

Predicting FPL Results

Using regression techniques



CLUSTER INNOVATION CENTRE,
UNIVERSITY OF DELHI

Semester Long Project

Submitted by:

Adit Negi (11702)

Mayank Malik (11722)

Under the mentorship of:

Dr. Nirmal Yadav

(January-May 2020)

Acknowledgement

We would like to express our special thanks of gratitude to our mentor Dr. Nirmal Yadav, who gave us the opportunity to work on this project report. She was a constant supporting presence during the process of working on this project. We are also thankful to our seniors, especially our alumni Mr. Vikas Balwada, who has been helpful constantly and has been extremely knowledgeable about our subject of work. Also, we are thankful to all our references from the web, especially some of the data sets. This report would not have been possible without them. Thanks to everyone, once again.

Adit Negi
Mayank Malik

Contents

| | |
|--|-----------|
| Acknowledgement | 2 |
| Contents | 3 |
| List of Figures and Tables | 4 |
| 1. Introduction | 5 |
| 1.1 Motivation | 5 |
| 2. Problem Statement: | 6 |
| 3. How does FPL work? | 6 |
| 3.1 A FPL Gameweek | 9 |
| 4. Predicting the best squad for the season | 10 |
| 4.1. Dividing into tiers | 10 |
| 4.2. Deciding on a Regression Model | 12 |
| 4.3. Relating Independent Variables | 13 |
| 4.4. Shortlisting | 14 |
| 4.5. Optimizing to form a squad | 15 |
| 5. Predicting the best squad for an upcoming gameweek | 18 |
| 5.1. Additional Factors | 18 |
| 5.2. Prediction | 18 |
| 5.3. The Right Comparisons | 21 |
| 6. Conclusion | 21 |
| 7. References | 22 |

List of Figures and Tables

| | |
|---|----|
| Figure 1: An FPL gameweek - twenty teams, ten fixtures..... | 7 |
| Figure 2: FPL players sorted by price (top 10 visible)..... | 7 |
| Figure 3: Point metric..... | 8 |
| Figure 4: A random squad and scoring for FPL gameweek 20..... | 9 |
| Figure 5: FPL scoring of one specific player - Kevin De Bruyne (GW 20)..... | 10 |
| Figure 6: Teams of the PL in order of their position in 18/19; goals scored in 17/18 and 18/19..... | 10 |
| Figure 7: The lack of relation between a team's ranking and their top player value | 11 |
| Figure 8: FPL history of one player - Aaron Lenon..... | 12 |
| Figure 9: Optimizing our squad..... | 16 |
| Figure 10: Predicted best squad for the entire season..... | 17 |
| Figure 11: Fixtures for Gameweek 20..... | 19 |
| Figure 12: Actual Best Squad for GW 20..... | 20 |
| Figure 13: Comparing the GW score and average score for our predicted squad for GW 20..... | 21 |
| | |
| Table 1: Predicted top scorers per team vs actual..... | 14 |
| Table 2: Our predicted best squad for the entire season..... | 16 |
| Table 3: Our predicted best squad for gameweek 20..... | 19 |

1. Introduction

Sports have always been largely ‘unpredictable’. They have carried with them the thrill of an unprecedented flow of events and scoring. Football - with its low point scoring (averaging close to 3 goals a game) is the underdog’s sport. Fantasy Premier League (FPL) is an online game based on real life top tier football league matches in England. It currently has over 7 million players playing it every week. The aim is to buy a set of players through a virtual budget which you think will best perform (according to metrics provided by FPL) in real life football over the next weekend.

Despite being a completely statistics based game, luck also plays a huge part in determining FPL results. For it is a direct consequence of the real-life event, the variables which determine results on a football pitch will also do so in FPL. However, this isn’t the same as predicting football game scores. We’ll get to the why of that in detail later.

1.1 Motivation

Our motivation to work on this came through our passion for it. We have been playing FPL since we were kids, and it only fascinated us how much of this was predictable. We have tried lots of different things manually - all within calculated risks. Every week, hundreds of forums tried various tactics to work their way through to the top of the rankings. The most absurd of decisions have worked wonders in the past. With over 7 million worldwide players, it has gone beyond just football fans trying their biased luck. Mathematicians and statisticians have also become a part of the race. Sports channels in Europe dedicate a lot of hours just analyzing which players to buy for the next gameweek.

But it was specifically this year, when Chess grandmaster and one of the living geniuses - **Magnus Carlsen broke into the top 10 of the season**. It became clear that FPL in a prolonged sense is much more about brains than simple luck. We knew we could work past most of the mysterious happenings on a football pitch, if we neutralize their probabilities just enough. That’s where machine learning came into the picture. and we decided to dig in.

We almost succeeded.

2. Problem Statement:

We have identified two major problem statements.

- 1) **What is the best squad for the entire season?**
- 2) **What is the best squad for an upcoming gameweek?**

Note: These two will require two completely different approaches.

