# HarvardX: PH125.9X - Data Science: Capstone
# Soccer - Goal Difference Prediction

Wilson Tan

9 Dec 2020

## Contents

# 1 Executive Summary

The objective of the project is to create a recommender system to predict matchday goal difference in the English Premier League using the football-data.co.uk dataset. The dataset is made up of data across 6,840 matches, with matchday goal difference ranging from -6 to 8 goals. These matches are played between 44 unique football clubs, stretching across 18 full seasons.

90% of the dataset is set aside as "model" to train the model while the remaining is used as "validation" to evaluate the proposed models. The Root Mean Square Error (RMSE) is used to evaluate the algorithm performance. RMSE measures the differences between predicted values and true values. This is regarded as a standard way to measure the model's accuracy. The RMSE of predicted values $\hat{y}$ versus true values $y$, for $N$ observations (for HomeTeam $h$, AwayTeam $a$ and Form $f$) is given by:

$$RMSE = \sqrt{\frac{1}{N}\sum(\hat{y}_{h,a,f} - y_{h,a,f})^2}$$

Considering $RMSE = 0$ would indicate a perfect fit to the data, a lower RMSE is generally desired over a higher one. The best performing model has registered a RMSE of 1.552, representing a substantial improvement from the RMSE of 1.702 based on the Naive Baseline Model. The major factors considered in this model are namely:

- Difference in the Strength between the Teams: Matchup (Hometeam, Awayteam) bias and
- Difference in the Performance between the Teams leading up to the matchday: Form (Normalised Goal Difference, Normalised Point Difference) bias

Regularization is further applied to the model to mitigate the risk of overfitting. Regularization serves to penalize on matchups with limited occurrences. This model is applied on the validation set to achieve a RMSE of 1.550. With the English Premier League regarded as the most competitive and unpredictable league in the world, a RMSE which falls close to ~10% of the range of the target value should be considered an acceptable error. Nonetheless, the model still has room for further improvement. For instance, game statistics (i.e. shots on target, possession, corners, etc), the formation deployed by the team and the ability of individual players (especially goal scoring prowess, injuries) can be included in future models. Unfortunately, due to the limitations on the data available, these models cannot be validated.

## 2 Introduction

Soccer is a physical sport played between 2 teams of 11 players, with a designated goalkeeper and 10 outfield players. The match is often played on a large grass field with each team attempting to score a goal by placing the ball across the line into the opposing team's goal. The team with the highest number of goals scored in a game wins the match.

Some of the acronyms used in this prediction model will be defined in the following table:

Table 2.1 Definition of Acronyms

| Acronym | Definition |
|---|---|
| FTHG | Full Time Home Goals Scored |
| FTAG | Full Time Away Goals Scored |
| FTR | Full Time Results (Final Scoreline) |
| HTFormPtsStr | Home Team Last 5 Home Games Results |
| ATFormPtsStr | Away Team Last 5 Away Games Results |
| HTGD | Home Team Normalized Goal Difference |
| ATGD | Away Team Normalized Goal Difference |
| GD | Normalized Goal Difference (HTGD - ATGD) |
| goal_diff | Matchday Goal Difference (FTHG - FTAG) |

Data processing is carried out in Section 2.3 to organize continuous data like the normalized goal difference and normalized point difference into discrete categories:

Table 2.2 Normalised Goal Difference Category

| Abb | Normalised Goal Diff Category | Definition |
|---|---|---|
| G1 | 1 - Overwhelming Disadvantage | GD <= -3.5 |
| G2 | 2 - Huge Disadvantage | -3.5 < GD <= -2.5 |
| G3 | 3 - Disadvantage | -2.5 < GD <= -1.5 |
| G4 | 4 - Slight Disadvantage | -1.5 < GD <= -0.5 |
| G5 | 5 - Neutral | -0.5 < GD < -0.5 |
| G6 | 6 - Slight Advantage | 0.5 <= GD < 1.5 |
| G7 | 7 - Advantage | 1.5 <= GD < 2.5 |
| G8 | 8 - Huge Advantage | 2.5 <= GD < 3.5 |
| G9 | 9 - Overwhelming Advantage | GD >= 3.5 |

Table 2.3 Normalised Point Difference Category

| Abb | Normalised Point Diff Category | Definition |
|---|---|---|
| P1 | 1 - Worst Form | PD <= -2.5 |
| P2 | 2 - Worse Form | -2.5 < PD <= -1.5 |
| P3 | 3 - Poor Form | -1.5 < PD <= -0.5 |
| P4 | 4 - Neutral | -0.5 < PD < 0.5 |
| P5 | 5 - Good Form | 0.5 <= PD < 1.5 |
| P6 | 6 - Better Form | 1.5 <= PD < 2.5 |
| P7 | 7 - Best Form | PD >= 2.5 |

As there is a strong correlation between normalized goal difference and normalized point difference, the two categories are combined to produce the GDPD Category:

Table 2.4 Form Difference (GDPD) Category

| GDPD | Goal Diff Category | Point Diff Category |
|---|---|---|
| G1P1 | 1 - Overwhelming Disadvantage | 1 - Worst Form |
| G1P2 | 1 - Overwhelming Disadvantage | 2 - Worse Form |
| ... | | |
| G5P4 | 5 - Neutral | 4 - Neutral |
| ... | | |
| G9P6 | 9 - Overwhelming Advantage | 6 - Better Form |
| G9P7 | 9 - Overwhelming Advantage | 7 - Best Form |

# 3 Preparation

This section elaborates on the steps taken from installing libraries through data processing to train-test split.

## 3.1 Prerequisites

The libraries required in this modeling are as follow:

```r
# Load installed libraries
library(tidyverse)
library(caret)
library(data.table)
library(recosystem)
library(kableExtra)
```

The operating system used in this modeling are as follow:

