Investigating the validity of FPL data in determining player performance and the most impactful players in the English Premier League teams

MSc Research Project
MSc Data Analytics

Nishant Meena
Student ID: x21221839

School of Computing
National College of Ireland

Supervisor: Vitor Horta

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	…Nishant Meena…………………………………………………………
Student ID:	…x21221839……………………………………………………………………………….….
Programme	……MSc Data Analytics……….. **Year:** …2023……….
Module:	………..MSc Research Project…………………………………………………….………
Supervisor:	……………Vitor Horta…………………………………………………………….………
Submission Due Date:	………..14-08-2023…………………………………………………….………
Project Title:	Investigating the validity of FPL data in determining player performance and the most impactful players in the English Premier League teams.
Word Count:	…7233……….. **Page Count**………………20……….……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ………..Nishant Meena…………………………………………

Date: ……………………….13/08/2023…………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Investigating the validity of FPL data in determining player performance and the most impactful players in the English Premier League teams

Student Name: Nishant Meena

Student ID: x21221839@student.ncirl.ie

MSCDADSEPT22B

August 2023

Abstract

The purpose of this research is to figure out how much information collected from the Fantasy Premier League (FPL) dataset can aid the real professional football managers to make accurate judgements of performance of players in English Premier League (EPL). The goal of this project is to find out the potential benefits for football managers/clubs using the FPL dataset to predict the performance of the players before the match and potentially selecting the best starting eleven for their club. In order to achieve this goal, this research aggregates and analyses FPL data from the previous EPL seasons and build Machine Learning models for predicting the performance of the players. This research incorporates the secondary data from other data sources for additional stats not present in the FPL dataset like xG (Expected Goals) and xA (Expected Assists) to check whether adding these secondary features to the FPL dataset improve the prediction results or not. The data is statistically analysed and machine learning models are used to build the models that forecast the performance of the players in an English Premier League club. The information derived from the FPL dataset and their dependency on the player performance can be used for better understanding of how the FPL data could be useful as a decision making tool for real world football managers. To find out the impact of secondary dataset, this study was conducted in two phases. In the first phase only FPL dataset was used to predict the points achieved by the players, and in the next phase the additional features from the secondary dataset were added to observe the change in prediction of the points of the players. The findings of this study suggests that the best performing model in both the experiments is Random Forest and it is determined that the secondary statistics such as xG and xA have minimal impact of prediction of the players' performances, which was concluded as the first experiment which includes only the data from FPL dataset performs slightly better as compared to the second experiment which included the secondary stats, xG and xA. This study will have an impact on team managers, allowing them to apply data-driven strategies to choose their best starting eleven with a much greater certainty of players who will perform better in the football match. Furthermore, it will allow FPL in game managers to better understand how will each player performs before the actual match, allowing them to get a much deeper understanding and use of the statistics available to them to achieve high points in the game and potentially winning the game.

1 Introduction

Football, or soccer (in western world), is universally regarded as the most popular sport in the world, with an expansive global fan base of over 3.5 billion individuals (Sport for

Business, 2017)[1]. Originating in England during the 19th century, modern football has gained immense popularity and now features professional leagues in numerous countries (Joy and Weil, 2019)[2]. In recent years, football has gone through a lot of changes, with the use of data analytics, data driven and statistical models being commonly used for better decision making. Data driven approach is being used in recruiting players, squad selection, and performance evaluation (AnalyiSport, n.d.)[3]. In this context, the use of data-driven insights produced from the FPL dataset gives an appealing avenue for evaluating individual player ratings in the English Premier League

Modern football, as it is called now a days has different complexities which necessitate exact evaluation of player performance to make judgements about the selection of the squad, player acquisitions, transfers, and negotiations of player contracts. In this regard, football teams and players have seen the significance of machine learning modelling and data analytics as important tools. Notably, Manchester City's Kevin De Bruyne avoided using an agent to secure an £83.2 million contract extension until 2025, instead relying on data analysts to assess his influence at the club (McDonnell, 2021)[4]. This instance highlights the football industry's growing dependence on data-driven insights to support discussions and decisions.

Fantasy Premier League (FPL) is an online fantasy football game which has gained significant amount of popularity among the football enthusiasts. It was launched during the season 2002-2003, and now it has over 11 million players (Šuľan, 2023)[5]. The players in the game are referred as “Managers”, as they assume the role of a virtual team manager and create their squad which consist of the real players from the EPL. Each participant is allocated with a designated budget at the start of the season, participants must strategically select a squad of 15 players, selecting 11 players (Fig 1) to compete each week and aiming to maximize points based on their chosen players' performances. The selection of the squad is quite challenging as the points allocated to players for each game week are based on their real life performance in a football match, which includes details such as their recent form, injuries and the difficulty of the upcoming fixture. The scoring system typically includes points for action such as: Saves made by goalkeepers, clean sheets by goalkeepers and defenders, Penalties won/saved, own goals, red/yellow cards, goals scored and assist provided by the players. The criterion followed to allocate the points to players are available on official website of FPL (www.premierleague.com,n.d.)[6]. The player’s performance is directly related to their performance in actual EPL match which means that the real matches have a direct impact of the team created by the FPL managers. The "Starting eleven" can be made up of any formation as long as it includes at least one goalkeeper, three defenders, three midfielders and one striker.

[1] Sport for Business. (2017). The Sporting Year Ahead 2023. [online] Available at: https://sportforbusiness.com/the-worlds-most-watched-sports/.

[2] Joy, B. and Weil, E. (2019). football | History, Rules, & Significant Players. In: Encyclopædia Britannica. [online] Available at: https://www.britannica.com/sports/football-soccer.

[3] AnalyiSport. (n.d.). *How Football Clubs Use Data to Sign Players*. [online] Available at: https://analyisport.com/insights/how-football-clubs-use-data-to-sign-players/.

[4] McDonnell, D. (2021). De Bruyne uses data analysts to broker £83m Man City contract without agent. [online] mirror. Available at: https://www.mirror.co.uk/sport/football/news/kevin-de-bruyne-uses-data-23870686.

[5] Šuľan, M. (2023). *When did FPL start? History of Fantasy Football*. [online] FPL reports. Available at: https://www.fantasyfootballreports.com/history-of-fantasy-football/

[6] www.premierleague.com. (n.d.). *FPL basics: Scoring points*. [online] Available at: https://www.premierleague.com/news/2174909.

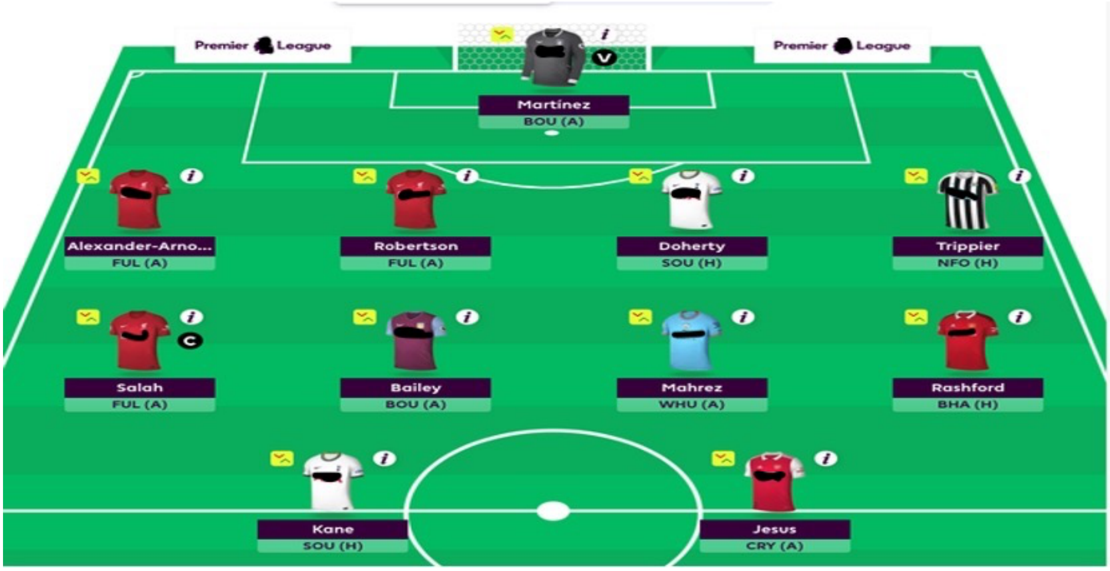


Fig 1 Sample FPL squad (Mirror.co.uk, 2023)

