Developing Machine Learning Models for Assessing Fantasy Premier League Player Performance

Andrew Jones

MSc Computing

Supervisor: Federico Liberatore

School of Computer Science and Informatics, Cardiff University

September 2022

Abstract

This study looks at utilising underlying metrics to assess player performance in Fantasy Premier League (FPL) using machine learning models. A model was created for each of the four positions in FPL; forward, midfielder, defender and goalkeeper. For each position, different regression models were validated and tested. As a result, the forward, defender and goalkeeper models are based on Random Forest Regression, whilst Gradient Boosting Regression performed best for midfielders. The models were trained, validated and tested on a real-world dataset of FPL and underlying metrics, specifically created for this project. The best model achieved greater than 96% precision on validation, with all models above 91% precision. Generally ensemble models outperformed the linear models investigated in this study. The forward and midfielder models were found to be effective for short term assessment of player performance.

Acknowledgements

I would like to thank my supervisor, Federico Liberatore, for his support and guidance throughout this project. I would also like to thank Vaastav Anand for creating and maintaining the open source GitHub repository that was essential to this project.

Contents

1 Introduction……3

2 Aims and Objectives……4

3 Background Material……5

3.1 Advanced and Underlying Metrics……5

3.2 Introducing Machine Learning……7

3.3 Related Work……8

4 Problem……9

5 Approach……11

6 Application of Chosen Approach……13

6.1 Gathering Data……13

6.2 Machine Learning……14

6.3 Model Selection……16

7 Analysis……17

7.1 Effect of Increasing Test Data……17

7.2 Analysing Distribution of Data……19

7.3 Ability of Models to Assess Performance……20

7.4 Difficulties Assessing Defenders and Goalkeepers……21

8 Conclusions……23

9 Further Work……24

10 Reflection……25

References……27

Appendix A – Hyperparameters Used by Final Models……30

Appendix B – Figures from Model Training……36

Appendix C – Figures from Model Testing……38

1 Introduction

Fantasy Premier League is the official fantasy football (soccer) game for the English Premier League. The game is popular, with over 9 million active accounts making it the biggest fantasy football game in the world (Premier League 2022). People playing FPL will be referred to as *managers* in this text. The objective for FPL managers is to score the highest number of points over the course of the season, either competing in the overall league table, or in smaller leagues set up with others. A manager can pick a squad of 15 players each week, and pick 11 players from that squad, whose points will count for that round of fixtures. Each round of fixtures is referred to as a *gameweek*. Points are scored per player based on their contributions during that gameweek in the Premier League.

Football is a low scoring game, where randomness is more significant to outcomes than in higher scoring sports. Yet outcome biased decisions are still highly prevalent in the sport. This short-sightedness is mirrored in football fans, where one unlucky, bad result can cause people to call for a manager to be replaced. Unsurprisingly, this kind of rash decision making is present in fantasy football too.

One way to address outcome bias is to use underlying or advanced metrics, these metrics are usually based on machine learning models produced by data science companies such as Opta (Stats Perform 2022) and, for example, look at the probability of a player scoring from a shot rather than whether or not they actually did.

Nowadays, these statistics are used by professional football teams when scouting and recruiting players, these are not solely relied on though. They make up a part of the picture that recruitment departments use, other methods such as video scouting are still integral, and the project will reflect this approach. Rather than making predictions based on only a handful of underlying statistics, the goal is to provide less biased information that could help inform an FPL manager in making decisions.

Chapter 2 will discuss the aims and objectives of the project, the following section will then explore some of the background to this project including a review of literature and an outline of some statistics used in this project, and why they are useful. Chapter 4 explains the problem this project aims to address, and provides context to it. Then the approach and

application sections are introduced, which discuss the methods used to tackle the problem and how they were implemented. Chapter 7 then provides analysis of the results and attempts to justify the value of the work. This is followed by the conclusion in Chapter 8.

Finally, the dissertation finishes by discussing the potential for further work, followed by a reflection on the project and its learning outcomes in Chapter 10.

2 Aims and Objectives

The primary aim of this project is to show that using advanced metrics as variables of a machine learning model can be an effective method of assessing a player's FPL performance and the likelihood that performance will be sustained.

This project aims to provide improved resources for decision making in FPL. This will be done using underlying metrics and data visualisation to evaluate a player's performance against their expected performance. Expected points totals will be calculated for each player, where outcome bias is largely removed, to offer an alternative to total points.

Players will be assessed based on a machine learning model trained on three previous seasons of data. The model will then be cross validated on the same dataset, the aim of the model is to have greater than 95% precision on the validation data set, since this is a large sample size, on the test data set the model should show some correlation but, due to the smaller sample a significantly lower precision is expected. Precision on the test data sets should improve as the data pool grows with passing gameweeks, for example tests on gameweek six (GW6) should be more precise than tests on gameweek four (GW4).

The project should be innovative in its methods, and contribute a unique solution to the assessment of fantasy football players based on their underlying metrics.

3 Background Material

3.1 Advanced and Underlying Metrics

Advanced/underlying metrics are numbers that are valued, "that give us a significant parameter" (Wyscout 2020). Examples of advanced metrics include:

- xG – the probability of a shot resulting in a goal.
- xA – the probability of a pass resulting in a goal.
- xT – expected threat, the probability of possession in that area of the pitch resulting in a goal.
- Post shot xG prevented – a metric to assess goalkeeper performance, looks at how many expected goals a goalkeeper has prevented.
- Defensive Coverage – the area covered in metres squared by a players defensive actions, such as tackles, blocks, interceptions and clearances.
- PPDA – Passes per defensive action, a measure of a team's pressing intensity.

(Werf 2020)

Other underlying metrics to be used in this project include:

- Non-penalty expected goals (npxG) – xG total with xG of penalties excluded.
- Expected goals against (xGA) – The total of probabilities of a goal from shots a team has faced.
- Non-penalty expected goals against (npxGA) – xGA with xGA of penalties excluded.
- Expected goal difference (xGD) – The difference between a team's total xG and their xGA.
- Non-penalty expected goal difference (npxGD) – xGD with xGD of penalties excluded.
- Expected goals chain (xG Chain) – Total xG of every possession a player is involved in.
- Expected goals build-up (xG Buildup) – Total xG of every possession a player is involved in where they did not contribute the shot or final pass.

(Understat 2022)

There is considerable research in the field of assessing team and player quality in football; football is a low scoring game and hence randomness often influences results.

A study in 2020 in the Journal of Sports Economics looked at how xG can be used to rethink performance evaluation in football (Brechot and Flepp 2020).

Brechot and Flepp (2020) conclude that "expected goals are a superior source of information to match outcomes". The study looked at over 7000 matches across four seasons, using *expected goal difference* (xGD) to rank teams, the case study of Swansea 2014/15 is particularly interesting. Swansea finished 8th in the Premier League, however Brechot and Flepp's (2020) model had them expected to finish 20th , in last place. As a result, the club extended the manager's contract and increased his wage significantly. The following season the club were placed 15th and terminated the manager's contract at greater expense due to the extension, this could have been avoided by making decisions based on statistics, such as Brechot and Flepp's (2020) model rather than being outcome biased.

It is worth saying that there are already many FPL managers who apply statistics in their decision making, in fact almost everyone will do so, even if they're not aware, typically a striker will be selected due to the number of goals they scored. But fewer people use advanced metrics and these metrics are not easily found for most football fans. In the recent past there has been the famous case of five time chess world champion Magnus Carlsen topping the global FPL leader board, an achievement he attributed partly to his use of advanced metrics (Ninan 2020). However, Carlsen never elaborated on which metrics he uses or how he uses them.

The obvious metric is xG, goals are the currency of football and this metric has long been the standout in football statistics.

Another useful metric is xA which can be used to evaluate the likelihood of players making assists. An example from the previous season would be Paul Pogba, who assisted four goals in the first game from an xA of just 0.72 (Understat 2022), a clear overperformance. However, many managers transferred him in as a reaction, for him to only make five more assists in the rest of the season.

Goalkeeper performance can be evaluated using xGA which looks at the probability of goals conceded from shots they have faced, this can then be used to predict clean sheets and goals conceded.

3.2 Introducing Machine Learning

The machine learning in this project will be conducted using the Python library Scikit-learn.

The models used for this project will be regression models, of which two main groups will be used, linear and ensemble.

The linear models used will be Elastic Net, Bayesian Ridge, Orthogonal Matching Pursuit and Automatic Relevance Determination (ARD) Regression. The ensemble models used will be Gradient Boosting Regression and Random Forest Regression.

Bayesian Ridge is a probabilistic linear regression method based on spherical Gaussian distribution, which in this case has the l_2 regularisation applied that is used in *Ridge* regression. This allows the model to adapt to the data at hand whilst regularisation parameters can still be included. (Scikit-learn 2022a)

ARD is very similar to Bayesian Ridge but rather than being based on a spherical Gaussian distribution a centred elliptical Gaussian is utilised (Scikit-learn 2022a). This particular ARD model is very similar to *Sparse Bayesian Learning* and is found to outperform *Lasso.* (Wipf and Nagarajan 2007)

Elastic Net is a linear regression model that works as a combination of *Lasso* and *Ridge* regression methods, using both the l_1 and l_2 penalties. This model combines the best of standard linear regression, lasso and ridge and thus outperforms these methods providing it is adequately tuned for the problem. (Giba 2021)

Gradient Boosting Regression works on the concept that, the best possible next model will have minimised prediction error when combined with previous models. The model works by attempting to reduce prediction error with each step, within the confines of all predictions possible for the training case, so that the newest prediction improves on the shortcomings of the boosted ensemble of previous iterations (Hoare 2022). This is done by looking at the residuals from the current prediction for each point, the next iteration then attempts to minimise these residuals and then combines this new prediction with the previous prediction. (Masui 2020)

Random Forest Regression works by creating decision trees based on the variables, the model will create multiple unique trees which then vote for the best prediction, based on

the fitting data. The regressor then calculates the mean of predictions from all estimators to give the final output. (Keboola 2020)

3.3 Related Work

There are several papers that use machine learning to assist with FPL decision making and predictions.

Khamsan and Maskat (2019) investigate using a classifier model to predict price changes of FPL players, the study looks at handling a highly imbalanced class label, and finds greater accuracy is achieved when class distribution is equalised in the dataset. This approach to imbalance could be applied to regression, when assessing the highest performing players, who are typically outliers.

Bonello et al. (2019) found that Gradient Boosting Machines were effective for predicting future player performance, particularly when human feedback was incorporated into their model. They tested two teams, one selected by the model that only used statistics, and another selected by the model that considered betting odds and human feedback. The team based only on statistics finished in the top 13% of all players, while the enhanced model finished in the top 0.5%, both models were gradient boosting machines. The statistics used in both models were limited to FPL statistics, no other metrics were used, hence there is a possibility of improving results using a wider range of statistics.

The work by Saifi (2020) is very similar in premise, using underlying statistics in machine learning models for FPL, ensemble machine learning methods were applied in this project. The study found that including underlying statistics to evaluate player performance increased “accuracy” by 0.19% and observed a 0.2% increase in F1 score. Saifi (2020) concludes that using underlying metrics improved the models, however, there were no further experiments that tested which underlying metrics had the most impact or which FPL statistics could be omitted. Saifi (2020) also aimed to predict player performance, whereas this study aims to show whether player performance is sustainable based on their underlying metrics.

As discussed above, several machine learning models have been created and applied to fantasy football, most often these models are predictive, using total points from the

previous gameweeks to select the best team or player for the next gameweek. In conducting research for this project however, no models have been found that exclude outcome biased statistics such as points, clean sheets, goals, assists; and instead use advanced metrics to create an expected points model, which can provide a less biased indication of points gained from previous gameweeks. That is where this project is unique.

4 Problem

Football in recent years has made huge leaps in the use of statistics and data to better inform decisions in the game, such that in professional football the use of analytics and data scouting is now common practice. FPL is played by millions of people and yet it has failed to keep pace with the Premier League's advances in decision making despite far more people being directly involved. There are advanced models and statistics for football analysis that are independent from outcome bias. For example statistics such as goals scored by a player in a small sample of 10 games are highly outcome biased. However, statistics such as xG examine how many goals this player is likely to have scored from the chances that they had. FPL points are based on these outcome biased statistics and therefore points scored are poor predictors of future performance. Many fantasy football managers are interested in the statistical analysis of football and will use this to inform their decision making, but there are no tools available for a more casual manager or a manager that does not know which statistics to look at.

Whilst statistics such as xG are freely available and easy to find alone, compiling a large set of relevant statistics matched to each player is a task that would take far too long for an average FPL manager. Secondly even if a person took the time to gather all the data needed, interpreting that data into reasonable predictions would be a very difficult task.

There are four position types in FPL: forward (FWD), midfielder (MID), defender (DEF) and goalkeeper (GK). Points for players in each of these positions are calculated using different formulae, shown in Equations 1, 2, 3 and 4, with different coefficients for statistics such as goals, assists and clean sheets, adding further complications to interpretation of data.