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## system         x86_64, mingw32
## status
## major          4
## minor          0.2
## year           2020
## month          06
## day            22
## svn rev        78730
## language       R
## version.string R version 4.0.2 (2020-06-22)
## nickname       Taking Off Again
```

## 3.2 Access to Data

The dataset can be downloaded from https://www.kaggle.com/saife245/english-premier-league. Individual datasets are available on https://www.football-data.co.uk/englandm.php but additional processing is required to produce the data required to evaluate the difference in performance between the teams (i.e. HTGD, ATGD, HTFormPtsStr, ATFormPtsStr).

## 3.3 Data Processing

A quick overview on the dimension of the dataset indicates that there 40 columns within the dataset. For simplicity, only selected features are extracted to be presented in the following table:

**Unprocesssed soccer dataset**

| | MatchId | Date | HomeTeam | AwayTeam | FTHG | FTAG | FTR |
|---|---|---|---|---|---|---|---|
| 6835 | 6835 | 2018/5/13 | Man United | Watford | 1 | 0 | H |
| 6836 | 6836 | 2018/5/13 | Newcastle | Chelsea | 3 | 0 | H |
| 6837 | 6837 | 2018/5/13 | Southampton | Man City | 0 | 1 | NH |
| 6838 | 6838 | 2018/5/13 | Swansea | Stoke | 1 | 2 | NH |
| 6839 | 6839 | 2018/5/13 | Tottenham | Leicester | 5 | 4 | H |
| 6840 | 6840 | 2018/5/13 | West Ham | Everton | 3 | 1 | H |

| | HTFormPtsStr | ATFormPtsStr | HTGD | ATGD |
|---|---|---|---|---|
| 6835 | DLWWL | WLDLL | 1.0263158 | -0.5000000 |
| 6836 | LLLLW | DWWWW | -0.2894737 | 0.7105263 |
| 6837 | WDWDL | WDWWW | -0.4736842 | 2.0526316 |
| 6838 | LLLLD | LDDDL | -0.7105263 | -0.8947368 |
| 6839 | WLWDL | WLLDL | 0.9736842 | -0.0789474 |
| 6840 | DWLLD | DWWDD | -0.5789474 | -0.3157895 |

From the above table, it is notable that some of the features can be further processed. These include:

1. Extract the `Year` and `Month` from the `date` feature;

2. Creating the matchday goal difference, `goal_diff`, by comparing the difference between the `FTHG` and `FTAG` features to represent the goal difference for the specific match. This is the target value that the model will be predicting;

3. Creating `GD` by comparing the difference between the `HTGD` and `ATGD` features before converting this continuous feature to discrete categories. The range of these categories is determined after a thorough analysis elaborated in Section 4.3. This conversion to discrete categories is necessary to avoid a situation where there are too many unique values which could not be matched between the model and validation datasets *(Note that the `HTGD` and `ATGD` features are normalized to per match basis in the original soccer dataset)*;

4. Creating `PD` by first assigning points to the `HomeTeamPtsStr` and `AwayTeamPtsStr` with:

- 3 points representing a win (W)
- 1 point for draw (D);
- 0 point for loss (L)

