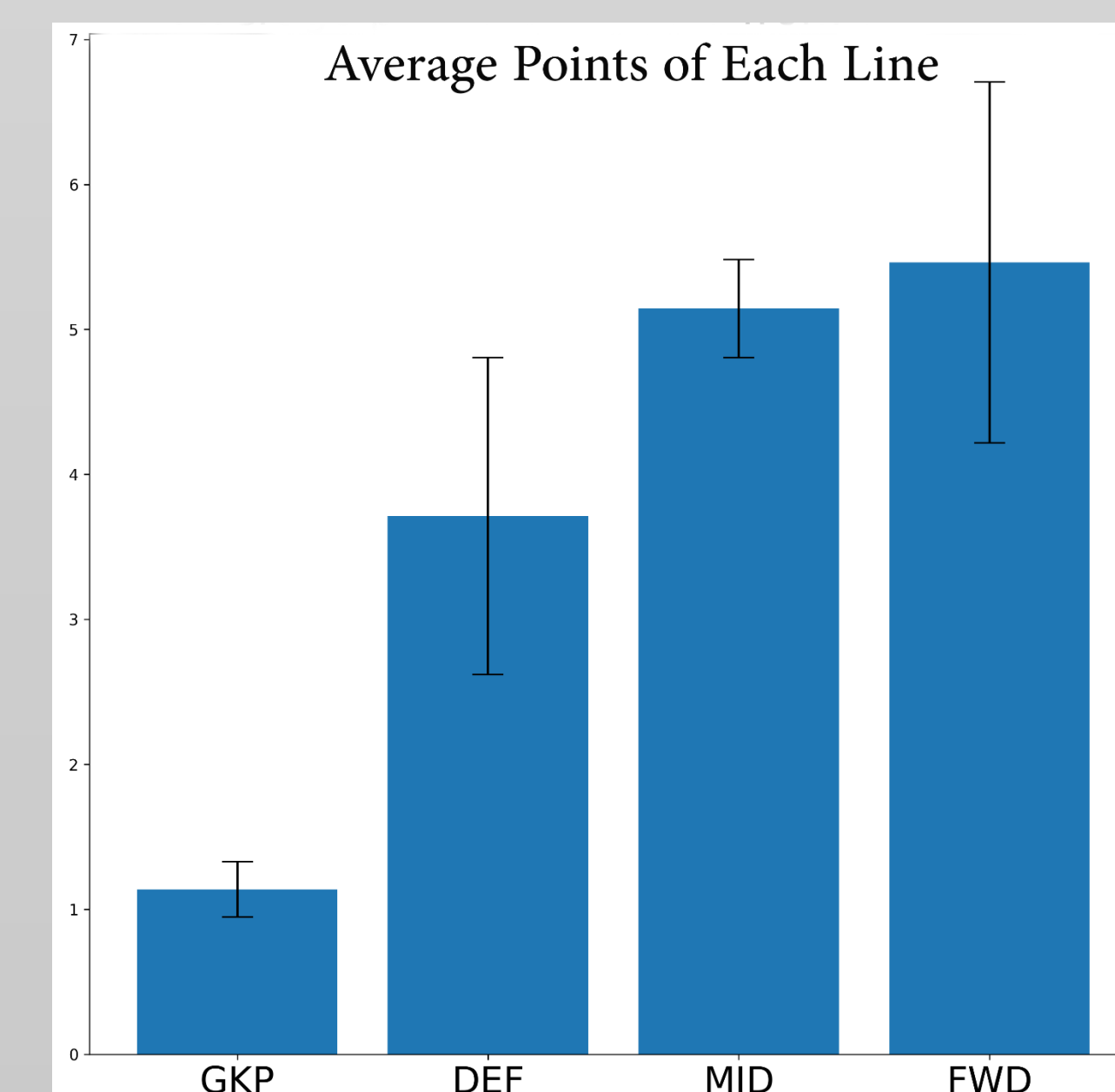
Also, based on the other constraints and players scores predicted from Random Forest model, we formed the line-up by trying all possible combinations of top predicted players and choose the highest one satisfying all constraints.