Points awarded in FPL for events depending on position are as follows in Table 1:

	GK	DEF	MID	FWD
Goal (g)	6	6	5	4
Assist (a)	3	3	3	3
Clean Sheet (cs)	4	4	1	0
Yellow Card (y)	-1	-1	-1	-1
Red Card (r)	-3	-3	-3	-3
Own Goal (og)	-2	-2	-2	-2
1-60 Minutes (m)	1	1	1	1
>60 Minutes (M)	2	2	2	2
Per 3 Shots Saved (s)	1	0	0	0
Penalty Saved (ps)	5	0	0	0
Penalty Missed (pm)	-2	-2	-2	-2
Per 2 Goals Conceded (gc)	-1	-1	0	0

Table 1 – Points awarded for events in FPL, with bonus points not considered. (The Scout 2022a)

Whilst data for all of these events is available, this data will not be used in any of the datasets. Instead less outcome biased data will be used, for example, xG in place of goals, xA in place of assists and xGA to help predict clean sheets and goals conceded. Other data less specific to the point scoring variables listed in Table 1 will also be used to provide context and give the machine learning models a larger sample to use. The only exceptions will be data for yellow cards, red cards and minutes, since these are not events with related advanced metrics.

Each game, players are also awarded bonus points based on the FPL *bonus points system* (The Scout 2022b). Three points are awarded to the highest scoring player, two to the second highest and one to the third highest. Since these points are scored relative to other players on a per game basis it is not practical to calculate them based on a dataset for the entire season, therefore the bonus points will not be considered in any models, instead the bonus points scored will be subtracted from the total points and all models will be trained with respect to this metric.

The exact equations for calculating player points, with bonus points removed in FPL are shown below. Where p represents the number of points and all other variables are defined in Table 1. (Note: These equations will not be used for machine learning, they are to demonstrate the linear nature of the problem)

$$p = 4g + 3a - y - 3r - 2og + m + 2M - 2pm$$

Equation 1 – Calculation of actual FPL points for a forward, with bonus points removed.

$$p = 5g + 3a + cs - y - 3r - 2og + m + 2M - 2pm$$

Equation 2 – Calculation of actual FPL points for a midfielder, with bonus points removed.

$$p = 6g + 3a + 4cs - y - 3r - 2og - gc + m + 2M - 2pm$$

Equation 3 – Calculation of actual FPL points for a defender, with bonus points removed.

$$p = 6g + 3a + 4cs - y - 3r - 2og - gc + m + 2M - 2pm + s + 5ps$$

Equation 4 – Calculation of actual FPL points for a goalkeeper, with bonus points removed.

These equations illustrate that forwards should be the simplest position to predict since their points are dependent on less variables, whilst goalkeepers will be the most complicated.

5 Approach

Four machine learning models will be developed, one for each of the FPL positions. The machine learning models for this project will be developed with Python using the Scikit-learn library, chosen due to its consistent user friendly interface and popularity in data science. Regression models will be used since they are most applicable to the aims of the project.

Since the equations for calculating a player's actual points is linear, such as Equation 1 in the previous chapter, one would expect the dependence on correlating advanced metrics to also be linear. Hence the starting point for the machine learning will be to train and validate linear models, such as Elastic Net, Bayesian Ridge and others.

Ensemble models were also effectively used by Bonello et al. (2019), in particular Gradient Boosting Machines. Whilst Random Forest Regression, Support Vector Regression and K

Neighbors Regression will also be investigated (Note: *Neighbors* will be spelt this way throughout the project since this is the spelling used in the Scikit-learn documentation).

Models will initially be scored based on the cross_val_score method of Scikit-learn. This function will be used to implement a three K Fold cross validation, and will return a score based on the default scorer of the estimator used. In this experiment all models have the same default scorer, R^2. "The coefficient of determination R^2 is defined as $(1 - \frac{u}{v})$, where u is the residual sum of squares ((y_true - y_pred)** 2).sum() and v is the total sum of squares ((y_true - y_true.mean()) ** 2).sum(). The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R2 score of 0.0." (Scikit-learn 2022b)

The models with the highest validation scores from initial experiments will then be improved by pre-processing data and tuning hyperparameters. Once the models are finalised the model with the lowest root mean squared error (RMSE) will be chosen for each position.

There will be two datasets for each position, the *train* and *test* datasets. Both datasets are made up of variables listed in Table 2. With the training data made up of players from the past three FPL seasons and the test data from the current season. The models will be trained using all the relevant independent variables for the position and fit to the *Total Points – Bonus Points* (tp-bp), the trained model will then use the test dataset to predict tp-bp.

The datasets will be made up of all the FPL statistics available, combined with the advanced metrics from Understat. In the datasets only underlying metrics will be used as the independent variables, with the only exceptions being minutes, yellow cards and red cards, which have no correlating underlying statistics. A breakdown of the variables used for each position is shown below in Table 2.

Variables	Position
Total Points – Bonus Points (tp-bp)	FWD, MID, DEF, GK
Minutes	FWD, MID, DEF, GK

Yellow Cards	FWD, MID, DEF, GK
Red Cards	FWD, MID, DEF, GK
Expected Goals (xG)	FWD, MID, DEF, GK
Expected Assists (xA)	FWD, MID, DEF, GK
Non-penalty Expected Goals (npxG)	FWD, MID, DEF
Games	FWD, MID, DEF, GK
Shots	FWD, MID, DEF
Key Passes	FWD, MID, DEF, GK
Expected Goals Chain (xG Chain)	FWD, MID, DEF
Expected Goals Buildup (xG Buildup)	FWD, MID, DEF
Non-penalty Expected Goal Difference (npxGD)	DEF, GK
Expected Goals Against (xGA)	DEF, GK
Non-penalty Expected Goals Against (npxGA)	DEF, GK
Expected Points (xPts)	DEF, GK

Table 2 – Variables used in respective datasets.

6 Application of the chosen approach

6.1 Gathering Data

While the plan initially was to gather data from the FPL website and understat.com, permission was not received, forcing pursuit of other data sources. The data source eventually found is a GitHub repository (Anand 2022) containing some FPL and Understat data.

The data source used has FPL data back to the 2016/17 season, however the advanced metrics data was only available from the 2019/20 season, reducing the pool of training data available to just three seasons.

Usually, the Premier League season only fits in four rounds of fixtures before 15 September which would leave a very small sample from which to make meaningful conclusions. Helpfully though, due to the World Cup being played during November and December this

season, there were instead six rounds of fixtures from which data could be gathered. This larger pool of data should improve confidence in testing the models created in the project.

All data used in this project was gathered using a Python script and stored in csv files. For each season from 2019/20 to 2021/22 the script first pulled the FPL data for each player, then the Understat data for each player and finally the Understat data for each team.

The FPL dataset was then filtered into four datasets, one for each position, and then further filtered to exclude players who had not played any minutes, since these players would all have zero points and skew the data set heavily toward zero. The FPL and Understat datasets were then merged together by matching player names in each set, where player matching via Python did not work due to the same player having inconsistencies between the two datasets; either due to additional names, nicknames or non-latin alphabet characters; the data was manually merged. For defenders and goalkeepers additional Understat team data was added for each player, to assist with determining the likelihood of cleansheets, particularly with use of the xGA metric. The now complete player datasets for the three seasons were then combined into a single training dataset across the three seasons for each position.

The test data gathered was from the current season, it was collected using the same steps as the above paragraph, this was done three times, after GW4, after GW5 and after GW6 in order to have multiple test sets for each position and show the effect of increasing the pool of test data. The complete datasets can be found at https://github.com/andthjones/FPLdatasets.

6.2 Machine Learning

Results of initial testing of the Scikit-learn machines are illustrated by Tables 3, 4, 5 and 6 showing the mean and standard deviation for each model and position, where *cv* is cross validations.

	Mean cv=3 Score	Standard Deviation of cv=3 scores
Elastic Net	0.96775	0.00832
Bayesian Ridge	0.96855	0.00796

Orthogonal Matching Pursuit	0.91984	0.00177
ARD Regression	0.96714	0.00758
Gradient Boosting Regressor	0.95614	0.00328
Random Forest Regressor	0.95816	0.00782
Support Vector Regressor	0.68868	0.05916
K Neighbors Regressor	0.89246	0.01902

Table 3 – Initial validation scores of models for forwards.

	Mean cv=3 Score	**Standard Deviation of cv=3 scores**
Elastic Net	0.95957	0.00152
Bayesian Ridge	0.96226	0.00173
Orthogonal Matching Pursuit	0.83808	0.01399
ARD Regression	0.96249	0.00228
Gradient Boosting Regressor	0.94357	0.00436
Random Forest Regressor	0.94554	0.00699
Support Vector Regressor	0.66904	0.00285
K Neighbors Regressor	0.82150	0.00800

Table 4 – Initial validation scores of models for midfielders.

	Mean cv=3 Score	**Standard Deviation of cv=3 scores**
Elastic Net	0.91943	0.01384
Bayesian Ridge	0.92293	0.01379
Orthogonal Matching Pursuit	0.81362	0.04637
ARD Regression	0.92211	0.01548
Gradient Boosting Regressor	0.91504	0.01733
Random Forest Regressor	0.90965	0.01710
Support Vector Regressor	0.73521	0.03977
K Neighbors Regressor	0.85975	0.03224

Table 5 – Initial validation scores of models for defenders.

	Mean cv=3 Score	Standard Deviation of cv=3 scores
Elastic Net	0.96607	0.00877
Bayesian Ridge	0.96603	0.00847
Orthogonal Matching Pursuit	0.95929	0.00534
ARD Regression	0.96167	0.00934
Gradient Boosting Regressor	0.96014	0.01705
Random Forest Regressor	0.96057	0.01237
Support Vector Regressor	0.69079	0.03634
K Neighbors Regressor	0.96066	0.00685

Table 6 – Initial validation scores of models for goalkeepers.

From the results in the tables it is clear that the Support Vector Regressor and Orthogonal Matching Pursuit were consistently outperformed by the other models, therefore they were not considered past this point.

Once the best candidate models had been found the next step was to optimise them, before choosing the best for each position, this was done by tuning the hyperparameters of the models with a grid search cross validation. The set of hyperparameters that scored highest for each model were saved, to be used for the final validation and testing.

The exact hyperparameters used for each model, and the ranges of values that were trialled can be found in Appendix A.

6.3 Model Selection

Once the models had been finalised the next step was to determine which was the best for each position, this was done by comparing the root mean squared error (RMSE) for each model, shown in Table 7, where a lower RMSE indicates a better model.

	Root Mean Squared Error			
	FWD	MID	DEF	GK
Elastic Net	8.085	8.608	10.626	9.580
Bayesian Ridge	7.806	8.225	10.302	9.617

	Root Mean Squared Error			
	FWD	MID	DEF	GK
ARD Regression	7.772	8.223	10.280	9.458
Gradient Boosting Regressor	5.523	4.519	6.671	5.209
Random Forest Regressor	4.373	4.993	4.938	4.777
K Neighbors Regressor	7.899	10.877	10.929	8.102

Table 7 – RMSE scores for the finalised models for each position.

It is clear from Table 7 that the linear models, Elastic Net, Bayesian Ridge and ARD Regression performed worse than the ensemble models Gradient Boosting Regressor and Random Forest Regressor, with K Neighbors Regression performing poorly except for goalkeepers.

Random Forest Regressor was the best model for forwards, defenders and goalkeepers, whilst it was second for midfielders and hence performed the best overall. Gradient Boosting Regressor performed best for midfielders and was second for the three other positions.

Hence the model selected for forwards, defenders and goalkeepers is Random Forest Regressor, and the model selected for midfielders is Gradient Boosting Regressor.

7 Analysis

7.1 Effect of Increasing Test Data

This project is slightly unusual for a machine learning experiment in terms of testing the final models. Typically the model that best predicts the test data would be considered the highest performing model. However this project is not aiming for perfect or near perfect prediction. It instead aims to show player performance against their expected performance, so whilst some correlation to the test data is desired, the evaluation of the model should not be solely based on its precision.

One important criterion for these models is that the tests should generally become more precise as the pool of data grows since this would reflect regression to mean, the principle fundamental to underlying metrics.

As shown in Table 8, as the pool of test data increases the RMSE decreases as a fraction of mean predicted value, showing the model improves with increased gameweeks.

	Percentage RMSE			
	FWD	MID	DEF	GK
Gameweek 4 (GW4)	57.75	38.00	79.96	60.99
Gameweek 5 (GW5)	56.71	31.64	56.54	44.18
Gameweek 6 (GW6)	41.82	29.59	43.24	41.41

Table 8 – RMSE as a percentage of mean predicted values for each model over gameweeks 4-6.

This is supported visually when observing scatter plots of the predicted values against the true values such as in Figure 1.

