EAST WEST UNIVERSITY

Project proposal:

Title: Football Squad Analysis using a Random Forest

Submitted by:

Tawfiqul Alam	2018-2-60-046
Md. Sajeeb Molla	2018-2-60-045
Md. Mostafijur Rahman	2018-2-60-042

Submitted to:

Amit Kumar Das

Senior Lecturer

Department of Computer Science & Engineering

Date of submission: 10 September, 2021

Football Squad Analysis Using Multiple Random Forest

Md. Mostafijur Rahman 2018-2-60-042 Department of CSE East West University	Md. Sajeeb Molla 2018-2-60-045 Department of CSE East West University	Tawfiqul Alam 2018-2-60-046 Department of CSE East West University

Abstract:

Squad analysis is a must to do task to observe the players performance and improve the statistics by taking necessary action. By reading and analyze these data anyone can create a fantasy squad by adding best players of the league. Also the league management can use this stats to analyze how the players under their league is improving. In this paper one of the well-known machine learning algorithm will be discussed. The FPL managers have a budget of 100M pounds to select 15 players (3 FWD, 5 MID, 5 DEF, 2 GK) out of all players, with a maximum of three players from one club. A ranked list will be created by this algorithm from past dataset and the selection committee will choose 15 players to play in fantasy squad matches and expect winning.

Introduction:

Football is a team sport played by two teams of 11 players. It is played by around 250 million players in over 200 countries and dependencies, making it the world's most popular sport. Fantasy Premier League (FPL) is an online football game in which we can choose a specific combination of actual Premier League players. Points will be earned based on the player's actual performance and their contribution on the match day. The performance of the players (forwards, mid-fielders, defenders, and keeper) is measured primarily on their contribution

Problem Description:

i) Selecting Machine Learning Algorithm: To create fantasy team, the algorithm is needed to analyze. Here the random forest regression, Linear regression is used to get optimal result.

ii) The aim is to use the data collected in game-weeks 33-37 to predict the scores in game-week 38. Therefore, 'data' table splitted into training and testing subsets based on these game-weeks.

Methodology:

A) Block Diagram:

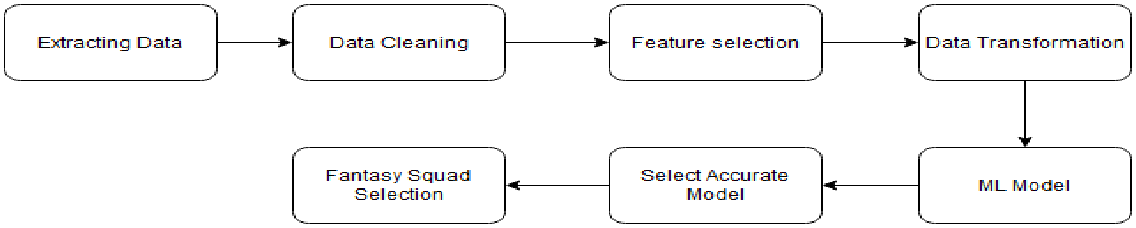
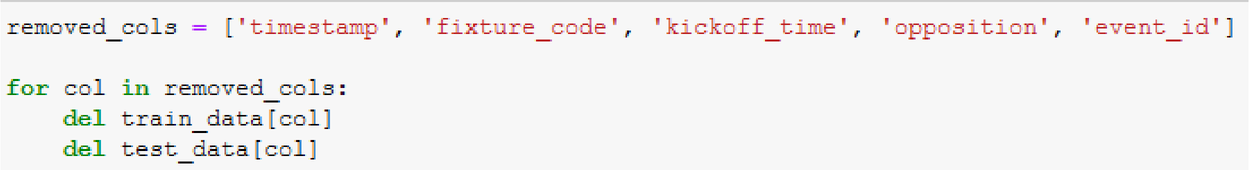


Fig: Block-Diagram of the system.

B) Data Extraction: The CSV dataset is collected from internet. There are 4 dataset: players, team, data, element type.

