

**BC2407** **Analytics II**

**Semester Project: Soccer Analytics Predictive Modelling for Performance and Outcome**

**Semester 2 AY 2020/2021**

**Seminar Group No. 1**

**Team No. 3**

**Group Members:**

NAMITA {Redacted}

JASON {Redacted}

YING FENG {Redacted}

AYUSH {Redacted}

YU TING {Redacted}

Table of Contents (Clickable Links)

Contents

[**1.** **Executive Summary** 4](#_Toc68391163)

[**2.** **Business Opportunity** 5](#_Toc68391164)

[**2.1 Main Issue** 5](#_Toc68391165)

[**2.2 Analysis of Business Opportunity** 5](#_Toc68391166)

[**2.3 Opportunity Statement** 6](#_Toc68391167)

[**2.4 Target Client Base** 6](#_Toc68391168)

[**2.5 Justification** 6](#_Toc68391169)

[**2.6 Proposed Business Solution** 7](#_Toc68391170)

[**3.** **Data Preparation** 7](#_Toc68391171)

[**3.1 Selected Dataset** 7](#_Toc68391172)

[**3.2 Introduction to the Dataset** 7](#_Toc68391173)

[**3.3 Data Cleaning** 8](#_Toc68391174)

[**3.3.1 Derived Fields** 8](#_Toc68391175)

[**3.3.2 Betting Odds** 9](#_Toc68391176)

[**3.3.3 Removing Missing Data Points** 9](#_Toc68391177)

[**Data Exploration** 9](#_Toc68391178)

[**4.1 Data Exploration on Match dataset** 10](#_Toc68391179)

[**4.2 Data Exploration on Player dataset** 10](#_Toc68391180)

[**5. Analytics Solution** 11](#_Toc68391181)

[**5.1 Match Outcome Prediction** 11](#_Toc68391182)

[**5.1.1 CART** 11](#_Toc68391183)

[**5.1.2 Random Forest** 12](#_Toc68391184)

[**5.1.3 Multinomial Regression** 12](#_Toc68391185)

[**5.1.4 Model Comparison** 13](#_Toc68391186)

[**5.2 Player** 14](#_Toc68391187)

[**5.2.1 Linear Regression** 15](#_Toc68391188)

[**5.2.2 Classification and Regression Tree (CART)** 15](#_Toc68391189)

[**5.2.3 Random Forest** 15](#_Toc68391190)

[**5.2.4 Multivariate Adaptive Regression Splines (MARS)** 16](#_Toc68391191)

[**5.2.5 XGBoost** 17](#_Toc68391192)

[**5.2.6 Model Comparison** 17](#_Toc68391193)

[**5.2.7 Quantile Regression** 18](#_Toc68391194)

[**6. Implementation and Implications** 18](#_Toc68391195)

[**6.1 Factors That Influence Match Outcome** 18](#_Toc68391196)

[**6.2 Factors That Influence Player Attributes** 19](#_Toc68391197)

[**6.3 Dashboard** 20](#_Toc68391198)

[**6.4 Limitations** 21](#_Toc68391199)

[**7. Conclusion** 22](#_Toc68391200)

[**8. Appendices** 23](#_Toc68391201)

# **1. Executive Summary**

This report analyses the use of predictive modelling techniques in professional soccer and suggests analytics tools that team managers, coaches, scouts and analysts can use to make well-informed choices regarding matchday tactics, training regiments and player recruitment. Analytics in soccer has not progressed as much as other sports due to its complexity of analysis but has the potential to grow. We propose a two-part analytics solution to address this gap.

Firstly, predictive modelling techniques are used to predict match outcome using a variety of match and team data attributes gathered from historical FIFA matches. Our findings show that team formation and betting probabilities are among the most significant factors influencing match outcome. Team managers can use this information to tailor their team training to maximise performance, and aid team selection on the actual matchday to play to the team’s strengths, and to the opponent’s vulnerabilities.

Secondly, modelling techniques are used to predict player performance using a range of player attributes. The player’s performance rating is taken from the FIFA games, which can be used as an indicator of the player’s performance in real life. Knowledge of the significant factors affecting player performance can be used by managers to recruit promising new players into the team.

Finally, our group has utilized the generated insights to develop a dashboard that presents key information about players and teams. This dashboard allows coaches and team managers to quickly view the statistics of the team and identify areas of improvement. They can also analyse the opponent team’s performance over historical matches and accordingly tweak their tactics for their next encounter.

# **2. Business Opportunity**

## **2.1 Main Issue**

Soccer is a sport that appeals to people across all nations. As FIFA (International soccer’s governing body) continues to support the growth of soccer and with the development of technology, tactical analysis in the sport is going beyond basic observational data. Analyses of team tactics now rely on detailed data including technical skill, individual physiological performance, and team formations to understand complex tactical behaviours.