These points are normalized to per match basis before being compared between the two teams. For the same reason as `GD`, this continuous feature is then converted to discrete categories, as explained in Section 4.4; 5. Since scoring goals, limiting goals conceded and winning games have a high correlation, combining `GD` and `PD` features categories to form up `GDPD` category would be rather intuitive.

Following the data processing, only features with significance are selected to remain in the table for subsequent analysis and modeling.

**Processed soccer dataset**

| | MatchId | Date | Year | Month | HomeTeam | HomeTeamId | AwayTeam | AwayTeamId |
|---|---|---|---|---|---|---|---|---|
| 6835 | 6835 | 2018/5/13 | 2018 | 5 | Man United | 26 | Watford | 40 |
| 6836 | 6836 | 2018/5/13 | 2018 | 5 | Newcastle | 29 | Chelsea | 13 |
| 6837 | 6837 | 2018/5/13 | 2018 | 5 | Southampton | 35 | Man City | 25 |
| 6838 | 6838 | 2018/5/13 | 2018 | 5 | Swansea | 38 | Stoke | 36 |
| 6839 | 6839 | 2018/5/13 | 2018 | 5 | Tottenham | 39 | Leicester | 23 |
| 6840 | 6840 | 2018/5/13 | 2018 | 5 | West Ham | 42 | Everton | 17 |

| | FTR | FTHG | FTAG | goal | goal_diff | GD_Category | PD_Category | GDPD |
|---|---|---|---|---|---|---|---|---|
| 6835 | H | 1 | 0 | 1 | 1 | 7 - Advantage | 5 - Good Form | G7P5 |
| 6836 | H | 3 | 0 | 3 | 3 | 4 - Slight Disadvantage | 2 - Worse Form | G4P2 |
| 6837 | NH | 0 | 1 | 1 | -1 | 2 - Huge Disadvantage | 3 - Poor Form | G2P3 |
| 6838 | NH | 1 | 2 | 3 | -1 | 5 - Neutral | 4 - Neutral | G5P4 |
| 6839 | H | 5 | 4 | 9 | 1 | 6 - Slight Advantage | 5 - Good Form | G6P5 |
| 6840 | H | 3 | 1 | 4 | 2 | 5 - Neutral | 3 - Poor Form | G5P3 |

### 3.4 Model-Validation Split

In order to evaluate the performance of the model, the soccer dataset is split into 2 subsets, “model” and “validation” while making sure the HomeTeam, AwayTeam and GDPD. Algorithm development will be carried out on the “model” subset while “validation” subset will be used to test the final algorithm.

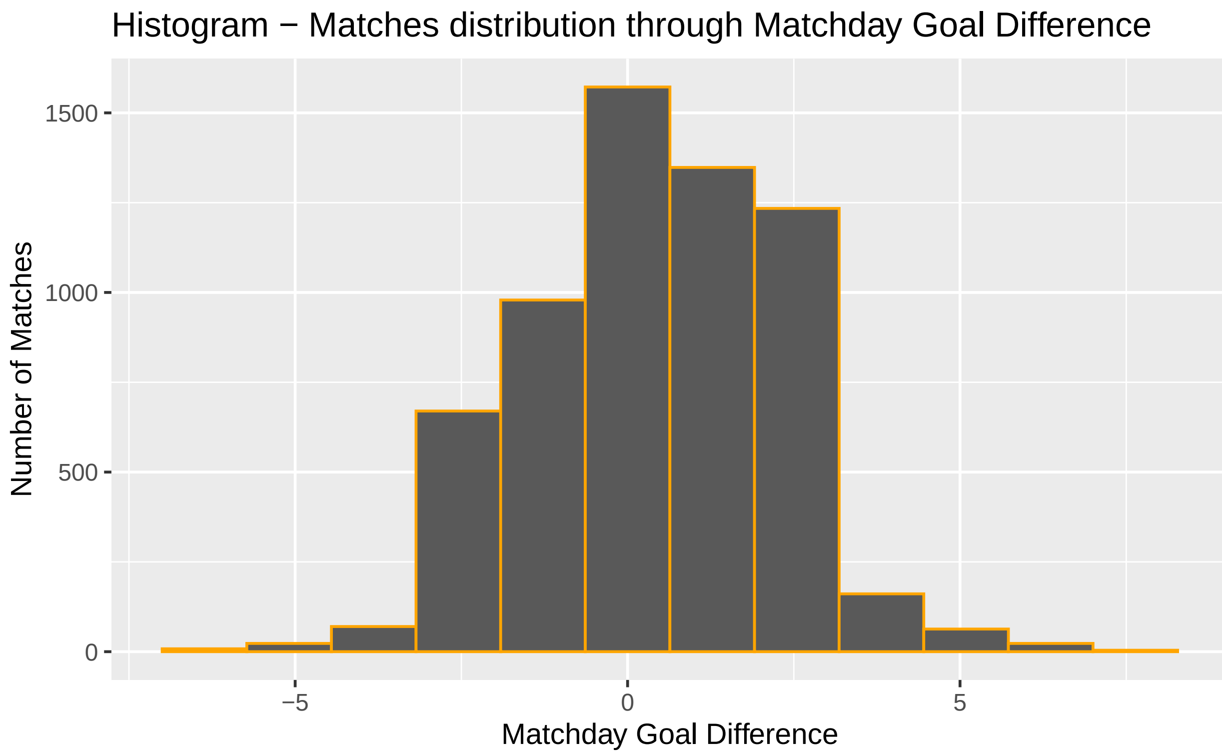
### 3.5 Train-Test Split

The model set is split further, with 90% of the data allocated to the “train” set and the remaining data allocated to the “test” set. The train set is a sample of data used to fit the model while the test set will used to provide an unbiased evaluation of a model fit on the train dataset while tuning the model parameters.

## 4 Data Analysis

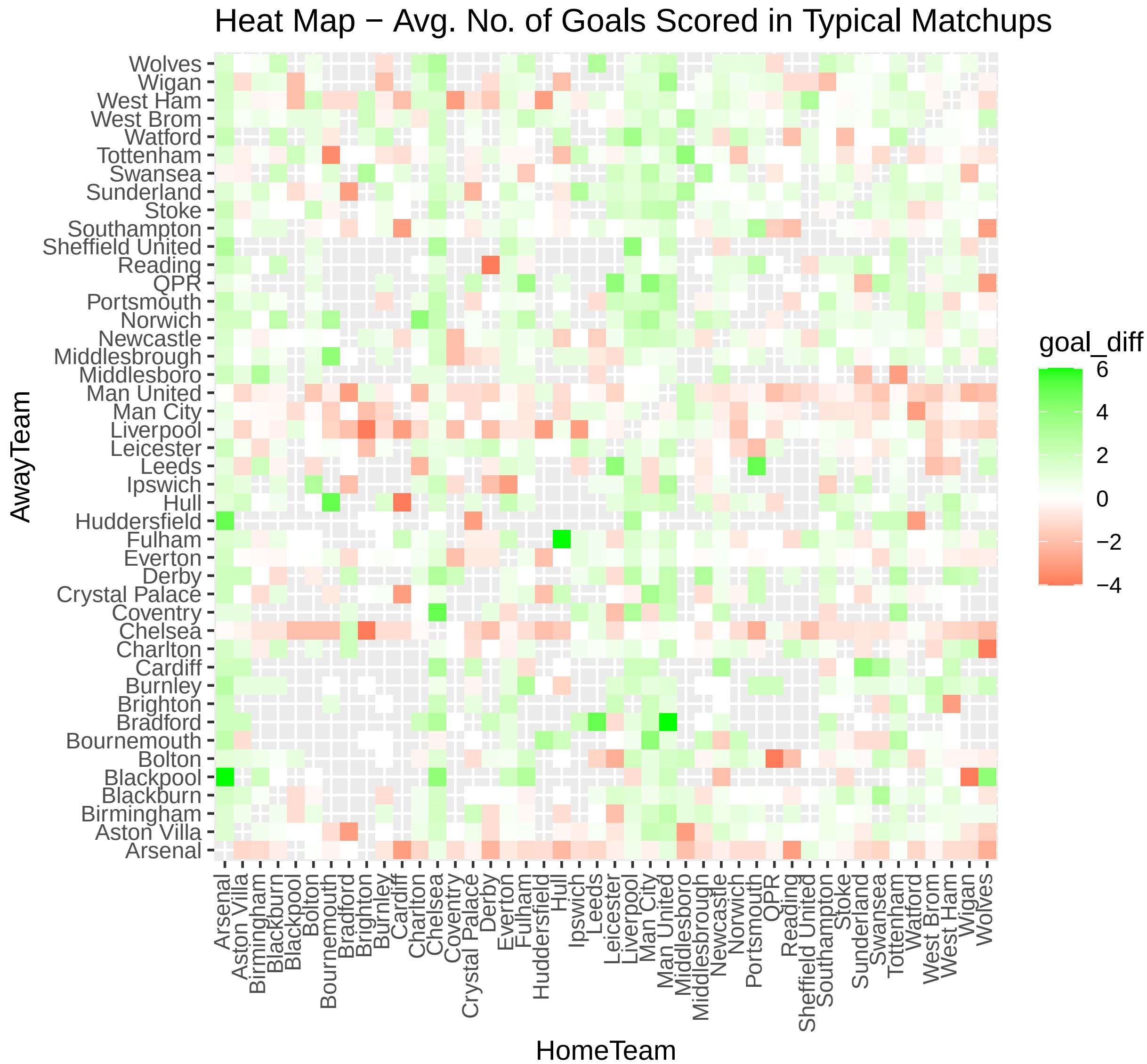
This section elaborates on the steps taken to identify notable trends and correlations between the rating and the features.

### 4.1 Number of Matches based on Matchday Goal Difference

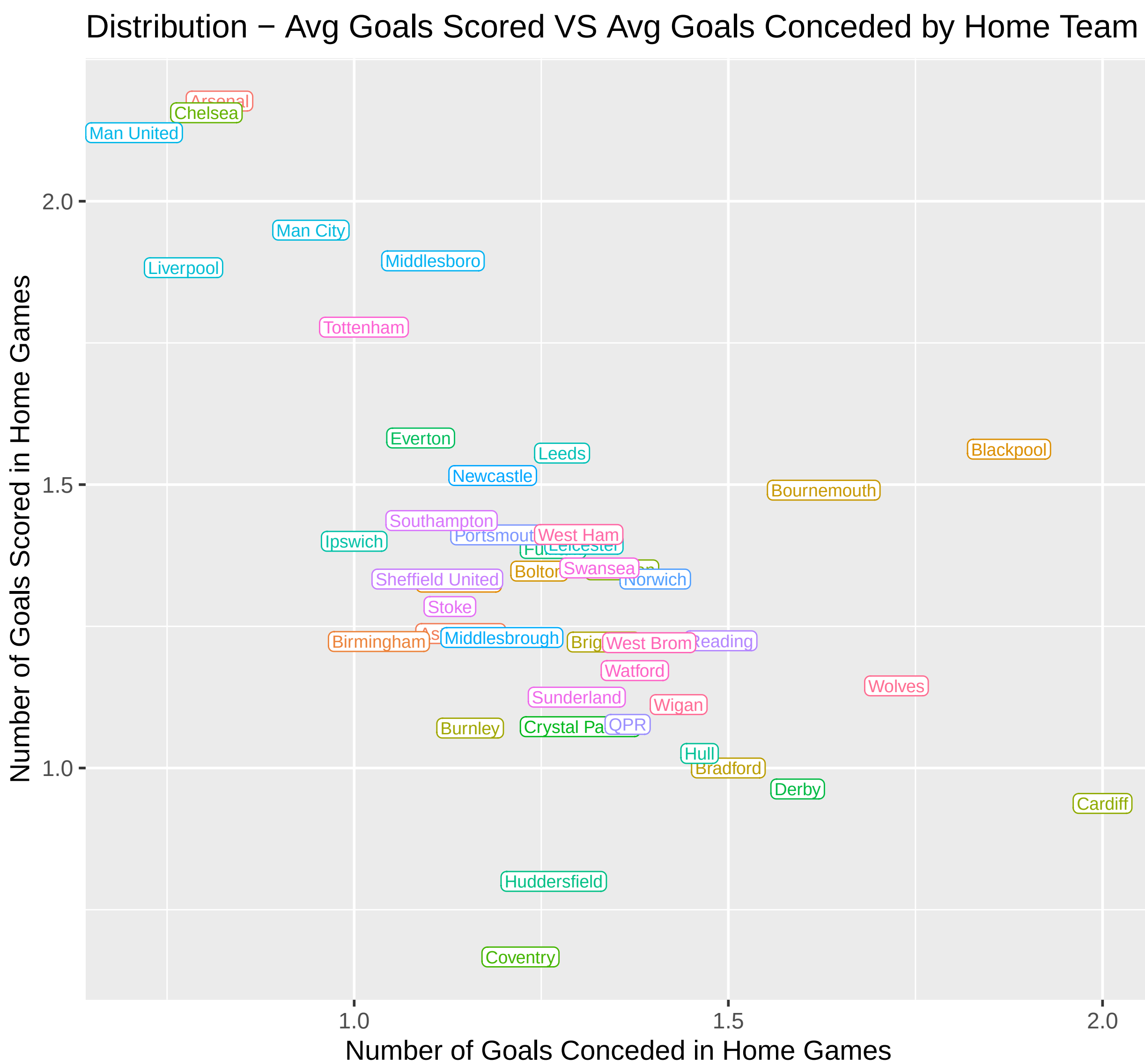



The histogram is a unimodal distribution with a single peak at 0 matchday goal difference. Another point to note is that the distribution is skewed left, meaning the majority of the observations above 0, with only a handful of observations being being negative.

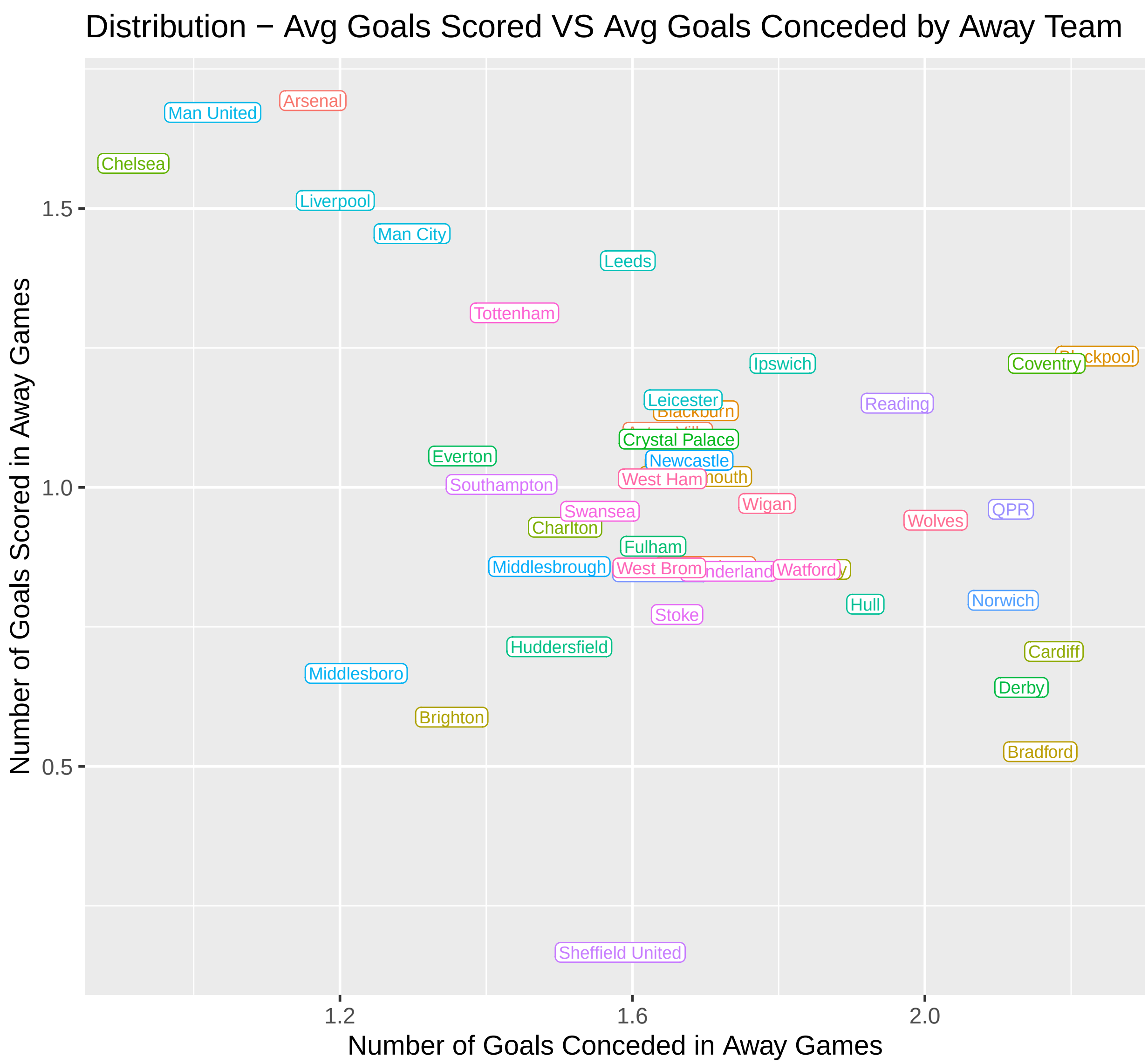
