Predicting Victory: A Logistic Regression Analysis of Manchester United’s 22/23 Season

Aiden Liu

Table of Contents

Introduction 1

Purpose 2

Preliminary Analysis 2

Data Manipulation 6

Model Building and analysis 12

Method and Assumption checks 22

Interpretation 30

Model 31

Explanation: 31

Key Predictors 31

Actionable Insights for Improvement 32

Reflection 32

Introduction

Manchester United Football Club is one of the most famous and successful football clubs in the world. Based in Manchester, England, the club was founded in 1878 as Newton Heath LYR Football Club before changing its name to Manchester United in 1902. The team plays its home matches at Old Trafford, a stadium with a seating capacity of over 74,000.

The club has won numerous domestic and international trophies, including 20 English league titles, 12 FA Cups, and 3 UEFA Champions League titles. Manchester United has been known for its attacking style of play and its legendary players, such as Cristiano Ronaldo, Bobby Charlton, George Best, Eric Cantona, and Ryan Giggs. The club has a huge global fanbase and is often considered one of the wealthiest and most commercially successful sports teams in the world.

Historically, the club enjoyed great success under manager Sir Alex Ferguson, who led Manchester United to numerous titles during his 26-year tenure before retiring in 2013. The club's fortunes have varied since then, with several managerial changes and attempts to return to its former glory.

This dataset includes the statistics of all Manchester United players in season 2022-2023. It includes statistics of all competitions and only players who have played. For more or detailed statistics, please refer to https://fbref.com/en/squads/19538871/2022-2023/all_comps/Manchester-United-Stats-All-Competitions.

This dataset also include match statistics of Manchester United in season 2022-2023. It's for highlighting which are the most important traits that contributes to win or lose matches.

Source: https://www.kaggle.com/datasets/hkwindvolder/manutd-2023-player-statistics?select=ManUtd+2023+Match+Statistics.csv

Purpose

Over the last 10 years, Manchester United has not been in contention for the Premier League. After £1 billion spent on players' transfer fees, some questions need to be asked on how the team plays. Personally, I believe the team has not been consistently playing well despite the amount of money injected into the squad. I aim to predict what leads Man Utd to win football matches based on match statistics, while providing some insights on the key attributes during football matches.

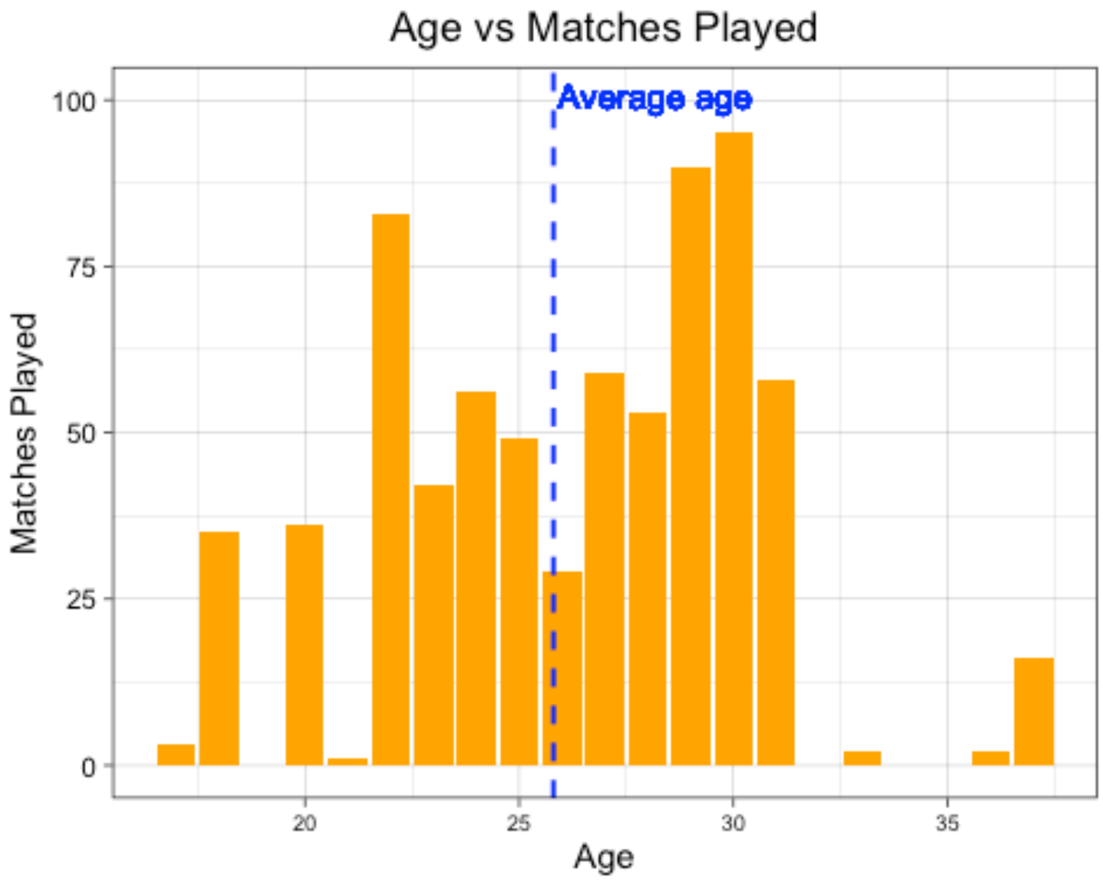
Preliminary Analysis

```r
#Importing required libraries
library(ggplot2)
library(tidyverse)
library(glmnet)
library(car)
library(detectseparation)
library(ResourceSelection)
library(pROC)
library(Metrics)
library(caret)
```

Reading in Man Utd player dataset

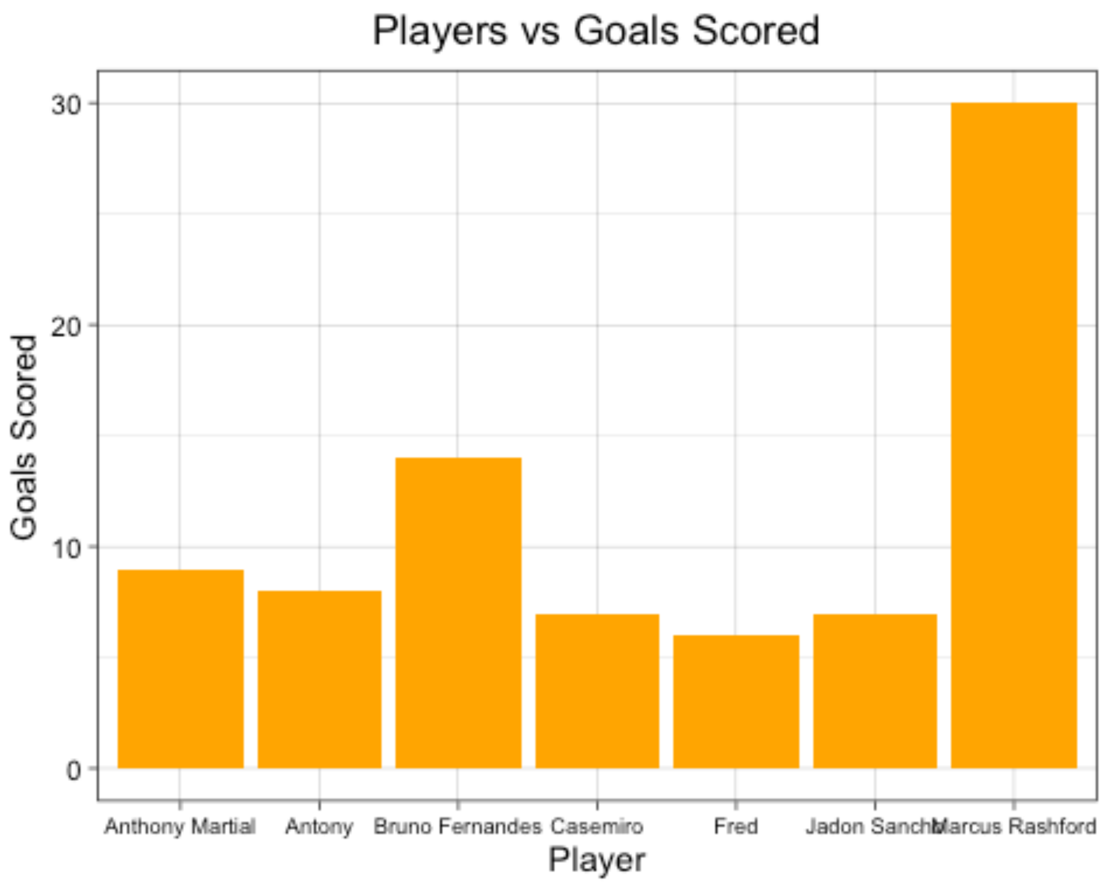
```r
manutd_players = read.csv("ManUtdplayer.csv", header=TRUE)

#Plotting Age vs Matches Played
average.age = mean(manutd_players$Age) #Average Age of Man Utd's squad
print(average.age) #25.8

## [1] 25.8
```

```
ggplot(data = manutd_players, mapping=aes(x=Age, y=Match.Played))+
  geom_col(fill = "orange")+
  geom_vline(xintercept = 25.8,                # Vertical line at x = 25
             color = "blue",                  # Line color
             linetype = "dashed",           # Line type (optional)
             size = 0.7)+
  geom_text(aes(x = 25.9, y = max(Match.Played) + 40, label = "Average age"),
            color = "blue",                   # Text color
            angle = 0,                        # Angle of text
            hjust = 0,                        # Horizontal alignment
            vjust = 0) +                      # Vertical alignment
  labs(x = "Age", y = "Matches Played", title = "Age vs Matches Played")+
  theme_linedraw()+
  theme(plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(size = 7) # Adjust text size
  )+ylim(0,100)
```



As per the plot above, Man Utd has a mixture of experience and young talent with two noticeable peaks at around ages 23 and 30. The average age of a player in the Man Utd squad is 25.8.