The FPL dataset consist of a large amount data on players performance which is gathered from actual EPL matches, making it a valuable source for analysing the player performance and determining their impact on a football match. In this study the focus is to predict the total points of the players in a game week as the total points given to a player in FPL dataset depends on their on-field performances. As mentioned earlier, if a player scored a goal, made an assist, saves a goal or saves a penalty they are allocated certain number of points based on the match played in the English Premier League (EPL). On the contrary, if players concede a goal, made a foul, receive a yellow card or red card etc. the points are deducted from the total points. This makes total points an important feature which can help in evaluating the performance of players. While previous researches have shown that FPL statistics have the potential to forecast player points, prior studies have primarily focused on predicting player points within the context of the FPL game, in order to gain the maximum points for "Managers" who are playing the FPL game. To analyse FPL data and generate projections, a variety of methodology like machine learning algorithms, ensemble models, statistical approaches, and optimisation strategies have been used. However, no consensus has been achieved if FPL dataset on its own without adding the additional features is enough to evaluate the performance of the players. In this study hypothesis is made that adding additional features such as xG (Expected Goals) and xA (Expected Assists) should improve the prediction of total points gained by a player in a given game week. This study also includes the five seasons as training dataset and one season dataset for testing in order to get the points of the players which is different from the previous work as the related work has been done with training dataset up to three seasons maximum.

This research project aims to address the following research question: *If the insights derived from the Fantasy Premier League (FPL) dataset alone can aid real professional football managers in accurately predicting individual player performance within the English Premier League?*

By addressing the research question and following this structured approach, this study aims to contribute our understanding of if and how the data from the Fantasy Premier League (FPL) can be effectively used to accurately predict the performance of individual players for English Premier League clubs. The conclusions drawn from this research will offer valuable insights for football club managers, enabling them to make data-driven decisions in team selection and helping the people who plays fantasy sports to help them understand the importance of getting insights from the data to win the game.

2. Literature Review

The review of earlier research that has been done in this field up to this point is included in this part. It is crucial to note that little research has been done on predicting players' total points for English Premier League clubs using the FPL dataset. However, some studies have predicted the players' point totals over the course of several Game Weeks (GW) using the FPL dataset.

2.1 Related work on the FPL points predictions

Stolyarov and Vasiliev (2017) conducted the first known attempt to predict player performance in FPL using fundamental machine learning techniques. The authors applied extreme gradient boosted trees, to forecast player performance. They then created an automated manager that makes use of the strategy to form teams for better outcomes. Raghunandh (2018) replicated XGBoost's methodology after a year to create a "FPL Captain Classifier" to forecast who should be the captain of an FPL team as they receive a point allocation that is twice as high as their game week (GW) score. However, Stolyarov and Vasiliev applied their predictions to more general circumstances without examining their accuracy. Raghunandh (2018), on the other hand, just created a model to choose the captain for an FPL manager, not the entire team or the top players in each position.

Saifi (2018) investigated prediction accuracy by employing ensemble-based classification modelling with extreme gradient boosted machines and random forests, as well as expected Goals (xG) and expected Assists (xA) that were added to the FPL dataset from other data sources. Similar to this, Bonello et al. (2019) used gradient boosted machines along with the examination of statistical data, including past performances, ratings of the difficulty of upcoming fixtures, betting market analyses, and opinions of public and experts via social media platforms and web publications. (Bonello et al.) were able to boost the FPL leader board ranks from 800,000 (top 13%) to 30,000 (top 0.5%) of the total 6.5 million participants by using this method of combining numerous datasets. Additional data features, according to Saifi (2018) and Bonello et al. (2019), increased prediction accuracy. This study, which was inspired by their work, makes use of the xG and xA data characteristics from the football statistics website understat.com. However, the problem statement in this study is considered as regression, and the dataset used for the model training and testing is made up of the preceding five seasons.

(Gupta, 2019) employed time series modelling to predict players' FPL points based on their performance in the previous season. This study set out to find out how effectively time series modelling methods could predict player performance in fantasy football leagues. In this study, player points are predicted over time utilising time series using a combination of Autoregressive Integrated Moving Average (ARIMA), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM), followed by the application of (Linear Programming) LPP to maximise overall points. Due to the FPL game's financial restrictions, he also employed linear optimisation to choose the players for the team. By providing a novel framework for predicting player performance in FPL using time series modelling techniques, this research was expected to make a contribution. The findings of this study demonstrated that the performance of the predicted team during the current season, exceeded the expectation. In order to anticipate the players' points for each game week and examine the performance of the players throughout the season, the methods from this study are applied in this paper's research question. The use of optimisation will not be taken into consideration because the budget constraint is not a part of the research issue for the current study. The study's drawback is that it only takes into account data from the 2016–2017 Fantasy Premier League season, which could not be indicative of player performance in previous seasons. Additionally, Gupta's best solution somewhat overfit the data, underestimating the overall team score by 87 points. The preceding five seasons' data will be taken into account in this study to get around this limitation and help the models learn more effectively from the training datasets.

In their research, (Bhatt et al., 2019) propose the use of a system dubbed Smart Crowd, which selects a small, diverse community of participants using tweets, to predict the best captains for an FPL team on weekly basis. By utilising social media-based variety, the approach samples more intelligent crowds than random crowds, single participants, and crowds made up of the top experts identified from historical performance data. The model outperformed the general players 93% of the time. The results and technique used in the Smart Crowd study may provide insight into the development of machine learning models for forecasting player performance in the FPL in the current study.

In order to predict football player performance in the English Premier League, Lindberg and Söderberg, n.d. (2020), applied three machine learning algorithms: Support Vector Mechanism (SVM), Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM) .The author looked at the issue from two angles: as a classification and as regression problem. They looked at the data from the last four FPL seasons, and the results showed that LSTM is the best model. They also found that the results are better when the problem is viewed as a regression problem rather than a classification one. Owing to their findings, this study focuses the problem as a regression problem instead of treating it as a classification problem. They also showed a bias in the data, which indicates that a player's recent form are probably going to be a good indicator of how successful they will be in the near future.

In a more recent study (Bangdiwala et al., 2022) proposes a method to predict how many points each football player would score during the course of the season using three machine learning models (Linear Regression, Decision Tree, and Random Forest). The models take into account a wide range of variables, such as fixture difficulty, team form, and the inventiveness and threat of specific individuals. The Linear Regression model performed the best out of the three, offering managers a useful tool to use when selecting players for their teams. The authors of the other study (Rajesh et al., 2022) created a recommendation system for the Fantasy Premier League that uses statistical analysis and machine learning algorithms to predict the expected points of a player. They compare their models to the 2021–22 English Premier League season using data from the FPL API. Both the FPL managers and the professional managers have preconceived beliefs about individual players, this recommender system addresses the problem of biases in selection of the player in starting eleven. One drawback of this model is that they only use data from a single season, which might not be sufficient to capture subtle differences in individual performance over time. The authors concur that including sentimental data may improve the accuracy of their models, although their current study does not explore this. To estimate player's points, the current study tests various machine learning methods on the FPL dataset together with additional variables from other data sources. The current effort differs from previous ones in that it focuses on determining how accurately the player performance can be predicted using the insights from FPL dataset only without tuning the models.

2.2 Related works to prediction of a football match

The purpose of the study was to see if Fantasy Premier League data could be utilised to forecast football game outcomes and improve existing networks (Kristinsson, 2022). The study's findings showed that an RNN's prediction accuracy was improved significantly with the addition of FPL data. For further enhancement of the effectiveness of this network, the authors proposed a number of solutions for future research, including using real-time line-ups, weather data, sentiment analysis, referee biases, and team historical results against the same opponent. They also suggested using the managers top players. They also recommended using a neural network while exploring with the transformer design. This study creates possibilities of the use case of FPL dataset for the prediction of football match outcome and optimisation of the network, which suggests that the use of FPL dataset is potentially reliable in predicting the players' performance too.