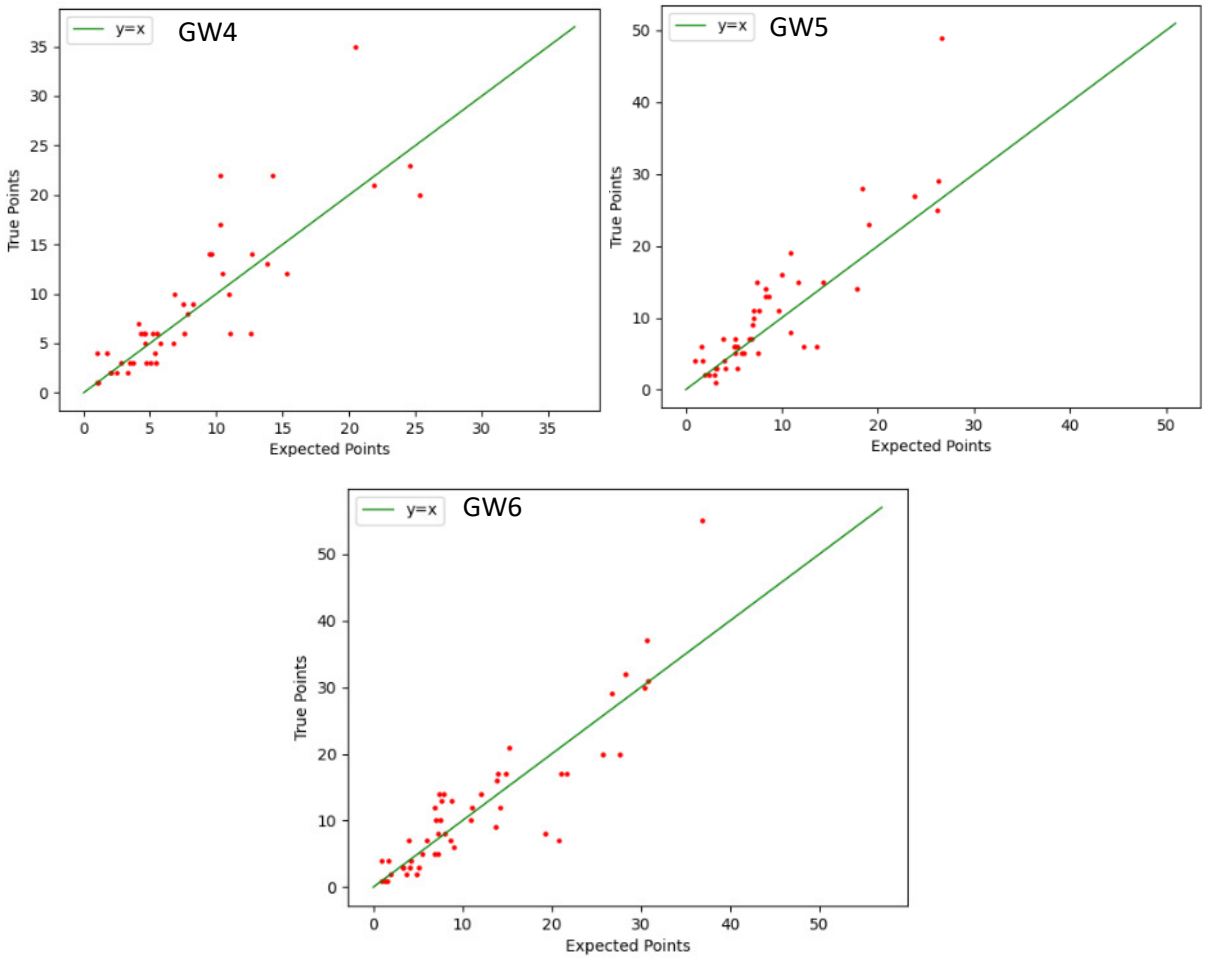

Figure 1 – Scatter plots of model predictions against actual points scored (bonus points excluded) for forwards from GW4 to GW6.

7.2 Analysing Distribution of Data

All the datasets used are unevenly distributed, there is a much higher density toward the origin this can be seen in the scatters of Figure 1 and more clearly in Figure 2.

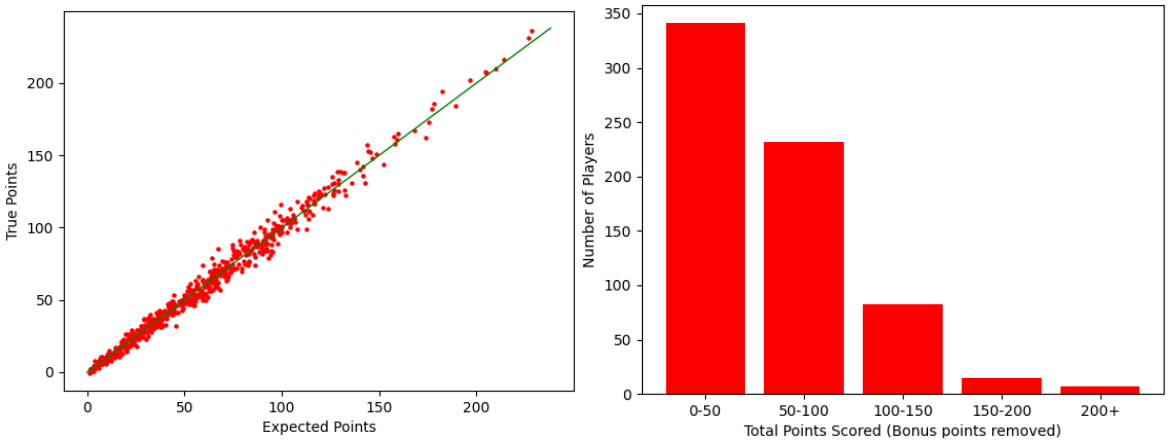

Figure 2 – Left: Scatter plot of true values against predicted values from the midfielder training dataset. Right: Histogram showing frequency of points scored by midfielders from the same dataset.

One solution to this skew of data would be to use a logarithmic transformation on the data so that it is distributed more evenly.

An alternative would be to simply filter out players who have played fewer minutes, for example if you were to filter out players who have played less than a quarter of the total minutes of the season then the distribution appears more Gaussian with a longer slope to the right as shown in Figure 3. Since the goal of the project is to assist in showing which players should be performing best, then removing players who are highly unlikely to have performed well is a logical process. This combined with further transformation would likely improve the models' ability to assess the higher scoring players. However, it is likely to compromise the accuracy of the model on smaller datasets such as the test data used in this project, since no player has played more than 540 minutes as of GW6, and this filter means the model is trained on players who have played at least 855 minutes.

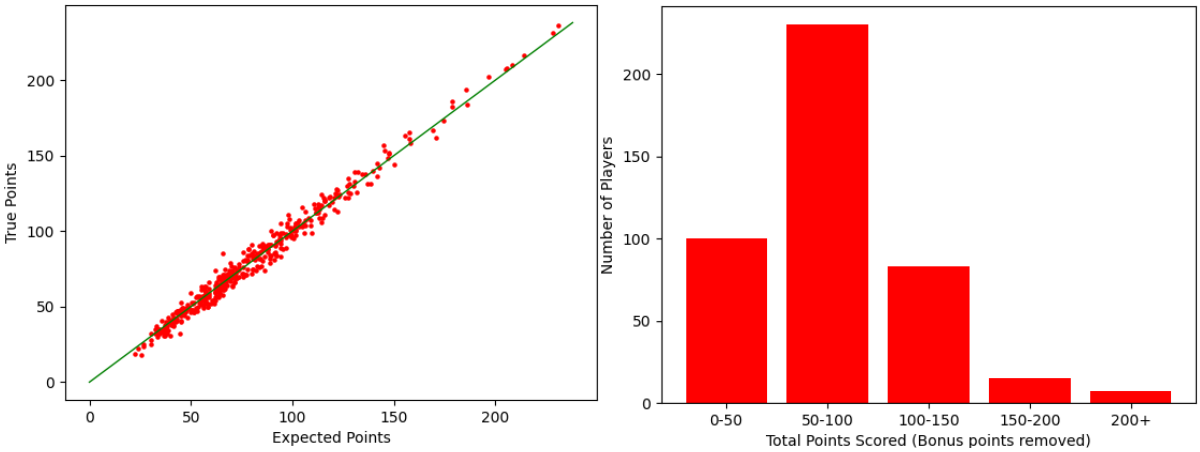

Figure 3 – Figure 2 filtered for players who have played at least 855 minutes, with the scatter plot from a retrained model.

7.3 Ability of Models to Assess Performance

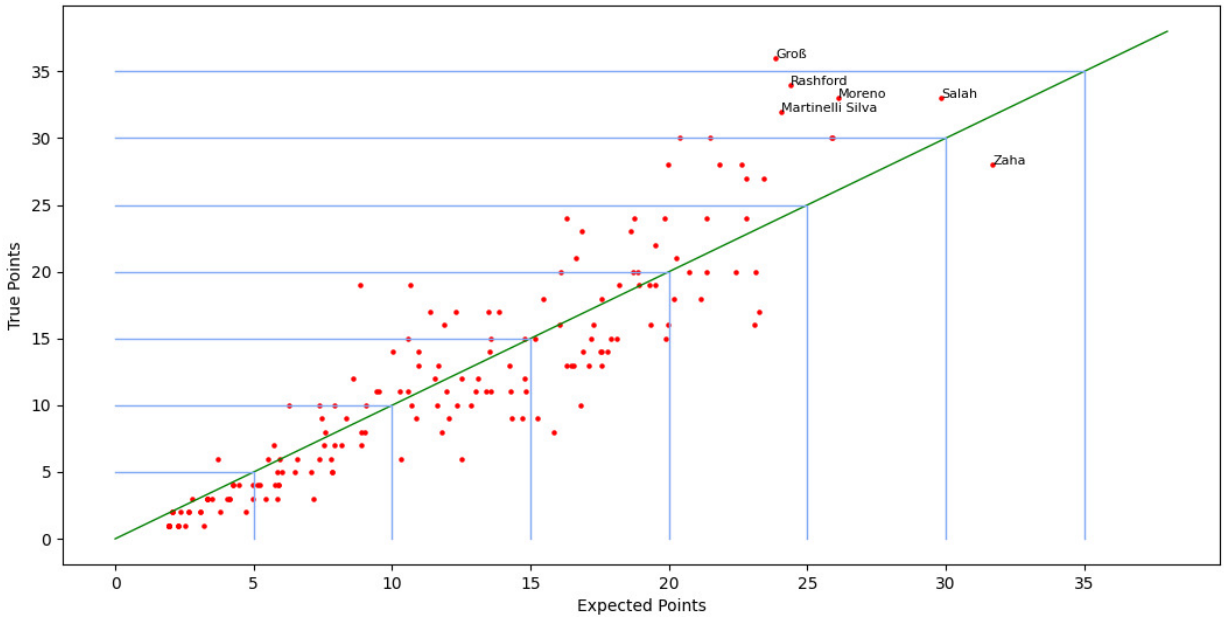

Figure 4 – Scatter plot of midfielders' predicted points against true points for GW6.

Another important criterion of this project is the ability to predict underperforming and overperforming players. A clear example of this can be interpreted from Figure 4. The highest scoring midfielder Pascal Groβ has 36 points, yet he is only predicted between 23 and 24 points, a clear overperformance. Zaha on the other hand has 28 points making him the joint 9th highest scoring midfielder, and therefore likely to be overlooked by many FPL managers; however, he has the highest expected performance, being the only player with over 30 expected points.

This principle can be seen by choosing players using the GW4 predictions, shown in Figure 5, and looking at Figure 4 to see if these choices are vindicated in GW6. From Figure 5 it is clear the player with the highest points is Moreno with 32 though he is only predicted 24.3, whilst the highest expected player is Salah with 26.5 expected points, although he only has 22 true points. However looking back at Figure 4, by gameweek 6 Salah's performance has converged to mean and he now has a total of 33 actual points, gaining 11 in two weeks. In contrast Moreno also has a total of 33 and has only gained one actual point in the same time.

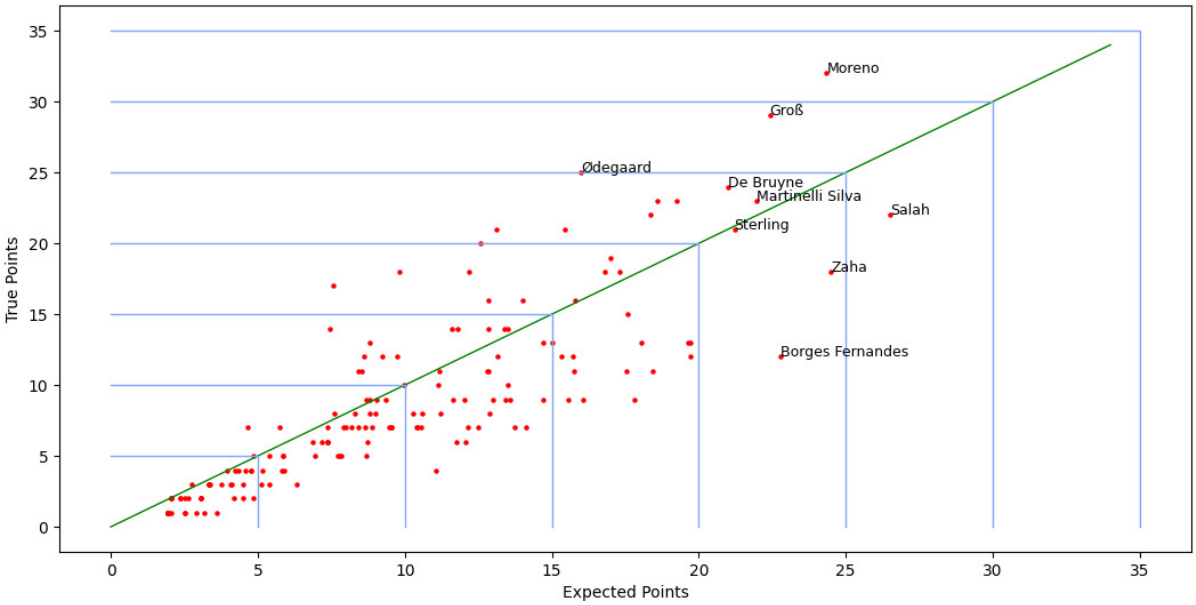

Figure 5 – Scatter plot of midfielders' predicted points against true points for GW4.