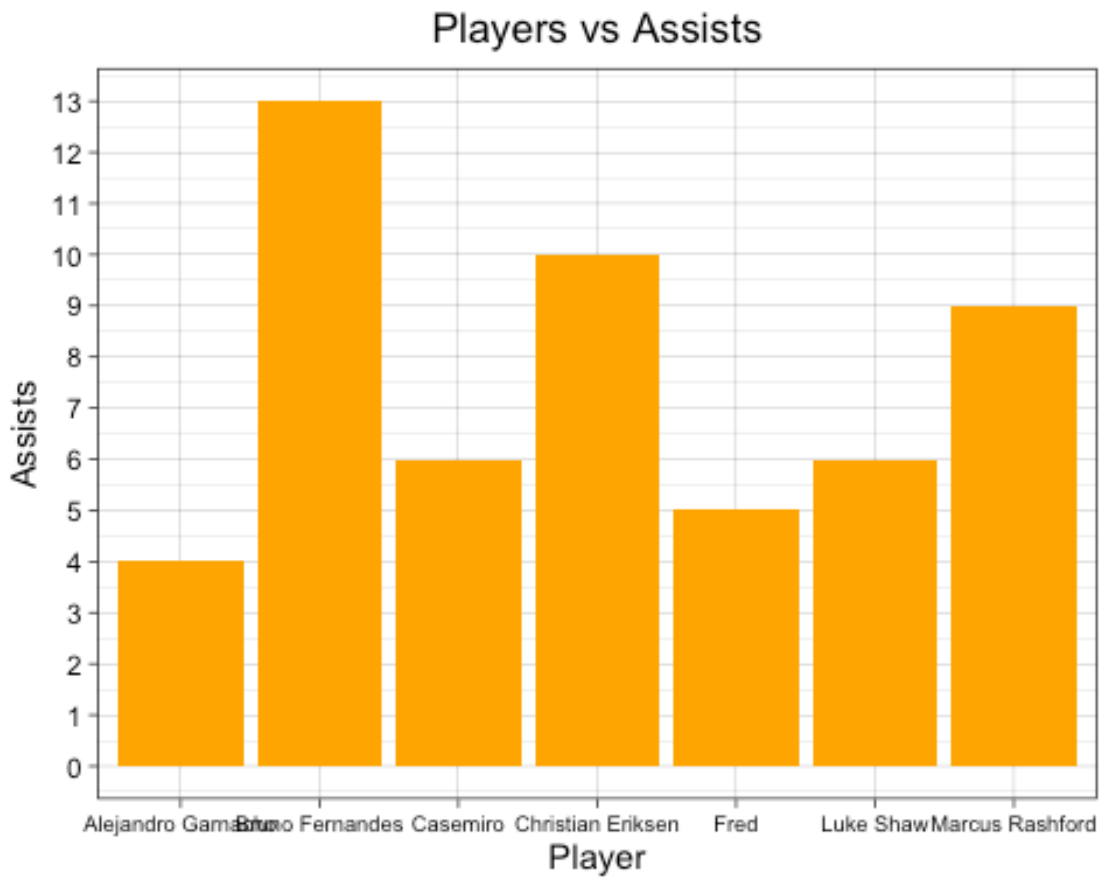
```
#Who has been scoring the goals?

# Filter the data and exclude players with less than 6 goals
manutd_scored = manutd_players %>%
  filter(Matches == Matches , Goals > 5)

ggplot(data=manutd_scored, mapping = aes(x=Player, y=Goals))+
  geom_col(fill = "orange")+
  labs(x = "Player", y = "Goals Scored", title = "Players vs Goals Scored")+
  theme_linedraw()+
  theme(plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(size = 7) # Adjust text size
  )+ylim(0,30)
```



During the 2022/23 campaign, Marcus Rashford (Forward) was the most prolific goalscorer across all competitions for Manchester United with 30 goals. Bruno Fernandes is second in that category, but is far behind with 14 goals; just below half of Rashford's goal tally.

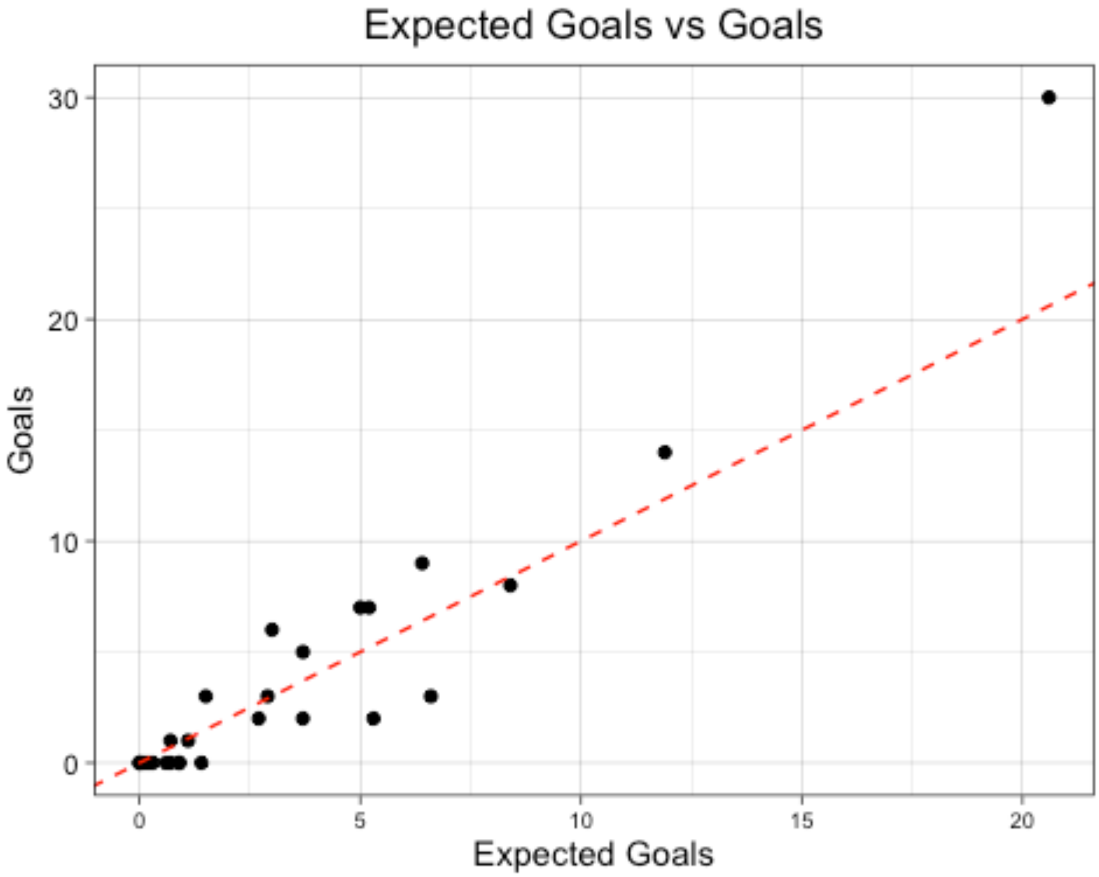
```
#Who has been creating goals?

# Filter the data and exclude players with less than 4 Assists
manutd_assist = manutd_players %>%
  filter(Matches == Matches , Assist >= 4)
```

```
ggplot(data=manutd_assist, mapping = aes(x=Player, y=Assist))+
  geom_col(fill = "orange")+
  labs(x = "Player", y = "Assists", title = "Players vs Assists")+
  theme_linedraw()+
  theme(plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(size = 7) # Adjust text size
  )+
  scale_y_continuous(breaks = seq(0, ceiling(max(manutd_assist$Assist)),
by=1))
```



Here, Bruno Fernandes (Midfielder) shines as he tops the assist charts for Manchester United which is a testament to his goal creation. Christian Eriksen (Midfielder) comes close in behind with 10 assists while Rashford comes 3rd best with 9.

```
#Has each player been clinical in front of goal?
ggplot(data=manutd_players, mapping = aes(x=Expected.Goals, y=Goals))+
  geom_point()+
  geom_abline(slope = 1, intercept = 0,       # y = x line
            color = "red", linetype = "dashed")+
  labs(x = "Expected Goals", y = "Goals", title = "Expected Goals vs Goals")+
  theme_linedraw()+
  theme(plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(size = 7)# Adjust text size
  )
```



Individually, Man Utd players are moderate in expected goals vs goals. In other words, there is not a consistent pattern indicating that players are scoring according to expected goals because around half of the goals are below the goals to expected goals ratio.

This plot identifies that Man Utd struggled to capitalise on their goal-scoring opportunities in the 22/23 season.

Data Manipulation

This dataset contains variables that are out of scope and domain knowledge of pattern of play. I have delete null irrelevant columns in Excel and renamed ‘Man Utd Statistics’ into ‘Man Utd Tidy’.

```
#Reading in the datasets
manutd.data = read.csv("ManUtdTidy.csv", header=TRUE)
manutd.data = manutd.data %>%
  select(-X, -X.1,-X.2,-X.3,-X.4,-X.5,-X.6,-X.7,-X.8,-X.9,-X.10,-X.11,-X.12,-
X.13,-X.14,-X.15,-X.16,-X.17,-X.18,-X.19,-X.20,-X.21,-X.22,-X.23,-X.24, -
Points, -Post.shot.Expected.Goals, -Crosses.Faced.By.Goalkeeper,-
```

```r
Crosses.Into.Penalty.Area, -Short.Passes.Completion.Percentage, -
Medium.Passes.Completion.Percentage, -Long.Passes.Completion.Percentage, -
Errors)

manutd.data = manutd.data[-62,]

#Result is binary so let a Win = 1, and not a win (D or L) = 0
manutd.data$Result[manutd.data$Result == "W"] = 1
manutd.data$Result[manutd.data$Result == "D"] = 0
manutd.data$Result[manutd.data$Result == "L"] = 0

#Possession inputs are in % terms, so let's convert to decimal form
manutd.data$Possession....=manutd.data$Possession..../100

#Same for venue: Let home = 1, and not home = 0
manutd.data$Venue[manutd.data$Venue == "Home"] = 1
manutd.data$Venue[manutd.data$Venue == "Away"] = 0
manutd.data$Venue[manutd.data$Venue == "Neutral"] = 0

manutd.data = as_tibble(manutd.data)
manutd.data$Result = as.numeric(manutd.data$Result)
manutd.data = data.frame(lapply(manutd.data, as.numeric))

manutd.data$Venue = as.factor(manutd.data$Venue)
manutd.data$Clean.Sheets = as.factor(manutd.data$Clean.Sheets)

#Replacing NA values with the mean of each respective column
manutd.data[] = lapply(manutd.data, function(x) {
  x[is.na(x)] = mean(x, na.rm = TRUE)
  return(x)
})
```