Moreover, like other professional sports, free agency has resulted in an exponential increase in player salaries and fees paid for player contract purchases. This has led to growing income disparities between wealthy and poorer soccer clubs, leading to the governing body of European club soccer (UEFA) establishing measures such as ‘Financial Fair Play (FFP)’. The vast difference in salaries between the new UEFA member clubs and the more established soccer ones has created a rapid outflow of talent to the higher paying soccer communities, increasing the player talent gap. This has increased the burden of certain soccer communities in creating efficient and effective training programs to retain their players. One way to address these concerns is to use analytics for player recruitment and training.

Sports analytics consists of identifying and acquiring insights about players’ and teams’ performances based on a variety of data sources such as game data and individual player performance data. Such analytics reveal actionable insights that coaches and managers can utilize to build a competitive, profitable team on a minimal budget.

## 

## **2.2 Analysis of Business Opportunity**

In recent years, many firms have started offering advanced analytics to sports clubs. Moneyball is the most widely recognised example of the application of analytics as it was the first sports team to successfully use analytics on a large scale. In baseball, sabermetrics (advanced analytics) has shown that home runs and strikeouts are of greater importance than stolen bases and bunts. Similarly in basketball, analytics has encouraged players to focus on three-pointers and close-range shots, rather than mid-range jumpers (Labaton, 2020).

With more advanced technological devices on the rise, several niche offerings are seen in the market. For example, Catapult offers devices that track training and fitness data to determine if a player’s routine is too strenuous and predict their likelihood of injury during the season. Therefore, the use of advanced analytics and predictive techniques in sport is a large market and has the potential to grow further.

In comparison to the analytics used in baseball and other professional sport, soccer analytics has not progressed much. Soccer has been analysed using many statistical techniques while predictive analysis techniques have focused on passing data or motion tracking (Camacho, 2020). However, in a sport with so many leagues, translating statistics across varying levels of competition could be useful, specifically for recruitment (Kidd, 2018), video scouting, and even performance analysis (Bogert, 2020). Thus, there exists a gap in the market for advanced analytics techniques to be utilised in professional soccer.

## **2.3 Opportunity Statement**

There is a business opportunity to understand what factors influence a match outcome and player performance. Team managers can incorporate the findings into their team strategies to improve their possibility of winning a match as well as saving costs. Through this report, we seek to answer the following questions:

* What is the outcome of a given match?
* What are the most important factors influencing match outcome?
* Which player attributes are most important to determine high-performing players?

## **2.4 Target Client Base**

The target clientele base for this analysis are the soccer clubs and national managers, as well as the executives of sports companies participating in the FIFA World Cup.

## **2.5 Justification**

The following are reasons why an analytics approach in soccer would be useful.

1. Identify factors that make a high performing player

Analytics can be used to identify the significant factors that make a high performing player. Managers can then use this information to identify areas of improvement for players and incorporate these factors into the players’ training regime. This was successfully done by the Red Sox, who used ball trackers to diagnose delivery flaws of their pitchers (Schrage, 2019). The information can also be useful for player recruitment. For example, Lorenzo Ebicilio was an unnoticed player who was recruited through analytics and went on to play a critical role in his team’s performance in the European Champions League (Labaton, 2020).

1. Reduce training cost

Training cost can be reduced as managers can direct their resources in a targeted manner for player development. Acquisition cost for new players may also be reduced if the analytics solution is used to identify and recruit undervalued players. Oakland Athletics started off with $41 million in funding and went on to make the playoffs 4 years in a row (2000-2003), beating the New York Yankees who had a budget of $125 million. The team identified several undervalued players and trained them using insights from their analytics solution.

1. Formulate better winning strategies

Analytics can be used to determine the significant factors that influence the match outcome and predict the outcome of the match. Managers and executives can accordingly change their strategies to improve their chances of winning.

## **2.6 Proposed Business Solution**

Our team’s proposed solution lies in constructing predictive models to identify significant factors influencing player performance and match outcome. The Random Forest, Classification Trees, and Multinomial Regression models are trained to predict match outcome and the best performing model is chosen, determined by model accuracy on a test dataset.

For prediction of player rating, we train Linear Regression, Regression Trees, Random Forest, Multivariate Adaptive Regression Splines (MARS) and XGBoost models. The results from these models are analysed, and the best performing model on the test set data will be used to predict match outcome. The best performing model is chosen, based on its Root Mean Squared Error (RMSE) on a test dataset. Additionally, Quantile Regression will be used to identify attributes that are characteristic of high performing players as compared to average players.

Lastly, Power BI software is used to visualise significant match, team and player attributes on a dashboard for easy interpretation and quick decision making by team managers. Our proposed solution is unique in that it allows team managers to directly use insights generated from the predictive models to improve their strategies. Analytics in soccer has mostly been confined to a research setting with little application of the findings to real life scenarios (Rein & Memmert, 2016). For example, traditional scouting and data-driven approaches have not been integrated together as it was done in baseball. Dashboards can capture the key findings of the complex models and present it in an easily accessible format, making it easy to incorporate into existing team strategies. This aspect differentiates our solution.