The first does not get specific to fixtures as all players will have played all teams by that time. It rather neutralizes the good and the bad gameweeks. It tells you which players will be performing well as a whole. To measure that we'll need to train data of past seasons.

The second demands emphasis on the specific fixture. How good is the team the player will be against? How is his past record against them? Is he value for money? What are the defensive and offensive ratings of the two teams involved in a fixture?

For that we'll need to train data of the past season and concurrently as the player plays in the current season.

3. How does FPL work?

English Premier League (EPL), on which FPL is based, contains twenty teams - which face each other twice each season. Most of the games are on weekends. That results in 38 games for each team, or 38 rounds of games for the Premier League. We call each round a gameweek - which is basically the weekend of fixtures.

Gameweek 19 - Thu 26 Dec 17:00

← Previous Next →

All times are shown in your local time





















| Thursday 26 December 2019 | | | | |
|---------------------------|---|-------|-------------|---|
| Spurs |  | 2 1 | Brighton |  Multiple Broadcasters |
| Aston Villa |  | 1 0 | Norwich |  hotstarVIP |
| Bournemouth |  | 1 1 | Arsenal |  Multiple Broadcasters |
| Chelsea |  | 0 2 | Southampton |  Multiple Broadcasters |
| Crystal Palace |  | 2 1 | West Ham |  hotstarVIP |
| Everton |  | 1 0 | Burnley |  hotstarVIP |
| Sheffield Utd |  | 1 1 | Watford |  hotstarVIP |
| Man Utd |  | 4 1 | Newcastle |  Multiple Broadcasters |
| Friday 27 December 2019 | | | | |
| Leicester |  | 0 4 | Liverpool |  Multiple Broadcasters |
| Saturday 28 December 2019 | | | | |
| Wolves |  | 3 2 | Man City |  Multiple Broadcasters |

Figure 1: An FPL gameweek - twenty teams, ten fixtures.

Let there be a participant of FPL, say X. X gets 100 million worth of virtual dollars to buy real-life players inside the game. The players are each allotted virtual cost prices by the game - depending on how good or consistent they are in real life. These prices increase/decrease according to the increase/decrease in the number of participants buying them. X needs to buy 15 such players.

All players ▼ Price ▼

Current buying price in the transfer market.











| | Player | Cost | Set. | Form | Pts. |
|---|---|------|-------|------|------|
| i |  Salah LIV MID | 12.7 | 43.6% | 0.0 | 186 |
| i |  Mané LIV MID | 12.5 | 25.1% | 0.0 | 175 |
| i |  Agüero MCI FWD | 11.8 | 17.5% | 0.0 | 124 |
| i |  Sterling MCI MID | 11.7 | 15.5% | 0.0 | 118 |
| i |  Aubameyang ARS FWD | 11.1 | 27.5% | 0.0 | 152 |
| |  Kane TOT FWD | 10.8 | 8.5% | 0.0 | 104 |
| i |  De Bruyne MCI MID | 10.6 | 45.7% | 0.0 | 178 |
| i |  Vardy LEI FWD | 9.7 | 29.8% | 0.0 | 167 |
| |  Son TOT MID | 9.7 | 4.4% | 0.0 | 122 |
| i |  Jesus MCI FWD | 9.6 | 3.2% | 0.0 | 97 |

Figure 2: FPL players sorted by price (top 10 visible)

However there are rules in the buying process.

1. X cannot buy more than three players from the same team.
2. X needs to buy exactly 2 goalkeepers, 5 defenders, 5 midfielders and 3 strikers.

Once bought, X decides which 11 of these 15 will play for the upcoming gameweek. The other 4 make for the 'bench' and play in case a player from the 11 does not play in real life.

Each of the players who play the real life game are scored based on their performance. The scoring system is provided by FPL. Its very basic and determines which type of players to aim for.

During the season, your fantasy football players will be allocated points based on their performance in the Premier League.

| Action | Points |
|--|--------|
| For playing up to 60 minutes | 1 |
| For playing 60 minutes or more (excluding stoppage time) | 2 |
| For each goal scored by a goalkeeper or defender | 6 |
| For each goal scored by a midfielder | 5 |
| For each goal scored by a forward | 4 |
| For each goal assist | 3 |
| For a clean sheet by a goalkeeper or defender | 4 |
| For a clean sheet by a midfielder | 1 |
| For every 3 shot saves by a goalkeeper | 1 |
| For each penalty save | 5 |
| For each penalty miss | -2 |
| Bonus points for the best players in a match | 1-3 |
| For every 2 goals conceded by a goalkeeper or defender | -1 |
| For each yellow card | -1 |
| For each red card | -3 |
| For each own goal | -2 |

Figure 3: Point metric

Having decided which 11 players to field for the next gameweek, X is all set now. Let us see how an FPL gameweek is scored.