The paper titled "Premier League Predictions Using Artificial Intelligence" by (Balawejder, 2022) examines the use of artificial intelligence to predict football games in the Premier League. Balawejder add up on the past work and integrates the Fantasy Premier League dataset to boost the model's accuracy. He compares the accuracy of the LSTM and GRU, two different types of recurrent neural networks, and concluded that GRU performs better than LSTM by 20% accuracy. The most effective network, as determined by Balawejder's investigation of the GRU's architectural layout, comprises three hidden layers, followed by dropout layers.

In their research, (Nazim Razali et al. 2017) employed Bayesian networks to forecast football game results. They utilised data from football-data.co.uk. They attained an average accuracy of 75% over the course of three seasons using their Bayesian networks. This study establishes that the performance of the model can be assessed using information from the official website which maintains the football stats. The use of performance analysis has gained importance, according to Wright, Carling, and Collins (2014), since professional football becomes more competitive and even small alterations can have a significant impact on results. They provided a full overview of the application of performance analysis in football coaching and emphasised the key elements coaches should take into account when utilising performance analysis technologies. This study shows that performance analytic techniques can be used by the football managers to discover the areas where the performance of the team could be improved.

2.3 Related work related to predicting players performance in football

(Pariath et al., 2018) in their research used the data from EA Sports FIFA game to predict the overall performance of the players (ratings given in FIFA game) and the market value of the players. In this study authors used Linear regression machine learning technique for both the prediction. For the performance model, the linear regression technique gives an accuracy of 84.34% and the standard deviation of 0.84. The study also provides the impact of market value on overall performance rating of the players, where the model is built with an accuracy of 91%. The overall performance rating used in the second model are prediction by the first model as the young players have very low market value at the beginning of their careers.

2.4 Related work which uses the FPL dataset

In the research paper titled "Identification of skill in an online game: The case of Fantasy Premier League" investigates whether the results of the competitions based on individual performance, such as online fantasy sports, are the results of skill or luck. (O'Brien, Gleeson and O'Sullivan, 2021) in their study focused on the FPL, in the study they addresses elements that lead to manager's(in-game managers) performance in the game and the steps they take to improve their chances of winning. The study shows that in FPL, results are impacted by union of skill and strategic decision-making. Managers that participate in effective long-term planning and make consistent, educated decisions outperform their peers throughout numerous seasons.

2.5 Related works to using of Fantasy Data in other sports

In his study (Robinson, 2020), works on fantasy football, a famous simulation game that allows the participants to become National Football League (NFL) team owners. It works similar to FPL where participants select the players for their team and points are allotted based on real life performance of players. The author used ARIMA models to predict the points of the players for the outcome of next 16 games and the accuracy of the model was conducted based on the Mean Absolute Percentage Error (MAPE) metric. The methodology's application of ARIMA models and evaluation metric provides a quantitative approach to enhance prediction accuracy.

2.6 Literature review conclusion

Based on the literature review, there have been different machine learning techniques and algorithms that have been used for the prediction of performance of the players using the FPL dataset. These Machine Learning techniques involves time series modelling, extreme gradient boosted trees, linear regression, random forest, SVM, MLP and LSTM, etc. It has been observed from the literature review that the prediction accuracy was improved minutely with the use of additional features such as xG and xA, past performances, the difficulty of the next features, and opinions of public and experts opinions. The literature review also identifies a bias that the recent form of the players affects their performance for the upcoming fixtures. Additionally, it has been observed that the accuracy of the models increases by the use of data from social media platforms and intelligent crowdsourcing. Overall, the literature review emphasises, how important it is to consider a variety of features and apply the right machine learning models in order to accurately predict the performance of the player using the FPL dataset.

3. Methodology

The methodology used in this study is Knowledge Discovery in Database (KDD). Fig 2 shows the basic understanding of this methodology. It is important to mentioned here that this study has been conducted in two phases, the first phase involves dataset only from FPL dataset and for phase two the secondary dataset is merged with the FPL dataset to see what impact it creates on the predictions.

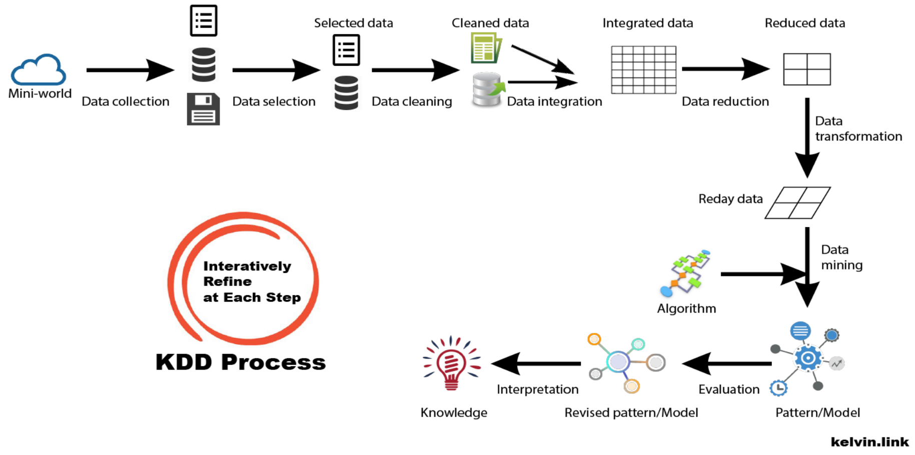


Fig 2 KDD methodology (Chehab, 2020)

The KDD methodology involves different steps before building a machine learning model. The steps used in this research paper are mentioned below:

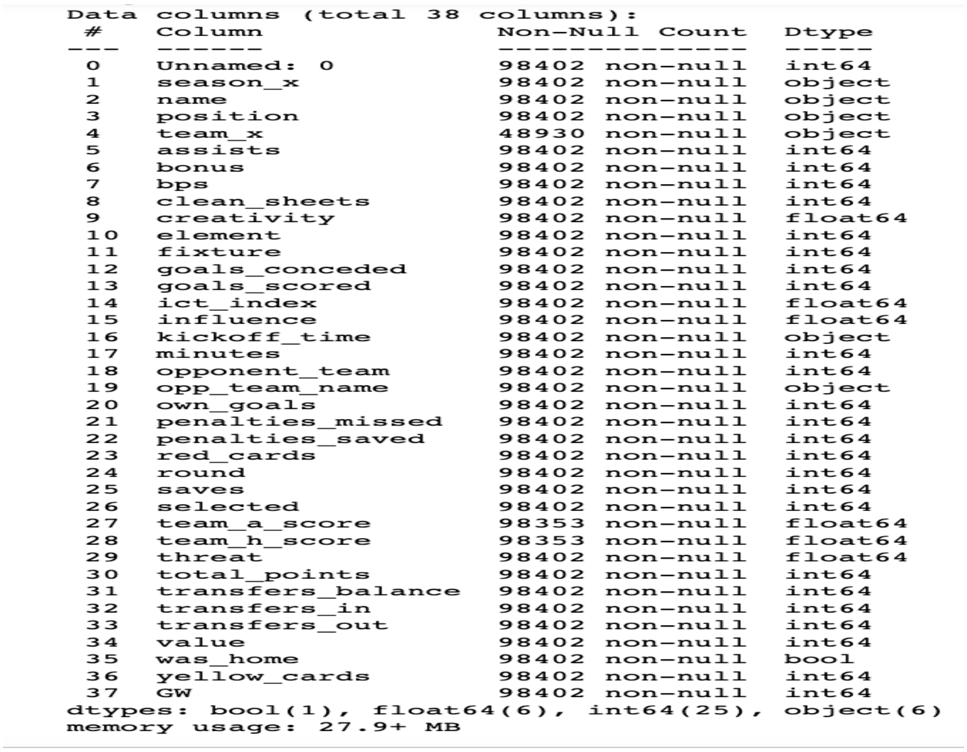
- **Data Collection/Selection**

This study involves data from two different sources as mentioned in the introduction section. The first phase of data extraction was completed from vaastav's github repository (*https://github.com/vaastav/Fantasy-Premier-League*). Which includes each game week data from 2016-2017 season to 2021-2022. The data used in this study of last six seasons (2021-2022 to 2016-2017. Where five season is used to train the model and the last season 2021-2022 is used as test dataset. Apart from this, for second phase the data has been used from *understat.com* to include additional features such as xG (expected goals), xA(expected assist) and players match ratings. xG is a measure of the quality of shots, it is a metric that estimates the probability of a shot which results into a goal. This takes into account various factors including the likes of distance from the goal, shot angle, the type of shot, and defensive pressure. xG helps to assess the quality of scoring opportunities created by a player (Williams, 2019)[7]. xA on other hand is expected assist which is a statistical measure that calculates the likelihood of a pass which leads to a goal. Like xG, xA also takes into account different factors such as pass location, type of pass, position of the player receiving the pass and the finishing ability of the player receiving the pass. xA helps to evaluate the player's creativity and ability to set up goal-scoring chances. The combination of python libraries and web scraping techniques were used to retrieve player's stats. To process and manipulate the data, the 'json' library was used to parse the JSON data and 'pandas' library to organize the data in structured manner. In order to extract data from HTML pages, 'BeautifulSoup' library was used. After the collection of data, as stated above the first experiment involves data only from FPL dataset and for the second phase both the datasets were merged on the basis of 'date' and the 'name' column in both datasets. The scraping of the data was a challenging part because the players id required to scrape the data, however in club football competitions, a player can be transferred to other club during or after the season, that is the reason it was important to join dataframes based on date and name of the players.

[7] Williams, J. (2019). *What is xG and xA? Expected goals and assists explained*. [online] Manchester Evening News. Available at: https://www.manchestereveningnews.co.uk/sport/football/xg-expected-goals-assists-explainer-15663493

- **Data Pre-Processing**

Phase 1 involves data only from the FPL dataset which contains a total of 37 features as seen in fig 3. For experiment one, in this step the variables which were not required such as unnamed:0 and ‘kickoff_time’ were removed as they were not required for our analysis. Another important criteria to judge a players’ performance is the number of minutes played by the player in a given match. As in some cases a player players only 5 minutes for example and scores which effects the points of the players. So in order to remove this biasness players who have played more than 20 minutes are considered and rest of the rows are deleted from the dataset. Missing values were also checked in the dataset and it was seen that from season 2016 to 2018 teams name were missing for the players in the dataset. This was inserted manually by cross verifying the teams of the player from official website of English premier league. Next it was observed that the data type of ‘season’ was in a string format and displayed as 2016-17 to improve the readability it was transformed to 2016 and so on.



```
Data columns (total 38 columns):
 #    Column               Non-Null Count  Dtype
---   ------               --------------  -----
 0    Unnamed: 0           98402 non-null  int64
 1    season_x             98402 non-null  object
 2    name                 98402 non-null  object
 3    position             98402 non-null  object
 4    team_x               48930 non-null  object
 5    assists              98402 non-null  int64
 6    bonus                98402 non-null  int64
 7    bps                  98402 non-null  int64
 8    clean_sheets         98402 non-null  int64
 9    creativity           98402 non-null  float64
 10   element              98402 non-null  int64
 11   fixture              98402 non-null  int64
 12   goals_conceded       98402 non-null  int64
 13   goals_scored         98402 non-null  int64
 14   ict_index            98402 non-null  float64
 15   influence            98402 non-null  float64
 16   kickoff_time         98402 non-null  object
 17   minutes              98402 non-null  int64
 18   opponent_team        98402 non-null  int64
 19   opp_team_name        98402 non-null  object
 20   own_goals            98402 non-null  int64
 21   penalties_missed     98402 non-null  int64
 22   penalties_saved      98402 non-null  int64
 23   red_cards            98402 non-null  int64
 24   round                98402 non-null  int64
 25   saves                98402 non-null  int64
 26   selected             98402 non-null  int64
 27   team_a_score         98353 non-null  float64
 28   team_h_score         98353 non-null  float64
 29   threat               98402 non-null  float64
 30   total_points         98402 non-null  int64
 31   transfers_balance    98402 non-null  int64
 32   transfers_in         98402 non-null  int64
 33   transfers_out        98402 non-null  int64
 34   value                98402 non-null  int64
 35   was_home             98402 non-null  bool
 36   yellow_cards         98402 non-null  int64
 37   GW                   98402 non-null  int64
dtypes: bool(1), float64(6), int64(25), object(6)
memory usage: 27.9+ MB
```

fig 3 Features in FPL dataset

For experiment two, the kickoff_time was an important feature as it contains the data on which the game was played. Using python code the date was extracted from this variable and stored in a new column as the date column is important in joining the FPL dataset with the secondary dataset which contains xG and xA values. After merging the data all the steps which were followed in phase one were cloned to clean the merged data were followed.

Out of these features there are few features which are pre-game features such as name, position, team, was_home etc. On the other hand other features such as goals_scored, goals_conceded, assist etc are the in-game features which are added to the dataset after a football match of each game week. The ‘total_point’ feature is the target variable in this study, total points feature emerges as a key and comprehensive metric for analysing players’ performances. Total points, are calculate using a predetermined scoring system that allocates values to various on-field actions such as goals, assists, and clean sheets. It provides an objecting and quantitative measure of players' contribution in a football match. It accurately reflects players' performance for each game week and overall season. Its importance extends beyond the fantasy team selection, as it provides a historical foundation for trend analysis and assists in evaluating performances.

- **Exploratory Data Analysis**

The next step followed was exploratory analysis as it provides insights in the data and is helpful in feature selection and model building phase[8]. Figure 4 shows the distribution of the points for the players which ranges between somewhere from -5 to 30. So the maximum points a player got in our dataset was 30 across all the game weeks available.

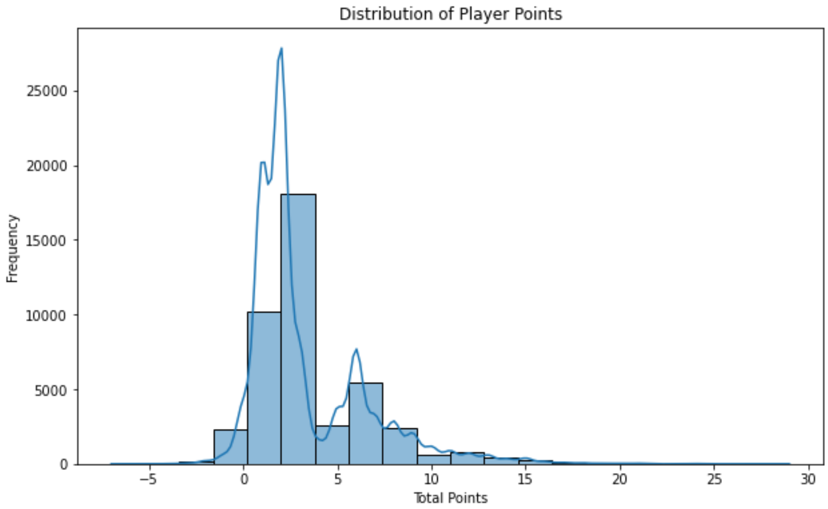


fig 4 Distribution of the players points

To check the correlation of the variables with the target variable which 'total_points' a correlation heatmap was plotted as seen in figure 5. In figure 5, it can be seen the variables are correlated with our target variables and which will be useful while building the model. As we know that football is a team sport and the performance of players depends on their fellow teammates as well. The big teams in EPL like Chelsea, Manchester United, Manchester City etc have some of the best players in the world and which can be seen in figure 6.