Checking the dimensions of the dataset

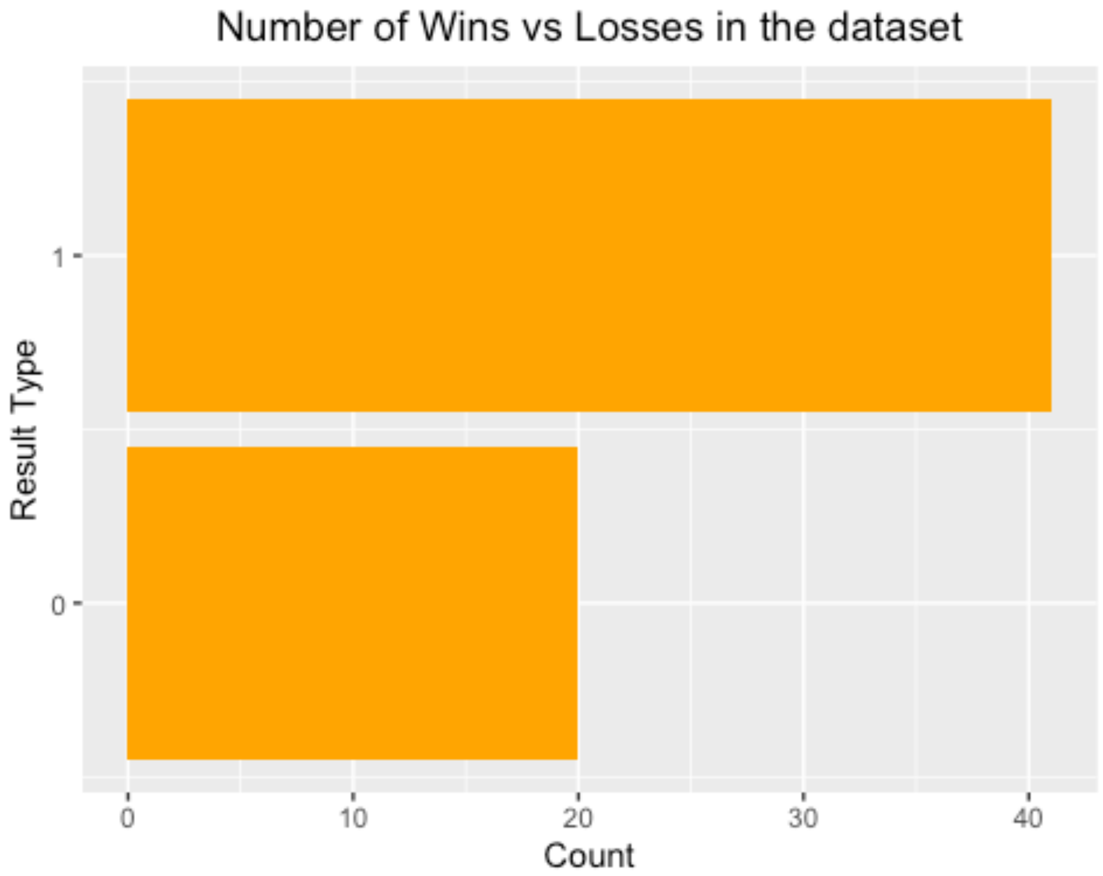
```r
print(dim(manutd.data))

## [1] 61 69
```

Initial look at the dataset

```r
#Initial look at the dataset
summary(manutd.data)

##  Venue      Result           Goals       Goals.Against       Shots
##  0:27   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   : 4.00
##  1:34   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:12.00
##         Median :1.0000   Median :2.000   Median :1.000   Median :16.00
##         Mean   :0.6721   Mean   :1.783   Mean   :1.017   Mean   :16.21
##         3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:19.00
##         Max.   :1.0000   Max.   :4.000   Max.   :7.000   Max.   :34.00
##  Shots.on.Target Average.Shot.Distance Shots.on.Target.Against
##  Min.   : 1.00   Min.   : 1.00         Min.   :0.000
##  1st Qu.: 4.00   1st Qu.:14.80         1st Qu.:2.000
##  Median : 6.00   Median :16.50         Median :3.000
##  Mean   : 5.82   Mean   :14.97         Mean   :3.475
##  3rd Qu.: 8.00   3rd Qu.:18.80         3rd Qu.:5.000
##  Max.   :13.00   Max.   :22.20         Max.   :9.000
##  Saves.By.Goalkeeper Clean.Sheets
##  Min.   :0.00        0:28
##  1st Qu.:1.00        1:33
##  Median :2.00
##  Mean   :2.18
##  3rd Qu.:3.00
##  Max.   :6.00
##  Passes.Completed..Longer.than.40.yards..By.Goalkeeper
##  Min.   : 0.000
##  1st Qu.: 2.000
##  Median : 4.000
##  Mean   : 4.246
##  3rd Qu.: 6.000
##  Max.   :15.000
##  Passes.Attempted..Longer.than.40.yards..By.Goalkeeper
##  Min.   : 0.00
##  1st Qu.: 8.00
##  Median :11.00
##  Mean   :19.62
##  3rd Qu.:22.00
##  Max.   :78.00
##  Passes.Attempted.By.Goalkeeper Average.Passes.Length.By.Goalkeeper
##  Min.   : 1.00                  Min.   : 0.00
##  1st Qu.:16.00                  1st Qu.:23.60
##  Median :25.00                  Median :30.20
##  Mean   :22.39                  Mean   :26.03
##  3rd Qu.:32.00                  3rd Qu.:34.70
##  Max.   :39.00                  Max.   :47.90
##  Goals.Kicks.Attempted.By.Goalkeeper
##  Min.   : 0.000
##  1st Qu.: 2.000
##  Median : 5.000
##  Mean   : 5.738
##  3rd Qu.: 9.000
##  Max.   :16.000
##  Goals.Kicks.Attempted..Longer.than.40.yards.
##  Min.   :  0.00
##  1st Qu.:  8.00
##  Median : 40.00
##  Mean   : 46.73
##  3rd Qu.: 87.50
##  Max.   :100.00
##  Average.Goal.Kicks.Length.By.Goalkeeper
##  Min.   : 6.00  
##  1st Qu.:19.00  
##  Median :36.30  
##  Mean   :38.56  
##  3rd Qu.:58.40  
##  Max.   :72.00  
##  Average.Distance.of.Defensive.Action.By.Goalkeeper Passes.Completed
##  Min.   :  3.00                                     Min.   : 78.4   
##  1st Qu.: 10.00                                     1st Qu.:389.0   
##  Median : 15.50                                     Median :440.1   
##  Mean   : 56.99                                     Mean   :440.1   
##  3rd Qu.: 19.00                                     3rd Qu.:512.0   
##  Max.   :535.00                                     Max.   :698.0   
##  Passes.Attemped  Passes.Completion.Percentage Total.Passing.Distance
##  Min.   :  77.2   Min.   :  64.3               Min.   :  189         
##  1st Qu.: 464.0   1st Qu.:  80.0               1st Qu.: 5949         
##  Median : 596.0   Median :  83.6               Median : 6820         
##  Mean   : 750.6   Mean   : 719.1               Mean   : 6820         
##  3rd Qu.: 750.6   3rd Qu.: 719.1               3rd Qu.: 8645         
##  Max.   :6837.0   Max.   :8197.0               Max.   :11752         
##  Progressive.Passing.Distance Short.Passes.Completed Short.Passes.Attemped
##  Min.   : 189                 Min.   : 85.5          Min.   : 82.7        
##  1st Qu.:2254                 1st Qu.:189.0          1st Qu.:189.0        
##  Median :2385                 Median :218.5          Median :229.8        
##  Mean   :2344                 Mean   :218.5          Mean   :229.8        
##  3rd Qu.:2842                 3rd Qu.:244.0          3rd Qu.:274.0        
##  Max.   :3595                 Max.   :367.0          Max.   :384.0        
##  Medium.Passes.Completed Medium.Passes.Attemped Long.Passes.Completed
##  Min.   : 54.0           Min.   : 35.0          Min.   :22.00        
##  1st Qu.:148.0           1st Qu.:159.0          1st Qu.:36.00        
##  Median :169.2           Median :182.1          Median :43.11        
##  Mean   :169.2           Mean   :182.1          Mean   :43.11        
##  3rd Qu.:195.0           3rd Qu.:220.0          3rd Qu.:51.00        
##  Max.   :272.0           Max.   :304.0          Max.   :68.00        
##  Long.Passes.Attemped Passes.Into.Final.Third Passes.Into.Penalty.Area
##  Min.   : 20.00       Min.   : 1.0            Min.   : 0.00           
##  1st Qu.: 60.00       1st Qu.:25.0            1st Qu.: 8.00           
##  Median : 67.97       Median :33.4            Median :11.00           
##  Mean   : 67.97       Mean   :33.4            Mean   :11.54           
##  3rd Qu.: 77.00       3rd Qu.:40.0            3rd Qu.:13.00           
##  Max.   :103.00       Max.   :91.0            Max.   :52.00           
##  Progressive.Passes    Tackles      Tackles.Won   
Tackles.in.Defensive.Third