3.1 A FPL Gameweek

This is Gameweek 20.

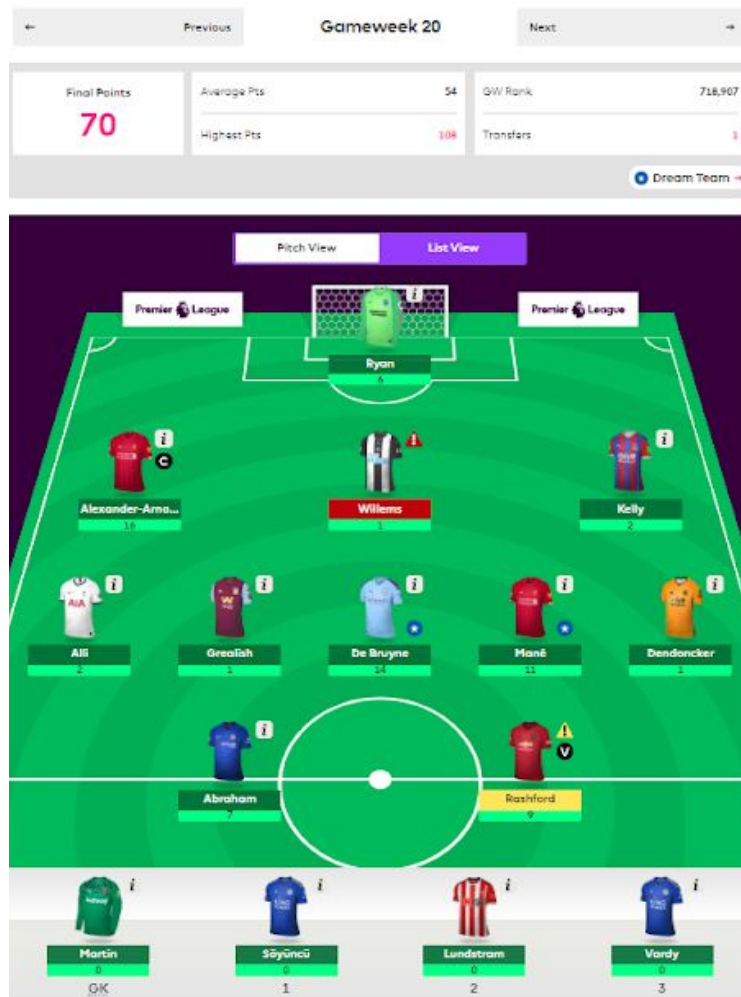


Figure 4: A random squad and scoring for FPL gameweek 20.

As we can see in Figure 4, X fielded 1 goalkeeper (compulsory), 3 defenders, 5 midfielders, and 2 forwards. He got a combined score of 70 points, which is pretty decent compared to the average that week (also mentioned in the figure): 54. He ranked below 1 million amongst 7 million participants with different combinations. X had captained Alexander Arnold whose 8 point return got doubled to a 16.

| Kevin De Bruyne | | |
|-----------------|-------|--------|
| MCI 2-0 SHU | | |
| Statistic | Value | Points |
| Minutes played | 90 | 2 |
| Goals scored | 1 | 5 |
| Assists | 1 | 3 |
| Clean sheets | 1 | 1 |
| Bonus | 3 | 3 |

Figure 5: FPL scoring of one specific player - Kevin De Bruyne (GW 20)

4. Predicting the best squad for the season

Determining the best 15 players to be fit into \$100 millions for the entire season would have us look at the data of previous seasons. Additionally, we would need to analyze the teams which weren't a part of the last few seasons.

4.1. Dividing into tiers

It isn't important that a player who plays well against the big teams should so against the small teams as well. Especially since we know the points aren't for playing well but doing the important things - scoring a goal, getting an assist, or getting a clean sheet (for defenders and goalkeepers).

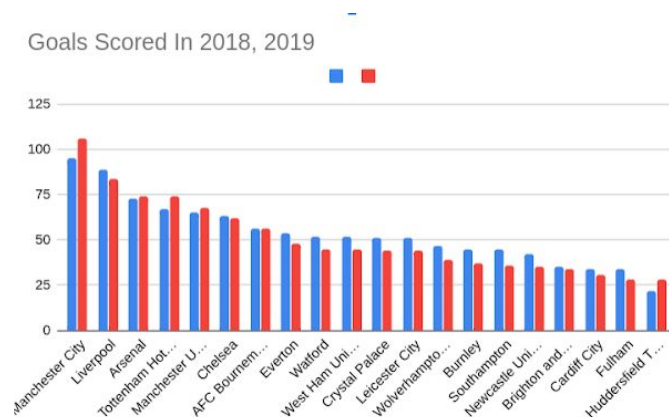


Figure 6: Teams of the PL in order of their position in 18/19 with bars of goals scored in 17/18 and 18/19

So to counter that, we divided the twenty teams into four tiers with assigned team values. The first four are tier one and get assigned five points. The next four are tier two with four points. The next four got tier three with three points. The next five got tier four and two points. The last three were tier five and zero points (as the last three don't play in next year's league).