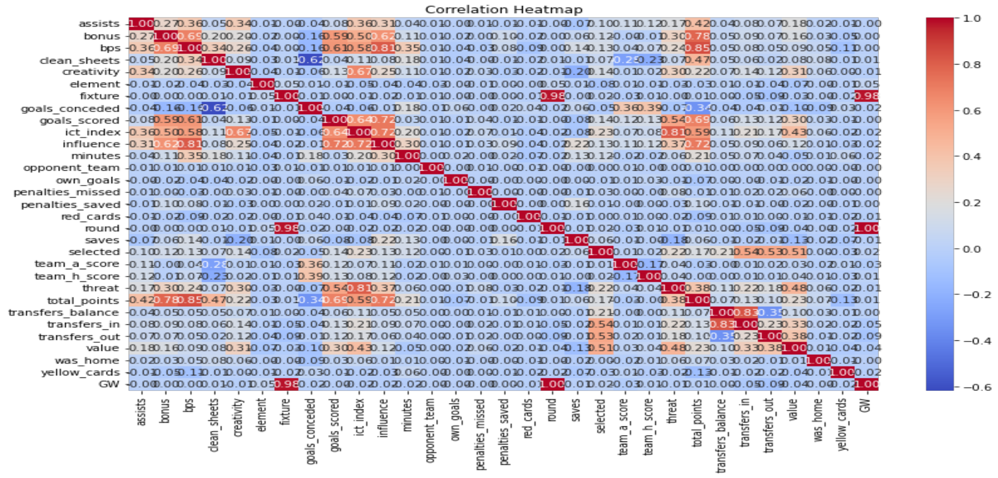


fig 5 Correlation heatmap

[8] Google Cloud Blog. (n.d.). *Building ML models with EDA, feature selection*. [online] Available at: https://cloud.google.com/blog/products/ai-machine-learning/building-ml-models-with-eda-feature-selection.

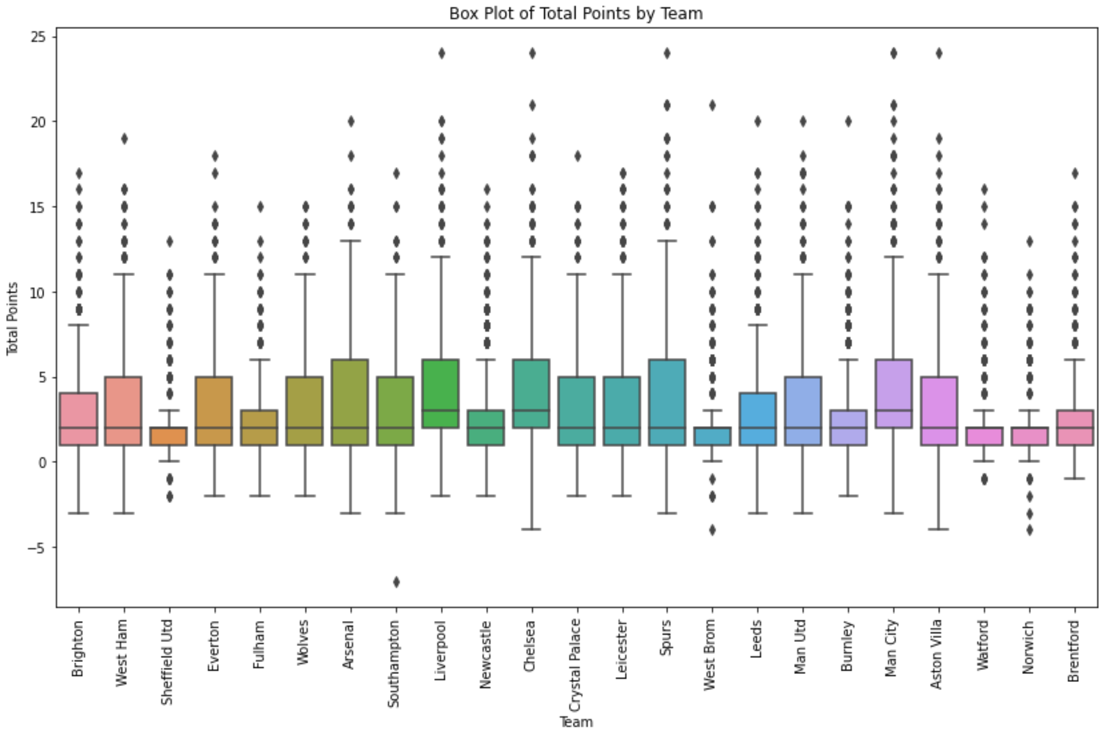


fig 6 Boxplot of the total points based on teams

- **Data Mining**

As mentioned in the related work section, there has been a limited work in terms of FPL data used for this problem statement. The solution of the problem statement in this paper has been carried out using two experiments as mentioned earlier. The first experiment uses the FPL data without the addition of secondary dataset and three machine learning models (Linear Regression, Random Forest, and XGBoost). In the second experiment, the secondary dataset was merged with the original dataset and applying the same models in order to determine whether FPL dataset alone can provide enough insights to evaluate the performance of the players. The below mentioned models have been considered in both the experiments:

1. **Linear Regression**
Because of its simplicity, interpretability, and broad applicability to a wide range of scenarios, linear regression is a widely used statistical technique in regression analysis[9]. Since this studies problem is related to regression a basic linear regression model is build to check the predictions.

2. **Random Forest**

As mentioned in the related work section, random forest have shown promising results for predicting the players' performance. Random Forest is a bagging method[10] that employs the Ensemble Learning technique. It grows as many trees as possible on the subset of data and

[9] Shin, T. (2021). *3 Reasons Why You Should Use Linear Regression Models Instead of Neural Networks*. [online] Medium. Available at: https://towardsdatascience.com/3-reasons-why-you-should-use-linear-regression-models-instead-of-neural-networks-16820319d644.

[10] Naresh Kumar (2019). *Advantages and Disadvantages of Random Forest Algorithm in Machine Learning*. [online] Blogspot.com. Available at: http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html.

then merges the results of all the trees. As a result, the overfitting problem in decision trees is reduced, as is the variance, which increases accuracy.

3. **XGBoost**

Bonello et al. (2019) , in their study achieved a great accuracy while using this Machine learning technique and were able to achieve a rank below 30,000 from 6.5 million players playing FPL which proves that XGBoost performs well on FPL dataset. Inspired from their work this study uses the XGBoost technique. It is also used because of its ability to handle the missing values in the data (FPL dataset has some missing values) and it handles large dataset efficiently[11].

4. Design Specification

In the first experiment only the FPL dataset was loaded into the jupyter notebook. After the necessary data cleaning and transformation of the dataset, Exploratory data analysis was performed to find insights from the dataset in order to help finding the important features for building the model. Before building the model a feature importance technique was applied called Permutation feature importance which measures the increase in the prediction error of the model after permuting the feature's value[12]. Then the ML models were build (Linear Regression, Random Forest and XGBoost). The evaluation of the models were based on the evaluation metrics such as Mean Square Error (MSE), R squared value (R2) etc. In case of second experiment the data was scraped from understat.com. However, minor changes were made to the name of the players in order to merge the scrapped data with the primary dataset of FPL. Then similar approach was followed as in the first experiment. Figure 7 shows the design specification of the work flow.

5. Implementation

The FPL dataset once loaded to the environment, the unnecessary columns were removed like Unnamed:0 and kickoff_time. It was also important to remove the players who have played less than 20 minutes in a game as it can be biased for a model. In the FPL dataset from season 2016 to 2019 teams name were missing which were manually entered in order to remove the missing team names as it is important to know which teams have the highest performing players. For better readability and model understanding season column was slightly altered and hyphen was removed. For example season value of 2016-17 was transformed to 2016. After which necessary EDA was performed to get the insights from the dataset. Before building the model the dataset was sliced into training and testing dataset where season from 2016 to 2020 were taken as training dataset and 2021 was considered as testing dataset. This study focuses on finding player performance for four positions Forward (FWD), Midfielder (MID), Defender (DEF) and Goalkeeper(GK). Players' playing position impact their influence in a football match as a striker is more expected to score a goal compared to a defender. A permutation feature importance was run to extract the relevant features before building the model and the

[11] apmonitor.com. (n.d.). *XGBoost Regressor*. [online] Available at: https://apmonitor.com/pds/index.php/Main/XGBoostRegressor#:~:text=One%20of%20the%20key%20advantages.