# **3. Data Preparation**

## **3.1 Selected Dataset**

The selected dataset from Kaggle (<https://www.kaggle.com/hugomathien/soccer>) contains records of at least 10,000 matches from 2008 to 2016 for us to predict future matches, as well as team and player attributes. This is useful to predict future match outcomes and determine the best performing player’s attributes, which aligns with our objectives. A brief data dictionary for the same has been included in [Appendix A](#A).

## **3.2 Introduction to the Dataset**

The dataset comes in a .sqlite database file with the following tables:

Match: “Match” is our core dataset, as it contains the information required for predictive modelling of the match outcome of the home team, including the participating teams, their leagues etc. The home team is the one who is playing the match in its own stadium.

Player: The “player” dataset offers information on the key biodata of each player, including their name and corresponding player code (the code has been used over the names to refer to players and teams in the other datasets), date of birth, and their height and weight.

Player Attributes: Player attributes contains a treatise of information on the various attributes that each player has. These attributes draw reference to both the player’s physical characteristics such as his preferred foot, stamina and physical strength, as well as his technical attributes like ball control, positioning etc. Each attribute is obtained from EA Sports’ FIFA video games.

Country & League: The league dataset elaborates on the name of the league of each country (Ex: English Premier League for England). Since the match dataset has taken just one league from each of the 11 countries, we obtain the same information from the country dataset as well.

Team: The team dataset offers information on the short and long names of the teams (Ex: Manchester United = MUN), as well as the codes corresponding to each team (Since the code has been used over the names to refer to teams and players in the other datasets).

Team Attributes: The team attributes offer qualitative and quantitative data on each team’s playing style, across a variety of domains. Some of this data includes the team’s ‘build up style’ (How quickly they move the ball from their goal to the opponent’s), which is further divided into constituents like how risky their passing is, how congested or widespread their players’ positions are etc. They also include qualitative information on the class of ‘defensive line’ that the team holds; do they engage in a rather risky ‘offside trap’ tactic in which they try to convert the opponent’s passes into ‘offsides’ (foul play), or do they defend the traditional way.

## **3.3 Data Cleaning**

This section covers the data cleaning stage of the datasets. The final datasets have been detailed in [Appendix B1](#B1) and [B2](#B2) respectively.

### 

### **3.3.1 Derived Fields**

Match Datatset

To get an estimate of a team’s performance, the number of goals scored and games won by the home team and away team in their last 10 matches were calculated. The number of wins and losses by the home team against the away team in the past 10 matches provides a comparison of the performance between the 2 teams. The last 10 matches give a more recent measure of team performance as their characteristics may change over time.

Each team’s overall rating takes the sum of individual player ratings from the player attributes data. Next, the team formation is derived using each player’s X and Y position on the football field, which is divided into a 9x11 grid. Each player’s position in the team is collated, and the formation is calculated from their y-positions. Home and away team formations are seen to be a large determiner of the team’s play style and match strategy. For example, a ‘4-4-2’ formation is perceived to be a high-risk, high-reward formation that focuses on attack.

Lastly, we incorporated the team and player names rather than their ID to provide more clarity and on occasion, context to the dataset.

### **3.3.2 Betting Odds**

Betting odds are typically used for match outcome prediction as they are indicative of public expectations. Betting agencies might also use a large variety of data and models to predict with a reasonably high degree. Historical data shows its high correlation with match outcome, thus it is included as a feature in our model. We converted betting odds into probabilities as shown:

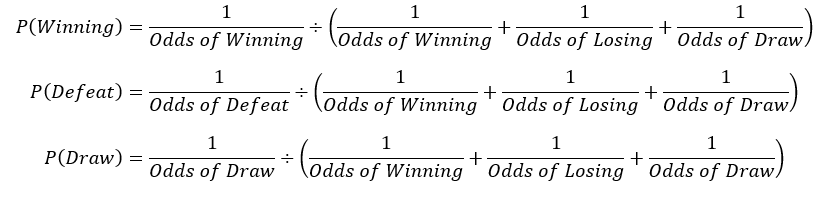


Fig 1. Conversion of betting odds into probabilities

### **3.3.3 Removing Missing Data Points**

A portion of the match data (24%) has missing features. These rows are removed due to the inability of certain models in handling missing values, and to provide a fair comparison of models. After cleaning, the dataset contains 19,585 records.