### 4.2 Matchday Goal Difference in Typical Matchup




From the heat map distribution, it can be observed that certain matchups tend to register a higher goal difference than the others. There is also strong evidence that when matches are played on the home ground of the “Big Six” (i.e. Arsenal, Chelsea, Liverpool, Man City, Man United or Tottenham), the matchday goal difference tend to end up positive, represented by the green tiles observed in these “Big Six” vertical column. In other words, these “Big Six” teams tend to score more goals than they concede in their respective home ground. The major contributing factor here is the difference in strengths between these “Big Six” teams and the AwayTeam. To further illustrate this difference in strengths, it is necessary to take a closer look at goals scored and goals conceded by the individual team.

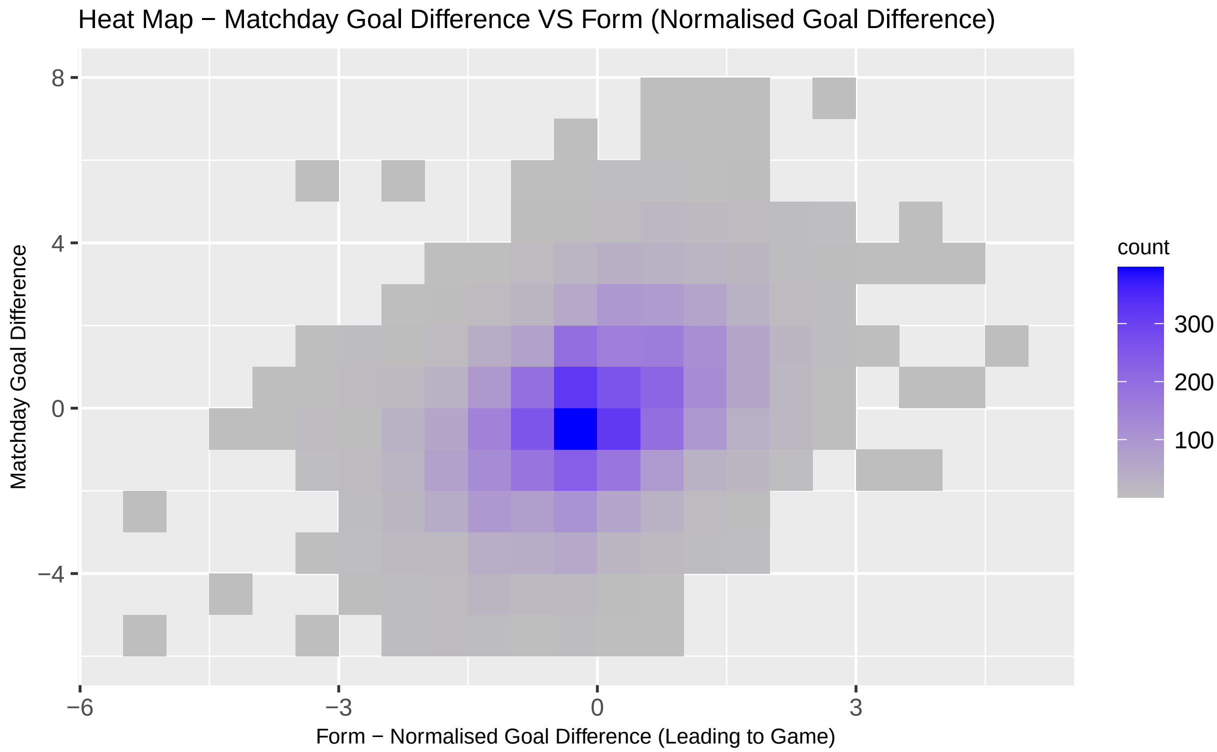



As expected, the "Big Six" teams are concentrated at the top left of the distribution, implying that these teams tend to score more and concede less goals in their home ground. From this distribution, it can also be inferred that other than a few outliers, there is an inverse relationship between goals scored and goals conceded. As explained, stronger teams normallyto score more and concede less when playing in their home ground whereas weaker teams not only find it more difficult to score but also tend to concede more. A similar trend would be expected if distribution is plotted from the perspective of the Away Team.




Similarly, the "Big Six" teams are concentrated at the top left of the distribution. One difference between the Home Team distribution and Away Team distribution is that in the Home Team distribution, majority of the teams are centered around 1.0 to 1.5 goals scored and 1.0 to 1.5 goals conceded whereas in the Away Team distribution, majority of the teams are centered around 0.75 to 1.25 goals scored and 1.4 to 1.8 goals conceded. This indicates that for majority teams outside of the "Big Six" playing at home, there is barely anything to separate goals scored from goals conceded. However, in away games, these team tend to concede more than scoring.

## 4.3 Matchday Goal Difference versus Form (Normalised Goal Difference)

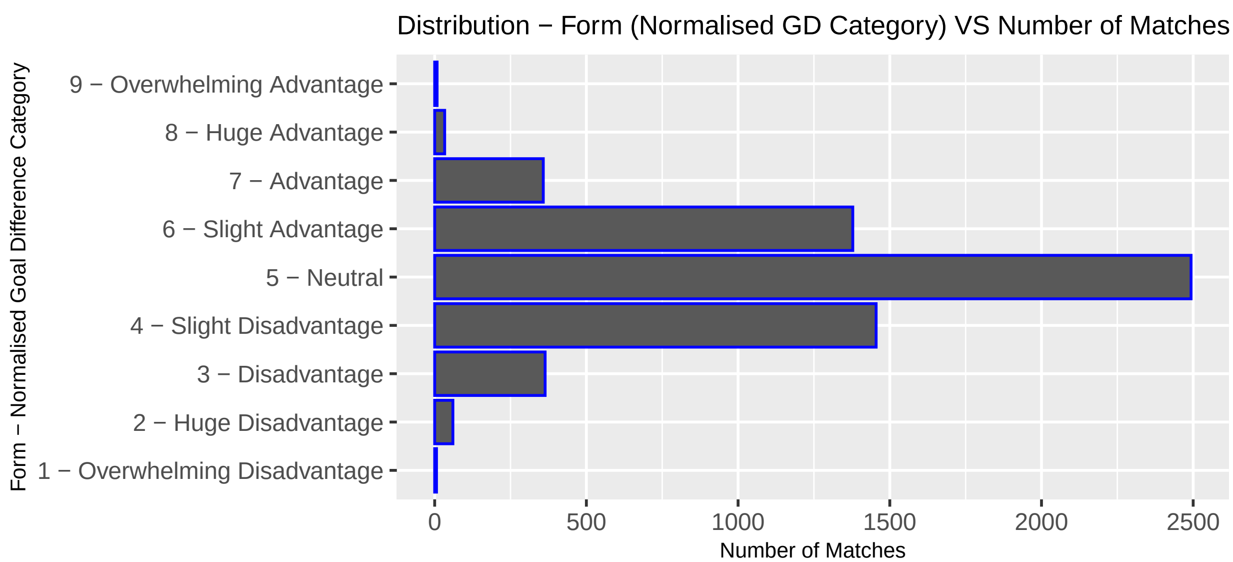


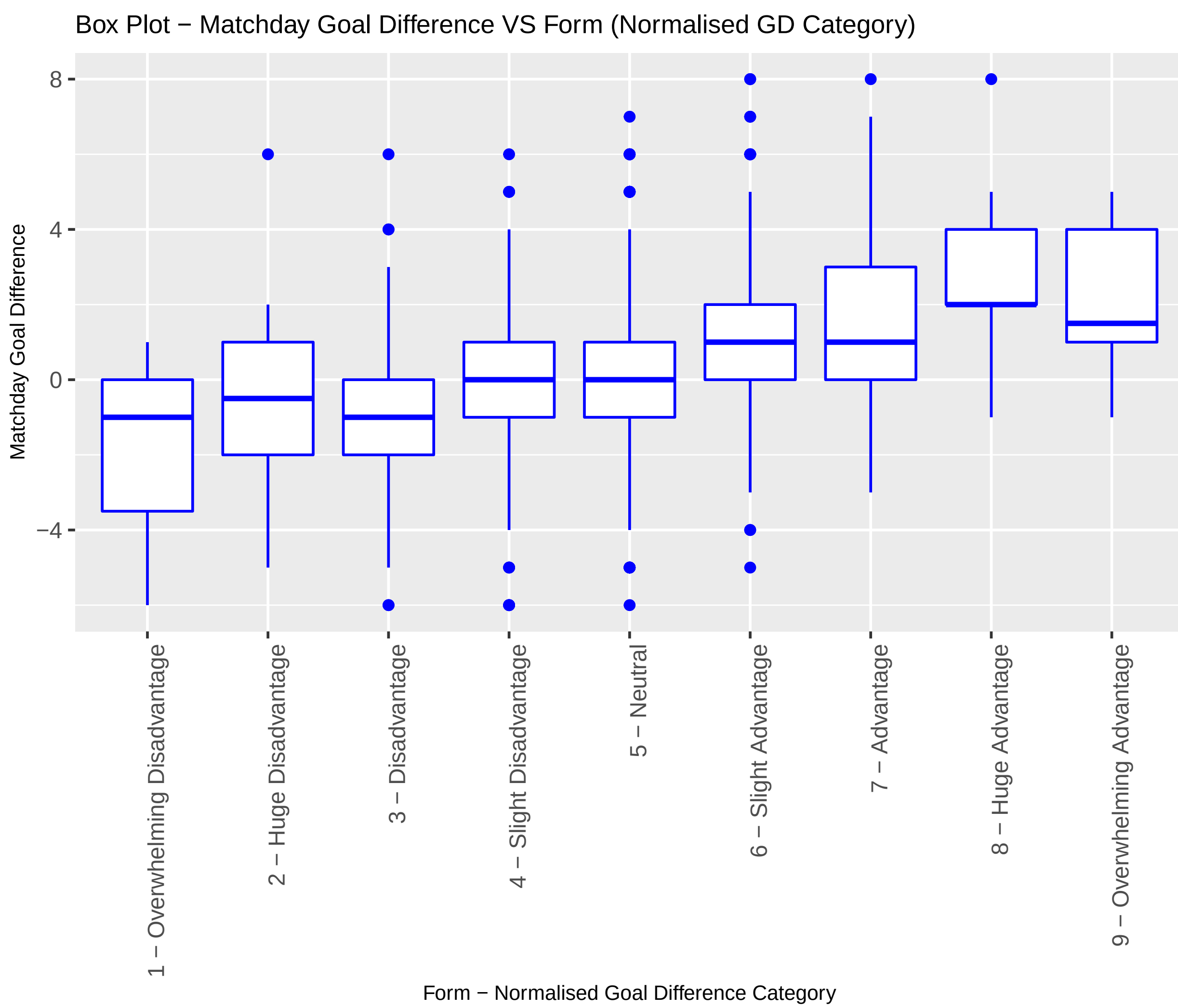

Normalized Goal Difference (GD) is determined by taking the difference between the normalized goal differences registered by the HomeTeam (HTGD) and the AwayTeam (ATGD). These figures, normalized to per match basis, are calculated from the Number of Goals Scored and Conceded by the respective team throughout the season, before being normalized to per match basis. From the heat map, GD appears to have a positive correlation with the the matchday goal difference, with a huge concentration of matches registering matchday goal difference between -2 and 2 with GD between -1.5 and 1.5. The positive correlation suggests that a matchup with higher GD would expect a higher matchday goal difference and vice versa.