A reason for doing that there is a high probability of the highest scoring players coming from the teams with more team value. A team with more points would have scored more and conceded less, so the players would have scored goals more and kept clean sheets. It's an obvious relation with few exceptions from the small teams. Figure shows the relation between the team's position with the goals it scored in two years. The position of the team in the table decreases from the left to the right.

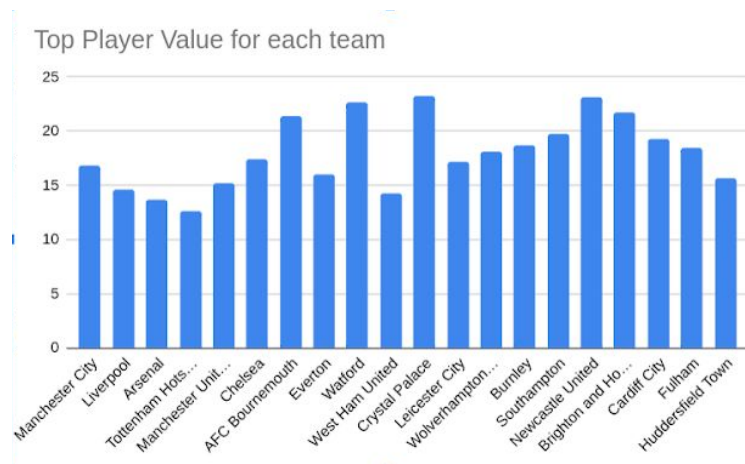


Figure 7: The lack of relation between a team's ranking and their top player value

So we get an interesting observation from the table above the top point scorer for each team in terms of value is visualised, where

$$value = \frac{Total\ points}{Cost}$$

So as we can see **the most valuable players don't belong to the top PL teams** in fact we can find great value in players who cost much less than the star players in the Premier League. Another point to note is that most of these top players for their respective teams have points in the range of 100-150. Thus if we can find multiple player's costing 5-6 million who give us over one hundred points we'll be pretty good to go. We start building our model and collecting the requisite data for it. We scraped all the data for each and every player according to game weeks and his career history to start building our model.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T |
|----|---------|-------|-----|--------------|------------|----------|--------------|----------|----------------|--------------|-----------|------|-----------|---------|-----------|------------------|-----------------|-----------|-------|--------|
| 1 | assists | bonus | bps | clean_sheets | creativity | ea_index | element_code | end_cost | goals_conceded | goals_scored | ict_index | id | influence | minutes | own_goals | penalties_missed | penalties_saved | red_cards | saves | season |
| 2 | 8 | 8 | 0 | 0 | 0 | 0 | 17349 | 74 | 36 | 3 | 0 | 393 | 0 | 1974 | 0 | 0 | 0 | 0 | 0 | 17 |
| 3 | 6 | 8 | 0 | 0 | 0 | 0 | 17349 | 74 | 32 | 2 | 0 | 1222 | 0 | 2129 | 0 | 0 | 0 | 0 | 0 | 21 |
| 4 | 6 | 16 | 0 | 0 | 0 | 0 | 17349 | 68 | 30 | 5 | 0 | 1852 | 0 | 2522 | 0 | 0 | 0 | 0 | 0 | 31 |
| 5 | 10 | 11 | 0 | 0 | 0 | 0 | 17349 | 79 | 25 | 3 | 0 | 2520 | 0 | 1801 | 0 | 0 | 0 | 0 | 0 | 42 |
| 6 | 3 | 10 | 0 | 8 | 0 | 0 | 17349 | 73 | 26 | 3 | 0 | 3155 | 0 | 2361 | 0 | 0 | 0 | 0 | 0 | 52 |
| 7 | 8 | 7 | 0 | 5 | 0 | 291 | 17349 | 72 | 21 | 3 | 0 | 3745 | 0 | 1571 | 0 | 0 | 0 | 0 | 0 | 62 |
| 8 | 8 | 10 | 0 | 11 | 0 | 469 | 17349 | 71 | 32 | 4 | 0 | 4460 | 0 | 2823 | 0 | 0 | 0 | 0 | 0 | 72 |
| 9 | 5 | 1 | 97 | 9 | 0 | 387 | 17349 | 73 | 40 | 1 | 0 | 5148 | 0 | 2177 | 0 | 0 | 0 | 0 | 0 | 82 |
| 10 | 3 | 4 | 210 | 6 | 0 | 225 | 17349 | 62 | 16 | 2 | 0 | 5839 | 0 | 1347 | 0 | 0 | 0 | 0 | 0 | 92 |
| 11 | 1 | 7 | 319 | 5 | 0 | 278 | 17349 | 54 | 25 | 5 | 0 | 6541 | 0 | 1538 | 0 | 0 | 0 | 0 | 0 | 102 |
| 12 | 1 | 0 | 104 | 2 | 93.4 | 0 | 17349 | 55 | 5 | 0 | 22.3 | 7171 | 78 | 509 | 0 | 0 | 0 | 0 | 0 | 112 |
| 13 | 4 | 0 | 204 | 6 | 203.2 | 0 | 17349 | 54 | 30 | 0 | 73.3 | 7844 | 232.2 | 1906 | 0 | 0 | 0 | 0 | 0 | 122 |
| 14 | | | | | | | | | | | | | | | | | | | | |

