2022 2nd Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 26-28, 2022

Using ML Models to Predict Points in Fantasy Premier League

Malhar Bangdiwala
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
malhar.bangdiwala@spit.ac.in

Rutvik Choudhari
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
rutvik.choudhari@spit.ac.in

Adwait Hegde
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
adwait.hegde@spit.ac.in

Abhijeet Salunke
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
abhijeet_salunke@spit.ac.in

***Abstract*—Fantasy Premier League is an ever-growing game, with millions of people playing the game. To outperform the rest, it is imperative for the players to accurately predict the expected points the footballer would earn over the course of the match. However, doing so is not easy as there are several aspects to consider as well as the human bias towards the players' favourite footballers and teams. This paper attempts to build and compare three machine learning models to accurately predict the number of points that each footballer would earn over the course of the season. For doing so, the Linear Regression, Decision Tree, and Random Forest algorithms have been leveraged. Features such as fixture difficulty, form of the two teams, creativity, and threat of the footballer have been considered. This would help the players of this game to make more informed decisions while making their respective teams.**

***Index Terms*—Linear Regression, Decision Tree, Random Forest, Fantasy Premier League**

I. Introduction

A fantasy sport is a form of game in which players assemble fictitious or virtual teams made up of proxies for real players in a professional sport. It is commonly played on the Internet. These teams compete based on their players' statistical performance in actual games. This performance is transformed into points, which are collated and totaled based on a roster chosen by the manager of each fantasy club. In fantasy sports, the owners, like in actual sports, select, trade, and remove players from their teams.

Fantasy Premier League, or FPL is a popular game that casts the player in the role of a Fantasy manager. The manager is given the task of picking a squad of fifteen players who would give the manager points based on their real life performances. This game is based on the English Premier League. The English Premier League (EPL) is the highest level in the English football league system. It is contested by 20 clubs and follows the English Football League's promotion and relegation system. Each team plays 38 matches during the season, which runs from August to May. This means that each team plays each other twice: once at home and once at the oppositions home ground. FPL is an ever growing community: from less than two million managers in the 2007/08 season, it has grown to around eight million for the 2020/21 season[1].

The managers must make a squad of fifteen players and would receive points for each player based on their real life performance for their clubs in the EPL matches. Each player has a certain price and the manager can not exceed the stipulated budget of 100 million while building the squad. Players earn points for goals, assists, saves, and clean sheets. As a reward for a solid performance in a match, players can receive additional bonus points. The starting XI will score the manager's team's points for the match round or "Gameweek". However, if a starting player does not appear for his club in that round of fixtures, the points earned by the first player off the bench will be tallied instead. If two or three starting players fail to show up, the same thing happens. Extra points can be earned by selecting a captain from the starting lineup. In that Gameweek, the captain's points are doubled. The exact breakdown of points for each action is mentioned in Table 1 [2].

Fantasy Premier League releases its own set of statistics for each player that includes, but is not limited to: form, influence, creativity, threat and opposition difficulty. With increasing number of statistics, it has become hard for the managers to keep track of which set of numbers play an important role in the squad making procedure. Further, guessing the number of a points a play would accumulate in the future is based on estimation and the bias of the managers.

This paper aims to eliminate the above-mentioned by building highly accurate models using the algorithms: Linear Regression, Decision Tree and Random Forest. The model has been validated by looking at each player's performance for

978-1-6654-6851-0/22/$31.00 ©2022 IEEE

TABLE I
POINTS FOR EACH ACTION

Sr. No.	Action	Points
1	Playing up to 60 minutes	1
2	Playing 60 minutes or more (excluding stoppage time)	2
3	Each goal scored by a goalkeeper or defender	6
4	Each goal scored by a midfielder	5
5	Each goal scored by a forward	4
6	Each assist for a goal	3
7	Clean sheet by a goalkeeper or defender	4
8	Clean sheet by a midfielder	1
9	Every 3 shots saved by a goalkeeper	1
10	Each penalty save	5
11	Each penalty miss	-2
12	Bonus points for the best players in a match	1-3
13	Every 2 goals conceded by a goalkeeper or defender	-1
14	Each yellow card	-1
15	Each red card	-3
16	Each own goal	-2

each gameweek of the 2020/21 season. For each gameweek, the model is fitted by using all historical data prior to that week, and then calculate the mean absolute error for the following 6 gameweeks. This historical data consists of all the data dating back to the 2016/17 season.