Since the GD involves continuous data, this could result in very distinct values of GD in different matchups. As such, there may be a significant number of GD values calculated in the model dataset not be repeated (distinct) in the validation dataset and vice versa. In order to avoid such a scenario, the proposal is to convert this continuous GD value to a discrete categorical class to avoid any mismatch in subsequent modeling. Observations from this heat map is used to determine the range of GD values in each category. For instance, for matchups where there is a GD of less than -3.5, the matchday goal difference is likely to be negative, whereas for matchups where there is a GD of more than 3.5, the matchday goal difference is likely to be positive. Categories between these two limits are specified with a span of 1. These categories are provided in Table 2.2.

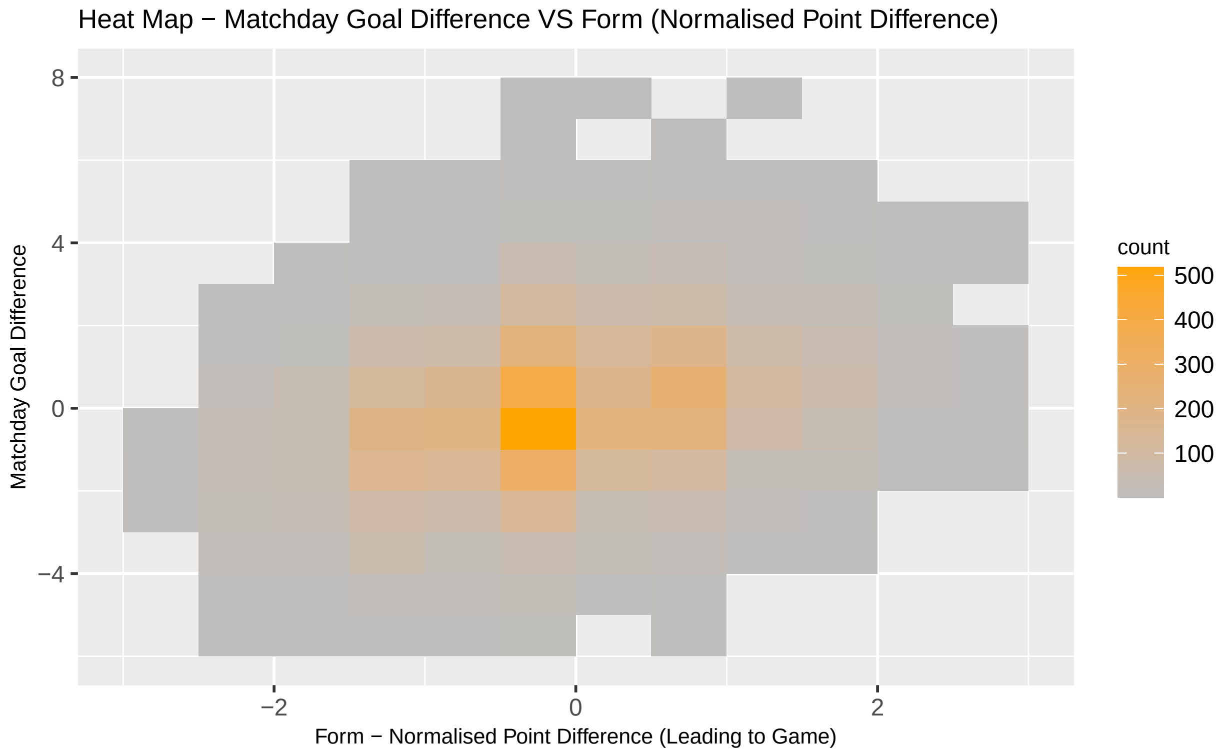
*Note: These considerations have been taken into account in Section 3.3 Data Processing .*

Further analyses on the distribution of these categories are carried out to substantiate the trend explained. As presented in the following charts, the observations are aligned with the initial understanding of the relationship between GD and matchday goal difference.


Distribution − Form (Normalised GD Category) VS Number of Matches
Form − Normalised Goal Difference Category
9 − Overwhelming Advantage
8 − Huge Advantage
7 − Advantage
6 − Slight Advantage
5 − Neutral
4 − Slight Disadvantage
3 − Disadvantage
2 − Huge Disadvantage
1 − Overwhelming Disadvantage
0
500
1000
1500
2000
2500
Number of Matches


Box Plot − Matchday Goal Difference VS Form (Normalised GD Category)
Matchday Goal Difference
8
4
0
−4
1 − Overwhelming Disadvantage
2 − Huge Disadvantage
3 − Disadvantage
4 − Slight Disadvantage
5 − Neutral
6 − Slight Advantage
7 − Advantage
8 − Huge Advantage
9 − Overwhelming Advantage
Form − Normalised Goal Difference Category

## 4.4 Matchday Goal Difference versus Form (Normalised Point Difference)

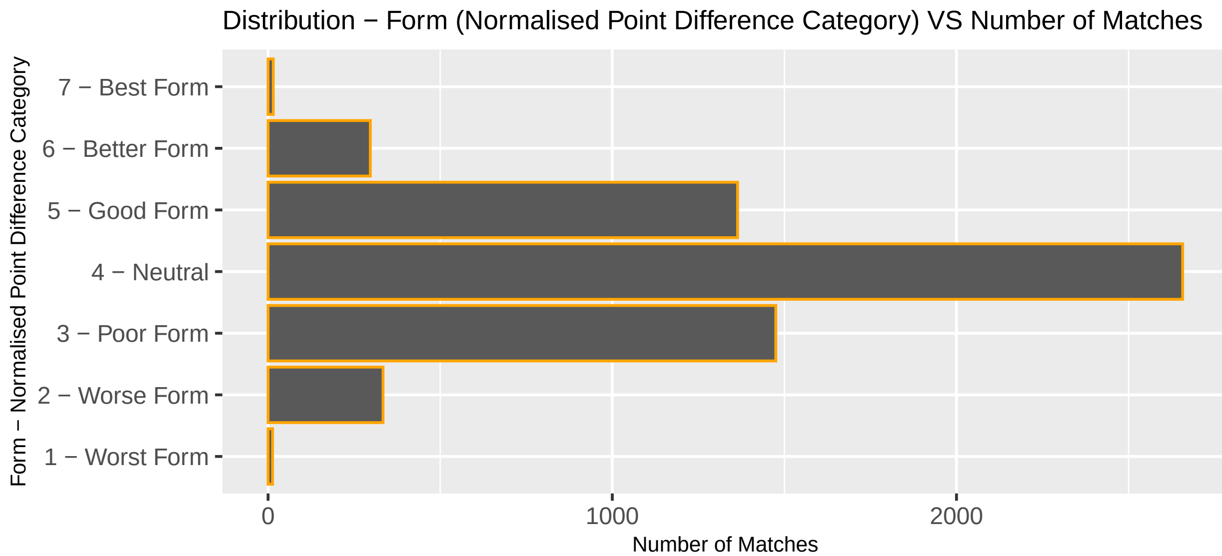


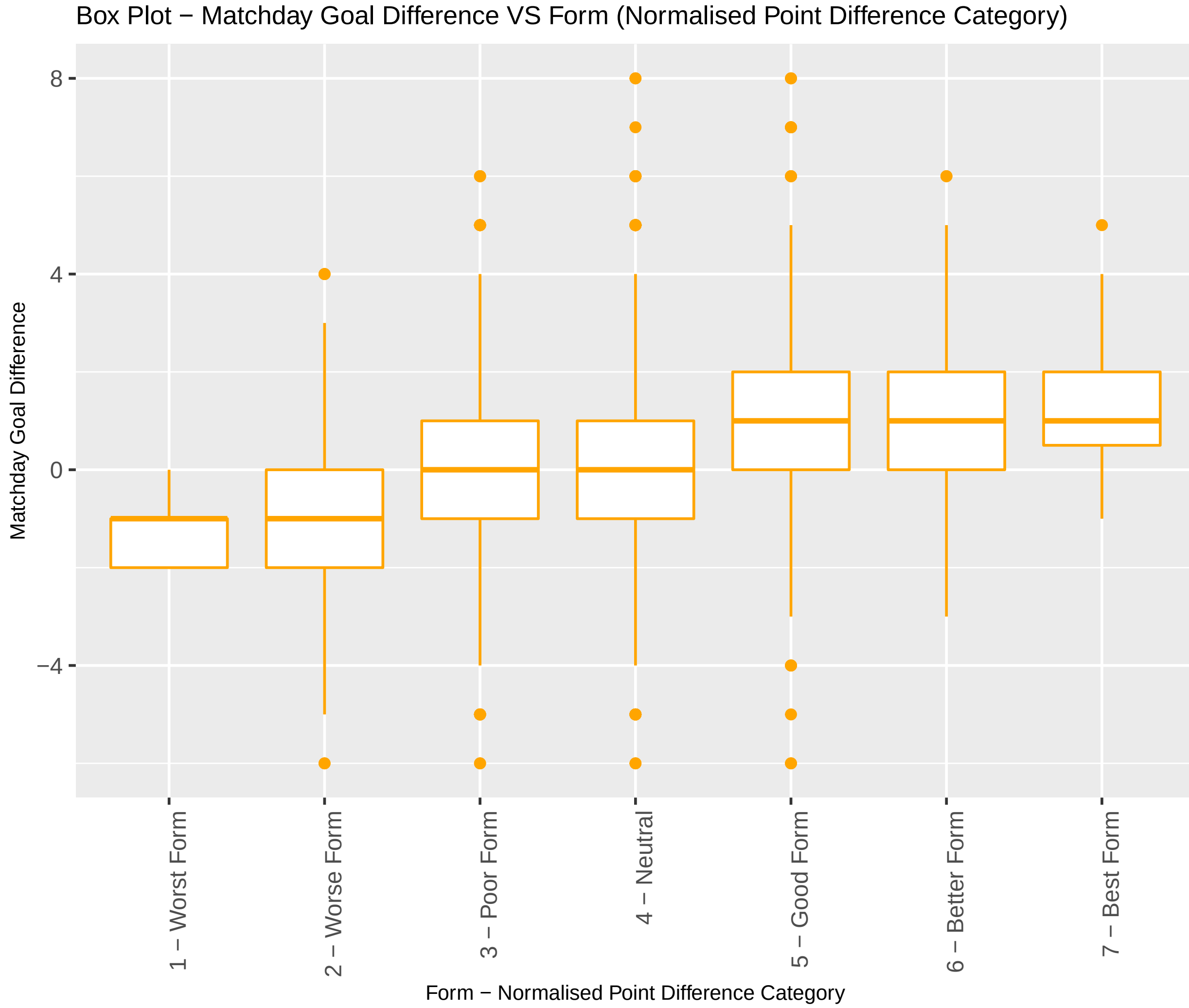

Normalized Point Difference (PD) is determined by taking the difference between the normalized point differences registered by the HomeTeam and the AwayTeam. These figures are calculated by first assigning 3 points to a win and 1 point to a draw in the last 5 games, before being normalized to per match basis. From the heat map, PD appears to have a positive correlation with the the matchday goal difference, with a huge concentration of matches registering matchday goal difference between -2 and 2 with PD between -1 and 1. Although not as apparent as GD, the positive correlation suggests that a matchup with higher PD would expect a higher matchday goal difference and vice versa.

A similar approach is adopted to convert the continuous PD values to discrete categorical class to avoid mismatch between the model and validation datasets. These categories are provided in Table 2.3.

*Note: These considerations have been taken into account in Section 3.3 Data Processing .*

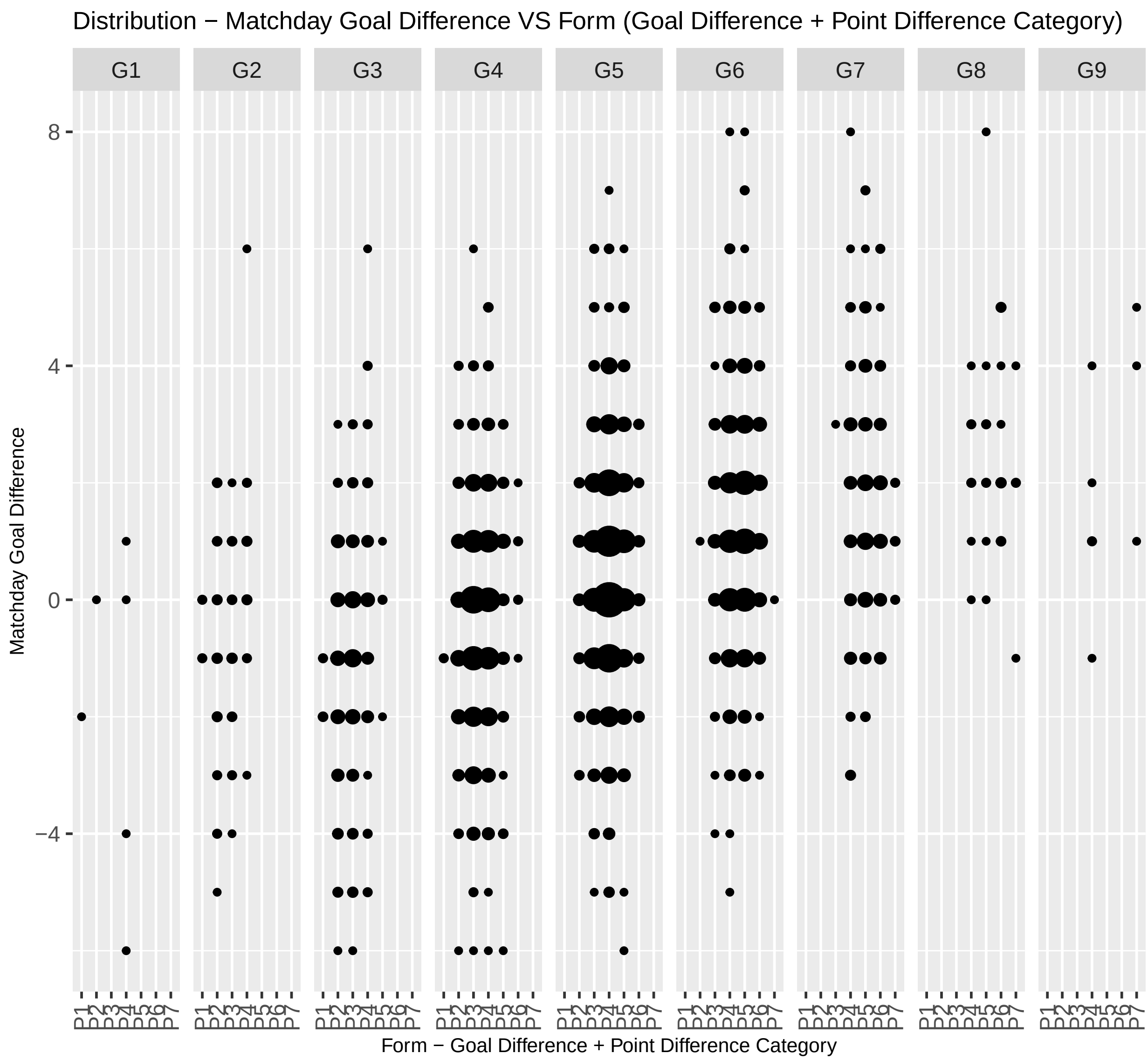
Further analyses on the distribution of these categories are carried out to substantiate the trend explained. As presented in the following charts, the observations are aligned with the initial understanding of the relationship between PD and matchday goal difference. .


Distribution − Form (Normalised Point Difference Category) VS Number of Matches
Form − Normalised Point Difference Category