Players Dataset

For the players dataset, player’s codes (api\_id) are substituted with player names. 836 rows with missing values were removed. Since player and team characteristics can change across seasons, the latest attributes are used for predictive modelling. Following the cleaning process, there are 10,582 rows and 38 columns. Specific columns are handled as follows:

* Attacking and Defensive work rates: The work rates typically take on values of ‘Low’, ‘Medium’ or ‘High’, but the dataset had many invalid values of work rate, thus the column was removed from the dataset.
* Birthday: The player age was extracted from his birthday using Python datetime library.
* Date: Since the player attribute ratings are obtained from the FIFA game, some players receive “special” cards to add a dynamic element to the game, and give players a chance to spend in-game currency to purchase “packs” in which these upgraded cards might be collected. Since these are generally given to players who exceed expectations over the course of a season, we took the most recent update to reflect an accurate picture of the player’s attributes temporally.

# **4. Data Exploration**

Prior to building the analytical models, the relationship between data attributes and the outcome variables of both datasets are analysed in order to get a better understanding of the dataset.

### **4.1 Data Exploration on Match dataset**

The categorical variable ‘label’ refers to the outcome of the match for the home team and can take on 3 possible values (‘Win’, ‘Defeat’, ‘Draw’). Table 1 and [Appendix C](#C) show that the outcome variable is well-balanced, and thus sampling is not required.

|  |  |  |  |
| --- | --- | --- | --- |
| Value | Win | Defeat | Draw |
| Percentage of label | 45.9 | 28.8 | 25.2 |
| Table 1 – ‘Label’ Distribution | | | |

Histograms are used to show the distribution of the home team’s overall rating and away team’s overall rating, grouped by the match outcome ([Appendices D1 and D2](#D)). As home team rating increases, the relative proportion of ‘Win’ increases as compared to ‘Defeat’ or ‘Draw’. On the other hand, lower values of away team ratings tend to have a higher proportion of ‘Win’. Similarly, the number of games won by the home team and away team are plotted on histograms, grouped by the match outcome ([Appendices E1 and E2](#E)). In general, the proportion of ‘Win’ is higher as the games won by the home team increases and games won by the away team decreases. Additionally, we examined if the betting agencies’ odds had any value. [Appendix G](#G) shows a clear pattern that the proportion of ‘Win’ increases with higher betting odds. These variables are likely to be a determining factor for match outcome.

The frequency of home team formation was also explored. From [Appendix F1](#F1), the formations of 4-2-3-1, 4-4-2 and 4-3-3 are most common among home teams. [Appendix F2](#F2) shows the match outcome for 3-3-3-1, 4-2-1-1-2 and 3-2-3-2 have the highest proportion of ‘Win’, but these formations are not commonly used by home teams. Hence, further analysis is required to determine if formation is a significant variable.

Home team attributes such as its class of build-up play speed, chance creation, passing and defensive line were also explored ([Appendix H1](#H1), [Appendix H2](#H2) and [Appendix H3](#H3)). The fast class was found to have the highest proportion of ‘Win’ and lowest proportion of ‘Defeat’, while the other attributes do not seem to influence match outcome.

Lastly, a correlation plot is used to determine if any of the numerical variables in the dataset are correlated. [Appendix I](#I) shows a high correlation between betting odds (B365\_Win, B365\_Defeat, B365\_Draw, BW\_Win, BW\_Defeat, BW\_Draw). A high correlation between predictor variables can lead to the problem of multicollinearity which affects model fit and performance. Therefore, only the BW\_Win variable is retained for further analysis while the other variables are removed.

### **4.2 Data Exploration on Player dataset**

The continuous variable ‘overall\_rating’ refers to the player rating and is the outcome variable of the Player dataset. [Appendix J](#J) shows the range and distribution of player ratings. A correlogram ([Appendix K](#K)) is used to determined correlation between ‘overall\_rating’ and other variables. The variables that might be significant in predicting ratings are short\_passing (0.41), volleys (0.3), dribbling (0.3), curve (0.33), long\_passing (0.39), ball\_control (0.38), reactions (0.79), shot\_power (0.35), long\_shots (0.33) and vision (0.41).

However, the correlation values for potential are similar to those of overall\_rating. Moreover, potential is dependent on the performance of the player and his team, as well as his age (Saunders, 2019). Due to multi-collinearity issues, it is a redundant variable for further analysis.

Bivariate analyses were conducted between overall\_rating and some of the variables listed above. Scatterplots are used to show the best-fit relationships of ratings with reactions ([Appendix L1](#L1)), long passing score ([Appendix L2](#L2)), as well as vision ([Appendix L3](#L3)). While Appendix L1 shows a clear linear relationship, L2 and L3 show minimal trends at the lower spectrum of long passing score and vision, followed by a non-linear increase.

After data exploration and visualisation, we proceed to the next stage of developing a predictive model on the matches and player attributes.

# **5. Analytics Solution**

In this section, the match and player datasets are used to construct a range of predictive models, before the best model is chosen for each use case. The models are first generated using a trainset (70% of dataset), then tested against a test set (30% of dataset). Model performance is based on either its accuracy or RMSE values, and a confusion matrix is used to examine its performance further.