II. LITERATURE SURVEY

There are not many papers when it comes to predicting points for Fantasy Premier League. Thus, the scope needs to be widened to prediction of points for players in any sport.

M. Pardee[3] presents and explores methods for predicting Quarterback Fantasy Football scores with minimal data. It also contains various recommendations for improving the data in order to attain better results. The author includes the game data from the previous six NFL in the dataset. Support Vector Regression (SVR) and Neural Networks were utilized by the author to predict the Fantasy Football scores of NFL players. The author states that the results of these models were promising given that the data used for training was limited

In 2013, Bonomo et al. [4] applied two mathematical programming models that operate as virtual coaches, selecting a virtual squad roster for each round of the real Argentinian soccer league. When the season is over and all the results are known, the a posteriori model determines what would have been the best lineup for the game. The apriori model was submitted to the fantasy game and received results that placed it among the top scorers. The authors state that further development of such models would provide important tools to assist real-world coaches and managers in making decisions. The authors created a posterior model which is a descriptive one because it addresses the question of what would have been the best team in each round of the tournament if the outcomes had been known in advance. As a result, it can discover an ideal set of teams that would have to be created in order to get the highest possible total points while staying within the game's limits. The second model which is the prescriptive one was created a priori and offers adjustments to the team lineup round by round in order to improve team robustness as the tournament advances without knowing future results. It accomplishes this by combining data on player performance in previous tournaments and past matches in the present one with data on the round's essential qualities.

Katara et al. [5] used deep neural networks to predict a team of eleven players for Dream11, the Indian Cricket version of Fantasy sports. They also compared the accuracy of the neural network against that of well-known machine learning approaches like k-nearest neighbours, logistic regression, naive bayes, random forest, support vector machines.

One of the earliest attempt on FPL prediction is that of Stolyarov et al. [6] in 2017, who predicted the players' performance with the help of gradient boosted trees. The author offers a model of optimal sequential decision-making in the Fantasy Premier League (FPL). The author uses XGBoost, one of the most powerful recent data science approaches, to forecast predicted points. The author then creates an automated manager that employs this strategy and produces decent team formation results.

Saifi et al. [7] combined random forests and gradient boosted machines to analyse the accuracy of the predictions. Delano[8] has made predictions using convolutional neural networks and has compared its accuracy with other models. More recently Gupta[9] combined Autoregressive Integrated Moving Average (ARIMA) and Long Short Term Memory for the predictions. Gupta went one step further by predicting the squad as well by using liner programming techniques. Although, Gupta expanded the horizon in this research direction, the model was overfitted. In 2020, Sonderberg et al.[10] used several machine learning models and compared their accuracy. However, in that paper, the subject of creating the squad was not touched upon. Similar to Sonderberg, Gunjan[11] created different models for each position: goalkeeper, defender, midfielder and attacker. This seemed to increase the accuracy of the predictions.

Nicholas Bonello et al.[12] states that the performance predictors for the Fantasy Premier League (FPL) are frequently based only on past statistical data. External factors such as injuries, managerial decisions, and other tournament match data can never be integrated into the final forecasts with this approach. The author suggests a new strategy for forecasting future player performances by automatically incorporating human feedback into the model. The author has described a novel strategy for selecting FPL player picks for upcoming gameweeks in this work that automatically incorporates human feedback into the algorithm via publicly available online data such as blogs published by subject experts. His recommender system can take into account expert and fan feedback by looking at news articles, fantasy football prediction blogs, and

even tweets utilizing fantasy premier league specific hashtags. These are employed as additional parameters in the predictive model, allowing it to automatically incorporate external factors such as midweek game lineups and performances, managerial decisions, rotations, and injuries, as well as unexpectedly good or dismal prior week results.