Whilst these interpretations on performance can be made with some confidence given the correlation of the forward and midfielder models, the defender and goalkeeper models do not perform so well, as explored in the next section.

7.4 Difficulties Assessing Defenders and Goalkeepers

There can be some confidence in the model for defenders since, observing Figure 6, there is still some correlation, though the model has a wider spread than is desirable. This is due to the different statistics on which defender points are based.

Defenders points are heavily reliant on clean sheets, which are much harder to predict than goals using underlying metrics, in this project xGA was used, but when four points are awarded for conceding zero goals, zero points for conceding one goal and one negative

point for conceding two or three goals it is clear that this relationship is not linear. A specific, expected clean sheets metric would be far better.

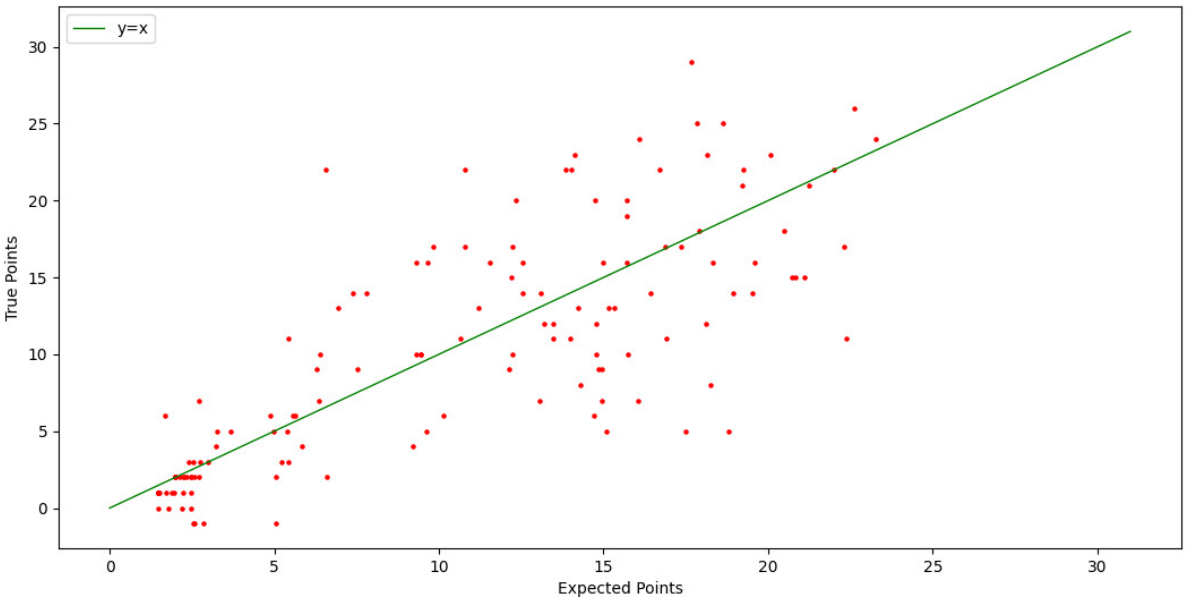

Figure 6 – Scatter plot of true points against points predicted by the model for defenders in GW6.

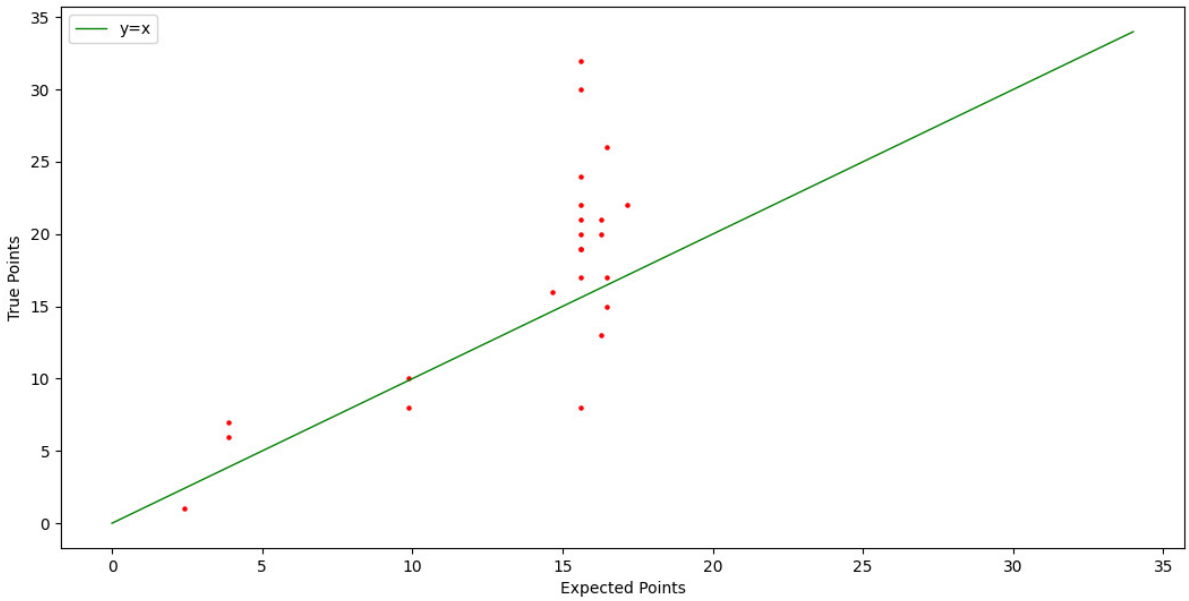

Figure 7 – Scatter plot of true points against points predicted by the model for goalkeepers in GW6.

The model for goalkeepers though correlates closer to a vertical line than y=x, as can be seen in Figure 7, this is due to the same factors that affected the defenders model, but this is amplified by the additional statistics that contribute to goalkeeper points.

Goalkeepers face a similar problem to defenders since their scores are also heavily based on clean sheets, however they are also affected by events such as penalty saves, which are very difficult to predict. Understat (2022) awards 0.76 xG to all penalties, so there must be a 24% chance of the penalty not being scored, this percentage also accounts for missed penalties though, for which points are not awarded to goalkeepers. Another significant portion of goalkeeper points is contributed by saves, there is an excellent metric for this, post-shot xG, however the project did not have access to this statistic.

8 Conclusions

The project clearly demonstrates that using only underlying metrics in a machine learning model is an effective method for assessing player performance, as evidenced by the high correlation achieved in the training datasets and the example of choosing Salah, an underperforming player in GW4, instead of Moreno an overperforming player. This decision would have led to a net gain of 10 FPL points over the two following gameweeks.

The aim of 95% precision in training was achieved in the models for forwards, midfielders and goalkeepers, with respective cross validation scores of 97%, 95% and 96%. The defenders model achieved a score of 91% in cross validation.

The best performing type of model overall was Random Forest Regression, as was found to be the case by Barbara (2021). Random Forest Regression proved the most precise model for forwards, defenders and goalkeepers as shown in Table 7. The Random Forest Regression models achieved an average RMSE of 4.770 across all four positions with a range of 0.62. Gradient Boosting Regression proved best for midfielders with an RMSE of 4.519 on training data. Generally, ensemble models significantly outperformed linear models, with ensemble models having an average RMSE of 5.125 across all positions on training data, compared to an average of 9.049 for linear models.

All models showed improvement in precision during testing as the pool of data from which to predict grew with passing gameweeks. There was an average improvement in percentage RMSE of 20.16% across the 4 positions from GW4 to GW6.

In testing, strong correlation was achieved for forwards and midfielders, due to fewer variables affecting their points, and the most significant variables having strongly correlating

advanced metrics to utilise. Predicting defender and goalkeeper performance proved more challenging, due to the increased number and complexity of variables that contribute to their points.

The forward and midfielder models would therefore be better for short term predictions, based on at least four gameweeks of data. The defender model would likely be better in the medium term with at least six gameweeks of data used for player assessment. The goalkeeper model showed no evidence of being a useful player assessment tool by GW6, it would likely perform better over the long term given the high correlation of the training data, however the minimum number of gameweeks for confident predictions cannot be ascertained by this study.

The models for defenders and goalkeepers could be significantly improved with access to additional metrics, for example using expected clean sheets rather than just xGA would be very useful. Additionally, the goalkeeper model would benefit from a metric to help predict the number of saves made, an excellent example is post-shot xG on target which looks at the likelihood of a goal after the shot has been taken, rather than before as xG does.

9 Further Work

As mentioned in the conclusions the models still have significant room for improvement. Creating a model specifically to produce an expected clean sheets metric for each player, which could then be utilised for the defender and goalkeeper models for potential significant improvement.

Retraining and tuning the models created in this project with larger datasets could also enhance performance, access to more metrics to use as independent variables and additional seasons of past data would improve confidence in the models.

The initial intention of this study was to create a web application utilising the machine learning models, however due to time constraints this was cut from the project. In future, creating an app for FPL managers to compare players based on their expected performance would be a logical extension of this study.

Neural Networks were not investigated in this project, but have also been shown to be effective by a few studies, most notably by Gupta (2019) and Ramdas (2021) where they were shown to be useful over the period of a season.

There is the intriguing possibility of creating a single model for all four positions, using the an encoder in pre-processing to differentiate between the positions. This would likely be done using Random Forest Regression given it was found to be the best and most consistent model in this project.

The models produced by this study could also be adapted for predicting high performing players, a similar project by Rajesh et al. (2022) did this also using Random Forests and Gradient Boosting Machines, however did not incorporate underlying metrics in their model.

10 Reflection

The initial intention of this project was to use the findings of the machine learning models to develop a sophisticated web app with a range of tools to help FPL managers in picking their teams; tools such as a draft squad builder and a page to compare key metrics of two or more players. However, due to problems accessing the data needed to undertake this project, starting data collection was significantly delayed and hence the focus of the project, creating models for expected points, took priority, leaving insufficient time to create a web app of acceptable quality. This delay demonstrated the value of having a backup plan, particularly when reliant on permission from another source.

The aspect I found most surprising during this project was how long it took to collect and prepare the data, this took up somewhere between 60-70% of my time working on this project. My initial proposal in hindsight was very ambitious. Producing a web application and possibly a mobile app on top of all the work I have carried was unrealistic and has taught me a lot about goal setting and time management. I do not regret my ambition however and intend to develop the web app in the future as a continuation of my work.

I feel that my research skills greatly improved through this project. Whilst I played FPL and had an interest in the use of advanced metrics in football through media such as podcasts and articles. I had not devoted time to deeply understand how these metrics were created

and appreciate their range of applications. Machine learning was completely new to me before this project, through my research I was able to gain an understanding of this incredibly useful area of computing. Furthermore I was able to find resources that taught me the basics of programming machine learning using Python packages with which I was very comfortable by the end of the project.

A skill I feel flourished in this project was my creativity in problem solving, I found a problem that others had attempted to address and took a unique approach to it in basing the models almost solely on underlying metrics. The world of football is beginning to widely embrace the use of data in decision making and thus there is growing room for creative solutions in this area.

The pool of test data was also unfortunately reduced. There were seven Premier League gameweeks scheduled initially before 15 September, however due to unforeseen circumstances all GW7 fixtures were cancelled. Whilst this meant I had less data available for testing I was able to adapt and still produced conclusive results.

References

Anand, V. 2022. *FPL Historical Dataset.* Available at: https://github.com/vaastav/Fantasy-Premier-League [Accessed: 15 August 2022].

Barbara, F. 2021. *GitHub - francescobarbara/FPL-point-predictor-via-random-forests: Using random forests to build a point predictor for Fantasy Premier League.* Available at: https://github.com/francescobarbara/FPL-point-predictor-via-random-forests [Accessed: 15 September 2022].

Bonello, N., Beel, J., Lawless, S. and Debattista, J. 2019. Multi-stream Data Analytics for Enhanced Performance Prediction in Fantasy Football. *arXiv preprint arXiv:1912.07441* doi: doi.org/10.48550/arXiv.1912.07441.

Brechot, M. and Flepp, R. 2020. Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals. *Journal of Sports Economics.* 21(4), pp. 335-362. doi: 10.1177/1527002519897962.

Giba, B. 2021. *Elastic Net Regression Explained, Step by Step.* Available at: https://machinelearningcompass.com/machine_learning_models/elastic_net_regression/ [Accessed: 6 September 2022].

Gupta, A. 2019. Time series modeling for dream team in fantasy premier league. *arXiv* [Preprint] doi: doi.org/10.48550/arXiv.1909.12938.

Hoare, J. 2022. *Gradient Boosting Explained – The Coolest Kid on The Machine Learning Block*. Available at: https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/ [Accessed: 29 August 2022].

Keboola. 2020. *The Ultimate Guide to Random Forest Regression*. Available at: https://www.keboola.com/blog/random-forest-regression [Accessed: 27 August 2022].

Khamsan, M.M. and Maskat, R. 2019. Handling highly imbalanced output class label: a case study on Fantasy Premier League (FPL) virtual player price changes prediction using machine learning. *Malaysian Journal of Computing*, 4(2), pp.304-316.