##  Min.   : 5.00      Min.   : 5.0   Min.   : 3.0   Min.   : 0.00             
##  1st Qu.:33.00      1st Qu.:15.0   1st Qu.: 8.0   1st Qu.: 7.00             
##  Median :41.04      Median :16.7   Median :10.3   Median : 8.78             
##  Mean   :41.04      Mean   :16.7   Mean   :10.3   Mean   : 8.78             
##  3rd Qu.:50.00      3rd Qu.:19.0   3rd Qu.:12.0   3rd Qu.:11.00             
##  Max.   :99.00      Max.   :30.0   Max.   :21.0   Max.   :15.00             
##  Tackles.in.Middle.Third Tackles.in.Attacking.Third     Blocks
##  Min.   : 0.00           Min.   : 0.0              Min.   : 1.00  
##  1st Qu.: 4.00           1st Qu.: 1.0              1st Qu.: 9.00  
##  Median : 6.08           Median : 3.0              Median :12.38  
##  Mean   : 6.08           Mean   : 3.1              Mean   :12.38  
##  3rd Qu.: 8.00           3rd Qu.: 3.1              3rd Qu.:15.00  
##  Max.   :14.00           Max.   :15.0              Max.   :23.00  
##  Shots.Blocked    Shot.Against    Interception     Clearance    
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00  
##  1st Qu.: 2.00   1st Qu.: 8.00   1st Qu.: 7.00   1st Qu.:13.00  
##  Median : 4.64   Median :12.42   Median : 9.28   Median :18.42  
##  Mean   : 4.64   Mean   :12.42   Mean   : 9.28   Mean   :18.42  
##  3rd Qu.: 6.00   3rd Qu.:15.00   3rd Qu.:10.00   3rd Qu.:24.00  
##  Max.   :22.00   Max.   :38.00   Max.   :29.00   Max.   :54.00  
##  Possession....     Touches      Touches.in.Defensive.Penalty.Area
##  Min.   :0.300   Min.   : 63.0   Min.   : 19.00                   
##  1st Qu.:0.520   1st Qu.:569.0   1st Qu.: 59.00                   
##  Median :0.610   Median :619.0   Median : 75.00                   
##  Mean   :1.036   Mean   :598.9   Mean   : 86.44                   
##  3rd Qu.:1.036   3rd Qu.:708.0   3rd Qu.: 86.44                   
##  Max.   :8.020   Max.   :915.0   Max.   :269.00                   
##  Touches.in.Defensive.Third Touches.in.Middle.Third
Touches.in.Attacking.Third
##  Min.   :106.0              Min.   : 25.0           Min.   : 12.0             
##  1st Qu.:189.0              1st Qu.:221.0           1st Qu.:125.0             
##  Median :209.9              Median :256.7           Median :165.7             
##  Mean   :209.9              Mean   :256.7           Mean   :165.7             
##  3rd Qu.:229.0              3rd Qu.:302.0           3rd Qu.:205.0             
##  Max.   :419.0              Max.   :440.0           Max.   :404.0             
##  Touches.in.Attacking.Penalty.Area Take.on..Dribble..Attempted
##  Min.   : 6.30                     Min.   : 11.00             
##  1st Qu.:21.00                     1st Qu.: 16.00             
##  Median :27.35                     Median : 21.00             
##  Mean   :27.35                     Mean   : 33.97             
##  3rd Qu.:33.00                     3rd Qu.: 33.97             
##  Max.   :53.00                     Max.   :383.00             
##  Take.on..Dribble..Success.Percentage    Carries      
Total.Carries.Distance
##  Min.   :  14.3                       Min.   : 178.0   Min.   :  15         
##  1st Qu.:  40.6                       1st Qu.: 327.0   1st Qu.:1411         
##  Median :  52.4                       Median : 441.0   Median :1751         
##  Mean   : 137.5                       Mean   : 545.4   Mean   :1751         
##  3rd Qu.: 137.5                       3rd Qu.: 545.4   3rd Qu.:2089         
##  Max.   :1799.0                       Max.   :2971.0   Max.   :3042         
##  Progressive.Carries.Distance Progressive.Carries Carries.into.Final.Third
##  Min.   :  11.0               Min.   : 7.00       Min.   : 2.00           
##  1st Qu.: 728.0               1st Qu.:15.00       1st Qu.:10.00           
##  Median : 898.4               Median :19.04       Median :13.14           
##  Mean   : 898.4               Mean   :19.04       Mean   :13.14           
##  3rd Qu.:1103.0               3rd Qu.:23.00       3rd Qu.:16.00           
##  Max.   :1957.0               Max.   :44.00       Max.   :31.00           
##  Carries.Into.Penalty.Area  Miscontrols     Dispossesed    Yellow.Card
##  Min.   : 2.00             Min.   : 1.00   Min.   : 0.0   Min.   : 0.00
##  1st Qu.: 5.00             1st Qu.:11.00   1st Qu.: 6.0   1st Qu.: 1.00
##  Median : 7.00             Median :12.58   Median : 7.0   Median : 2.14
##  Mean   : 7.18             Mean   :12.58   Mean   : 7.1   Mean   : 2.14
##  3rd Qu.: 9.00             3rd Qu.:15.00   3rd Qu.: 8.0   3rd Qu.: 3.00
##  Max.   :15.00             Max.   :22.00   Max.   :16.0   Max.   :11.00
##    Red.Cards      Fouls.Commited    Fouls.Drawn      Crosses
Interception.1
##  Min.   : 0.00   Min.   : 3.00   Min.   : 3.0   Min.   : 4.00   Min.   :
1.0
##  1st Qu.: 0.00   1st Qu.: 9.00   1st Qu.: 7.0   1st Qu.: 9.00   1st Qu.:
7.0
##  Median : 0.00   Median :10.94   Median : 8.7   Median :13.88   Median
:10.0
##  Mean   : 1.26   Mean   :10.94   Mean   : 8.7   Mean   :13.88   Mean
:11.1
##  3rd Qu.: 1.26   3rd Qu.:13.00   3rd Qu.:10.0   3rd Qu.:18.00   3rd
Qu.:11.1
##  Max.   :13.00   Max.   :16.00   Max.   :16.0   Max.   :30.00   Max.
:63.0
##  Tackles.Won.1   Ball.Recovery  Aerial.Duel.Won Aerial.Duel.Lost
##  Min.   : 4.00   Min.   : 6.0   Min.   : 4.00   Min.   : 3.00
##  1st Qu.: 9.00   1st Qu.:51.0   1st Qu.:11.00   1st Qu.: 9.00
##  Median :12.00   Median :52.0   Median :12.71   Median :11.32
##  Mean   :14.72   Mean   :51.2   Mean   :12.71   Mean   :11.32
##  3rd Qu.:14.72   3rd Qu.:59.0   3rd Qu.:14.00   3rd Qu.:13.00
##  Max.   :91.00   Max.   :74.0   Max.   :26.00   Max.   :23.00
```