All of this work done in earlier times is a testament that the prediction of points is possible by leveraging several statistical and machine learning models.

III. Methodology

A. Dataset

The data for training the models are taken from Vaastav's Github repository[13]. Here, there is complete information about every player and team season-wise, starting from the 2016/17 session. The repository is being updated every week based on the real-life matches that take place in the Premier League. For the purpose of this paper, player and team data from the 2016/17 season and onwards has been considered, upto the 2020/21 season. The 2021/22 season has not been used since that is the ongoing season and complete data for the same is not available.

The available data can be broadly classified into two categories: player-specific data and team-specific data. Team-specific data is the data which is common to all the players of that specific team. This includes the following data:

1) Difficulty rating of the opposition team relative to their team.
2) Whether the match is being played at home or away.
3) Strength of the various departments (attack, midfield and defense) at home and away.

Some of the player-specific data would include:

1) Goals scored
2) Assists
3) Minutes Played
4) Clean sheets kept

B. Data Cleaning

From all the available data, only the important features are used for training the models. Such important features include the goals, shots, blocks, tackles etc. Certain predictive statistics which are provided officially by the FPL platform such as form, ict(innovation, creativity, threat) index, expected goals, expected assits, probability of playing the next match etc. have also been used. For each gameweek, a new dataset is curated that contains all the data from the beginning of the 2016/17 season until the last completed gameweek. Certain columns which contain data of a specific window have also been added. This includes columns such as minutes played in the last three matches, points scored in the last five matches etc. The exact features to be selected was based on Fig.1. This is done because these parameters will be used as features to predict the points of future game weeks.

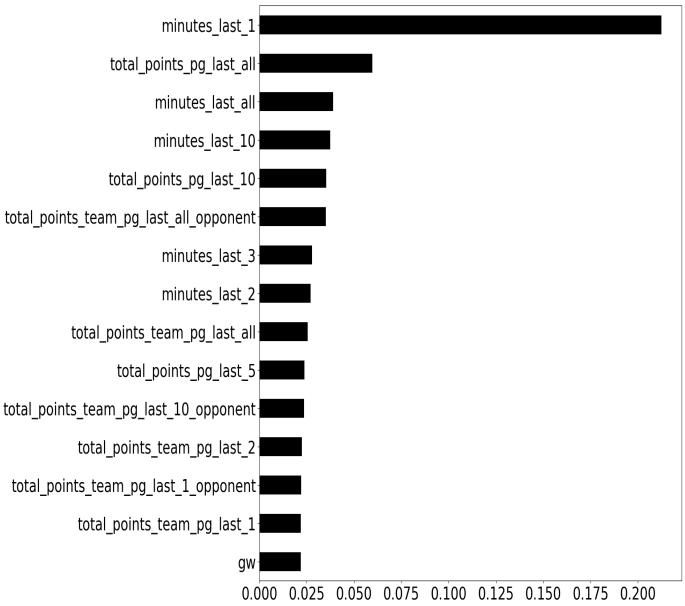

Fig. 1. Feature vs importance graph

The columns where data is not available are replaced with a zero. The reason for this absence of data can be either due to a new player being added to a squad during the transfer window or if a player has been sold to another team during the same. Major injuries can also lead to a player missing out on many of the matches in that season. Further, encodings have been carried out to convert the season and playing position of the player into numerical data. This cleaned data is now fed to the models for training, testing and validation.

C. Models

For the purpose of this paper, three regression models were used: Linear Regression model, Decision Tree Regression model, and Random forest Regression model.

1) Linear Regression Model: A linear model is one in which the input variables (x) and the single output variable (y) are assumed to have a linear relationship. That y can be determined using a linear combination of the input variables is more detailed (x). The procedure is known as simple linear regression when there is only one input variable (x). Multiple linear regression is the term used in statistics literature when there are multiple input variables. In our case, we will use the multiple regression model as there are multiple features that may affect the points of the player. Some of the independent variables are the number of assists, goals, clean sheets, goals conceded, goals scored, minutes played, red cards, yellow cards, etc. The dependent variable in this scenario is the points of the player.