## **5.1 Match Outcome Prediction**

The analytics solution for the ‘Match’ dataset consists of two parts. Firstly, predicting a match outcome is modelled as a 3-part classification problem. The model should be able to predict the match outcome with the highest possible accuracy. Secondly, the most significant factors affecting the match outcome are determined. Black-box models such as neural networks are not useful in this case, as they do not give information about which features are important. Random Forest, CART and Multinomial Regression models are constructed and compared.

### **5.1.1 CART**

CART (Classification and Regression Tree) models are widely used in decision-making because of their interpretive nature and flexibility for various business situations. The model constructs a decision tree using a recursive partitioning algorithm. At each node of the tree, it determines the binary split of a variable that results in the most homogenous division of the outcome variable in child nodes.

The CART model is constructed on the trainset by first growing the decision tree to the maximum depth. The tree is then pruned to get the simplest tree with a low 10-fold cross validation (CV) error. [Appendix M1](#M1) shows the CV error for each sub-tree in the growing phase, where the fourth tree has the lowest CV error. The third tree is optimal as it is the simplest tree that falls below the CV error cap (denoted by the dotted line). The cutoff complexity penalty (CP) value is used to prune the decision tree to its optimal size ([Appendix M2](#M2)). The model is tested against the test set and the resulting confusion matrix is shown in [Appendix M3](#M3).

The variable importance in a CART model is determined by the number of times each feature is used as the splitting criteria. The top 5 most important variables in our CART model are listed as follows:

1. BW\_Win
2. away\_team\_goals\_difference
3. home\_team\_goals\_difference
4. home\_team\_overall\_rating
5. away\_team\_overall\_rating

### **5.1.2 Random Forest**

Random Forest is a type of supervised machine learning algorithm based on ensemble learning, where multiple models are trained on a training dataset and their individual outputs are combined to derive the final output. Random Forest constructs multiple CART trees using subsets of the dataset then combines the predictions. It is more stable and provides better predictive accuracy compared to CART. However, like CART, it has issues with variable selection, missing values and outlier handling, non-linear relationships, and variable interaction detection (Das, 2017).

The train set data was fitted using a Random Forest model with 500 maximal trees and a random subset feature size of 4. This means that 500 bootstrap samples are taken from the train set data, and a maximal tree is grown using each bootstrap sample, with 4 random features selected for consideration at each split point to de-correlate the trees in the forest. The random forest model achieved an out-of-bag (OOB) error of 0.4861. The confusion matrix on the test set data is shown in [Appendix N1](#N1).

The importance of a feature is calculated by first removing the feature from the dataset then training the model on that dataset. The features are ranked by how much the average accuracy decreases when the feature is removed from the dataset. A feature that reduces the accuracy by a greater amount would be considered more important. The variable importance is plotted in [Appendix N2](#N2) and the top 5 most important variables are listed as follows:

1. BW\_Win
2. home\_team\_overall\_rating
3. away\_team\_overall\_rating
4. home\_team\_goals\_difference
5. away\_team\_goals\_difference

### **5.1.3 Multinomial Regression**

We also trained a multinomial regression model on the train set data. The multinomial regression model in this case involves training 2 logistic regression on the baseline outcome of “Defeat”. The first logistic regression model predicts whether the match will end in a draw or defeat, while the second logistic regression model predicts whether the match will end in a win or defeat. Categorical variables are converted into dummy variables before training the model. The model equations for the 2 logistic regression models will look like this:

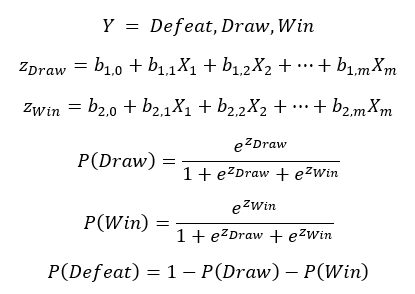


Fig 2. Model equations of multinomial regression model

The model performance on test set data is represented by the confusion matrix as shown in [Appendix O](#O). The coefficient for each variable represents the change in the log odds of ‘Draw’ or ‘Defeat’ with a unit change of the variable. This is used as a measure of variable importance. To determine if the match will end in a draw or defeat, the following variables have the highest coefficients:

1. Home\_formation (5-3-1-1)
2. BW\_Win
3. Home\_formation (3-4-3)
4. Home\_formation (5-3-2)
5. League (Portugal Liga ZON Sagres)

To determine if the match will end in a win or defeat, the following variables have the highest coefficients:

1. BW\_Win
2. Home\_formation (3-3-3-1)
3. Away\_formation (3-2-4-1)
4. Home\_formation (5-3-1-1)
5. Away\_formation (5-3-1-1)