Figure 8: FPL history of one player - Aaron Lenon

All our scrapped data is available on Adit's github. [1]

4.2. Deciding on a Regression Model

As we hypothesised earlier - if we can find the top two-three point scorers from each team we can build a pretty solid team for the year. Here we start building our model this was regression model and was implemented on the three models -

1. Linear Regression - A linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.
2. Random Forest - Random forest uses the idea of decision trees, it is an ensemble model made of many decision trees using bootstrapping, random subsets of features, and average voting to make predictions.
3. Naïve Bayes - Naïve Bayes has a naive assumption of conditional independence for every feature, which means that the algorithm expects the features to be independent which not always is the case.

Let us first discuss the features used for each of the models and then we can talk about the accuracy factors. The features/attributes include:

1. Total goals scored in the past season
2. Assists scored last season
3. Clean sheets last season
4. Form past season (ICT)
5. Is he a starter? (yes or no)
6. Bonus points last season (relates directly to amount of games impacted)
7. Player rank in the team last season (Was he the top in points scored?)
8. Team tier

9. Predicted form this season (ICT taken from the FPL website)

For there are a large number of features included, which give rise to further more decision nodes: **Random Forest gave us the most precise results.** We chose to go ahead with that in both evaluations.

4.3. Relating Independent Variables

All these features were narrowed down after deeply analysing the data, **multicollinearity** is one of the factors we used to narrow down the features as we found that a players offensive rating was almost completely dependent on the number of goals the player scored. Calculating VIF in python, to determine the correlation between seeming independent variables.

VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable; or simply - VIF score of an independent variable represents how well the variable is explained by other independent variables.

```
# Calculating VIF
vif = pd.DataFrame()
vif["variables"] = X.columns
vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

print(vif)
```

$$VIF = \frac{1}{1-R^2}$$

After calculating VIF we removed the variables with high scores as that was affecting the overall result of the model causing it to behave erratically.

Missing values

We used the 3 industry standard methods to deal with missing values which were-

1. Mean
2. Median
3. KNN imputations

We found **the median value to be the best choice** after examining the data as the stats for most of the top players were available pretty easily and median values were low enough to not cause a significant change in our predictions.

We used the data from the 2017-18 season as the training set, and then the 2018-19 data to predict values for the 2019-20 FPL season.

Note: All implementations were done using TensorFlow 2.1 and Google Colab.

We established a 10-fold-cross-validation on our data - which is nothing but a k-fold-cross-validation where we provide data in bunches so that the model doesn't overfit and remains as unbiased as possible. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. We formed this validation in terms of epochs in Tensorflow which are provided with the library itself.

4.4. Shortlisting

Having flirted with the ideas of multicollinearity, we set out to find out the top two players from each club. We combined to form a list of forty players (20 teams * 2 players per team). We compared it to the top scoring players from each team as of 20th May 2020 i.e. two months since football and FPL stopped due to COVID -19.

| TEAMS | PLAYERS PREDICTED VS ACTUAL |
|-------------------|--|
| Manchester City | Bruyne, Sterling vs Bryune, Aguero |
| Liverpool | Salah, Mane vs Salah, Mane |
| Arsenal | Aubameyang, Luiz vs Aubameyang, Leno |
| Tottenham Hotspur | Son, Kane vs Son, Aurier |
| Manchester United | Rashford, Martial vs Rashford, Martial |
| Chelsea | Abraham, Mount vs Abraham, Willian |

| | |
|-----------------|--|
| Brighton | Ryan, Moupay vs Ryan, Moupay |
| Burnley | Pope, Wood vs Pope, Wood |
| Sheffield | Lundstram, Mousset vs Lundstram, Henderson |
| Southampton | Ings, Ward Prowse vs Ings, Ward Prowse |
| Wolves | Jimenez, Doherty vs Jimenez, Doherty |
| Norwich | Pukki, Cantwell vs Pukki, Cantwell |
| Newcastle | Dubravka, Almiron vs Dubravka, Hernandez |
| AFC Bournemouth | Wilson, Frazer vs Wilson, Frazer |
| Everton | Richarlison, Lewin vs Richarlison, Lewin |
| Watford | Foster, Sarr vs Foster, Sarr |
| West Ham United | Haller, Snodgrass vs Haller, Snodgrass |
| Crystal Palace | Guaita, Ayew vs Guaita, Ayew |
| Leicester City | Vardy, Pereira vs Vardy, Pereira |
| Aston Villa | Grealish, Heaton vs Grealish, El Ghazi |

Table 1: Predicted top scorers per team vs actual

After predicting these values/ players, this prediction model was reduced to an optimization problem for which we built a simple algorithm. Not very large scale data was being processed and thanks to our machine learning model - we just used conditional statements to finish our predictive analysis.