Masui, T. 2020. *All You Need to Know about Gradient Boosting Algorithm – Part 1. Regression*. Available at: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502 [Accessed: 29 August 2022].

Ninan, S. 2020. *League of extraordinary gentlemen: How Carlsen positioned his way to FPL stardom.* Available at: https://www.espn.com/chess/story/_/id/28482047/how-carlsen-positioned-way-fpl-stardom [Accessed: 5 September 2022].

Premier League. 2022. *Fantasy Premier League, Official Fantasy Football Game of the Premier League.* Available at: https://fantasy.premierleague.com/ [Accessed: 4 September 2022].

Rajesh, V., Arjun, P., Jagtap, K.R., Suneera, C.M. and Prakash, J. 2022. 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. doi: 10.1109/JCSSE54890.2022.9836260.

Ramdas, D. 2021. *Using Convolution Neural Networks to Predict the Performance of Footballers in the Fantasy Premier League.* Availbale at: https://www.researchgate.net/profile/Delano-Ramdas/publication/360009648_Using_Convolution_Neural_Networks_to_Predict_the_Performance_of_Footballers_in_the_Fantasy_Premier_League/links/625c73fc4173a21a0d1a9543/Using-Convolution-Neural-Networks-to-Predict-the-Performance-of-Footballers-in-the-Fantasy-Premier-League.pdf [Accessed: 15 September 2022].

Saifi, M. 2020. *Implementation of Machine Learning Techniques to Predict Player Performance using Underlying Statistics.* MSc thesis, Dublin, National College of Ireland.

Scikit-learn. 2022a. *1.1. Linear Models — scikit-learn 1.1.2 documentation.* Available at: https://scikit-learn.org/stable/modules/linear_model.html [Accessed: 27 August 2022].

Scikit-learn. 2022b. *sklearn.linear_model.BayesianRidge — scikit-learn 1.1.2 documentation.* Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html [Accessed: 13 September 2022].

Stats Perform. 2022. *Opta data from Stats Perform*. Available at: https://www.statsperform.com/opta/ [Accessed: 15 September 2022].

The Scout. 2022a. *FPL Basics: Scoring Points.* Available at: https://www.premierleague.com/news/2174909 [Accessed: 8 September 2022].

The Scout. 2022b. *How the FPL Bonus Points System works.* Available at: https://www.premierleague.com/news/106533 [Accessed: 8 September 2022].

Understat. 2022. *EPL xG Table and Scorers for the 2021/2022 season | Understat.com*. Available at: https://understat.com/league/EPL/2021 [Accessed: 30 July 2022].

Werf, M.v.d. 2020. *An Overview of Advanced Metrics in Football Analysis | by Max van der Werf | Medium.* Available at: https://medium.com/@max_vander_werf/an-overview-of-advanced-metrics-in-football-analysis-4e75fd82bef8 [Accessed: 29 July 2022].

Wipf, D and Nagarajan, S. 2007. A New View of Automatic Relevance Determination. *Advances in neural information processing systems,* 20.

Wyscout. 2020. *Analytics – Wyscout FootballData.* Available at: https://footballdata.wyscout.com/advanced-metrics/ [Accessed: 30 July 2022].

Appendix A – Hyperparameter Tuning and Hyperparameters Used by Final Models

Note: Parameters will be default unless specified.

Range of parameters for all positions during tuning

BayesianRidge:

- alpha_1: [1e-07,1e-06,1e-05,1e-04]
- alpha_2: [1e-07,1e-06,1e-05,1e-04]
- compute_score: [True, False]
- fit_intercept : [True, False]
- lambda_1 : [1e-07,1e-06,1e-05,1e-04]
- lambda_2 : [1e-07,1e-06,1e-05,1e-04]
- n_iter : [100,500,1000,5000,10000,50000]
- tol : [0.01,0.005,0.001,0.0005,0.0001]

ElasticNet:

- fit_intercept: [True, False]
- l1_ratio : [0.2,0.3,0.4,0.5,0.6,0.7,0.8]
- max_iter : [100,500,1000,5000,10000,50000,100000]
- tol : [0.1,0.05,0.01,0.005,0.001,0.0005]

OrthogonalMatchingPursuit:

- fit_intercept: [True,False]

ARDRegression:

- alpha_1: [1e-07,1e-06,1e-05,1e-04]
- alpha_2: [1e-07,1e-06,1e-05,1e-04]
- compute_score: [True, False]
- fit_intercept : [True, False]
- lambda_1 : [1e-07,1e-06,1e-05,1e-04]
- lambda_2 : [1e-07,1e-06,1e-05,1e-04]
- n_iter : [100,500,1000,5000,10000,50000]
- tol : [0.01,0.005,0.001,0.0005,0.0001]

GradientBoostingRegressor:

- max_depth: [2,3,4]
- alpha: [0.75,0.8,0.85,0.9,0.95,1.0]
- ccp_alpha: [0.0,0.1,0.2]
- learning_rate: [0.1]
- min_samples_leaf: [2,3,4]
- min_samples_split: [2,3]

- n_estimators: 1 – 200 with step of 1

KNeighborsRegressor:

- leaf_size: [10,20,30,40,50]
- n_neighbors: [2,3,4,5,6,7,8,9,10]
- p: [1,2,3]

RandomForestRegressor:

- max_depth: [None]
- ccp_alpha: [0.0,0.1,0.2]
- min_samples_leaf: [2,3]
- min_samples_split: [2,3]
- n_estimators: 50 – 129 with step of 1

Forwards

BayesianRidge:

- alpha_1=0.0001
- alpha_2=1e-07
- compute_score=False
- fit_intercept=False
- lambda_1=1e-07
- lambda_2=0.0001
- n_iter=100
- tol=0.01

ElasticNet:

- fit_intercept=False
- l1_ratio=0.2
- max_iter=500
- tol=0.001

ARDRegression:

- alpha_1=1e-07
- alpha_2=0.0001
- compute_score=False
- fit_intercept=False
- lambda_1=0.0001
- lambda_2=0.0001
- n_iter=500
- tol=0.01

GradientBoostingRegressor:

- max_depth=2
- alpha=0.85
- ccp_alpha=0.2
- learning_rate=0.1
- min_samples_leaf=4
- min_samples_split=3
- n_estimators=54

KNeighborsRegressor:

- leaf_size=10
- n_neighbors=2
- p=1

RandomForestRegressor:

- max_depth=None
- ccp_alpha=0.0
- min_samples_leaf=2
- min_samples_split=2
- n_estimators=70

Midfielders

BayesianRidge:

- alpha_1=0.0001
- alpha_2=1e-07
- compute_score=True
- fit_intercept=True
- lambda_1=1e-07
- lambda_2=0.0001
- n_iter=100
- tol=0.01

ElasticNet:

- fit_intercept= False
- l1_ratio= 0.6
- max_iter=100
- tol=0.001

ARDRegression:

- alpha_1=1e-07
- alpha_2=0.0001

- compute_score=True
- fit_intercept=True
- lambda_1=0.0001
- lambda_2=0.0001
- n_iter=100
- tol=0.01

GradientBoostingRegressor:

- max_depth=4
- alpha=0.9
- ccp_alpha=0.0
- learning_rate=0.1
- min_samples_leaf=2
- min_samples_split=3
- n_estimators=58

KNeighborsRegressor:

- leaf_size=10
- n_neighbors=3
- p=1

RandomForestRegressor:

- max_depth=None
- ccp_alpha=0.2
- min_samples_leaf=2
- min_samples_split=3
- n_estimators=95

Defenders

BayesianRidge

- alpha_1=0.0001
- alpha_2= 1e-07
- compute_score=True
- fit_intercept=True
- lambda_1=1e-07
- lambda_2=0.0001
- n_iter=100
- tol=0.01

ElasticNet:

- fit_intercept=True
- l1_ratio=0.2

- max_iter=100000
- tol= 0.0005

ARDRegression:

- alpha_1=1e-06
- alpha_2=0.0001
- compute_score=False
- fit_intercept=True
- lambda_1=0.0001
- lambda_2=0.0001
- n_iter=100
- tol=0.01

GradientBoostingRegressor:

- max_depth=2
- alpha=0.9
- ccp_alpha=0.0
- learning_rate=0.1
- min_samples_leaf=2
- min_samples_split=3
- n_estimators=130

KNeighborsRegressor:

- leaf_size=40
- n_neighbors=6
- p=1

RandomForestRegressor:

- max_depth=None
- ccp_alpha=0.1
- min_samples_leaf=2
- min_samples_split=3
- n_estimators=117

Goalkeepers

BayesianRidge:

- alpha_1=0.0001
- alpha_2=1e-07
- compute_score=True
- fit_intercept=True
- lambda_1=1e-07
- lambda_2=0.0001

- n_iter=100
- tol=0.01

OrthogonalMatchingPursuit:

- fit_intercept=True

ElasticNet:

- fit_intercept=False
- l1_ratio=0.2
- max_iter=500
- tol=0.01

ARDRegression:

- alpha_1=1e-07
- alpha_2=0.0001
- compute_score=False
- fit_intercept=False
- lambda_1=0.0001
- lambda_2=0.0001
- n_iter=50000
- tol=0.01

GradientBoostingRegressor:

- max_depth=2
- alpha=0.9
- ccp_alpha=0.2
- learning_rate=0.1
- min_samples_leaf=2
- min_samples_split=2
- n_estimators=64

KNeighborsRegressor:

- leaf_size=30
- n_neighbors=4
- p=1

RandomForestRegressor:

- max_depth=None
- ccp_alpha=0.1
- min_samples_leaf=2
- min_samples_split=2
- n_estimators=80

Appendix B – Figures from Model Training

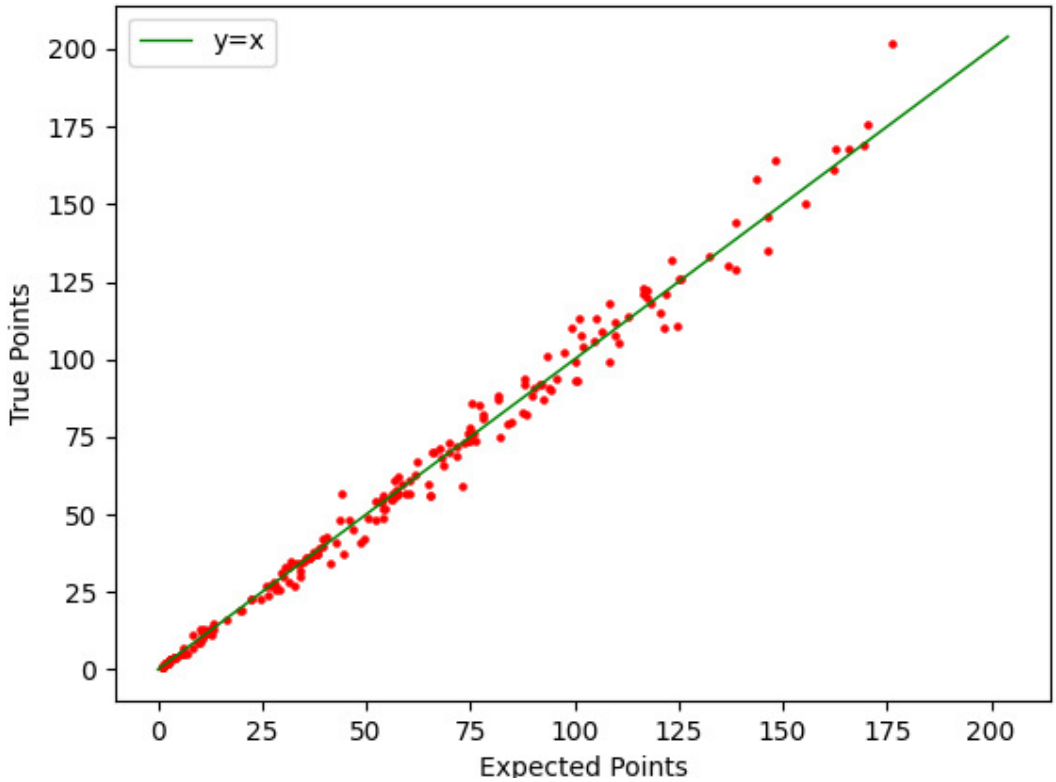

Figure B1 – Scatter of true points against points predicted by the Random Forest Regression model for Forwards