### **5.1.4 Model Comparison**

The 3 models are compared by their prediction metrics on the test set, as it shows how the models would perform on unseen data. Table 2 shows that Random Forest had the lowest accuracy while Multinomial Regression performed slightly better than CART. Multinomial Regression also has higher precision and recall. The accuracy is significantly higher than random guessing (33.33%) and just predicting win on all data points (45.93%). The 5 most important variables for each model are listed in Table 3.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Precision | Recall |
| Random Forest | 51.18% | 51.07% | 51.07% |
| Pruned CART | 52.89% | 52.89% | 52.89% |
| Multinomial Regression | 52.97% | 53.41% | 53.41% |
| Table 2 – Model Performance Comparison on test set | | | |

|  |  |
| --- | --- |
| Model | Top 5 significant variables |
| Random Forest | BW\_Win, home\_team\_overall\_rating, away\_team\_overall\_rating, home\_team\_goals\_difference, away\_team\_goals\_difference |
| CART | BW\_Win, away\_team\_goals\_difference, home\_team\_goals\_difference, home\_team\_overall\_rating, games\_won\_home\_team |
| Multinomial regression | (Draw/Defeat): Home\_formation (5-3-1-1), BW\_Win, Home\_formation (3-4-3), Home\_formation (5-3-2), League (Portugal Liga ZON Sagres) |
| (Win/Defeat): BW\_Win, Home\_formation (3-3-3-1), Away\_formation (3-2-4-1), Home\_formation (5-3-1-1), Away\_formation (5-3-1-1) |

Table 5. most importance variables for each model

Apart from accuracy, the ability to draw actionable insights is a major consideration for choosing the model. The CART model has simplified the decision-making process to just two factors, the home team formation and betting odds of winning (BW\_Win). Multinomial regression, on the other hand, considers the betting odds, home and away team formations, and the league played in. In this case, the regression model may be more useful for decision making as it gives information about specific variables that may affect match outcome. For example, the model coefficient of ‘Home formation (3-3-3-1)’ can be used to compute the odds of achieving a ‘Win’ outcome over a ‘Defeat’ outcome with such a formation. The coach can then change the team’s formation to get a favorable outcome.

The regression model may be harder to visualise and interpret from a business perspective, therefore a dashboard is used in this project to hide the complexity of interpreting a regression model. Additionally, CART is much more computationally intensive for large datasets involving multiple teams and leagues’ data.

## **5.2 Player**

The player dataset is used to construct five predictive models, namely Linear Regression, CART, Random Forest, Multivariate Adaptive Regression Splines (MARS) and XGBoost, to predict overall rating of a player. As we are interested in the variable importance of player attributes, we did not use any black box models that give limited information on feature importance. Root Mean Square Error (RMSE) was used to compare model performance, where a lower RMSE indicates better model fit. RMSE is the standard deviation of the residuals, or the prediction errors between the model-predicted values and actual values. In line with our objectives, we are interested to find the most significant determinants of player performance, therefore Quantile Regression was used to identify the gap in attributes of players at different quantiles.

### **5.2.1 Linear Regression**

A common regression analysis method conducted to predict a continuous y variable is linear regression. Backward elimination, an automatic filtering algorithm, was first used to eliminate the least contributing variable at each step until a statistical criterion is met. Manual filtering is used to further filter variables with p-values exceeding 5%; the variables dribbling and penalties were removed in this step. The regression model is then trained and statistically insignificant variables (p-values > 5%) are removed, such as agility and long passing. These steps remove any non-significant variables from the model.

The final model had a RMSE of **2.98** (detailed in [Appendix P1](#P1)). Variable importance was computed by taking the absolute of t-statistic values. T-statistic is the model coefficient divided by estimated standard deviation of the error. Since t-statistic is standardised and the variables use the same scale, the t-statistic coefficients can be interpreted as the mean change in overall rating, given one standard deviation shift in the independent x-variable (Frost, 2021). [Appendix P2](#P2) shows the plot for variable importance. The top 5 variables are: reactions (52.42), ball control (28.66), heading accuracy (23.31), short passing (11.46) and positioning (11.00).

### **5.2.2 Classification and Regression Tree (CART)**

To construct the CART decision tree, it is first grown to the maximum size to determine the best X variables and their splitting criteria. The complexity penalty (CP) was changed from default 0.01 to 0 to grow the full tree, but the maximum depth was capped at 4. Given a likely overfitted and large tree, the second step was to prune the tree by finding the simplest tree (smallest number of terminal nodes) within 1 standard error of the minimum CV error.

The pruned CART tree is shown in [Appendix Q1](#Q1) and achieved an RMSE of **3.36**. Variable importance is computed using the sum of the goodness of split measures for each split for which it was the primary variable, and the goodness metric in this case would be the deviance, or node Sum of Squared Errors, from the actual value of overall rating. The more times an X variable is listed in the top primary split criteria, and reduces prediction error in the child nodes, the higher its variable importance.