Now, is the dataset balanced or not?

```
ggplot(data = manutd.data, mapping = aes(y=Result))+
  geom_bar(fill="orange")+
  labs(x="Count", y="Result Type", title="Number of Wins vs Losses in the
dataset")+
  scale_y_continuous(breaks = c(0, 1))+
    theme(plot.title = element_text(hjust = 0.5))
```



As per the plot above, there appears to be more observations for a Man Utd win than a Man Utd loss.

Model Building and analysis

Since the desired outcome is to predict a win or loss and the fact that Result is binary, a logistic model seems an appropiate fit to the data.

Fitting the Logistic Regression model

```
#Fitting the model with all possible variables.
#Included weights to the data as there is an imbalance in the dataset
model.fit = glm(Result~.,data=manutd.data, family = "binomial"(link =
logit),weights = ifelse(manutd.data$Result == 0, 20, 1))
summary(model.fit) #A bit dreadful at the moment

##
## Call:
## glm(formula = Result ~ ., family = binomial(link = logit), data =
manutd.data,
##     weights = ifelse(manutd.data$Result == 0, 20, 1))
##
```

```
## Coefficients: (8 not defined because of singularities)
##                                                         Estimate Std.
Error
## (Intercept)                                           -1.177e+02
5.139e+06
## Venue1                                                 2.611e+01
6.903e+05
## Goals                                                  7.801e+01
7.521e+05
## Goals.Against                                         -5.059e+01
6.082e+05
## Shots                                                 -4.046e+00
5.516e+04
## Shots.on.Target                                       -8.135e-01
2.011e+05
## Average.Shot.Distance                                 -5.583e+01
6.991e+05
## Shots.on.Target.Against                                5.414e+01
6.655e+05
## Saves.By.Goalkeeper                                   -6.624e+01
9.936e+05
## Clean.Sheets1                                         -4.269e+02
6.130e+06
## Passes.Completed..Longer.than.40.yards..By.Goalkeeper -1.484e+00
7.431e+04
## Passes.Attempted..Longer.than.40.yards..By.Goalkeeper -1.892e+00
3.503e+04
## Passes.Attempted.By.Goalkeeper                         1.193e+01
1.805e+05
## Average.Passes.Length.By.Goalkeeper                    2.652e+00
1.069e+05
## Goals.Kicks.Attempted.By.Goalkeeper                   -1.833e+00
1.329e+05
## Goals.Kicks.Attempted..Longer.than.40.yards.           1.865e+00
8.249e+04
## Average.Goal.Kicks.Length.By.Goalkeeper                7.184e+00
7.840e+04
## Average.Distance.of.Defensive.Action.By.Goalkeeper    -9.194e-01
6.066e+04
## Passes.Completed                                      -2.641e+01
5.806e+05
## Passes.Attemped                                        1.673e-01
6.538e+03
## Passes.Completion.Percentage                          -1.126e+00
1.507e+04
## Total.Passing.Distance                                 1.195e+00
1.590e+04
## Progressive.Passing.Distance                          -1.650e-02
2.280e+03
## Short.Passes.Completed                                 4.986e+01
```

```
9.501e+05
## Short.Passes.Attemped                                 -1.642e+01
2.832e+05
## Medium.Passes.Completed                               -3.684e+00
2.930e+05
## Medium.Passes.Attemped                                 9.836e+00
1.084e+05
## Long.Passes.Completed                                  2.538e+01
6.056e+05
## Long.Passes.Attemped                                  -3.704e+00
9.614e+04
## Passes.Into.Final.Third                               -6.965e+01
1.006e+06
## Passes.Into.Penalty.Area                              -6.781e+01
1.094e+06
## Progressive.Passes                                     3.341e+01
4.977e+05
## Tackles                                               -1.476e+02
1.864e+06
## Tackles.Won                                           -1.442e+01
3.793e+05
## Tackles.in.Defensive.Third                             1.567e+02
2.122e+06
## Tackles.in.Middle.Third                                1.781e+02
2.205e+06
## Tackles.in.Attacking.Third                             1.412e+02
1.899e+06
## Blocks                                                 2.512e+01
4.347e+05
## Shots.Blocked                                         -1.963e+01
3.501e+05
## Shot.Against                                           9.658e+00
8.393e+04
## Interception                                           8.632e+00
1.834e+05
## Clearance                                              2.952e+01
3.732e+05
## Possession....                                         3.189e+03
4.543e+07
## Touches                                               -1.459e+01
2.127e+05
## Touches.in.Defensive.Penalty.Area                     -8.682e+00
1.415e+05
## Touches.in.Defensive.Third                             3.720e+00
1.015e+05
## Touches.in.Middle.Third                                4.141e+00
9.321e+04
## Touches.in.Attacking.Third                             1.481e+01
2.353e+05
## Touches.in.Attacking.Penalty.Area                      1.848e+01
```

```
3.341e+05
## Take.on..Dribble..Attempted                            3.599e+01
5.423e+05
## Take.on..Dribble..Success.Percentage                  -1.542e+00
3.200e+04
## Carries                                               -5.488e+00
8.983e+04
## Total.Carries.Distance                                -1.214e-01
5.293e+03
## Progressive.Carries.Distance                           9.256e-01
2.085e+04
## Progressive.Carries                                   -5.270e-01
1.248e+05
## Carries.into.Final.Third                              -4.511e+01
6.941e+05
## Carries.Into.Penalty.Area                             -5.571e+01
9.023e+05
## Miscontrols                                           -6.245e+00
2.401e+05
## Dispossesed                                            9.251e+00
2.012e+05
## Yellow.Card                                            8.056e+01
9.534e+05
## Red.Cards                                             -2.594e+02
3.740e+06
## Fouls.Commited                                                NA
NA
## Fouls.Drawn                                                   NA
NA
## Crosses                                                       NA
NA
## Interception.1                                                NA
NA
## Tackles.Won.1                                                 NA
NA
## Ball.Recovery                                                 NA
NA
## Aerial.Duel.Won                                               NA
NA
## Aerial.Duel.Lost                                              NA
NA
##                                                       z value Pr(>|z|)
## (Intercept)                                                 0        1
## Venue1                                                      0        1
## Goals                                                       0        1
## Goals.Against                                               0        1
## Shots                                                       0        1
## Shots.on.Target                                             0        1
## Average.Shot.Distance                                       0        1
## Shots.on.Target.Against                                     0        1
## Saves.By.Goalkeeper                                         0        1
## Clean.Sheets1                                               0        1
## Passes.Completed..Longer.than.40.yards..By.Goalkeeper       0        1
## Passes.Attempted..Longer.than.40.yards..By.Goalkeeper       0        1
## Passes.Attempted.By.Goalkeeper                              0        1
## Average.Passes.Length.By.Goalkeeper                         0        1
## Goals.Kicks.Attempted.By.Goalkeeper                         0        1
## Goals.Kicks.Attempted..Longer.than.40.yards.                0        1
## Average.Goal.Kicks.Length.By.Goalkeeper                     0        1
## Average.Distance.of.Defensive.Action.By.Goalkeeper          0        1
## Passes.Completed                                            0        1
## Passes.Attemped                                             0        1
## Passes.Completion.Percentage                                0        1
## Total.Passing.Distance                                      0        1
## Progressive.Passing.Distance                                0        1
## Short.Passes.Completed                                      0        1
## Short.Passes.Attemped                                       0        1
## Medium.Passes.Completed                                     0        1
## Medium.Passes.Attemped                                      0        1
## Long.Passes.Completed                                       0        1
## Long.Passes.Attemped                                        0        1
## Passes.Into.Final.Third                                     0        1
## Passes.Into.Penalty.Area                                    0        1
## Progressive.Passes                                          0        1
## Tackles                                                     0        1
## Tackles.Won                                                 0        1
## Tackles.in.Defensive.Third                                  0        1
## Tackles.in.Middle.Third                                     0        1
## Tackles.in.Attacking.Third                                  0        1
## Blocks                                                      0        1
## Shots.Blocked                                               0        1
## Shot.Against                                                0        1
## Interception                                                0        1
## Clearance                                                   0        1
## Possession....                                              0        1
## Touches                                                     0        1
## Touches.in.Defensive.Penalty.Area                           0        1
## Touches.in.Defensive.Third                                  0        1
## Touches.in.Middle.Third                                     0        1
## Touches.in.Attacking.Third                                  0        1
## Touches.in.Attacking.Penalty.Area                           0        1
## Take.on..Dribble..Attempted                                 0        1
## Take.on..Dribble..Success.Percentage                        0        1
## Carries                                                     0        1
## Total.Carries.Distance                                      0        1
## Progressive.Carries.Distance                                0        1
## Progressive.Carries                                         0        1
## Carries.into.Final.Third                                    0        1
## Carries.Into.Penalty.Area                                   0        1
## Miscontrols                                                 0        1
## Dispossesed                                                 0        1
## Yellow.Card                                                 0        1
## Red.Cards                                                   0        1
## Fouls.Commited                                             NA       NA
## Fouls.Drawn                                                NA       NA
## Crosses                                                    NA       NA
## Interception.1                                             NA       NA
## Tackles.Won.1                                              NA       NA
## Ball.Recovery                                              NA       NA
## Aerial.Duel.Won                                            NA       NA
## Aerial.Duel.Lost                                           NA       NA
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2.7285e+02  on 60  degrees of freedom
## Residual deviance: 4.9881e-10  on  0  degrees of freedom
## AIC: 122
## 
## Number of Fisher Scoring iterations: 25
```

Trying Backwards Stepwise Regression

```
backward_model = step(model.fit, direction = "backward", trace = 1)
```

Inspecting the Stepwise model plus an interaction between Shot.Against and Blocks (Based on AIC)

```
model = glm(Result ~ Venue + Clean.Sheets+Goals +
Passes.Attempted.By.Goalkeeper +
    Medium.Passes.Attemped + Tackles + Tackles.in.Defensive.Third +
    Tackles.in.Middle.Third + Blocks + Shot.Against + Touches
+Shot.Against*Blocks, family = "binomial"(link = logit),
    data = manutd.data)

vif(model, type="predictor")
```

```
##                           Venue                   Clean.Sheets
##                        9.149248                      22.735538
##                           Goals Passes.Attempted.By.Goalkeeper
##                       15.590206                      17.181876
##          Medium.Passes.Attemped                        Tackles
##                       62.704391                      86.798798
##      Tackles.in.Defensive.Third        Tackles.in.Middle.Third
##                       66.070992                      82.694777
##                          Blocks                   Shot.Against
##                       52.664070                      43.000093
##                         Touches            Blocks:Shot.Against
##                       47.392682                      83.462955
```

High Multicollinearity in the model.

Checking for perfect seperation

```
regressors = manutd.data %>%
  select(Goals , Passes.Attempted.By.Goalkeeper ,
    Medium.Passes.Attemped ,Tackles , Tackles.in.Defensive.Third ,
    Tackles.in.Middle.Third , Blocks , Shot.Against ,Touches
,Shot.Against:Blocks)


seperation.result = detect_separation(y = manutd.data$Result, x =regressors
, family = binomial())
print(seperation.result)
```

```
## Implementation: ROI | Solver: lpsolve
## Separation: FALSE
## Existence of maximum likelihood estimates
##                           Goals Passes.Attempted.By.Goalkeeper
##                               0                              0
##          Medium.Passes.Attemped                        Tackles
##                               0                              0
##      Tackles.in.Defensive.Third        Tackles.in.Middle.Third
##                               0                              0
##                          Blocks                  Shot.Against
##                               0                              0
##                         Touches                 Shots.Blocked
##                               0                              0
## 0: finite value, Inf: infinity, -Inf: -infinity
```

```
table(manutd.data$Venue, manutd.data$Result) #No perfect seperation between
Venue and Result
```

```
##
##      0  1
##   0 14 13
##   1  6 28
```

```
table(manutd.data$Clean.Sheets, manutd.data$Result) #No perfect seperation
between Clean sheet and Result
```

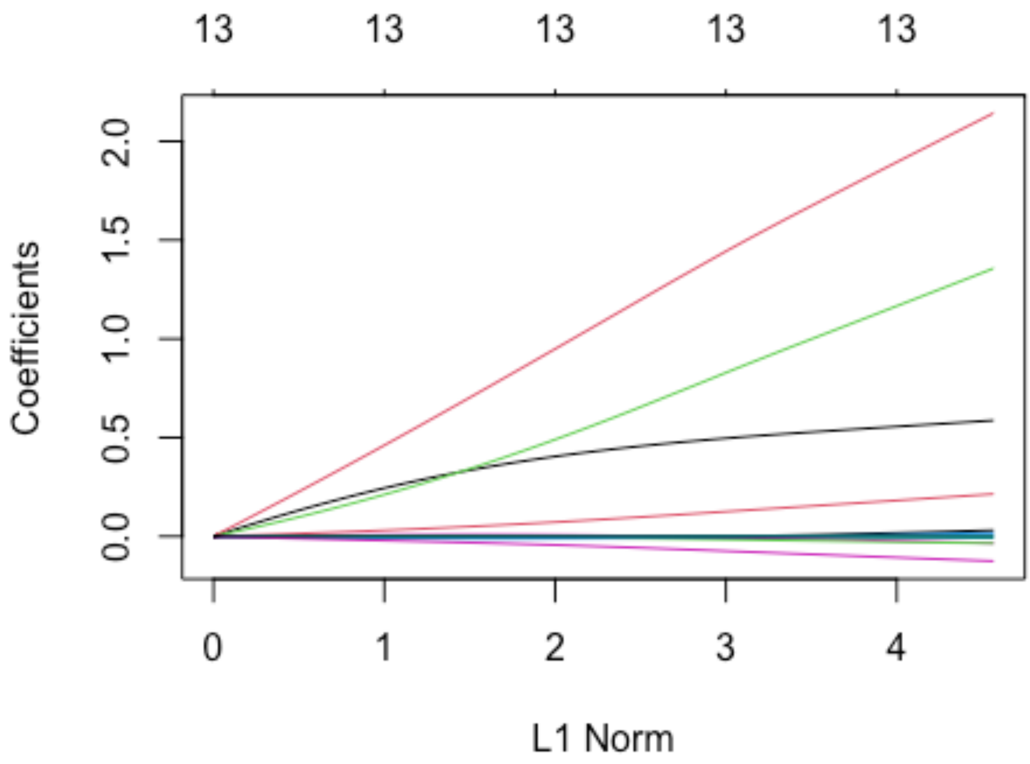
```
##
##      0  1
##   0 17 11
##   1  3 30
```

There is no perfect seperation within the model.