4.5. Optimizing to form a squad

From the 40 top players predicted for the season, our model got 33 right. That's an accuracy of 82.5%. But that's also an almost irrelevant number to us. We use these 40 to get the top 15 for the season under a budget of 100 million. We optimize.

```
def get_money_team_objects(budget = 100, star_player_limit = 3, gk = 2, df = 5, md = 5, fwd = 3):
    money_team = []
    star_player_limit = star_player_limit
    budget = budget
    injured = players_by_status('injured')
    positions = {'Goalkeeper': gk, 'Defender': df, 'Midfielder': md, 'Forward': fwd}
    for player in points_top_players():
        if len(money_team) < star_player_limit and player not in injured and budget >= player.cost and
           positions[player.position] > 0:
            money_team.append(player)
            budget -= player.cost
            positions[player.position] = positions[player.position] - 1
        else:
            for player in roi_top_players():
                if player not in money_team and budget >= player.cost and player not in injured and
                   positions[player.position] > 0:
                    money_team.append(player)
                    budget -= player.cost
                    positions[player.position] = positions[player.position] - 1
    final_team = [(item.name, item.position, item.cost) for item in money_team]
    total_points = sum([item.total_points for item in money_team])
    return money_team
```

Figure 9: Optimizing our squad

(selecting 15 players in a 100M budget)

We got the following as the result.

| Position | Player | Team | Price (in million \$) |
|----------|---------------------|------------------|-----------------------|
| GK | Ederson | Manchester City | 6.0 |
| GK | N. Pope | Burnley | 4.5 |
| DEF | J. Lundstrum | Sheffield United | 4 |
| DEF | R. Pereira | Leicester City | 6 |
| DEF | Y. Mina | Everton | 5.5 |
| DEF | C. Cathcart | Watford | 4.5 |
| DEF | T. Alexander Arnold | Liverpool | 7 |
| MID | Son | Tottenham | 9.5 |
| MID | Ricarlison | Everton | 8 |
| MID | H. Barnes | Leicester | 6 |
| MID | D. Luiz | Aston Villa | 4.5 |

| | | | |
|-----|-----------------|-------------|-----|
| MID | K. De Bruyne | Liverpool | 9.5 |
| FWD | P.E. Aubameyang | Arsenal | 11 |
| FWD | T. Pukki | Norwich | 6.5 |
| FWD | D. Ings | Southampton | 6 |

Table 2: Our predicted best squad for the entire season

Total Team Value = \$98.5 millions

The team picked by our model for the entire season has done pretty well till the time PL worked in February, just before COVID related lockdown. It even saved us a very important \$1.5 million in the bank. How it eventually racks up at the end of season remains a question unknown, but till now most of the players have seen massive increases in prices.



Figure 10: Predicted best squad for the entire season

Figure 10 shows how the squad looks just before the lockdown began. The figures mentioned below the player names are their updated prices through the season. As we can see this team amounts to a whopping \$103.4 millions. That is \$4.9 millions of price increases - a good indicator of how well the team has done in the season.

All but four (Ederson, Mine, Cathcart, Douglas Luiz) have managed over a 100 points already. It is imaginable to place these lower priced players on the bench for most gameweeks.

5. Predicting the best squad for an upcoming gameweek

While we can use the same model for this problem as well, its usage, factors, and requirements differ greatly. That is entirely our point of keeping two different problem statements. **While picking the best of a season or predicting the best cost effective squad before a season begins is a testament of how well we understand players' statistics; a gameweek-specific makes matters further interesting.** While predicting for a gameweek, we need to understand the mystery that goes into sports. We need to attach variables for certain players who perform well only against weaker oppositions, or for those who only play well at home venues. Random logics make up sports-related statistics, and we as fellow statisticians can only buy into that.

5.1. Additional Factors

So in order to counter that, we added a bunch of very specific factors to the list already curated above. We also attempted to build an algorithm which provides us with the best player at certain price points each week. This model was trained over 10 game weeks in the season using the same features but now we added fixture strength to optimize it for our needs.

The features added are:

1. Team home offensive rating
2. Team home defensive rating
3. Opponent home offensive rating
4. Opponent home defensive rating
5. Venue: Home/ Away
6. Tier difference between the teams

Factors like understanding the dimensions of the pitch at home venues, fan support and psychological comfort at home to more bizarre factors like proving home fans wrong in away fixtures - all work in mysterious ways. That is why we landed upon the above-mentioned six factors.