The following bar chart shows **predicted final team score Vs average score obtained by fantasy users**:

Predicted Team Score Vs Average Fantasy Users Score in Season 2018/19
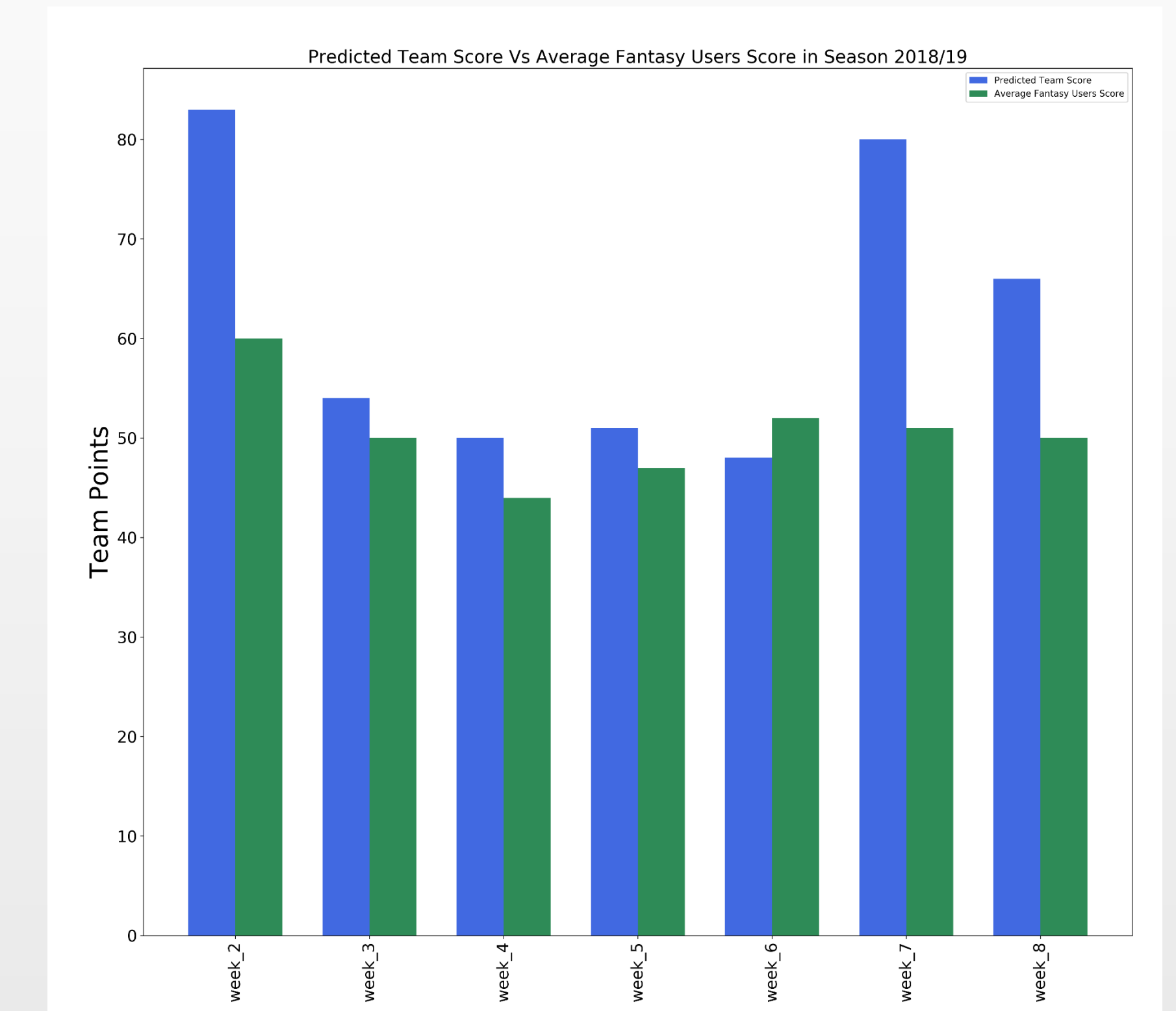
Here are some examples of the predicted team line-ups for two weeks of current season:
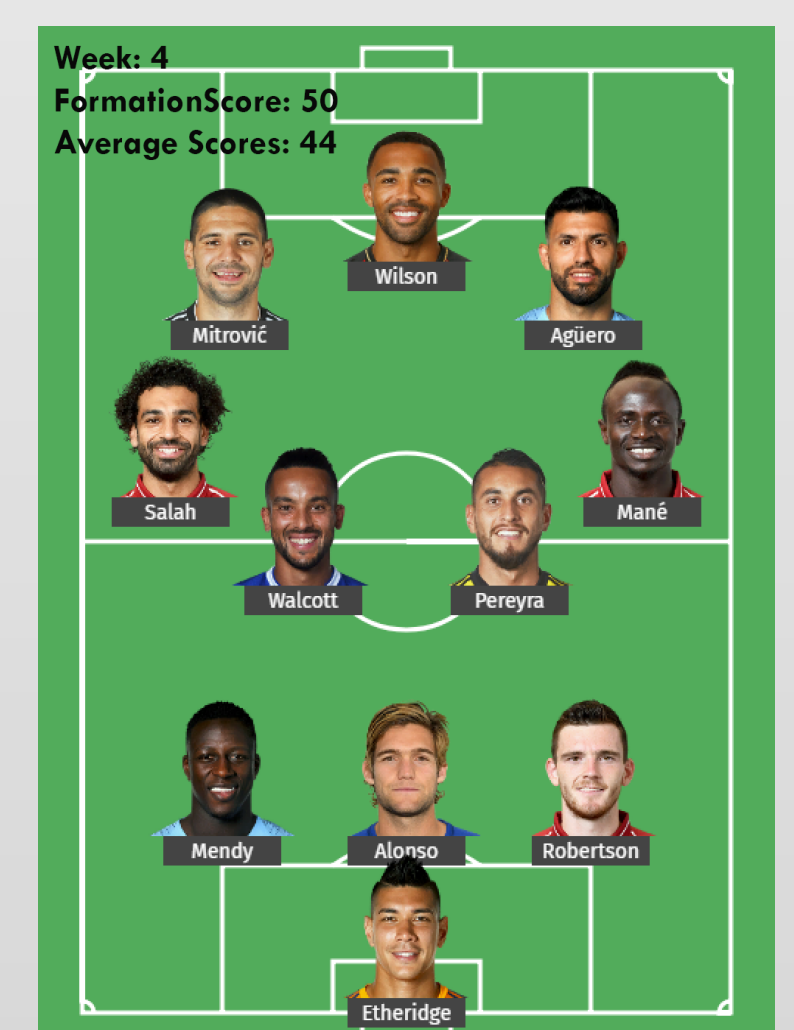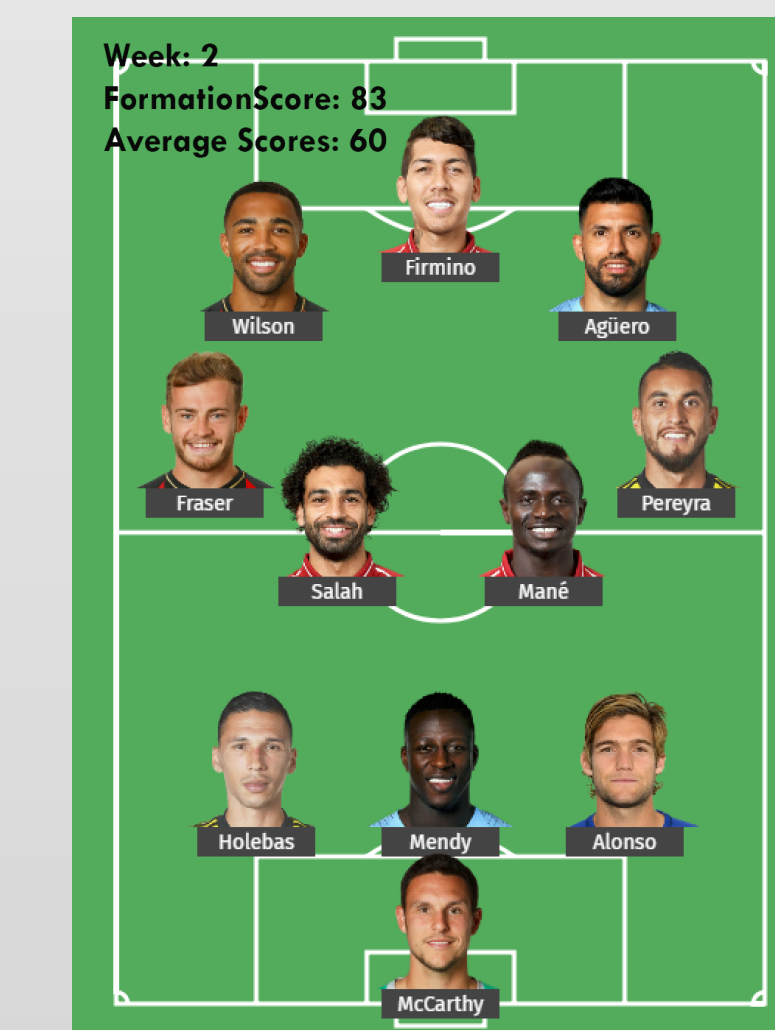
Conclusion

To sum up, we have extracted some features from players, and teams historical records, and we managed to build a regression model that can predict players scores every week. Then, we used these scores to form the time-lineup satisfying the fantasy rules and constraints. We have chosen the 3-4-4 formation based on our analysis as we have found that forwards usually have average scores higher than midfielders, and defenders.

Based on model interpretability made, we concluded that player average of points in last 3 weeks, opponent team rank, and player cost are the main factors in determining the model prediction.

Future Work

As a future work, we suggest to:
- Try more different features about either team, or player performance history.
- Include more information about player injury state.
- Add features about match importance to each team, and which team is the home team and which one is the away team.
- Use more seasons data in training can improve the model performance.
- Develop a web application and deploy the recommendation system to be used by all fantasy users.