2) Decision Tree Regression Model: In the form of a tree structure, a decision tree creates regression or classification models. It incrementally breaks down a dataset into smaller and smaller subsets while also developing a decision tree. A tree with decision and leaf nodes is the end result. The Root Node is the first node in the tree, which represents the

complete sample and can be further divided into nodes. The branches indicate decision rules, whereas the inside nodes reflect data set features. Finally, the outcome is represented by the Leaf Nodes. This algorithm is extremely beneficial for handling problems involving decisions. Decision trees offer the advantages of being simple to grasp, requiring minimal data cleansing, non-linearity having no effect on the model's performance, and having a small number of hyper-parameters to modify. It may, however, have an issue with over-fitting.

3) Random Forest Regression: Random Forest Regression is a supervised learning approach for regression that uses the ensemble learning method. The ensemble learning method combines predictions from several machine learning algorithms to produce a more accurate forecast than a single model. During training, a Random Forest constructs many decision trees and outputs the mean of the classes as the prediction of all the trees. Random Forest Regression is a powerful and precise model. It usually works well on a wide range of issues, including those with non-linear relationships. However, there are some drawbacks: there is no interpretability, overfitting is a possibility, and we must choose the number of trees to include in the model. In our case we will have 7 decision tress in the Random Forest Regressor.

D. Comparison Parameters

To compare the above mentioned models, Root Mean Square Error (RMSE) and Mean Absolute Error(MAE) have been used.

1) Root Mean Square Error: The root mean square error (RMSE) is the residuals' standard deviation. The RMSE is a measure of how spread out the residuals are in comparison to the regression line data points. In other words, it indicates how tightly the data is clustered around the line of best fit. It is calculated according to the formula:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{1}$$

Where,
i = Summation variable
n = Total number of observations
x_i = Actual value of observation
$\hat{x}_i$ = Predicted value of observation

2) Mean Absolute Error: The Mean Absolute Error (MAE) is a regression model evaluation statistic. The mean absolute error of a model with regard to a test set is the average of all individual prediction errors on all instances in the test set. The difference between the true value and the anticipated value for each instance is the prediction error. It is calculated using the formula:

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}|y_i - x_i| \tag{2}$$

Where,
i = Summation variable
n = Total number of observations
x_i = Actual value of observation
y_i = Predicted value of observation

E. Results

After the models are trained, RMSE and MAE are calculated. RMSE for training dataset for Linear model, Decision Tree, and Random Forest Regressor are 2.172224, 0.260842, 0.260842. It is observed that the RMSE of the Linear model, Decision Tree, and Random Forest Regressor are 2.159776, 2.987197, and 2.145552 respectively for the test dataset. This has been visualized in Fig.2 and 3 respectively.