5.2. Prediction

Having added these features, we used the same model as before to build week-wise teams and player suggestions customized for individuals. We also tried working on a web-app which gives you the most valuable player to add to your team at the user given price point.

Not to our surprise, our newly added features of offensive and defensive ratings got big coefficients. We decided to run our model for the fixtures of Gameweek 20. Figure 11 shows the real-life results of those fixtures.

Fixtures

Gameweek 20 - Sat 28 Dec 17:00

← Previous Next →

All times are shown in your local time

Saturday 28 December 2019

| | | | | |
|-------------|---|---|----------------|-----------------------|
| Brighton | 2 | 0 | Bournemouth | Multiple Broadcasters |
| Newcastle | 1 | 2 | Everton | Multiple Broadcasters |
| Southampton | 1 | 1 | Crystal Palace | Multiple Broadcasters |
| Watford | 3 | 0 | Aston Villa | hotstar VIP |
| Norwich | 2 | 2 | Spurs | Multiple Broadcasters |
| West Ham | 1 | 2 | Leicester | Multiple Broadcasters |

Sunday 29 December 2019

| | | | | |
|-----------|---|---|---------------|-----------------------|
| Burnley | 0 | 2 | Man Utd | Multiple Broadcasters |
| Arsenal | 1 | 2 | Chelsea | Multiple Broadcasters |
| Liverpool | 1 | 0 | Wolves | Multiple Broadcasters |
| Man City | 2 | 0 | Sheffield Utd | Multiple Broadcasters |

Figure 11: Fixtures for Gameweek 20

| Position | Player | Team | Points |
|----------|-------------|-------------------|--------|
| GK | V. Guaita | Crystal Palace | 3 |
| GK | N. Pope | Burnley | 1 |
| DEF | J. Tomkins | Crystal Palace | 10 |
| DEF | V. Lindelof | Manchester United | 7 |
| DEF | S. Duffy | Brighton | 6 |
| DEF | C. Cathcart | Watford | 8 |
| DEF | C. Chambers | Arsenal | 4 |
| MID | A. Mooy | Brighton | 11 |
| MID | A. Doucoure | Watford | 6 |

| | | | |
|-----|-----------------|-----------------|----|
| MID | J. Henderson | Liverpool | 3 |
| MID | M. Vrancic | Norwich | 7 |
| MID | K. De Bruyne | Manchester City | 14 |
| FWD | P.E. Aubameyang | Arsenal | 9 |
| FWD | T. Abraham | Chelsea | 7 |
| FWD | I. Sarr | Watford | 7 |

Table 3: Our predicted best squad for gameweek 20

The highlighted players form the starting 11. They get a combined score of 89, which is much higher than the average of 54 that gameweek.