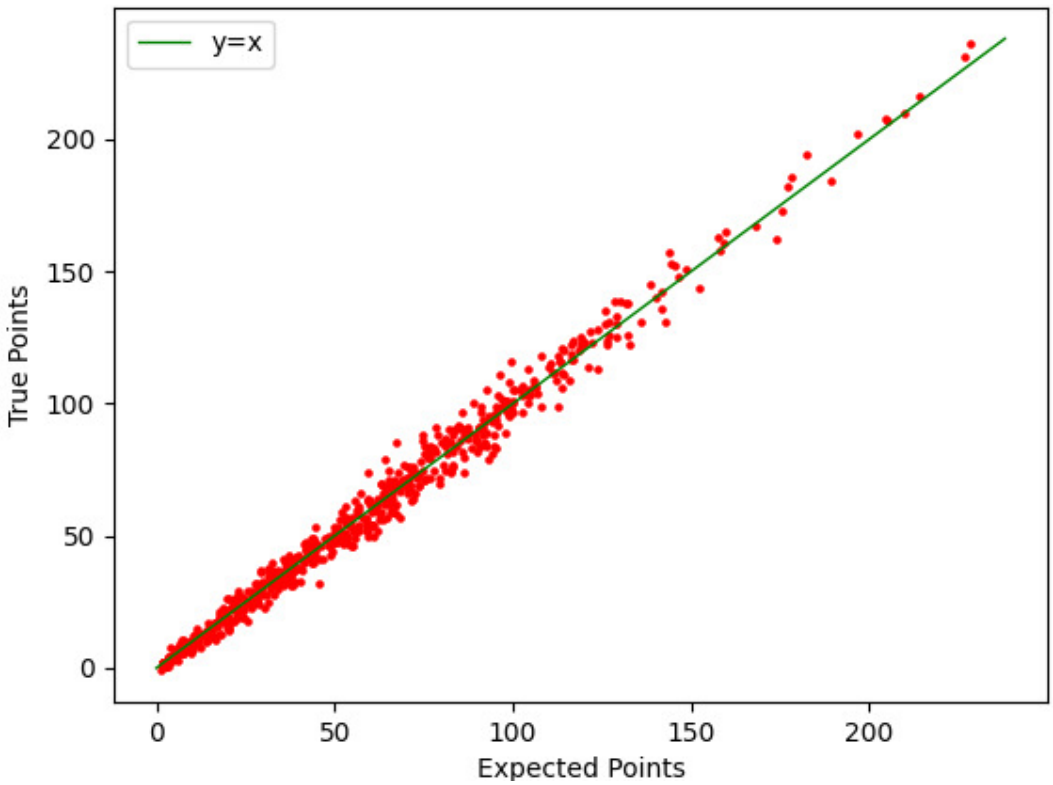

Figure B2 - Scatter of true points against points predicted by the Gradient Boosting Regression model for Midfielders

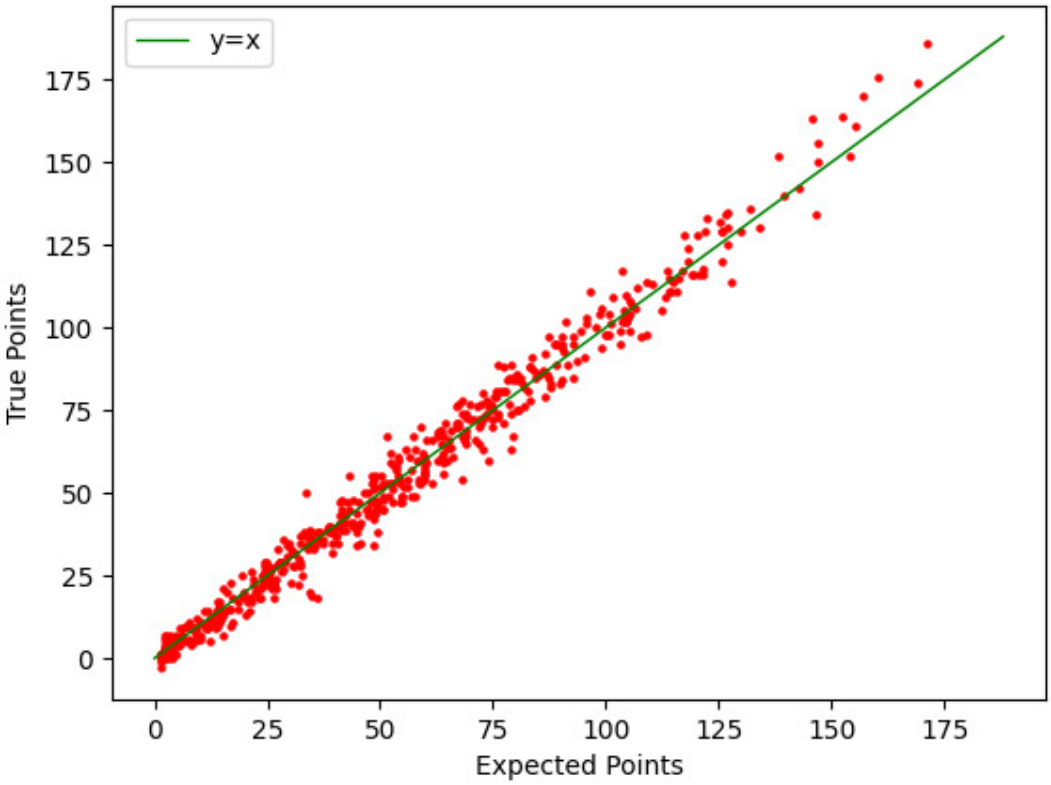

Figure B3 - Scatter of true points against points predicted by the Random Forest Regression model for Defenders

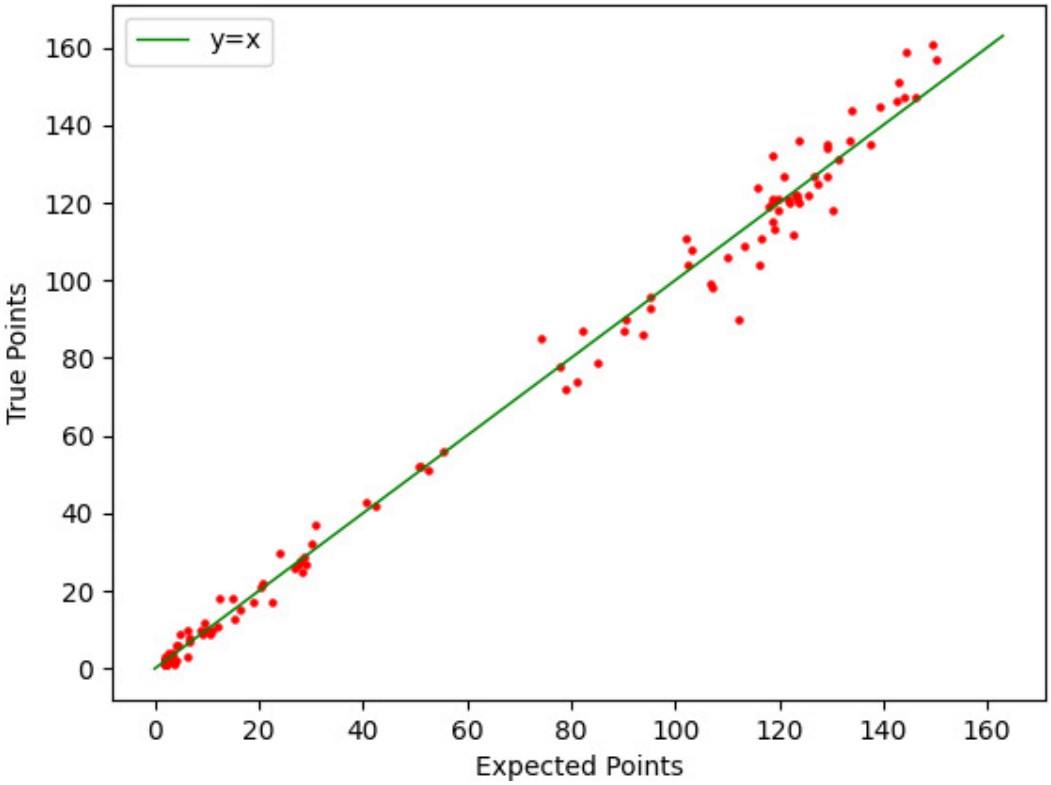

Figure B4 - Scatter of true points against points predicted by the Random Forest Regression model for Goalkeepers

Appendix C – Figures from Model Testing

Forwards

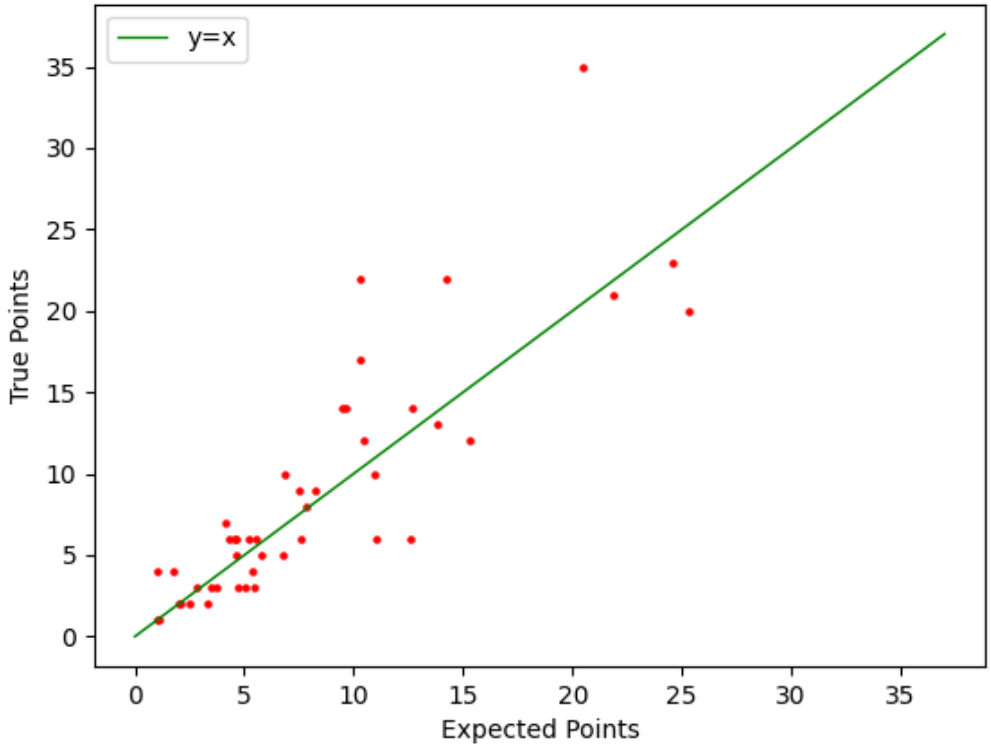

Figure C1 – Scatter plot of true points against points predicted by the Random Forest Regression model for Forwards in GW4

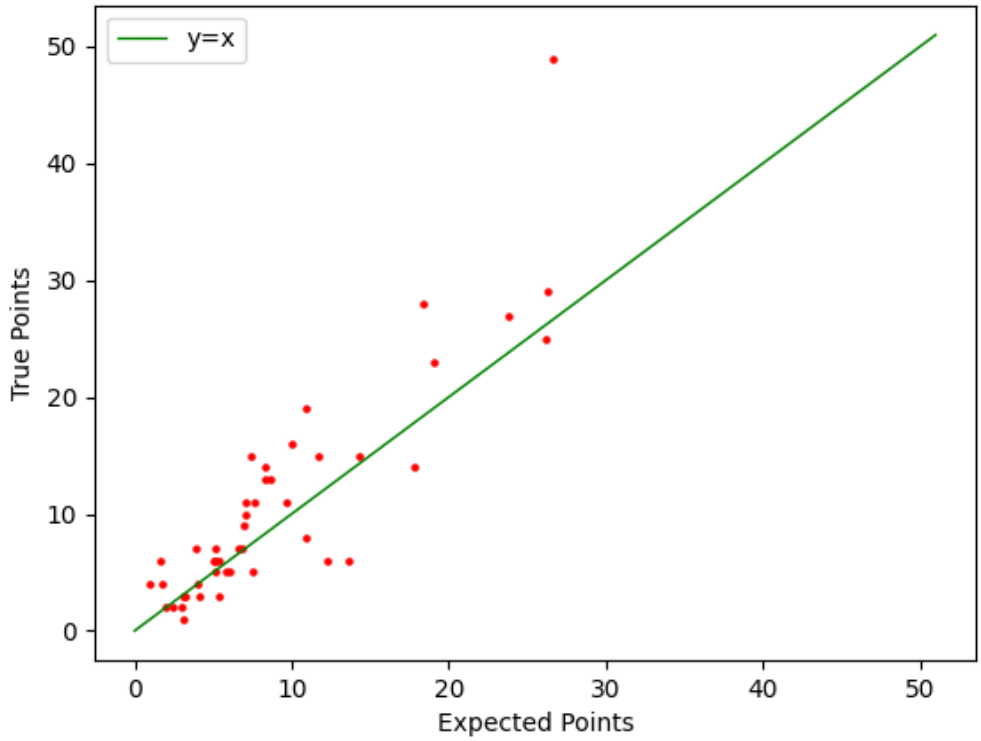

Figure C2 – Scatter plot of true points against points predicted by the Random Forest Regression model for Forwards in GW5