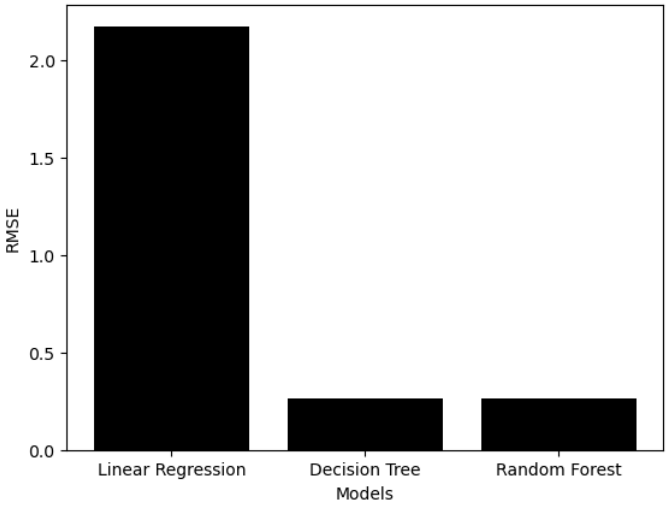

Fig. 2. RMSE for test dataset

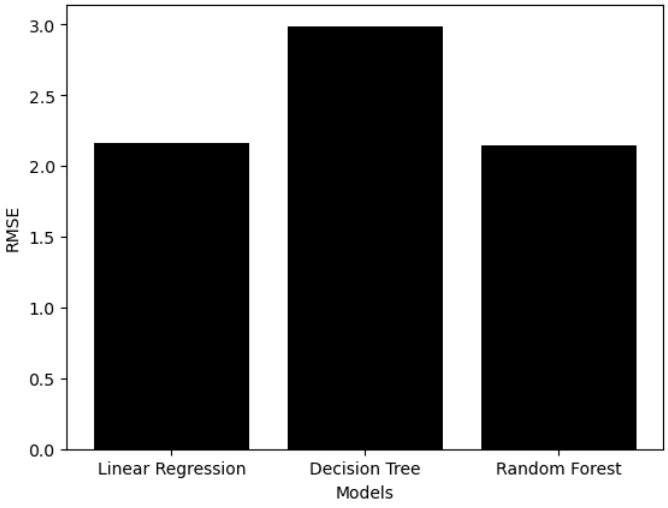

Fig. 3. RMSE for training dataset

The MAE for the training dataset is 1.289776, 0.024551, and 0.024551 in the same order, and for the test dataset 1.264294, 1.527351, 1.214767. This can be seen above in Fig. 4 and 5 respectively.

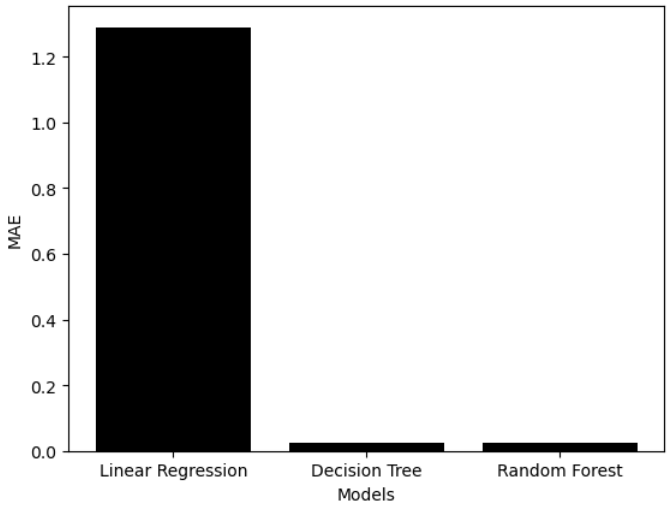

Fig. 4. MAE for test dataset

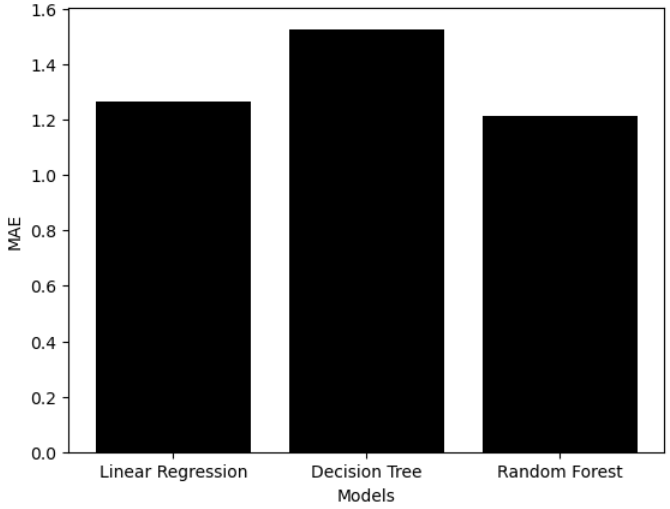

Fig. 5. MAE for training dataset

Model-wise accuracy has been tabulated in the tables 2, 3 and 4.