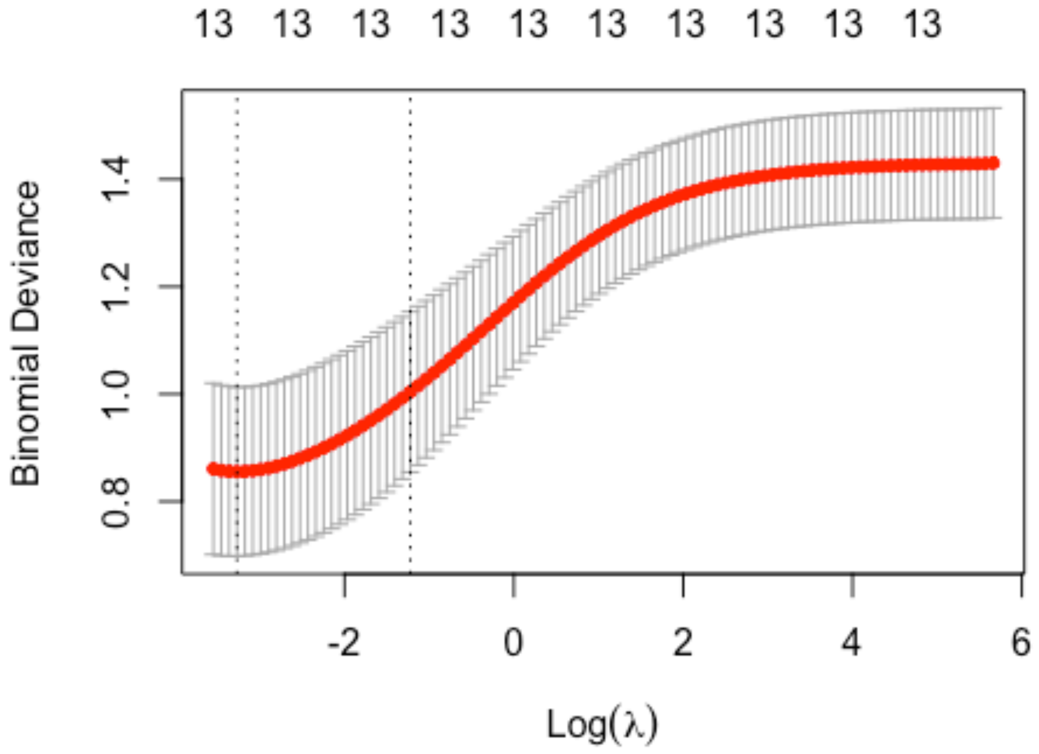
Performing a Ridge logistic regression to reduce the effect of multicollinearity

```
#Creating new Dataset for Ridge Regression, only including variables from
backwards stepwise regresison
ridge_dataset = manutd.data %>%
```

```r
  select(Venue, Clean.Sheets,Goals, Passes.Attempted.By.Goalkeeper ,
    Medium.Passes.Attemped , Tackles , Tackles.in.Defensive.Third ,
    Tackles.in.Middle.Third , Blocks , Shot.Against , Touches, Result,
Shot.Against:Blocks)

ridge_dataset = ridge_dataset %>%
  mutate(Shot.Against.Blocks =
ridge_dataset$Shot.Against*ridge_dataset$Blocks)

# Dataset with only predictors
ridge_regressors = ridge_dataset %>%
  select(-Result)

#Creating a matrix for regressors
x = as.matrix(ridge_regressors) # Predictors (all columns except the
response)
y = ridge_dataset$Result

#Defining weights for imbalanced data
weights = ifelse(y == 0, length(y) / sum(y == 0), 1)  # Inverse class
frequencies


#Fitting Ridge model
ridge_model = glmnet(x, y, family="binomial",alpha = 0,weights=weights)
plot(ridge_model)    # Draw plot of coefficients
```



```r
#performing LOOCV for the Ridge model as dataset is relatively small
cv_loocv = cv.glmnet(x, y, alpha = 0, family = "binomial", nfolds = 5,
weight= weights)
best_lambda = cv_loocv$lambda.min

ridge_final = glmnet(x, y, family = "binomial", alpha = 0,
weight=weights,lambda = best_lambda)


#Plot of the lambdas
plot(cv_loocv)
```



```
#Coefficients of Ridge logistic model
coef(ridge_final, s = "lambda.min")

## 14 x 1 sparse Matrix of class "dgCMatrix"
##                                        s1
## (Intercept)                  -3.008000572
## Venue                         0.563134764
## Clean.Sheets                  1.954872459
## Goals                         1.212004716
## Passes.Attempted.By.Goalkeeper  0.011617989
## Medium.Passes.Attemped       -0.007036086
## Tackles                      -0.025515798
## Tackles.in.Defensive.Third    0.020525667
## Tackles.in.Middle.Third       0.189536346
## Blocks                       -0.030095005
## Shot.Against                  0.007207645
## Touches                       0.001632173
## Shots.Blocked                -0.111691633
## Shot.Against.Blocks          -0.001823470
```

Method and Assumption checks

Null Model (Baseline Model) for Comparision

```
# Fit null logistic model
null_model = glm(manutd.data$Result ~ 1, family = "binomial",
data=manutd.data)
summary(null_model)
```

```
##
## Call:
## glm(formula = manutd.data$Result ~ 1, family = "binomial", data =
manutd.data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.7178     0.2727   2.632  0.00849 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 77.184  on 60  degrees of freedom
## Residual deviance: 77.184  on 60  degrees of freedom
## AIC: 79.184
##
## Number of Fisher Scoring iterations: 4
```

The baseline model predicts a constant probability for all observations. Note, that the p-value for $\beta_0 < 0.05$ which rejects the null hypothesis that the intercept equals 0.

Extracting the intercept value

```
# Extract the intercept
beta_0 = coef(null_model)[1]

# Compute constant predicted probability
p_hat = 1 / (1 + exp(-beta_0)) #Inverse of logit

# Predicted classes based on threshold
predicted_class = ifelse(p_hat >= 0.5, 1, 0)

#Calculating predicted values of the logistic model with ridge penalisation
predictions = predict(ridge_final,newx = x, s=best_lambda, type="response")

#Transforming predictions into binary outcomes
predicted_classes = ifelse(predictions >= 0.5, 1,0)
```

Confusion Matrix

```
# Confusion matrix
Confusion.matrix = table(Predicted = predicted_classes, Actual = y)
print(Confusion.matrix)
```

```
##          Actual
## Predicted  0  1
##         0 19  4
##         1  1 37
```

```
#Estimated specificity, sensitivity, and prediction error
specificity = 19/(19+4); sensitivity = 37/(37+1)
print(specificity); print(sensitivity) #Specificity and sensitivity
respectively
```

```
## [1] 0.826087
```

```
## [1] 0.9736842
```

```
prediction.error = (1+4)/(19+4+1+37)
print(prediction.error)
```

```
## [1] 0.08196721
```

From the Confusion matrix, the probability that the model predicts a win, given the Result is a win is 0.97. Moreover, the probability that the model predicts a loss, given the Result is a loss is 0.83.

Calculating Accuracy, Precision, Recall, and F1-Score of the model from the Confusion Matrix

```
accuracy = (19+37)/(19+4+37+1) #proportion of correct predictions (both true
positives and true negatives) out of all predictions made
precision = 19/(19+1) #accuracy of the positive predictions made by the model
recall = 37/(37+4) #how well the model identifies positive instances

#Computing the F1-score since dataset is imbalanced
f1.score = 2*(precision*recall)/(precision+recall) #the harmonic mean of
precision and recall

print(precision)
```

```
## [1] 0.95
```

A precision of 95% indicates that the model’s predictions of a win matched the actual values 95% of the time

```
print(recall)
```

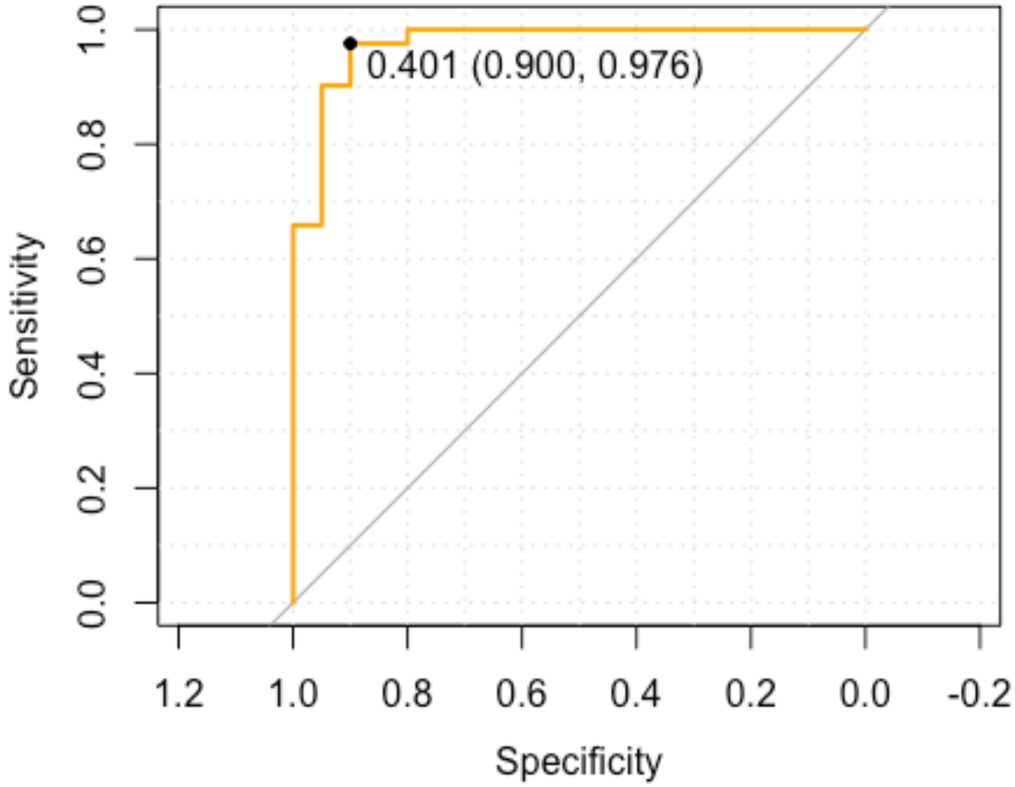
```
## [1] 0.902439
```

A recall value of 0.90 indicates the model is strong in identifying whenever Man Utd would win

```
print(f1.score)

## [1] 0.925609
```

F1-score indicates the model achieves a good balance between identifying most wins (recall) and not over-predicting (precision) Results.

```
# Estimate of AUC
roc_curve = roc(y, predictions) # ROC curve
plot(roc_curve, grid=TRUE, col="orange",print.thres = "best") #Plot ROC curve
alongside the point that maximises both sensitivity and specificity
```



Area under ROC curve

```
auc(y, predictions)

## [1] 0.9756098
```

Predictors are excellent in the model in predicting Man Utd's Result for a imbalanced and moderate sized dataset

Hosmer and Lemeshow GOF test

```
hoslem.test(y, predict(ridge_final, newx = x, type = "response"))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  y, predict(ridge_final, newx = x, type = "response")
## X-squared = 12.45, df = 8, p-value = 0.1322
```

P-value > 0.05 which indicates a failure to reject the null hypothesis that there is a lack of fit of the model to the data.