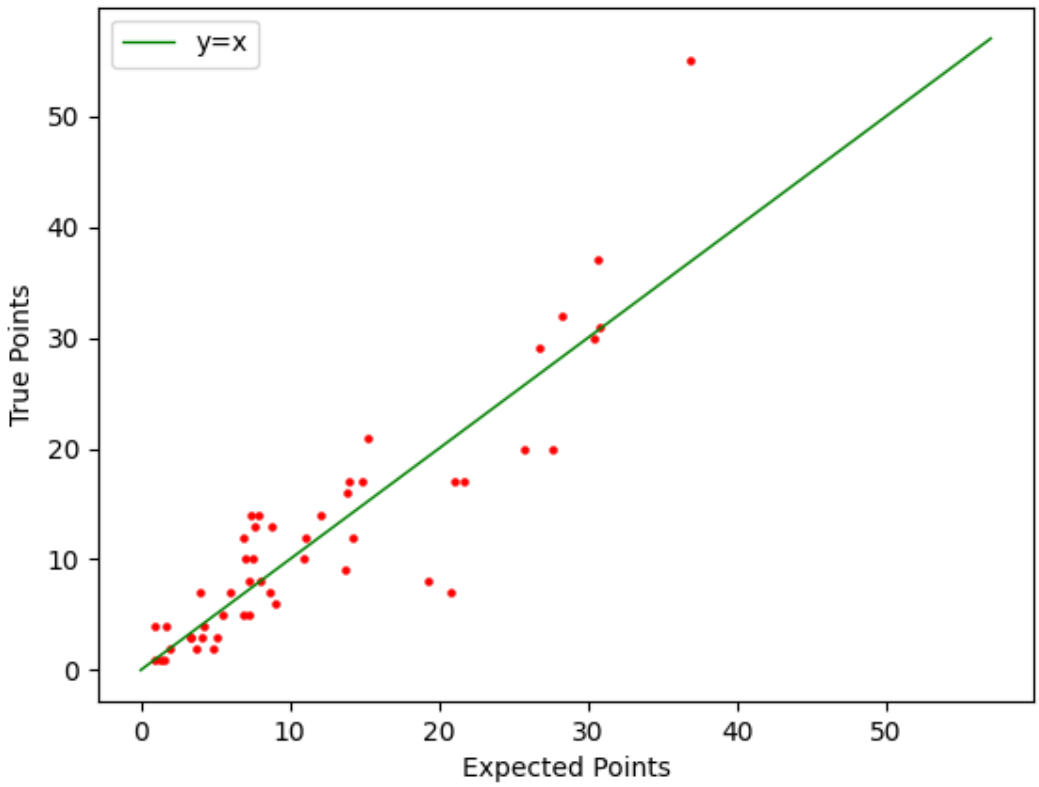

Figure C3 – Scatter plot of true points against points predicted by the Random Forest Regression model for Forwards in GW6

Midfielders

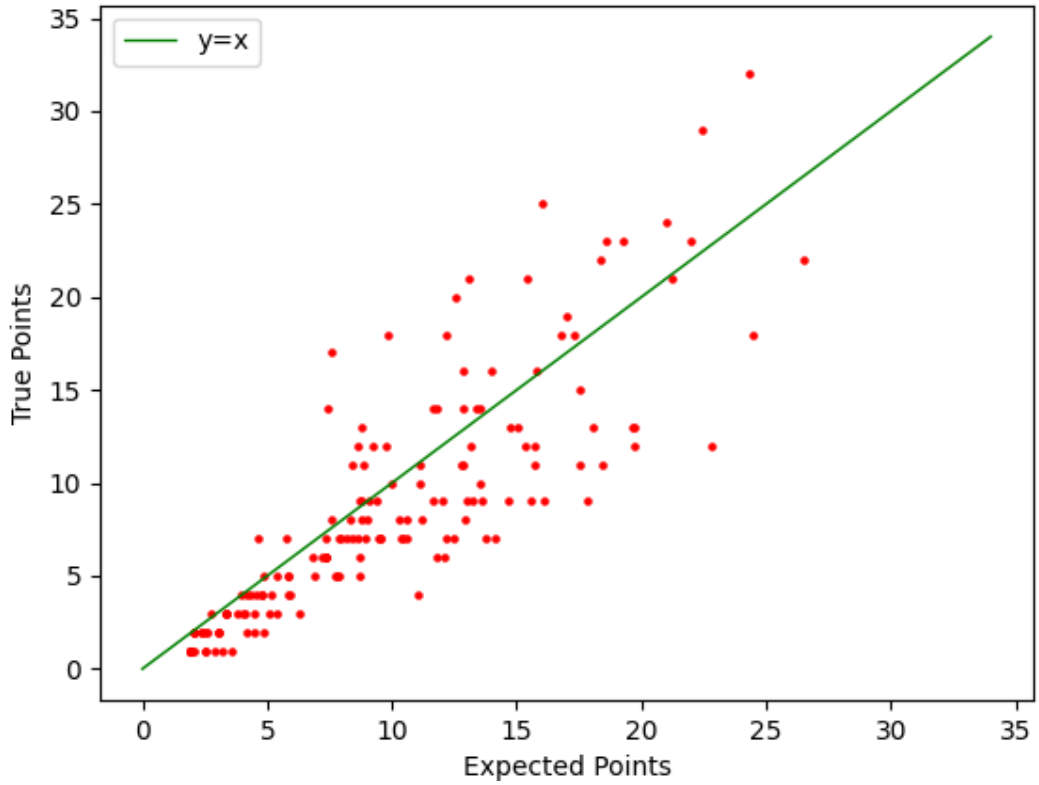

Figure C4 – Scatter plot of true points against points predicted by the Gradient Boosting Regression model for Midfielders in GW4

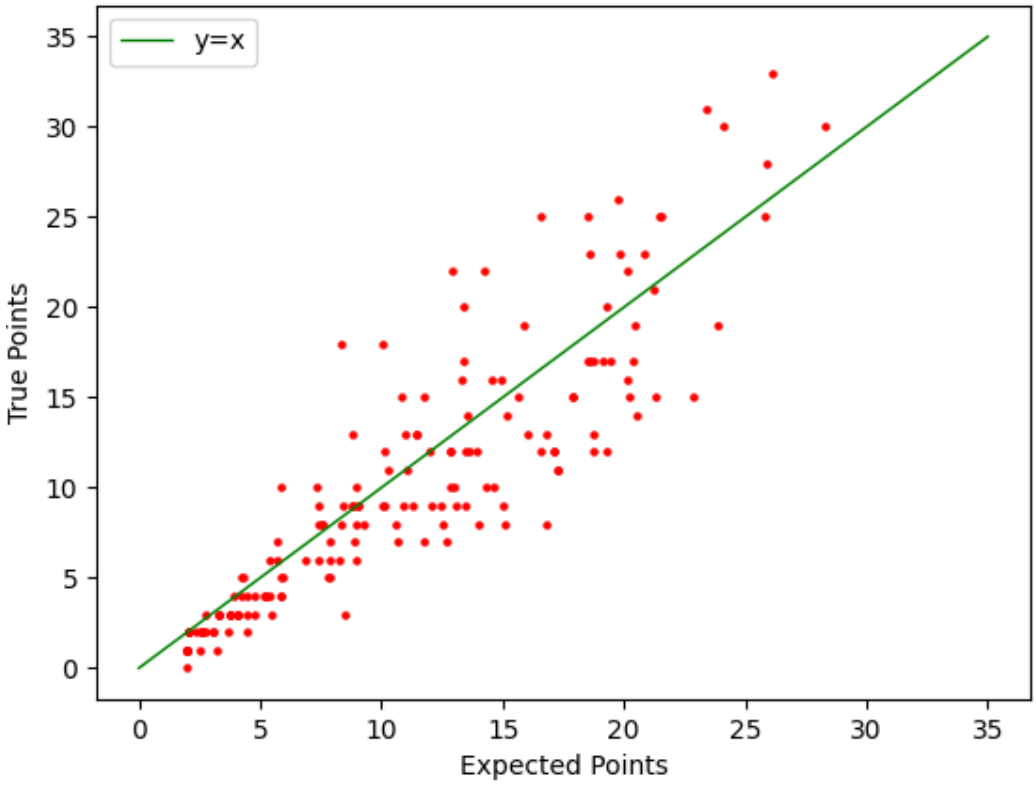

Figure C5 – Scatter plot of true points against points predicted by the Gradient Boosting Regression model for Midfielders in GW5

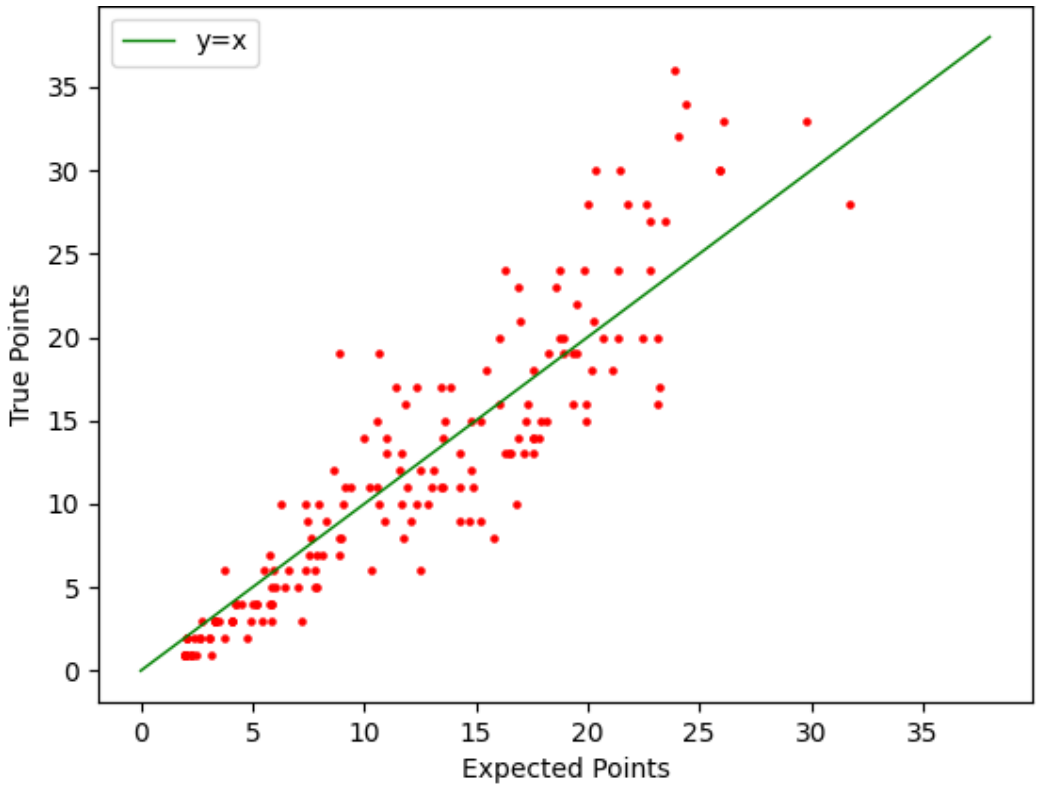

Figure C6 – Scatter plot of true points against points predicted by the Gradient Boosting Regression model for Midfielders in GW6

Defenders

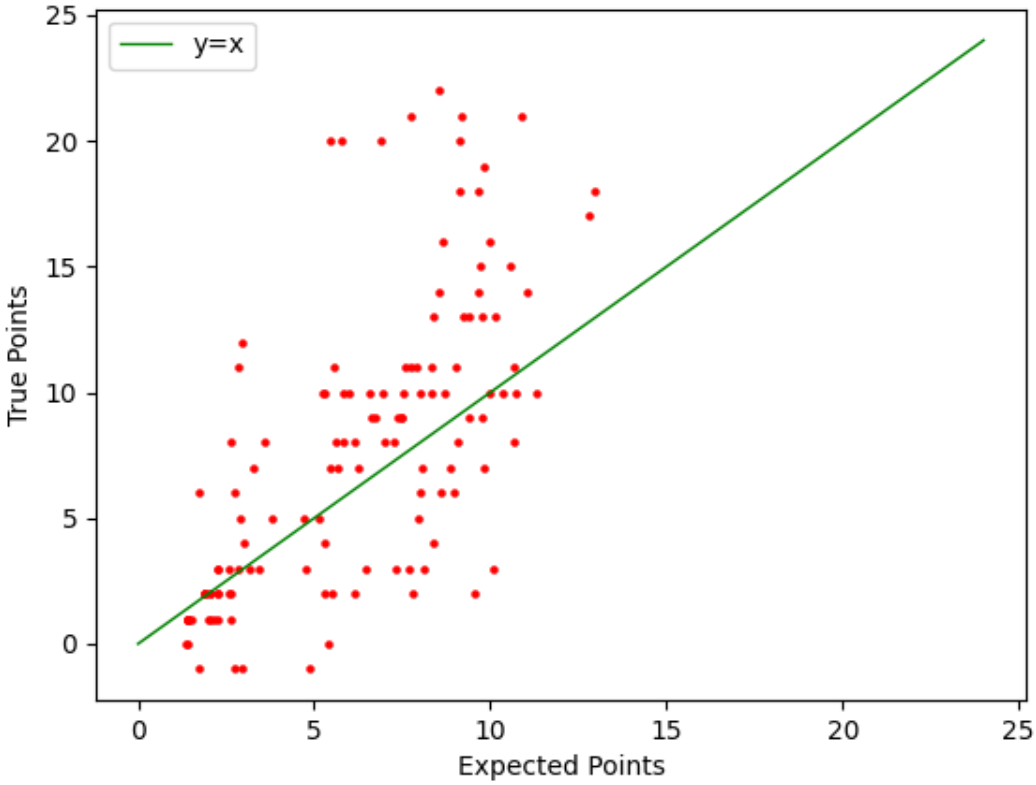

Figure C7 – Scatter plot of true points against points predicted by the Random Forest Regression model for Defenders in GW4