[12] Molnar, C. (n.d.). *5.6 Permutation Feature Importance | Interpretable Machine Learning*. [online] *christophm.github.io*. Available at: https://christophm.github.io/interpretable-ml-book/feature-importance.html.

result of which can be seen in figure 8 below. The important features extracted seems relevant to every position as it cover all the important stats for all the positions (FWD, MID, DEF, GK).

For the second experiment the data was scraped from understat.com website which contains the xG (expected goals) and xA (expected assists) and was saved to a data frame. However, the names of the player on the understat website had special character (numerical, spaces and hyphen, accented characters) and they were converted to English character in order to merge both the data frames. The data frames were merged on full name and the date column available in both the dataset. After merging the data frames redundant columns were dropped from the merged data frame.

Opta, a sports analytics company, has developed a set of performance indicator known as xG and xA, which helps in better understanding of football stats and providing high accuracy in evaluating a performance of a team or player. The xG and xA numbers have also received public attention after the media houses such BBC and Monday Night Football (Sky Sports) have started extensively using them to asses a team or player's performance[13]. Opta defines expected goals (xG) as "Expected goals (xG) assesses a shot's quality based on numerous factors, including assist type, shooting angle and distance from goal, whether it was a headed shot, and if it was classified as a major chance." On the other hand expected assist is defined as "Expected assists (xA) calculates the probability that a particular pass will result in a goal. It takes various things into account, including the type of pass, pass end-point, and pass length." Both these parameters gives a projection on how many goals scored or assist could have been provided by the player in a match. After the merging of the dataset, the three models were trained using 2016 to 2020 seasons as the training dataset and 2021 as the test dataset. This approach was same for both the experiment with the only difference that expected goals and expected assists features were added in the second experiment. After training the models total points of the players' were predicted using all three models. The evaluation of these models were based on evaluation metrics such as Mean Square Error (MSE), R square value and Root Mean Squared Errors (RMSE).

[13] Gosavi, A. (n.d.). *Expected Goals explained: What are xG and xA and why are they a good measure of player performances*. [online] www.sportskeeda.com. Available at: https://www.sportskeeda.com/football/what-are-expected-goals-xg-and-expected-assists-xa-why-are-they-a-good-measure-of-player-performances.

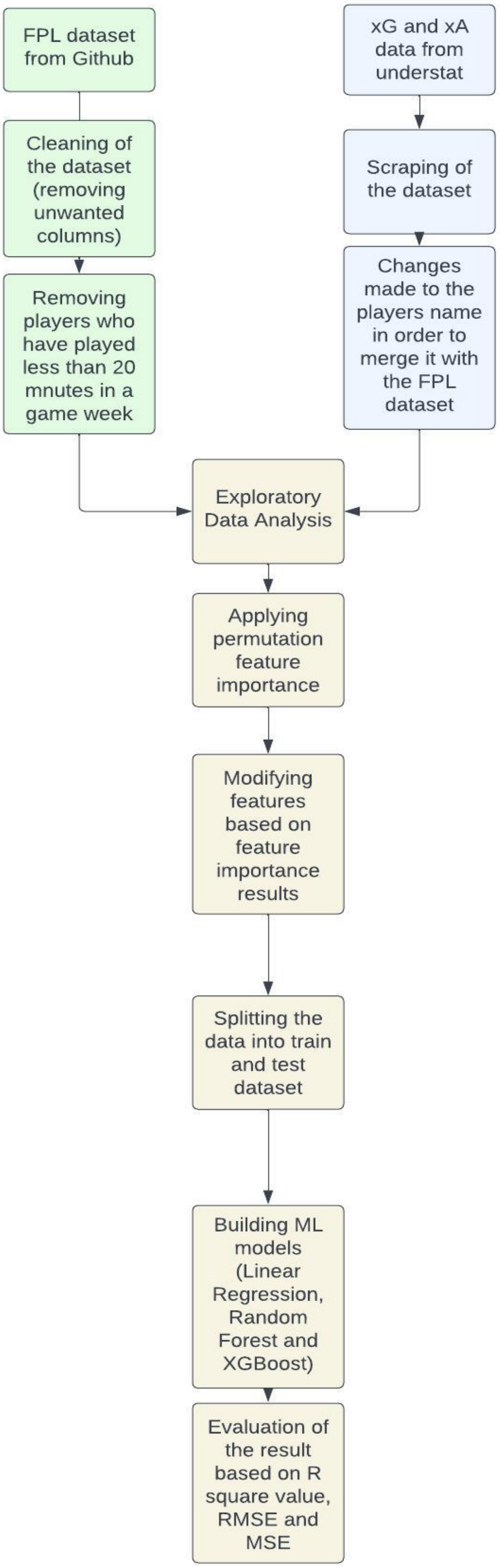

FPL dataset from Github
Cleaning of the dataset (removing unwanted columns)
Removing players who have played less than 20 mnutes in a game week
xG and xA data from understat
Scraping of the dataset
Changes made to the players name in order to merge it with the FPL dataset
Exploratory Data Analysis
Applying permutation feature importance
Modifying features based on feature importance results
Splitting the data into train and test dataset
Building ML models (Linear Regression, Random Forest and XGBoost)
Evaluation of the result based on R square value, RMSE and MSE

fig 7 Design Specification work flow

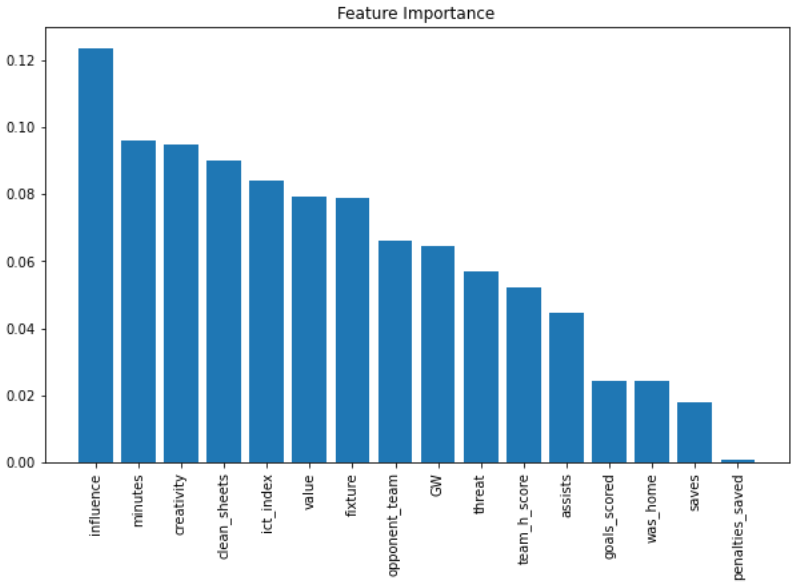


fig 8 Permutation Feature Importance

Sl. No.	Feature	Description
1	influence	Shows the influence of players in a football match
2	minutes	Number of minutes played by the player in a match
3	Creativity	The creativity of the player to find a pass to other teammate
4	clean_sheets	Shows if the team didn't concede a single goal
5	ict_index	this is a combination of influence creativity and threat of a player in a given game
6	value	the value of the player in the FPL game
7	fixture	it is the value given to a match between two teams in a given season
8	opponent_team	Shows the opponent team against which the player is playing
9	GW	Game week of the season
10	Threat	value assigned to players based on their threat level of creating chances or creating goals
11	team_h_score	goals scored by the home team
12	assists	number of assist by the player in a particular game week
13	goals_scored	goals scored by the player in given game week
14	was_home	whether the player was playing in his home ground or away
15	saves	the number of saves by the goalkeeper
16	penalties_saved	the number of penalties saved by the goalkeeper

Table 1 Important feature description

6. Evaluation

The intention of this study is to find out if the insights derived by the FPL dataset can help in predicting the performance of players. To evaluate the performance of models built the evaluation metrics are key to answer the research question. The evaluation metrics for both the experiment can be seen in section 6.1 and 6.2.

6.1 Experiment One

The evaluation metrics for the first experiment can be seen in the table 2.