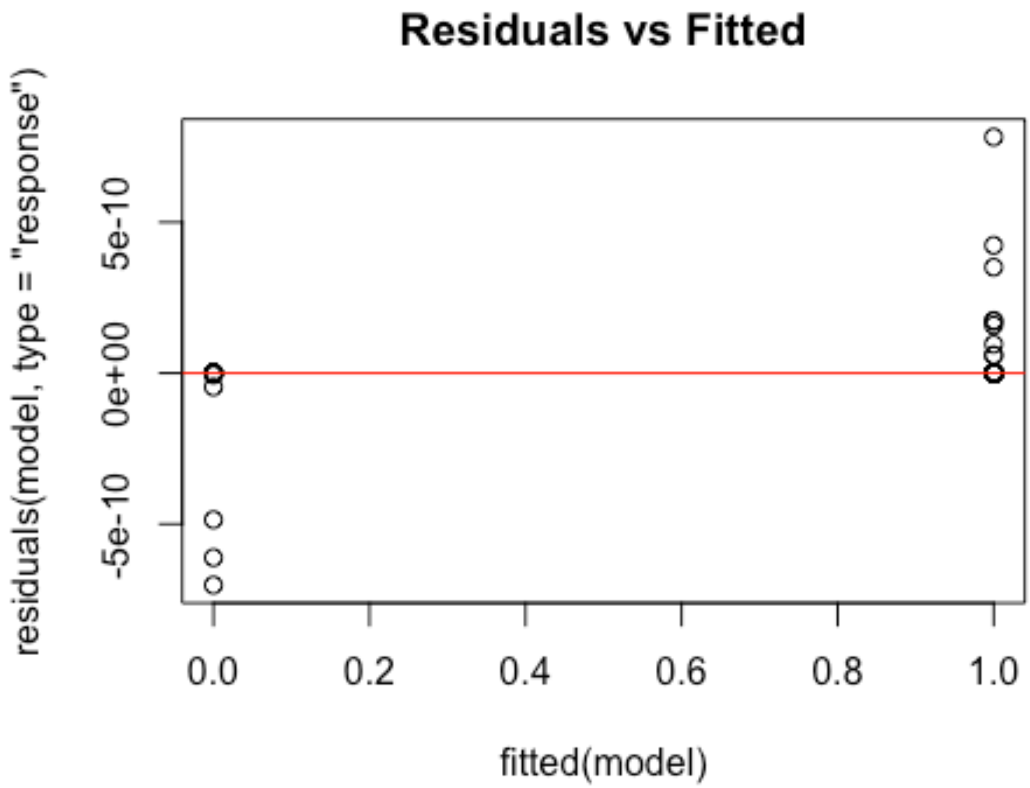
Residual vs Fitted values plot of Logistic model without Ridge penalisation

```
residuals.deviance = residuals(model, type = "deviance")
print(mean(residuals.deviance)) #Mean of deviance residuals close to 0

## [1] 1.102836e-06

residuals.pearson = residuals(model, type="pearson")
print(mean(residuals.pearson)) #Mean of pearson residuals close to 0

## [1] 7.798223e-07
```

Means of both Deviance and Pearson residuals are close to 0, indicating a good fit.

Residual vs Fitted values plot of Logistic model without Ridge penalisation

```
plot(fitted(model), residuals(model, type="response"), main = "Residuals vs
Fitted")+
  abline(h = 0, col = "red")
```



```
## integer(0)
```

Bands of residuals are forming above and below 0 at the tails of residuals = 0. Some residual observations hover around 0, which could indicate a possible overfit.

Deviance of the baseline model

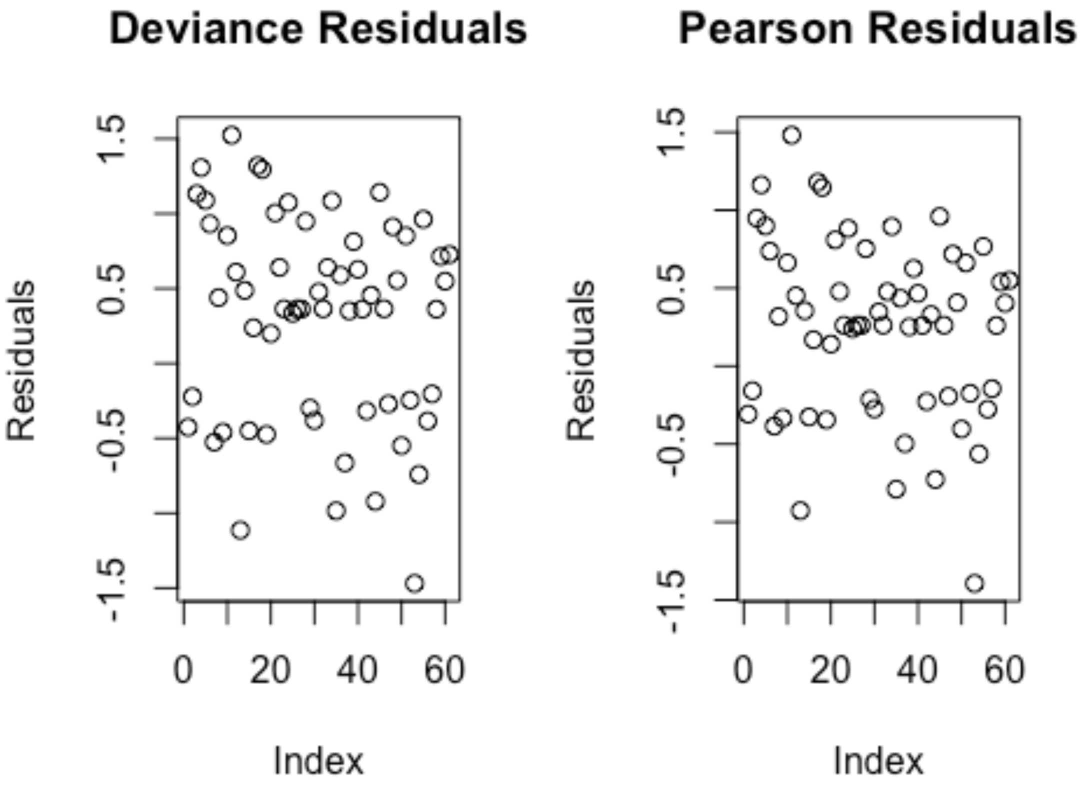
```
deviance(model)

## [1] 8.256683e-09

1 - pchisq(deviance(model), 48)

## [1] 1
```

Pretty small deviance which indicate the number of predictors close to saturation. Large p-value under chi-square distribution indicates a failure to reject the null hypothesis that the model is correct.

Checking the Deviance and Pearson residuals of the final model (Ridge)

```
# Deviance residuals
deviance_residuals = sign(y - predictions) *
  sqrt(-2 * (y * log(predictions) + (1 - y) * log(1 - predictions)))
```

```r
summary(deviance_residuals) #Residuals are between -2 and 2

##        s1         
##  Min.   :-1.4692  
##  1st Qu.:-0.2984  
##  Median : 0.3644  
##  Mean   : 0.3004  
##  3rd Qu.: 0.8521  
##  Max.   : 1.5238  

# Pearson residuals
pearson_residuals = (y - predictions) /
  sqrt(predictions * (1 - predictions))


# Plot residuals of ridge penalised model
par(mfrow = c(1, 2))
plot(deviance_residuals, main = "Deviance Residuals", ylab = "Residuals",
xlab = "Index")+
plot(pearson_residuals, main = "Pearson Residuals", ylab = "Residuals", xlab
= "Index")
```



```
## integer(0)
```

Residuals for the case of both deviance and Pearson scatter around zero, with no noticeable trend or pattern. Moreover, residuals are between -2 and 2, indicating no outliers. As the respective index increases, variability in the residuals does not appear to increase.