[Appendix Q2](#Q2) shows the scaled variable importance of X-variables from the CART model. The top 5 most important variables are: reactions (34), ball control (15), short passing (11), positioning (9) and vision (9).

### **5.2.3 Random Forest**

After training the Random Forest model and applying it on the test set, the RMSE was discovered to be **1.40**. In comparison with Appendix P1 on the distribution of residuals from the linear regression model, [Appendix R1](#R1) shows a lower deviation of the predicted value from the actual value of rating.

Lastly, variable importance was determined using %IncMSE, which is the increase in Mean Squared Error of predicted value as a result of that variable being randomly permuted. [Appendix R2](#R2) shows that the top 5 important variables are: reactions, finishing, ball control, sprint speed and standing tackle.

### **5.2.4 Multivariate Adaptive Regression Splines (MARS)**

From the previous stage of data exploration, some variables were identified to have a non-linear relationship with overall rating. Thus, we have decided to implement the MARS algorithm, which takes into account multivariate nonlinearities in the data and essentially creates a piecewise linear model.

The first MARS model was created using the train set with degree = 1, which built an additive model with no interaction terms among the variables. The MARS model was grown by adding the hinge functions in reflected pairs with one pair per step by the largest increase in R2 to the current model, and stopped when the increase in R2 was less than 0.001 or no new terms contributed to the increase in R2. The MARS model was then pruned to remove the weakest term, whose removal resulted in the smallest increase in residual sum-of-squares. The RMSE of this MARS model was **2.08** and the variable importance plot is shown in [Appendix S1](#S1). The ‘nsubsets’ criterion counts the number of model subsets that include the variable. Thus, variables that are included in more subsets are considered more important.

A second MARS model was created to assess the potential interactions between different hinge functions. The argument of degree = 2 was set to account for the assumption that candidate hinge functions may be multiplied to existing hinge functions during the MARS model growth phase. The RMSE of second MARS model was **1.54**, which was lower than the previous MARS model with degree 1. [Appendix S2](#S2) shows a comparison in the distributions of residuals between the 2 MARS models. The model performance had improved after considering the interactions between hinge functions using degree = 2, showing a lower variance of residuals.

[Appdendix S3](#S3) shows the variable importance plot for the MARS model of degree = 2. Compared to the first MARS model, dribbling had lower importance while ball control had higher importance.

### **5.2.5 XGBoost**

XGBoost is an implementation of gradient boosted decision trees. Gradient boosting is an approach where new models are created to predict the residuals of prior models and it minimizes the loss when adding new models (Brownlee, 2016).

The objective of the model was set to be regression with squared loss. There would be 2 passes on the train set, in which the second pass would enhance the model by further reducing the difference between ground truth and prediction. This model achieved a RMSE of **2.78**.

Additionally, feature importance was plotted as a bar graph ([Appendix T](#T)). The variables were ranked in terms of Gain, which represents the fractional contribution of each feature to the model based on the total gain of that feature’s splits. A higher proportion suggests a more important predictive feature. From the plot, ‘reactions’ (0.81) is the most important feature, followed by ball control (0.54), gk diving (0.34), standing tackling (0.32), and heading accuracy (0.26).

### **5.2.6 Model Comparison**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Linear Regression** | **CART** | **Random Forest** | **MARS**  **Degree = 1** | **MARS**  **Degree = 2** | **XGBoost** |
| **RMSE** | 2.98 | 3.36 | 1.40 | 2.08 | 1.54 | 2.78 |

Table 4. RMSE of each model on Player dataset

The RMSE on the test set data is compared in Table 4. In maximising the accuracy and predictive power of the analytical model for player rating, we have decided to choose the model with the lowest RMSE of the test set - Random Forest. Besides high predictive accuracy, one of the advantages of using Random Forest as a model is the stability of the model. With more data points to be introduced into the dataset in the future, it is important that a small change does not greatly affect the entire model. Another advantage is that Random Forest reduces the overall inherent biasedness in the dataset since multiple trees are created and each tree is trained on a subset of data. Therefore, this allows team managers and other users of the model to make an objective evaluation of every player in the team.

Under the Random Forest model, the top 5 important variables are: reactions, finishing, ball control, sprint speed and standing tackle. These variables would be included in the dashboard to provide soccer team managers insights on player attributes.