7 − Best Form
6 − Better Form
5 − Good Form
4 − Neutral
3 − Poor Form
2 − Worse Form
1 − Worst Form
0
1000
2000
Number of Matches

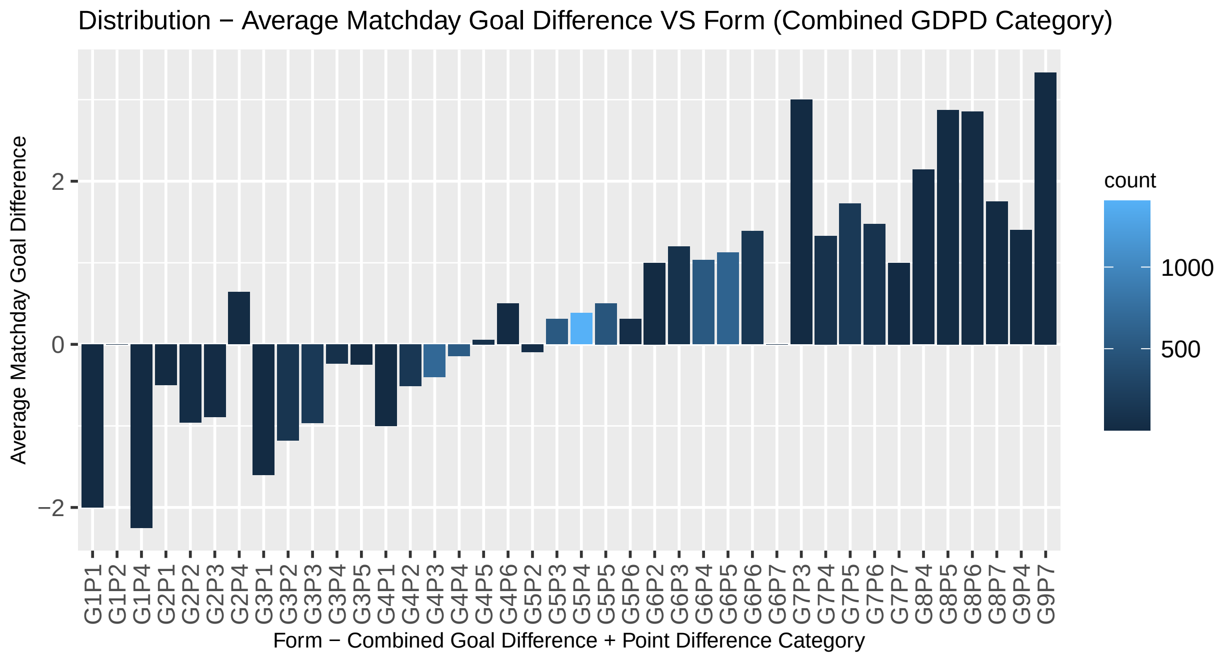

Box Plot − Matchday Goal Difference VS Form (Normalised Point Difference Category)
Matchday Goal Difference
8
4
0
−4
1 − Worst Form
2 − Worse Form
3 − Poor Form
4 − Neutral
5 − Good Form
6 − Better Form
7 − Best Form
Form − Normalised Point Difference Category

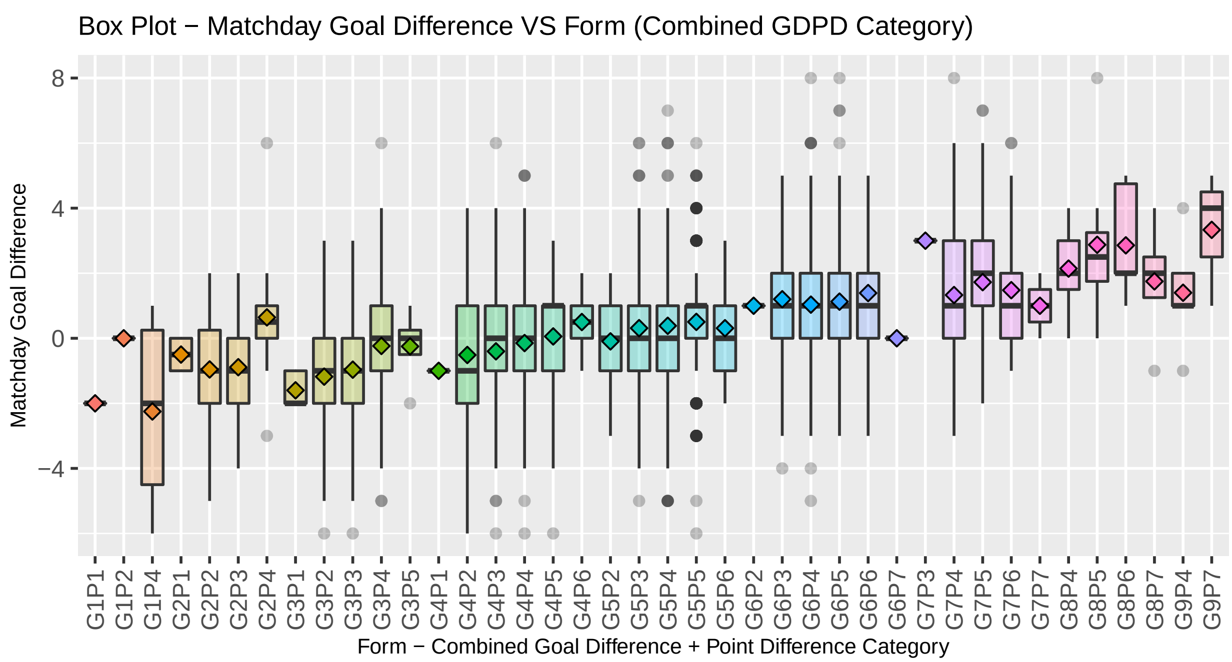
## 4.5 Matchday Goal Difference versus Form (Combined Normalised Point and Goal Difference)

Scoring more goals than conceding will win games, resulting in more points. Under this logic, combining GD and PD categories would be intuitive and is expected to produce more accurate predictions for the matchday goal difference than the individual categories. A facet grid is employed to allow multiple axes (GD, DD, matchday goal difference) to be reflected on the same plot for a more comprehensive study of the relationship between GD and PD. Number of matches is represented by the size of the point.



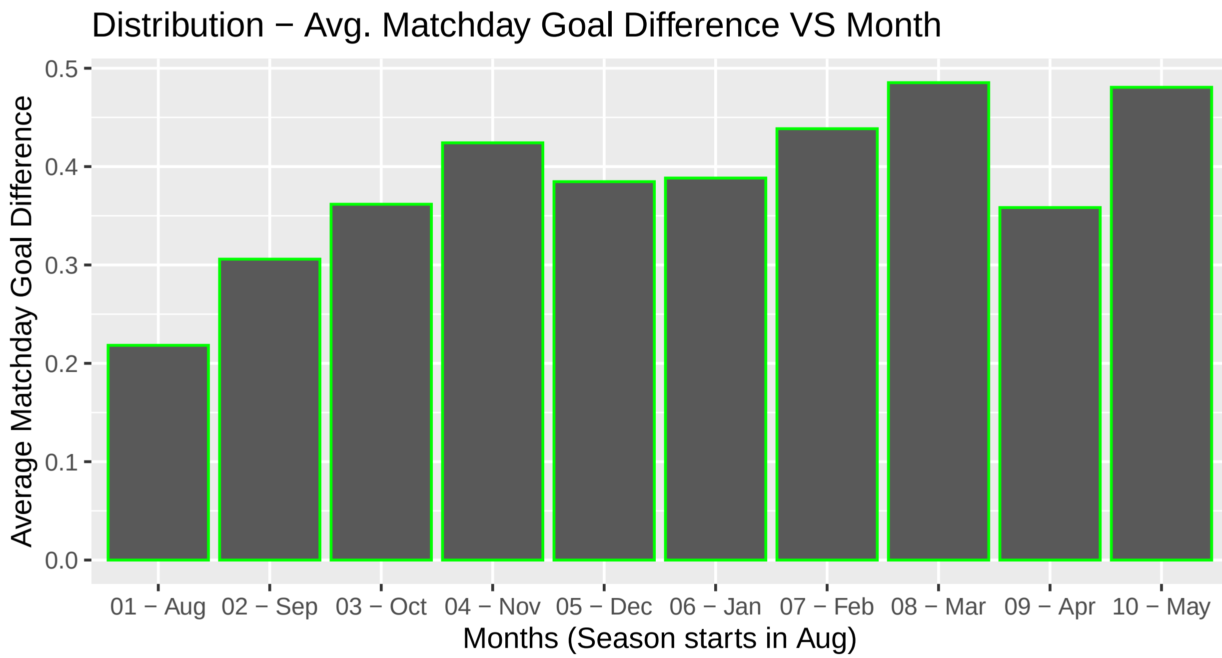

As expected, moving across the plot, there is a shift in the number of matches towards the right, both in GD and PD Category. A HomeTeam playing with a better GD and PD is expected score more and concede less goals. At the same time, it can be inferred that between GD and PD, GD seemed to have a stronger influence on the matchday goal difference. This explains for the GD Category being placed in front of the combined GDPD category to reflect its importance. The following plots will be better representations of this trend.



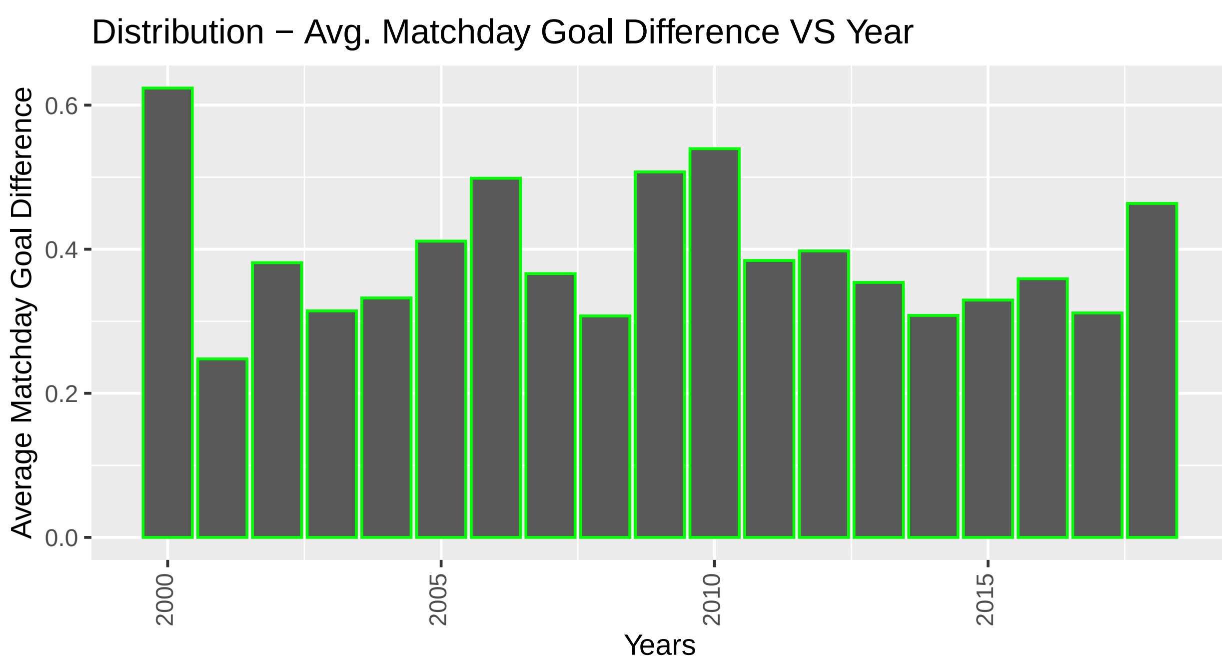




From both plots, it can be observed that both the mean and median matchday goal difference increases moving across the plots. It is also worth noting that some categories happen less compared to others and there are even cases where the category did not occur in the model dataset.

### 4.6 Matchday Goal Difference versus Month & Year




From the distribution, it can be observed that as the season progresses, there is a increase in the average matchday goal difference. This observation could be a result of teams gaining goal scoring prowess as players become more familiar with the team's formation and playing style.




Unlike the month feature, the significance of year remains inconclusive.

# 5 Modeling Approach

## 5.1 Naive-Baseline Model

As the name suggests, the Naive-Baseline (Simple Average) Model will be used as the reference model for measuring the performance of subsequent models. The Naive-Baseline Model assumes that matchday goal difference across all matchups (HomeTeam and AwayTeam) will be the same. In this model, the mean value, which is approximated to be ~0.3826, will be used as the predicted rating for all reviews, regardless of movie or user. The formula of this Naive-Baseline Model can be represented by:

$$Y_{h,a} = \hat{\mu} + \varepsilon_{h,a}$$

where $\hat{\mu}$ refers to the mean and $\varepsilon_{h,a}$ refers to the independent error sampled from the same distribution centered at 0.