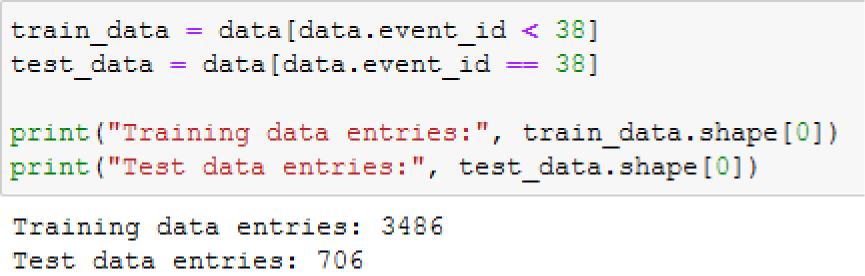
C) Data Cleaning: In these dataset, the corrupted, unnecessary, NULL data's. As for example 'timestamp', 'fixture_code', 'kickoff_time', 'opposition', 'event_id' is unnecessary for this analysis.



```
removed_cols = ['timestamp', 'fixture_code', 'kickoff_time', 'opposition', 'event_id']

for col in removed_cols:
    del train_data[col]
    del test_data[col]
```

D) Preparing data: The aim isto use the data collected in game-weeks 33-37 to predict the scores in game-week 38. Therefore, the data table is splitted into training and testing subsets based on these game-weeks. Here the data is trained and tested.



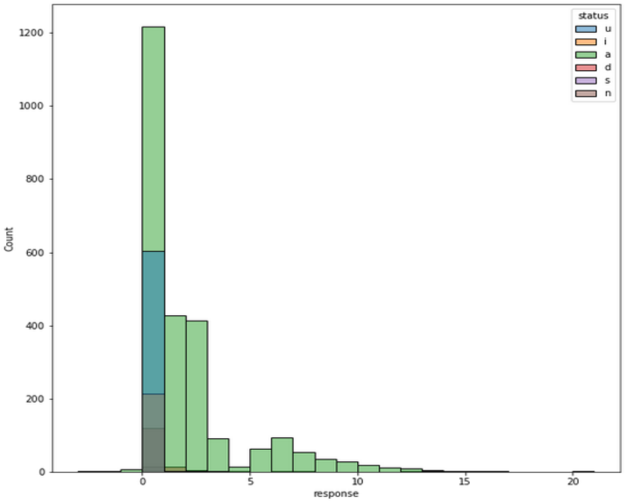
```
train_data = data[data.event_id < 38]
test_data = data[data.event_id == 38]

print("Training data entries:", train_data.shape[0])
print("Test data entries:", test_data.shape[0])
```

```
Training data entries: 3486
Test data entries: 706
```

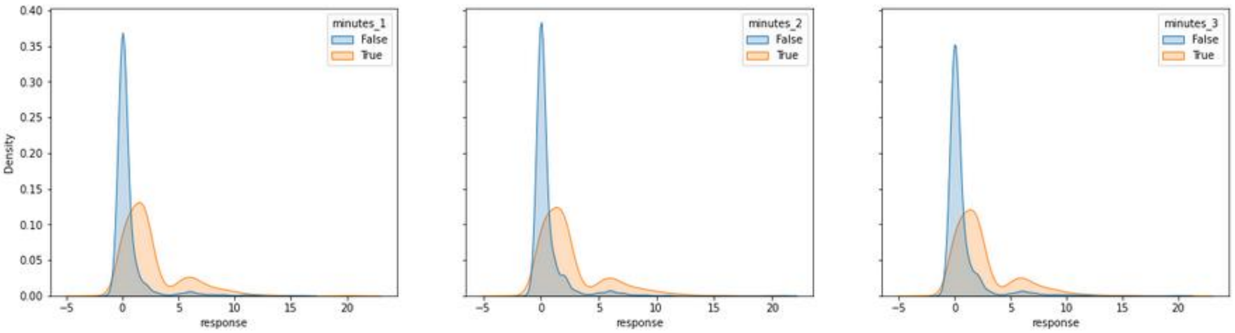
E) Data analysis:

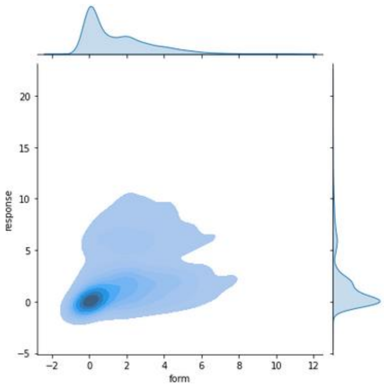
The 'response' of the player, i.e. how many points they went on to score in that game-week, will be the goal column to predict. If a player is wounded or unavailable, they are very certain to receive zero points, however if they are completely fit and available, they are far more likely to receive points.