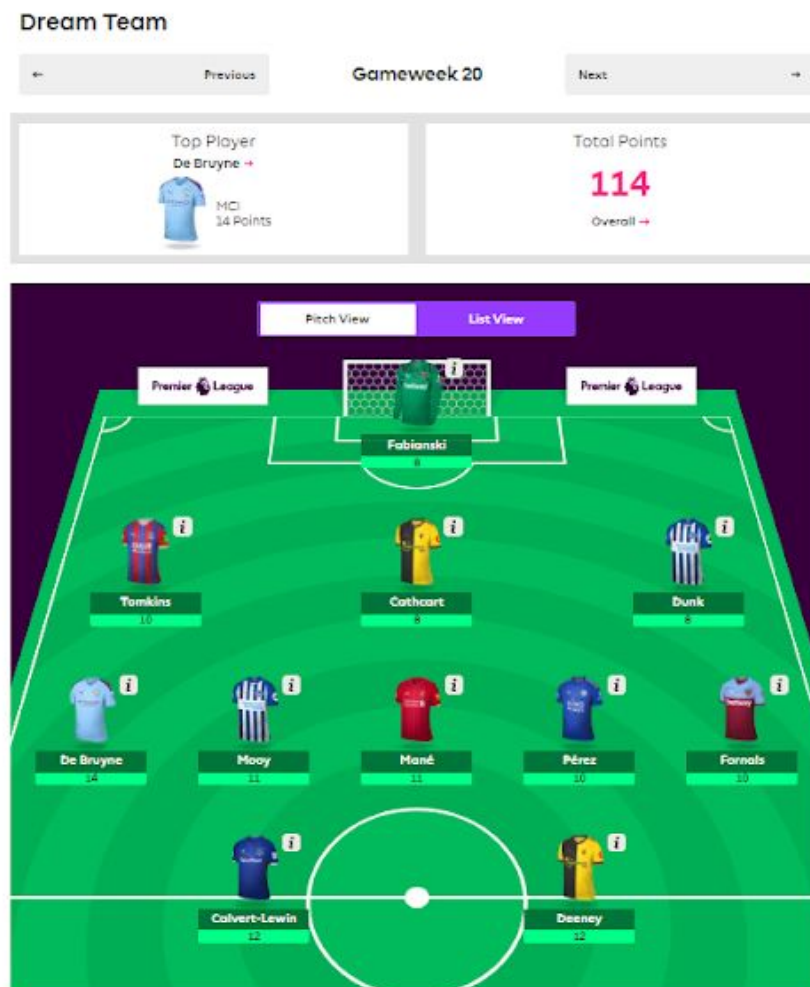


Figure 12: Actual Best Squad for GW 20

5.3. The Right Comparisons

Having established that, only 4 of our players were in the actual best eleven of the gameweek (also called Dream Team). While it appears like a bothering issue, it is not. Most of our players just miss out. A better indicator of whether or not our model worked or not would be to compare the players we selected's points to their averages.

Game week Points and Avg Points

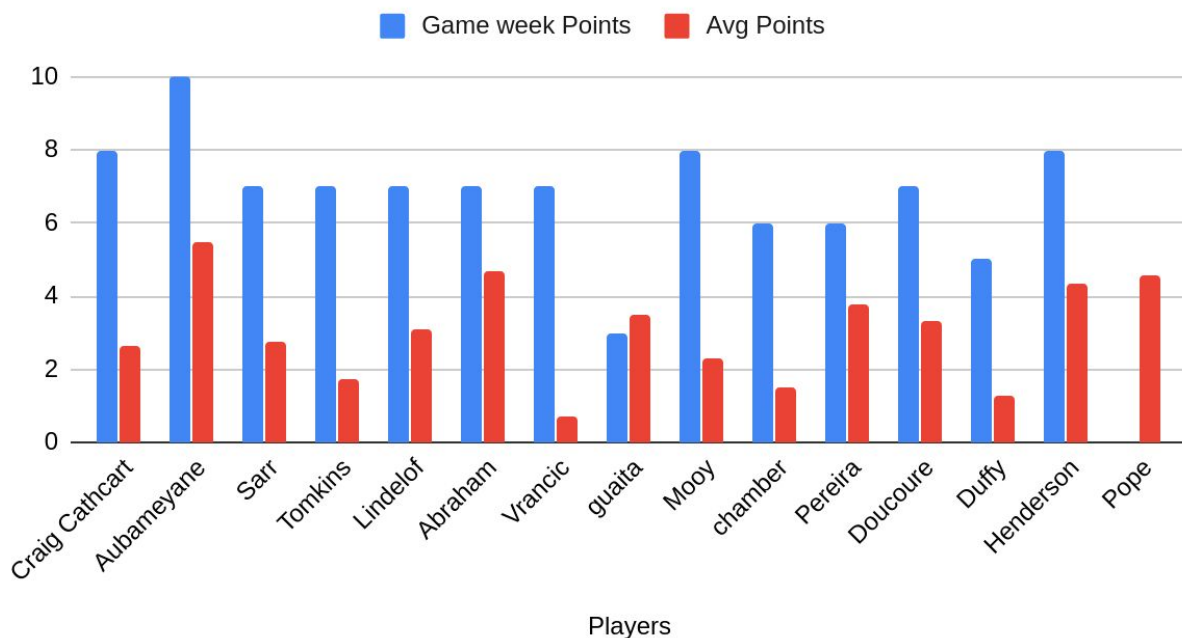


Figure 13: Comparing the GW score and average score for our predicted squad for GW 20

Having done that in Figure 13, we realized all but two of our players got more points than their average. One of them (N. Pope) did not start the game due to injury. All the other players outscored their averages multiple times. Our highlighted take from that would be the case of Vrnancic. The player scored just 20 points all season, but got 7 in the week we selected him. **This shows how dependent a player is on the team he is about to face.**

6. Conclusion

Both our problem statements were answered. While the first produced a result well on path of an almost-perfect finish, come the end of the season, the second showed just how specific our model got.

Our venture into the deepest corners of FPL through regression techniques went successfully. While we learnt that sports will remain unpredictable events, we also learnt that we can de-mystify most of it through correctly placed variables and well-thought of features which can influence the sporting event.

7. References

1. Adit-Negi. "Adit-Negi/FPL." *GitHub*, github.com/adit-negi/FPL.
2. Boldrin, Brienne. "Predicting the result of English Premier League soccer games with the use of Poisson models." Master's thesis, Stetson University, DeLand, FL (2017).
3. Herbinet, Corentin. "Predicting football results using machine learning techniques." MEng thesis, Imperial College London (2018).
4. Ganiyu, Mubarak. "Who Will Win the EPL Golden Boot in the 2019/20 Season." *Medium*, Towards Data Science, 31 July 2019, towardsdatascience.com/who-will-win-the-epl-golden-boot-in-the-2019-20-season-5e8a3a45deaa.