The RMSE of the Naive-Baseline Model on the `test` dataset is **1.702** with an accuracy of predicting home win at **54.1%**.

## 5.2 Matchup Effect

As pointed out in the Data Analysis Section 4.2, both the HomeTeam and AwayTeam, collectively known as **Matchup** has a strong influence in the matchday goal difference. Therefore, it would be sensible to include the HomeTeam effect, $b_h$, and the AwayTeam effect, $b_a$ to enhance the model. The resulting formula that represents the Matchup Effect Model is given by:

$$Y_{h,a} = \hat{\mu} + b_h + b_a + \varepsilon_{h,a}$$

where:

- $b_h$ refers to the HomeTeam effect or bias for HomeTeam $h$ and
- $b_a$ refers to the AwayTeam effect or bias for AwayTeam $a$

The resulting RMSE of the Matchup Effect Model on the `test` dataset is **1.578** with an accuracy of predicting home win at **63.9%**. This represents an improvement in the RMSE of ~7.3% when compared with the Naive-Baseline Model. Comparing the RMSE with the actual matchday goal difference which spans from -6 to 14, this error is ~11.3%.

## 5.3 Matchup & Form Effect

Another feature that has a significant influence on the matchday goal difference is the **Form** difference of the HomeTeam and AwayTeam. For simplicity, this difference in Form is represented by discrete categories GDPD which is a combination of the Normalised Goal Difference and Normalised Point Difference. The reason for this combination is elaborated in the Data Analysis Section 4.5. Adding the Form effect, $b_f$, the resulting Matchup & Form Effect Model formula is given by:

$$Y_{h,a,f} = \hat{\mu} + b_h + b_a + b_f + \varepsilon_{h,a}$$

where $b_f$ refers to the Form (GDPD) effect or bias $f$

The resulting RMSE of the Matchup & Form Effect Model on the `test` dataset is **1.573** with an accuracy of predicting home win at **64.8%**. This represents an improvement in the RMSE of ~7.6% when compared with the Naive-Baseline Model. Comparing the RMSE with the actual matchday goal difference which spans from -6 to 14, this error is ~11.2%. However, this is only a slight improvement from the Matchup Effect Model.

## 5.4 Matchup & Form & Month Effect

The last feature to exhibit a significant influence on the matchday goal difference is the **Month** of the matchday. The impact of the month on the matchday goal difference is discussed in the Data Analysis Section 4.6. The inclusion of the Month effect, $b_m$, would result in the following formula:

$$Y_{h,a,f,m} = \hat{\mu} + b_h + b_a + b_f + b_m + \varepsilon_{h,a}$$

where $b_m$ refers to the Month effect or bias $m$

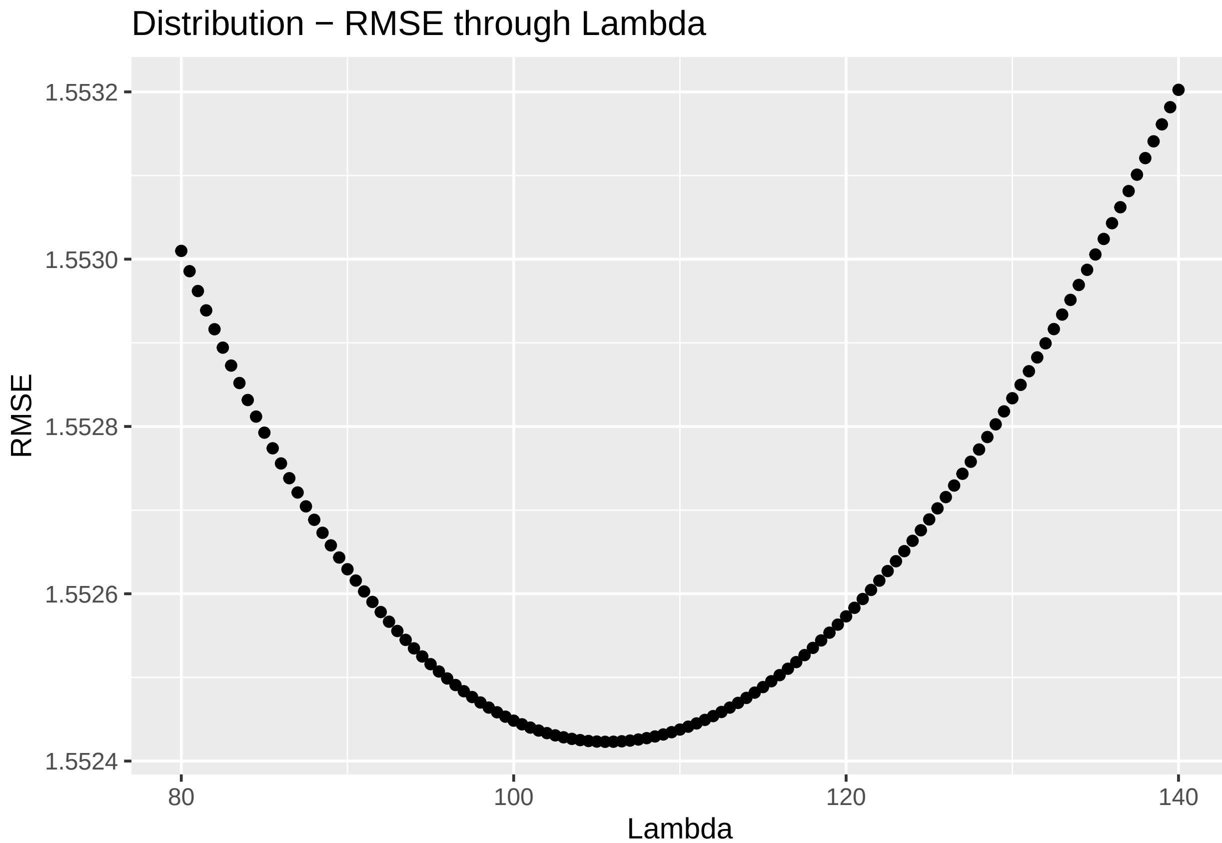
The resulting RMSE of the Matchup & Form & Month Effect Model on the `test` dataset is **1.574** with an accuracy of predicting home win at **64.0%**. This represents an improvement in the RMSE of ~7.5% when compared with the Naive-Baseline Model. Comparing the RMSE with the actual matchday goal difference which spans from -6 to 14, this error is ~11.2%. However, when compared to the Matchup & Form Effect Model, both RMSE and accuracy are worse of. As a result, the previous model, the Matchup & Form Effect Model, will be selected for further optimization.