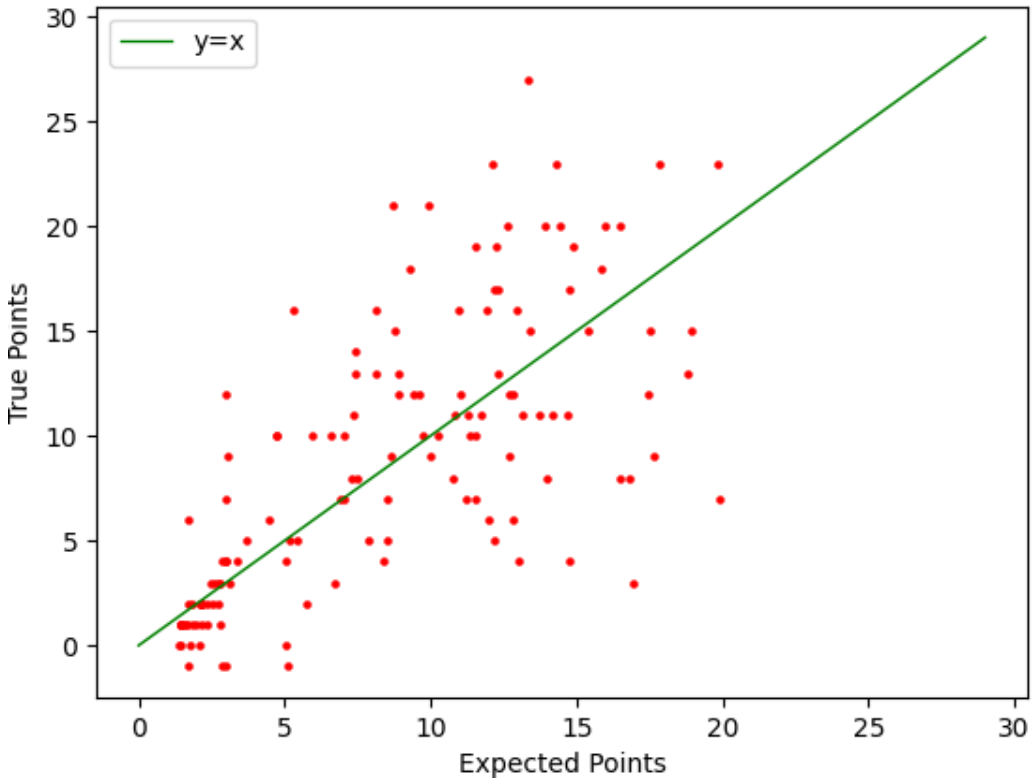

Figure C8 – Scatter plot of true points against points predicted by the Random Forest Regression model for Defenders in GW5

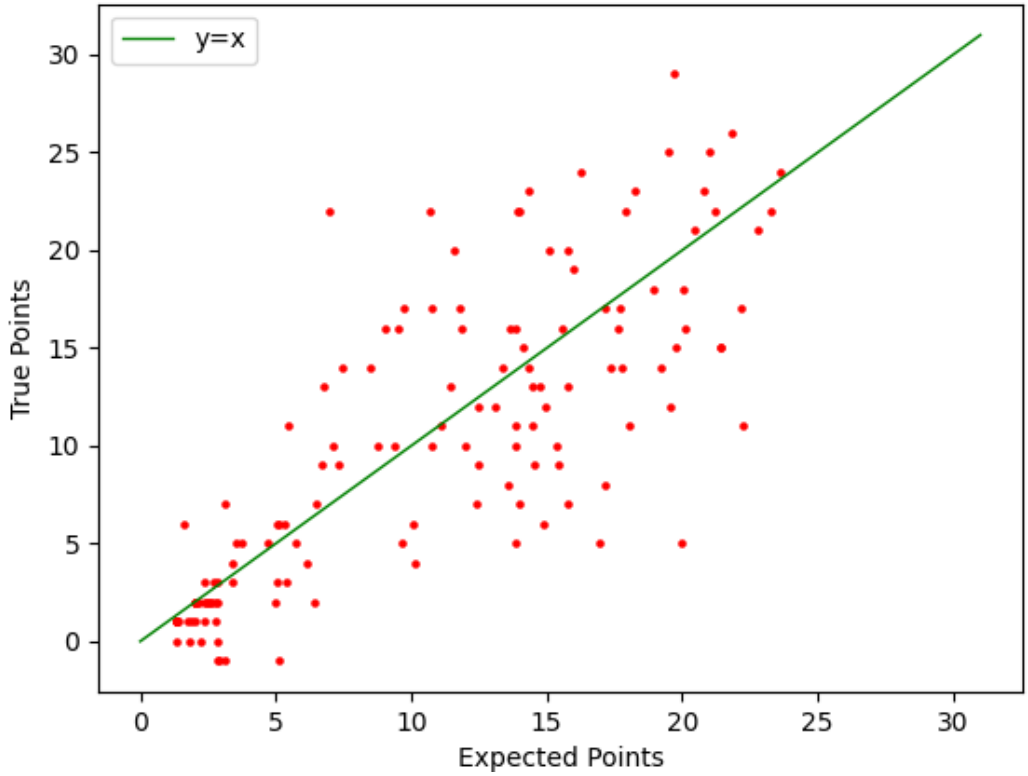

Figure C9 – Scatter plot of true points against points predicted by the Random Forest Regression model for Defenders in GW6

Goalkeepers

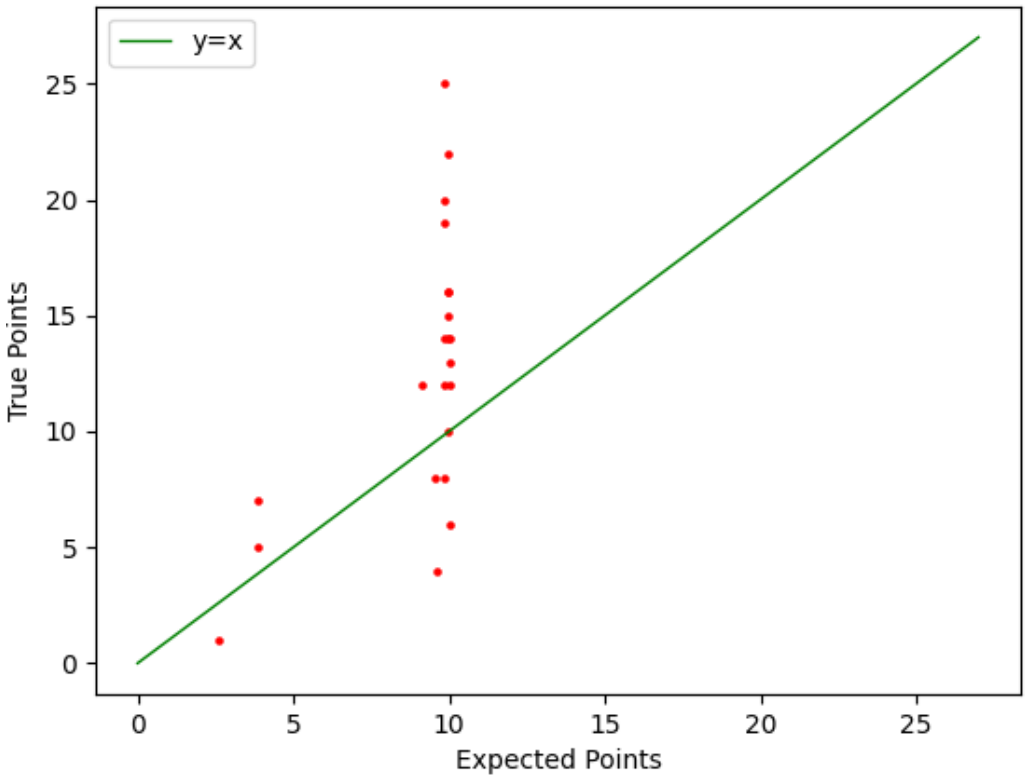

Figure C10 – Scatter plot of true points against points predicted by the Random Forest Regression model for Goalkeepers in GW4

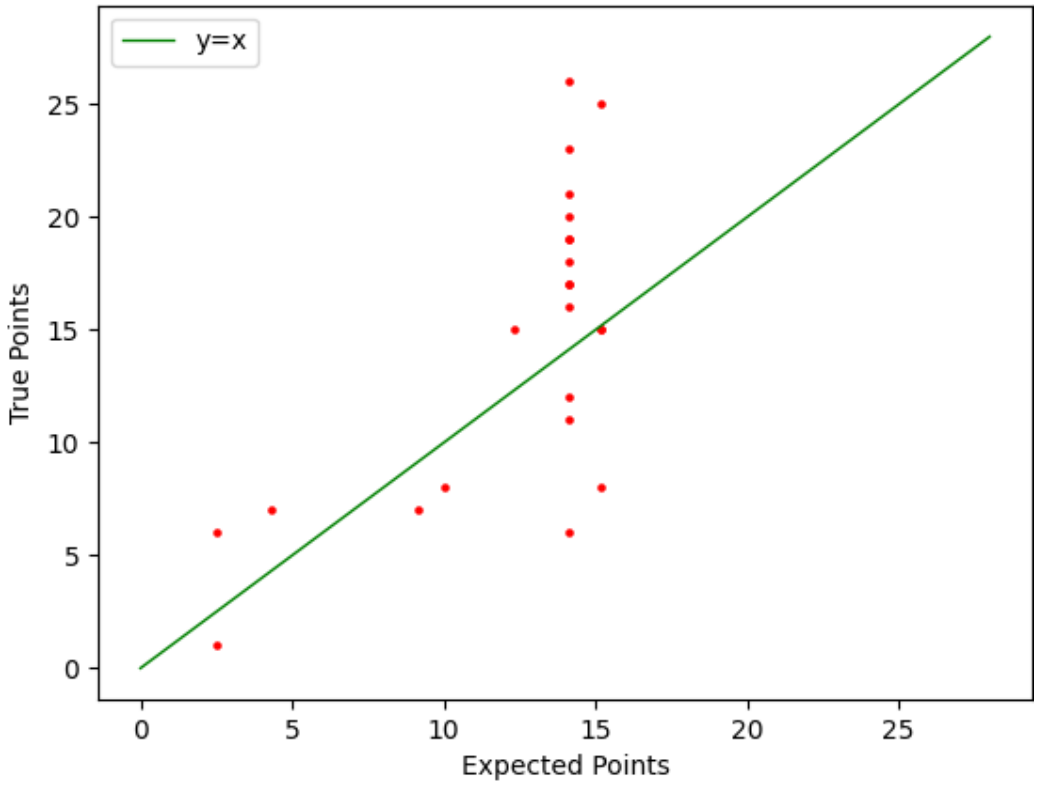

Figure C11 – Scatter plot of true points against points predicted by the Random Forest Regression model for Goalkeepers in GW5

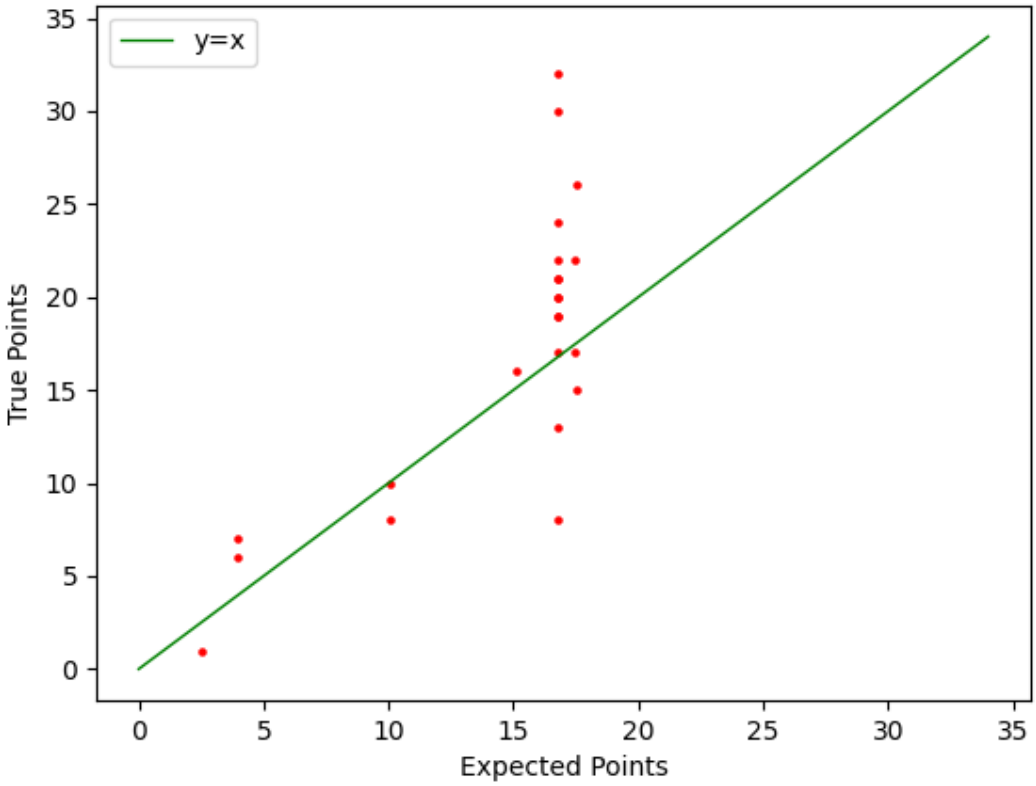

Figure C12 – Scatter plot of true points against points predicted by the Random Forest Regression model for Goalkeepers in GW6