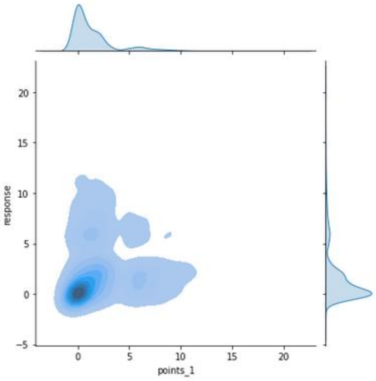


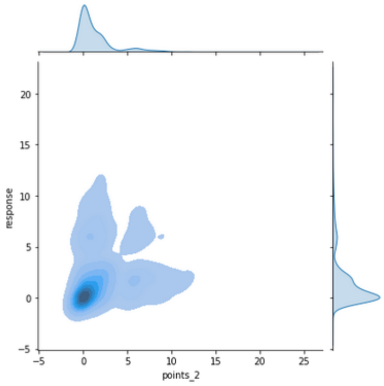
As this shows that if a player's status is not 'available', then they are very likely to score 0 points, it is therefore only interesting to explore the data for those whose status is available. This is based around the number of minutes a player plays - if a player is not available, they will play 0 minutes and therefore score 0. How else can we predict minutes? Well, we have the data for minutes for the last three matches.

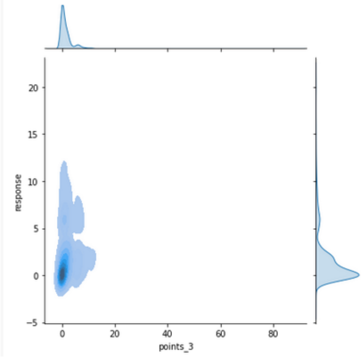
```
Out[11]: <AxesSubplot:xlabel='response', ylabel='Density'>
```

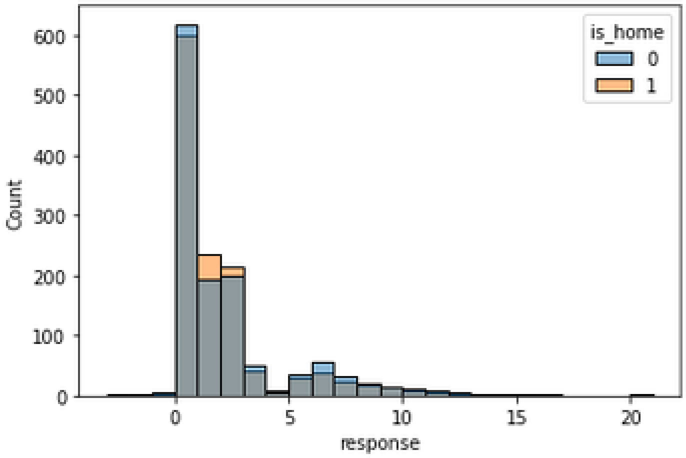






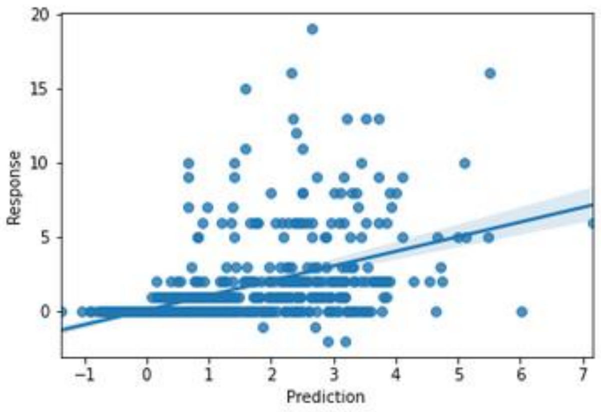
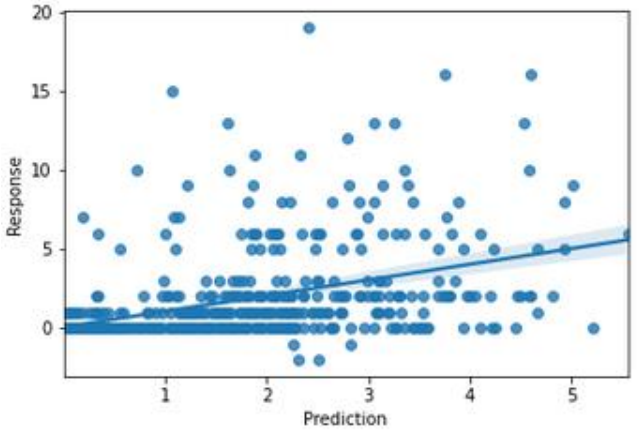