TABLE II
LINEAR REGRESSION MODEL ACCURACY SUMMARIZATION

Sr. No.	Metric	Train	Test
1	RMSE	2.172224	2.159776
2	MAE	1.289776	1.264294

TABLE III
DECISION TREE REGRESSION MODEL ACCURACY SUMMARIZATION

Sr. No.	Metric	Train	Test
1	RMSE	0.260842	2.987197
2	MAE	0.024551	1.527351

TABLE IV
RANDOM FOREST REGRESSION MODEL ACCURACY SUMMARIZATION

Sr. No.	Metric	Train	Test
1	RMSE	0.260842	2.145552
2	MAE	0.02455	1.214767

Thus we can observe that the decision tree and random forest regressor are overfitted as the error metrics for the training dataset are low but are high for the test dataset. The linear regression model and random forest model perform similarly for the training dataset but we can notice that the linear regression model is not overfitted to the training dataset.

IV. CONCLUSION

In conclusion, this paper has successfully predicted the points a player would accumulate in the upcoming match on the FPL platform. Three models have been created and their accuracy has been compared. It can be inferred that the Linear Regression Model performed the best among the three. Predicting the points a player would earn is the key decision any manager on FPL has to make. The entire game revolves around this sole aspect. This decision is rather daunting given the amount of data and statistics that is available. With the help of our models, this task has been made easier for all the managers on FPL. This paper also contributes to the work that has taken place in this domain by validating three models that can be used for making predictions.

V. FUTURE WORK

This line of work can be extended by creating the entire squad of fifteen players based on the points predictions. Simple mathematical and statistical models such as Linear Programming can be used for the same. Further selecting the starting eleven from the fifteen can also be done. The team can then be used on the platform to check how well the models perform compared to the real world managers.

REFERENCES

[1] J. Gatsman, "Fantasy Premier League grows to +7 million users this season" fanarena.com. https://fanarena.com/fantasy-premier-league-growth-users-per-year/ (accessed: May 8,2022)
[2] The Scout, "FPL basics: Scoring points" premierleague.com. https://www.premierleague.com/news/2174909 (accessed: May 8,2022)
[3] Pardee, M. An Artificial Neural Network Approach to College Football Prediction and Ranking. 1999
[4] Bonomo, Flavia, Guillermo Duran, and Javier Marenco. "Mathematical programming as a tool for virtual soccer coaches: A case study of a fantasy sport game". In: International Transactions in Operational Research 21. DOI: 10.1111/itor.12068. 2013
[5] Katara, Karthik Krishnan, Gokul Shetty, Shashank Bankapur, Sanjay Kolkar, Ranjit T S, Ashwin Vanahalli, Manjunath. (2020). Analysis and Prediction of Fantasy Cricket Contest Winners Using Machine Learning Techniques. 10.1007/978-981-15-5788-0_43.
[6] Stolyarov, Arseniy and Gleb Vasiliev. "Predict To Succeed: Optimal Sequential Fantasy Football Squad Formation Using Machine Learning Tools". 2017
[7] Saifi, Murtaza. "Implementation of Machine Learning Techniques to Predict Player Perfor-mance using Underlying Statistics". PhD thesis. Dublin, National College of Ireland. 2018

[8] Ramdas, Delano. Using Convolution Neural Networks to Predict the Performance of Footballers in the Fantasy Premier League. 10.13140/RG.2.2.10010.72645/2. 2022
[9] Gupta, Akhil. Time Series Modeling for Dream Team in Fantasy Premier League. 2019
[10] Lindberg, Adrian and David Sonderberg. “Comparison of Machine Learning ApproachesApplied to Predicting Football Players Performance”. English. MA thesis. Gothenburg: ChalmersUniversity of Technology and University of Gothenburg. 2020
[11] Gunjan, Kumar. “Machine learning for soccer analytics”. In: University of Leuven. 2013
[12] Bonello, Nicholas, Joeran Beel, Seamus Lawless, and Jeremy Debattista. “Multi-streamData Analytics for Enhanced Performance Prediction in Fantasy Football”. In: arXiv preprintarXiv:1912.07441 2019.
[13] Anand, Vaastav. ”Fantasy-Premier-League” github.com. https://github.com/vaastav/Fantasy-Premier-League (accessed: May 8,2022)