Sl. No.	Evaluation Metric	Linear Regression	Random Forest	XGBoost
1	Mean square error	1.67	**1.17**	1.26
2	Root mean square error	1.29	**1.08**	01.12
3	R square value	0.83	**0.882**	0.8730
4	Mean absolute error	0.79	**0.71**	0.98

Table 2 Evaluation metric for experiment one

This comparison of models can also be seen in figure 9, which clearly shows that our best performing model is Random forest model with the least RMSE and MSE values and maximum R squared value. XGBoost model is also similar with a minute difference and the Linear Regression Model is the worst performing model out of all three models.

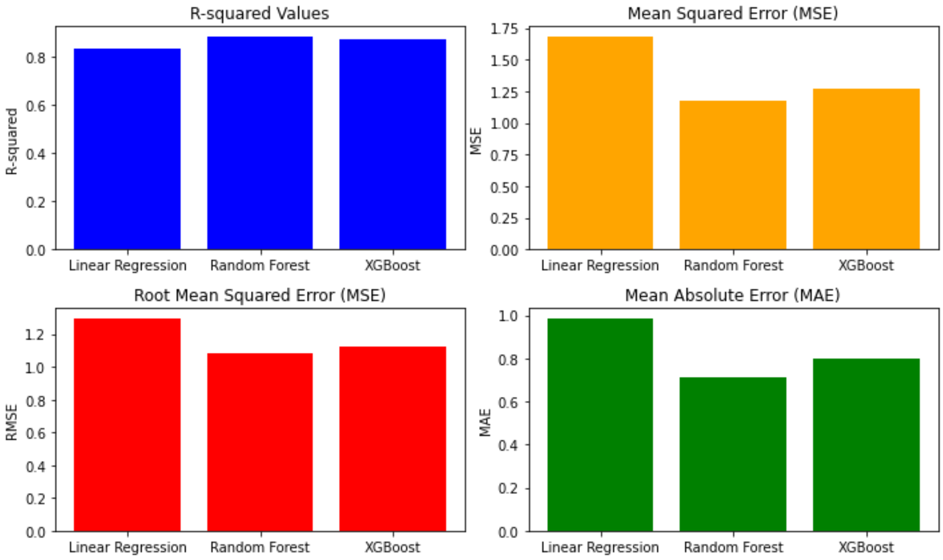


fig 9 Comparison of Evaluation Metrics of all three models (Experiment 1)

6.2 Experiment Two

Similarly, for second experiment the evaluation metrics can be seen in table 3 below

Sl. No.	Evaluation Metric	Linear Regression	Random Forest	XGBoost
1	Mean square error	1.708	**1.20**	1.283
2	Root mean square error	1.306	**1.095**	1.132

3	R square value	0.8348	**0.8840**	0.8758
4	Mean absolute error	1.015	**0.74**	0.820

Table 3 Evaluation metric for experiment two

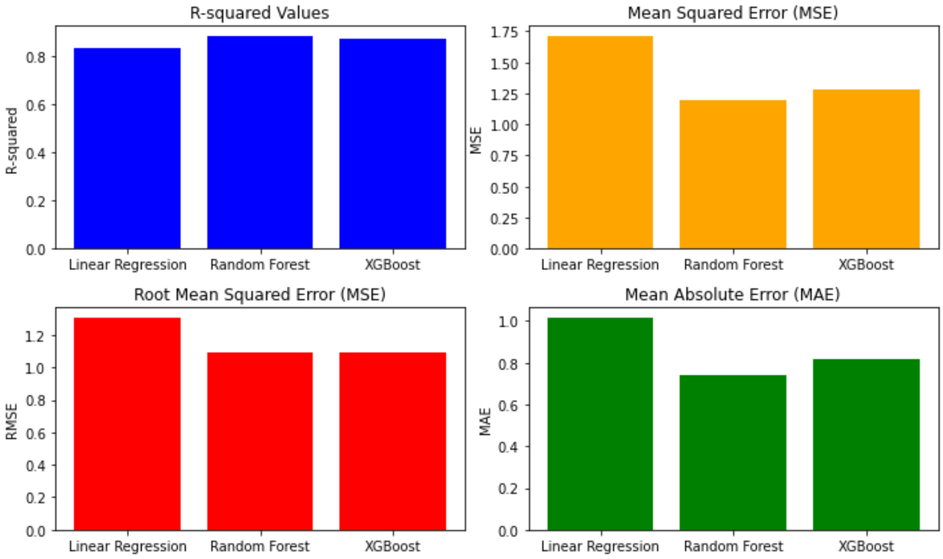
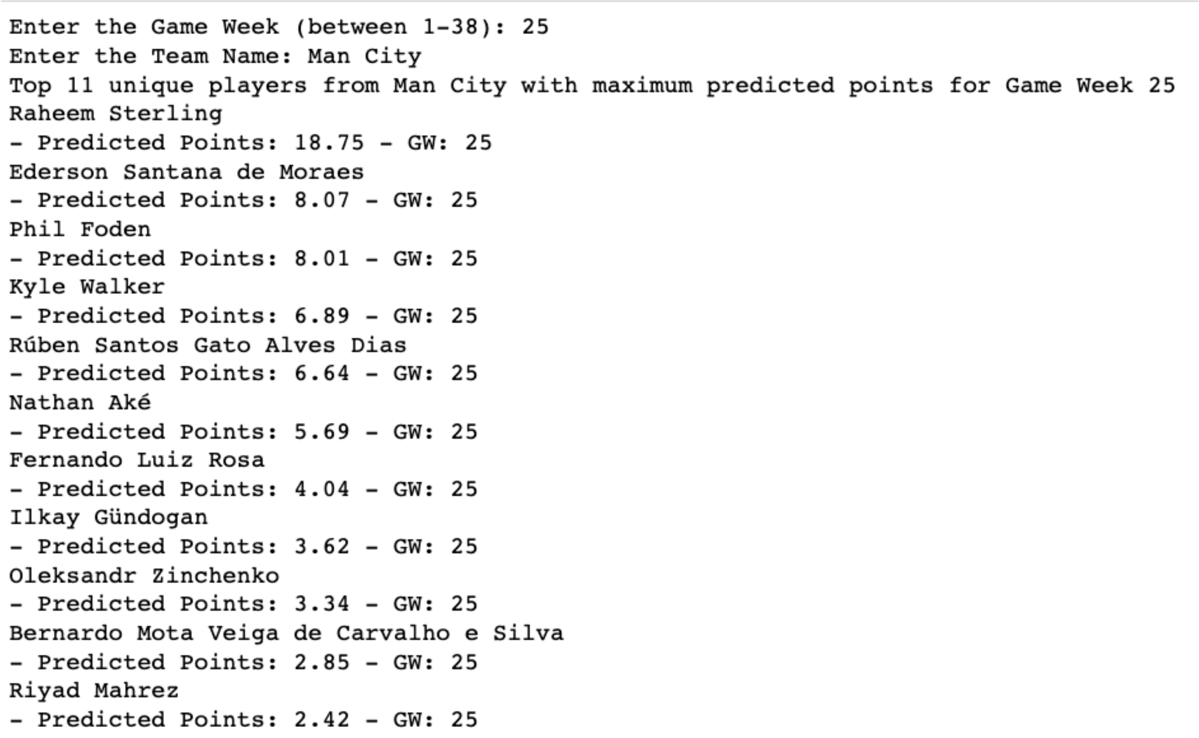


fig 10 Comparison of Evaluation Metrics of all three models (Experiment 2)