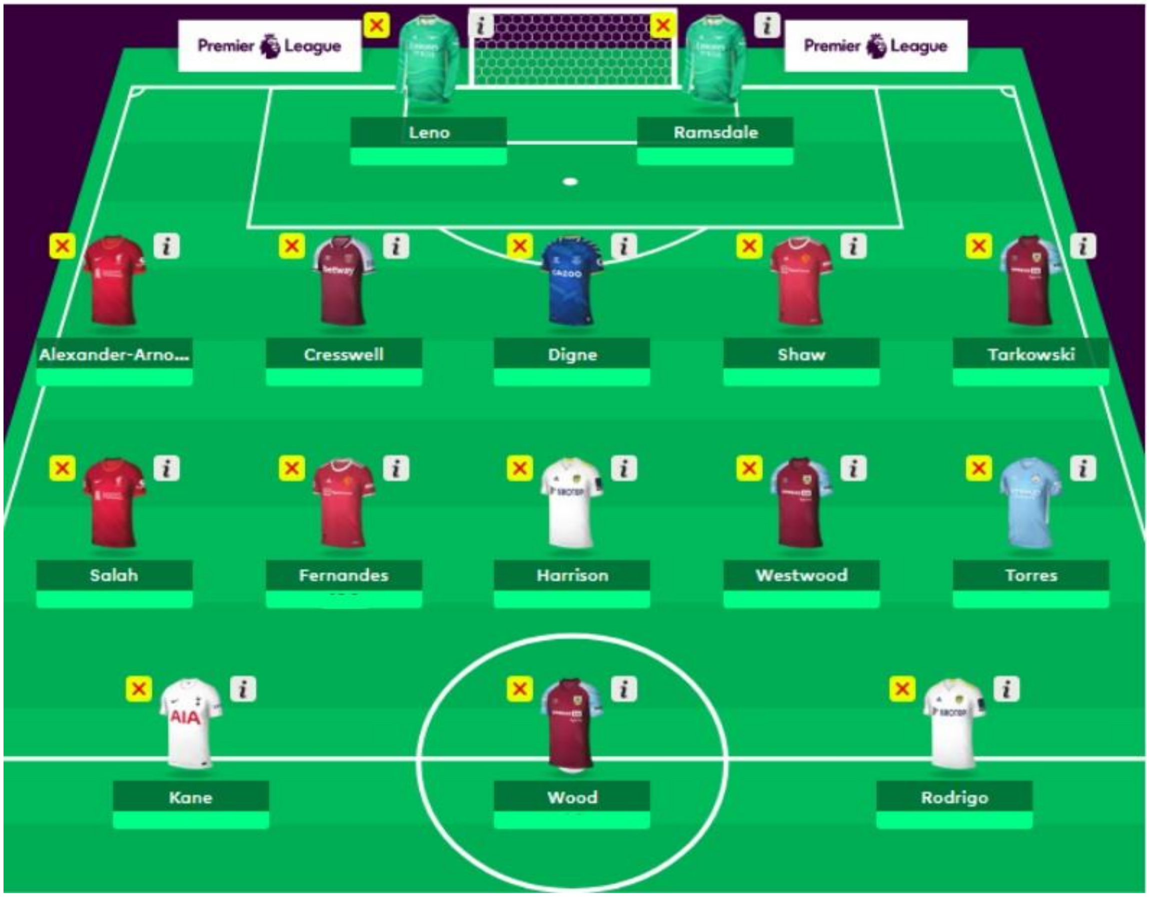


F) Visualization of Machine Learning Algorithm Comparison:

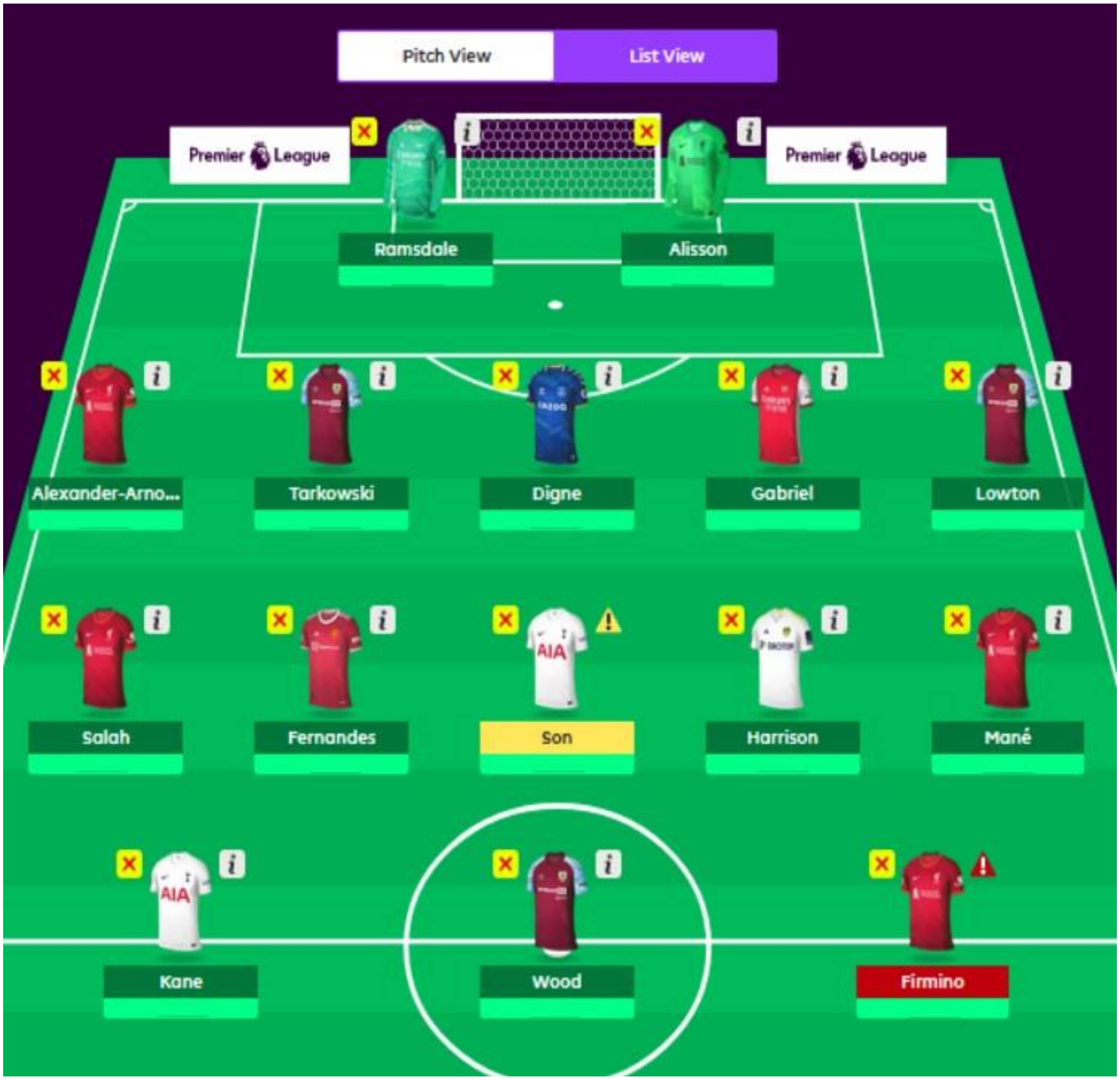
Linear Regression R^2coefficient of 26 % of prediction vs actual scores and Random forest Regression gives 25.6% of prediction**.**

Linear Regression	Random Forest Regression
 	

G) Fantasy Squad Created by Random Forest Regression:



H) Fantasy Squad Created by Linear Regression:


Pitch View
List View
Premier League
Premier League
Ramsdale
Alisson
Alexander-Arno...
Tarkowski
Digne
Gabriel
Lowton
Salah
Fernandes
Son
Harrison
Mané
Kane
Wood
Firmino