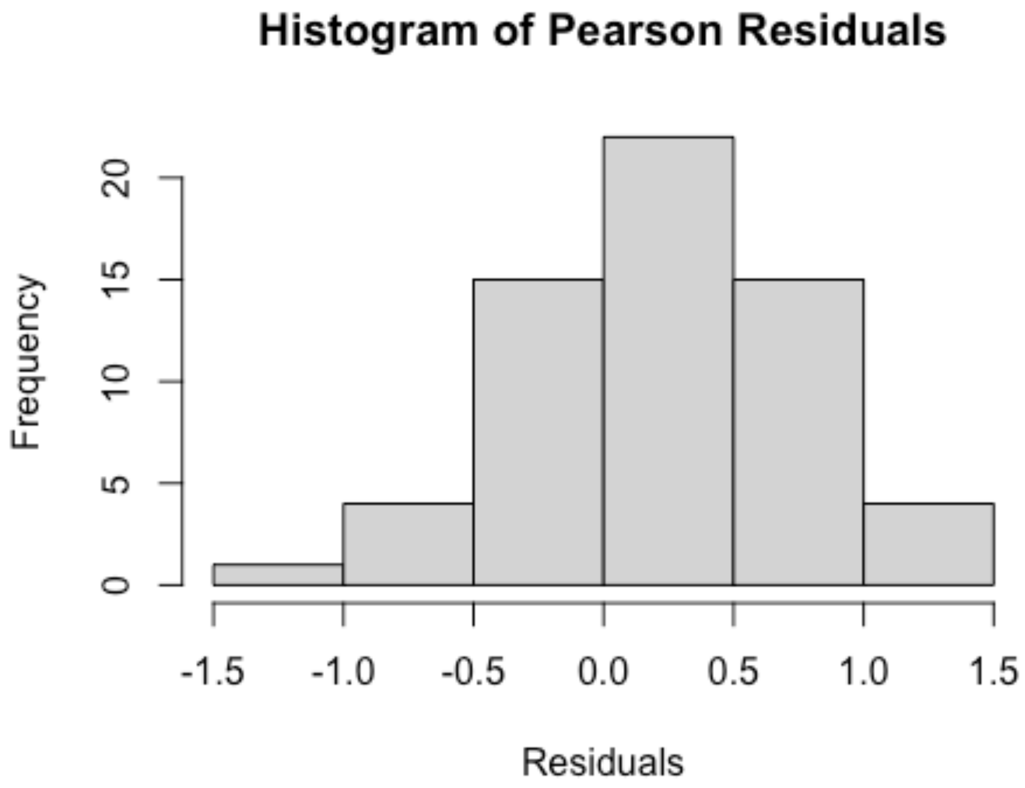
Log-loss of the baseline model vs the fitted Ridge logistic model

```
# Calculate log-loss of fitted model
log_loss.model = logLoss(actual = y,predicted = predictions)

# Print log-loss
print(log_loss.model)

## [1] 0.2804768

# Calculate log-loss of logistic null model (baseline model)
log_loss_baseline = logLoss(manutd.data$Result, predicted = p_hat)
print(log_loss_baseline)

## [1] 0.6326591
```

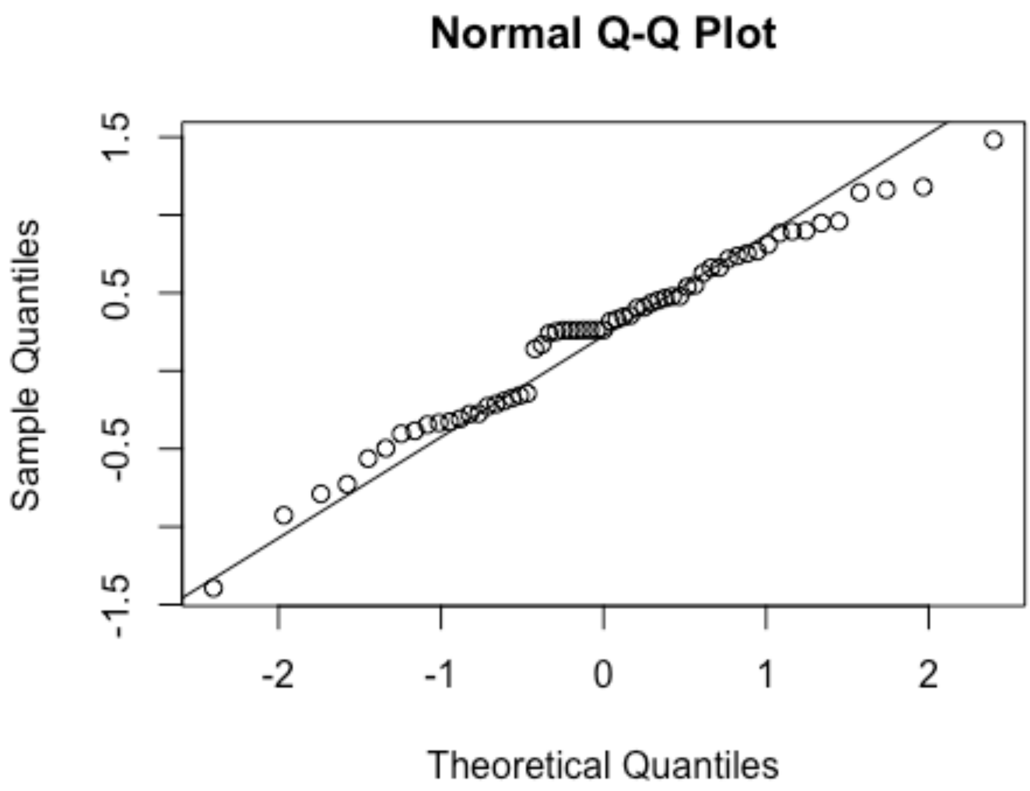
The log-Loss of the fitted model is less than the log-Loss of the baseline logistic model, indicating that the ridge model is a better predictor than the baseline model and regularisation via Ridge improved the model's performance.

Histogram of residuals

```
hist(pearson_residuals, main = "Histogram of Pearson Residuals", xlab =
"Residuals")
```



```
qqnorm(pearson_residuals); qqline(pearson_residuals)
```



Residuals are centered around 0.0 - 0.5 with an approximate normal distribution. Q-Q plot is almost on the 45 degree line understandibly as the dataset is quite small

Interpretation

Coefficients of the model

```
#Coefficients of the model
coefs = coef(ridge_final)
print(coefs)

## 14 x 1 sparse Matrix of class "dgCMatrix"
##                                          s0
## (Intercept)                    -3.008000572
## Venue                           0.563134764
## Clean.Sheets                    1.954872459
## Goals                           1.212004716
## Passes.Attempted.By.Goalkeeper  0.011617989
## Medium.Passes.Attemped         -0.007036086
## Tackles                        -0.025515798
## Tackles.in.Defensive.Third      0.020525667
## Tackles.in.Middle.Third         0.189536346
## Blocks                              -0.030095005
## Shot.Against                         0.007207645
## Touches                              0.001632173
## Shots.Blocked                       -0.111691633
## Shot.Against.Blocks                 -0.001823470
```

Model

The equation for the logistic regression model is:

$$\begin{aligned}\text{logit}(p_i) = {} & -3.32 + 0.59\text{Venue} + 2.14\text{CleanSheet} + 1.36\text{Goals} + 0.02\text{GKpasses} \\ & - 0.03\text{Tackles} + 0.03\text{TacklesDefensiveThird} + 0.21\text{TacklesMiddleThird} \\ & - 0.04\text{Blocks} - 0.13\text{ShotsBlocked}\end{aligned}$$

Explanation:

- p_i is the predicted probability for observation i.
- $\text{logit}(p_i)$ is the log-odds of the predicted probability.
- Each coefficient corresponds to the associated predictor variable (e.g., Venue,Clean Sheet,Goals, etc.).

The coefficients indicate the impact of each predictor on the log-odds of the a Man Utd win.

The loss function for logistic regression with ridge penalisation, where $\lambda = 0.03$, is:

$$\mathcal{L}(\beta) = -\frac{1}{62}\sum_{i=1}^{62}[\text{Result}_i\log\hat{p}_i + (1-\text{Result}_i)\log(1-\hat{p}_i)] + \frac{0.03}{2}\sum_{j=1}^{101}\beta_j^2$$

Key Predictors

The intercept value indicates that, without any influence from the predictors, the odds of the outcome (likely a win, goal, or other event) are less than 1. negative. Moreover, whenever Man Utd keeps a clean sheet in a match, the log-odds of them Winning increases. Similarly, the more goals Man Utd scores in a match, the greater the log-odds of them Winning. Moreover, if Man Utd is at home, then the log-odds of them Winning increases.

Defensive Actions

Tackles in Defensive Third and Tackles in Middle Third are positively associated with the log-odds of a Win. This suggests that successful tackles, especially in the middle third and defensive zones, are beneficial. The fact that tackles in the middle third have a relatively larger coefficient indicates that they may be more impactful than tackles in other areas, potentially because they break up opposition attacks before they reach the defensive zone.

Blocks and Shots Blocked have negative coefficients, implying that blocking shots and making defensive interventions might not always be favorable for the outcome. It could be due to the nature of these defensive actions (e.g., blocking shots might indicate a defensive team under pressure, or blocked shots could lead to dangerous rebounds).

Tackling in the defensive third is slightly positive, indicating that strong defense in the team's own half is beneficial. Tackling in the middle third is more strongly positive, suggesting that controlling the midfield and breaking up opposition play in the center of the field is particularly important for the outcome.

Shots Blocked and the interaction between Shot Against and Blocks are negatively correlated with the outcome, suggesting that increased defensive actions like blocking shots are associated with a lower likelihood of a positive outcome. This could be due to increased pressure on defense, possibly indicating that the team is underperforming in a more vulnerable position.

A team that maintains clean sheets, scores more goals, and makes tackles in the middle third is likely to be more successful, according to the model. This suggests a playing style focused on defensive stability (clean sheets, tackles) and attacking efficiency (goals, passes).

Actionable Insights for Improvement

Improvement through offense: The model suggests focusing on improving goal-scoring ability and maintaining clean sheets as these variables have large positive coefficients. Potentially, improving current attacking positions' goal output or buying a goalscorer (Striker or Winger or Attacking Midfielder) with consistent goal output would improve the log-odds of Man Utd winning matches.

Improvement through defense: Tackling in the middle third has a relatively strong positive effect, so improving midfield defense could also be beneficial. Conversely, reducing shots blocked might also be something to focus on, as these are negatively correlated with the outcome. Recommended that Man Utd improve defensively through a shot-stopping-focused goalkeeper (Keeping De Gea) and/or buying ball-winning midfielders as tackling in the middle third is strongly positive and could be interacting with the negatively correlated 'Shots-blocked' and keeping a clean-sheet.

Reflection

I am proud of being able to apply statistical knowledge to Manchester United, and provide some insight on how they could improve based on their 22/23 statistics. Specifically, I am proud of being able to handle multicollinearity through adding a Ridge penalisation; a concept that I have never known previously. Moreover, I am proud of utilising Leave One Out Cross-Validation when assessing my model as I had also never known this method of cross-validation previously. I had only known cross validation with an 80/20 split of my data into a training and test set. Other methodologies I learned through this project include: adding weights to my glm() due to the imbalanced dataset, checking for perfect seperation, and model metrics such as precision, recall, and F-score.

One thing I want to improve on is bootstrapping. Since the dataset is relatively small and my football domain knowledge that Man Utd performed consistently in 22/23, then I improve on non-parametric bootstrapping in order to simulate more observations. At the moment, I was content with keeping real results from the season.