7. Discussion

The above tables and graphs show that the xG and xA stats (from understat.com) minimally affect the model and the results are marginally better in case of FPL dataset. The accuracy of the model in this study is measured by the R squared value as it is a regression problem. R squared is a statistical measure commonly used in regression analysis. In regression model, it represents the amount of the change in the dependent variable (‘total_points’ in our case) that can be explained by the independent variables(selected features in our case). R squared measures how well the independent variables explain the variation in the dependent variable. In this study the measure of R squared value for first experiment for random forest is 0.882 which means that 88.2% of the variability in the dependent variable is explained by the independent variables used in the model. The remaining 11% of the variability is not accounted for by the model and may be attributed to other factors. For XGBoost model in first experiment this value takes a slight dip to 0.873 which explains that 87.3% of the variability in the dependent variable is explained by the independent variables, a total difference of 1.1%. On the contrary, the R-squared value for second experiment for random forest and XGBoost is slightly increased to 0.884 which means that in terms of accuracy of the model without any tuning of the hyperparameters the FPL dataset with addition of secondary features (xG and xA) gives slightly better results. The other two evaluation metrics are MSE (Mean squared error) and RMSE (Root mean squared error). MSE is a metric that calculates the average squared difference between expected and actual values in a dataset. The average amount of errors between anticipated and observed values is quantified, with greater errors penalised more heavily due to the squaring process. On the other hand, The square root of the MSE is used in RMSE, a form of MSE. It is used to calculate the average difference between expected and actual values using the same units as the target variable. RMSE is a more interpretable measure of error that is frequently used when communicating the model's performance. In both the experiments the value of RMSE and MSE is better in random forest compared to the rest of the models. The MSE and RMSE values are better in first experiment compared to the second experiment. The results obtained after the evaluations show that xG and xA have shown minimal impact on prediction of the players performance, indicating that FPL dataset alone is enough to determine the performance of the players. Hence, the hypothesis made earlier that adding the additional parameters(xG and xA) to FPL dataset will improve the prediction of player performance is rejected. After selecting the final model as to be Random Forest in experiment one, predicted points for Manchester City for game week 25 can be seen below in figure 11.



```
Enter the Game Week (between 1-38): 25
Enter the Team Name: Man City
Top 11 unique players from Man City with maximum predicted points for Game Week 25
Raheem Sterling
- Predicted Points: 18.75 - GW: 25
Ederson Santana de Moraes
- Predicted Points: 8.07 - GW: 25
Phil Foden
- Predicted Points: 8.01 - GW: 25
Kyle Walker
- Predicted Points: 6.89 - GW: 25
Rúben Santos Gato Alves Dias
- Predicted Points: 6.64 - GW: 25
Nathan Aké
- Predicted Points: 5.69 - GW: 25
Fernando Luiz Rosa
- Predicted Points: 4.04 - GW: 25
Ilkay Gündogan
- Predicted Points: 3.62 - GW: 25
Oleksandr Zinchenko
- Predicted Points: 3.34 - GW: 25
Bernardo Mota Veiga de Carvalho e Silva
- Predicted Points: 2.85 - GW: 25
Riyad Mahrez
- Predicted Points: 2.42 - GW: 25
```

fig 11 Best players for Manchester City for game week 25

8. Conclusion and Future work

Sports industries, specifically the sport of football has now been using data driven techniques and approaches to thoroughly evaluate the player performance. For example advanced tracking technologies like GPS and RFID are used to collect data on players' movements, distance covered, speeds, accelerations, and more. In this study a comprehensive study is conducted to investigate the potential benefits of using FPL data to evaluate the performance of the players in the English Premier League which involved two experiments one with only FPL dataset and other with secondary attributes like xG and xA. The findings of the experiments suggested that the xA and xG statistics have minimal impact on predicting the performance of the player, and FPL data alone is effective in determining player performance. The best performing model in both the experiment is found out to be Random Forest, providing the most accurate predictions in terms of Mean Square Error (MSE), Root Mean Square Error (RMSE), and the value of R squared value was increased by on 0.2% in the second experiment. The study concludes that the insights derived from the FPL dataset can be used by the real world managers in order to evaluate and predict the performance of their player in order to help them in predicting the best starting line-up of eleven for their team. In terms of future work, as this study focuses only on testing the validity of FPL dataset with or without the additional parameters using the basic machine learning models, the accuracy of the prediction could further be improved by model tuning techniques. Different ensemble methods or advanced deep learning architectures might provide more insights and more accurate predictions. More detailed work could be done to predict the real-time performance of the players that continuously updates the player performance based on the ongoing matches and provides managers with dynamic insights.

References:

1. Pariath, R., Shah, S., Surve, A. and Mittal, J. (2018). Player Performance Prediction in Football Game. *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. doi:https://doi.org/10.1109/iceca.2018.8474750.
2. Balawejder, M. (2022). *Premier League Predictions Using Artificial Intelligence*. [online] Nerd For Tech. Available at: https://medium.com/nerd-for-tech/premier-league-predictions-using-artificial-intelligence-7421dddc8778.
3. Bangdiwala, M., Choudhari, R., Hegde, A. and Salunke, A., 2022, August. Using ML Models to Predict Points in Fantasy Premier League. In 2022 2nd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-6). IEEE.
4. Bonello, N., Beel, J., Lawless, S. and Debattista, J. (2019). Multi-stream Data Analytics for Enhanced Performance Prediction in Fantasy Football. *arXiv:1912.07441 [cs, stat]*. [online] Available at: https://arxiv.org/abs/1912.07441.
5. Bhatt, S., Chen, K., Shalin, V.L., Sheth, A.P. and Minnery, B. (2019). Who Should Be the Captain This Week? Leveraging Inferred Diversity-Enhanced Crowd Wisdom for a Fantasy Premier League Captain Prediction. Proceedings of the International AAAI Conference on Web and Social Media, 13, pp.103–113. doi:https://doi.org/10.1609/icwsm.v13i01.3213.
6. Gupta, A. (2019). Time Series Modeling for Dream Team in Fantasy Premier League. [online] International Conference on Sports Engineering ICSE. Available at: https://arxiv.org/pdf/1909.12938.pdf.
7. GS, R. (2018). Building an FPL Captain Classifier. [online] DataComics. Available at: https://medium.com/datacomics/building-an-fpl-captain-classifier-cf4ee343ebcc
8. Sveinn, H. and Kristinsson (2022). *Predicting Football Match Outcomes with Fantasy League Data and Deep Learning*. [online] Available at: https://skemman.is/bitstream/1946/43314/1/RUThesisProjectSveinnHenrik.pdf.
9. Lindberg, A. and Söderberg, D. (n.d.). Comparison of Machine Learning Approaches Applied to Predicting Football Players Performance. [online] Available at: https://odr.chalmers.se/server/api/core/bitstreams/c7d1c22f-c8e5-4dd9-b07c-cf57733f1592/content
10. Rajesh, V., Arjun, P., Jagtap, K.R., Suneera, C.M. and Prakash, J., 2022, June. Player Recommendation System for Fantasy Premier League using Machine

Learning. In 2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-6). IEEE.

11. Razali, N., Mustapha, A., Yatim, F.A. and Ab Aziz, R., 2017, August. Predicting football matches results using Bayesian networks for English Premier League (EPL). In Iop conference series: Materials science and engineering (Vol. 226, No. 1, p. 012099). IOP Publishing.

12. Saifi, M. and Milosavljevic, V. (n.d.). *Implementation of Machine Learning Techniques to Predict Player Performance using Underlying Statistics MSc Research Project Programme Name*. [online] Available at: https://norma.ncirl.ie/4270/1/murtazatasadduqhusainsaifi.pdf.

13. Stolyarov, A. and Vasiliev, G. (2017). *Predict To Succeed: Optimal Sequential Fantasy Football Squad Formation Using Machine Learning Tools*. [online] Available at: https://events-files-bpm.hse.ru/files/35D87EBC-3E72-4B9F-8003-36751B5B90B3/Stolyarov_Vasiliev_April_HSE_Conference.pdf

14. Wright, C., Carling, C. and Collins, D. (2014). The wider context of performance analysis and it application in the football coaching process. International Journal of Performance Analysis in Sport, 14(3), pp.709–733. doi:https://doi.org/10.1080/24748668.2014.11868753.

15. O’Brien, J.D., Gleeson, J.P. and O’Sullivan, D.J.P. (2021). Identification of skill in an online game: The case of Fantasy Premier League. *PLOS ONE*, 16(3), p.e0246698. doi:https://doi.org/10.1371/journal.pone.0246698.

16. Robinson, C. (2020). ‘The Prediction of Fantasy Football’. *Mathematics Senior Capstone Papers*. [online] Available at: https://digitalcommons.latech.edu/mathematics-senior-capstone-papers/20/