## 5.5 Matchup & Form Effect + Regularization

To mitigate the risk of overfitting, **regularization** is applied to the selected model. The use of regularization penalizes on matchups or Form Categories with low occurrences. As explained in Section 4.5, there are some Form Categories where there is only 1 data point. The tuning parameter, lambda, resulting in the smallest RMSE will be used to shrink the HomeTeam, AwayTeam and Form effect for the test set. The formula that represents the Movie & User Effect + Regularization Model is:

$$\frac{1}{N}\sum_{h,a,f}(y_{h,a,f} - \mu - b_h - b_a - b_f)^2 + \lambda(\sum_h b_h^2 + \sum_a b_a^2 + \sum_f b_f^2)$$

where $\lambda$ is the tuning parameter applied to the movie and user effect.




From the plot, it can observed that the lambda value that corresponds to lowest RMSE of 1.552 in the train set is 105.5.

Applying this lambda value onto the the test set, the resulting RMSE (of the Matchup & Form Effect + Regularization Model) is **1.552** with an accuracy of predicting home win at **63.0%**. This represents an improvement in the RMSE of ~8.8% when compared with the Naive-Baseline Model. Comparing the RMSE with the actual matchday goal difference which spans from -6 to 14, this error is ~11.1%.

With the resulting RMSE for the Matchup & Form Effect + Regularization Model on the `test` dataset being the lowest at 1.552, the Matchup & Form Effect + Regularization Model will be selected as the final algorithm to be applied on the `validation` dataset.

Table 5.1: RMSE Results for All Models

| Model | RMSE | Accuracy |
|---|---|---|
| [Test] Naive Baseline (Mean) Model | 1.702022 | 0.5413290 |
| [Test] MatchUp Effect Model | 1.578257 | 0.6385737 |
| [Test] MatchUp & Form Effect Model | 1.572987 | 0.6482982 |
| [Test] MatchUp & GDPD & Month Effect Model | 1.574032 | 0.6401945 |
| [Test] Matchup & Form Effect + Regularization Model | 1.552423 | 0.6304700 |
| [Validation] Matchup & Form Effect + Regularization Model | 1.549620 | 0.6335766 |

As reflected in the table, the RMSE for the final Matchup & Form Effect + Regularization Model is **1.550** with an accuracy of predicting home win at **63.3%**. Comparing the RMSE with the actual matchday goal difference which spans from -6 to 14, this error is ~11.1%.

## 5.6 Conclusion

Characterized by the lowest RMSE value, the Matchup & Form Effect + Regularization Model is regarded as the optimal model for predicting matchday goal differences. From the executive summary, it is noteworthy to point out that there could be other features like game statistics, formation deployed and the ability of individual players that could be applied to further improve on the model's prediction accuracy. However, due to the limitations of the data available online, these models cannot be validated. Overall, a RMSE of **1.550** with an accuracy of predicting home win at **63.3%** should be regarded an acceptable prediction model, considering the English Premier League to be the most unpredictable soccer leagues in the world.