For other models, the top 5 significant variables could also be examined. Reactions seems to be the most significant variable due to its recurrence in every model, while ball control, standing tackle and finishing are also important variables identified during the modelling process. Besides the 5 significant factors under the Random Forest model, heading accuracy, short passing, and positioning could also be used as additional information for soccer team managers in determining the overall rating of a player.

|  |  |
| --- | --- |
| Model | Top 5 significant variables (In descending order) |
| Random Forest | reactions, finishing, ball control, sprint speed, standing tackle |
| Linear Regression | reactions, ball control, heading accuracy, short passing, positioning |
| CART | reactions, ball control, short passing, positioning, vision |
| MARS Degree = 1 | reactions, standing tackle, gk diving, finishing, dribbling |
| MARS Degree = 2 | standing tackle, reactions, gk positioning, ball control, finishing |
| XGBoost | reactions, ball control, gk diving, standing tackling, heading accuracy |

Table 5. Top 5 most important variables for each model in descending order of performance

### **5.2.7 Quantile Regression**

Lastly, we ran a quantile regression model on the top five most important attributes, or characteristics as identified by our chosen model – Random Forest, to gauge how significant the gap is between the top, average and sub-average players across the most important features identified.

Different quantiles are calculated for overall rating and regression lines are plotted for these quantiles to find the variance in median of the quantiles. Quantile regression was run for different levels of tau: 0.1, 0.25, 0.5, 0.75, 0.9 and 0.99, to gauge the gap between the best of the best (99th percentile) and the rest. Our findings are graphically depicted in [Appendix U1](#U1) through [U5](#U5).

At a glance, we notice that each variable depicts a positive (linear) relationship with the overall rating. Reactions have the steepest slope, whereas standing tackles have the flattest. Further, these figures usually converge as we move from a lower variable score to a higher variable score. The convergence is most striking for ball control, suggesting that it’s the most difficult skill to truly stand out in; in contrast sprint speed shows borderline divergence as we move from left to right. For the 99th percentile of players, a steeper gradient between sprint speed and overall rating is observed, implying that higher ranked players work on their physical capabilities like sprint speed that can further improve their overall rating.

Lastly, we notice that across each of these five attributes, there is a significant gap between the 99th percentile and the 90th percentile of overall rating of players, with this particular gap for most variables exceeding those of a much wider percentile range. This reflects that while it is not easy to get to the top 10%, it takes a lot more rigour, dedication, and discipline to improve even further, and penetrate the top one percentile of players.

# **6. Implementation and Implications**

## **6.1 Factors That Influence Match Outcome**

Our model shows that the formations of home and away teams are among the most significant factors determining match outcome. A home team formation of 5-3-1-1 has the largest influence on a draw outcome while the home team formation of 3-3-3-1 has the second largest influence on a win outcome.

Formations typically have to be changed during the game depending on how the team is playing. For example, an attacking formation like 4-3-3 is used for a team that is on the losing end while a defensive formation like 4-5-1 may work better for a winning team, to try and preserve the lead they have (Camacho, 2020). The uncertainty of changing formations can be reduced if the coach can plan the formation in advance using a predictive model, considering the characteristic and formation of the opposing team. Furthermore, professional players and teams require extensive work on the training group (typically at least a week) to become well-versed in a different formation; so planning it in advance will also help the coaching staff to prepare the training regime accordingly, and ensure the team is ready for the match.

The choice of formation also has implications beyond match outcome, for example, to select players for each role in the formation. Bradley et al. (2011) found that certain formations required more high intensity running abilities while other formations required players to cover longer distances on average. This also varied across player roles, such as defender or midfielder. Such information can be combined with our insights from the individual player dataset for a more holistic analysis.

Betting odds of winning is the most significant factor to determine if the match is won or lost. The betting odds is from a 3rd party that uses their own proprietary features to calculate winning odds. This variable does not have much implication to the team itself and is mostly used to improve the model’s predictive performance by borrowing the betting agencies’ capabilities.

Other team related factors, such as the build-up play styles do not have a significant effect on the match outcome but can still be considered when planning the team strategy. For example, our model shows that a slower play style results in greater odds of winning than a fast play style, where the ball is moved around rapidly and with urgency. Some managers prefer to set their team up differently depending on who the opponent is. They might deliberately choose a slower playing style to suffocate their fast-moving opponent, and instead attack them at a critical moment. Therefore, the model results can be used to supplement the managers knowledge. Both sets of managers can know what to expect from their counterpart going into a game and can tweak their own tactics accordingly.

Admittedly, these methods do not always yield dividends (Panch, R. 2018), as information is one thing, and actual execution is another. Therefore, it is equally important to choose the right players with the skill to effectively utilise the information gained. The next section looks at the player attributes that determine high performing players.

## **6.2 Factors That Influence Player Attributes**

Our models have identified several factors that are useful in predicting a player’s overall rating, with the most significant factors being reactions, finishing, ball control, sprint speed and standing tackle. The direct implications of the predictive analysis majorly impact two domains: recruitment and training.

Recruitment scouts can pay special attention to a player’s ball control, reactions, finishing ability etc. Modern-day scouting and analytics techniques leverage on technology to get video data of “every pass attempt” by a certain player to save performance analysts the trouble of manually sifting through a 90-minute game to identify the player’s passing attempts. Similarly, scouts can pay attention to “every first touch”, or even “every touch in the opposition half” data to gauge these specific characteristics and make data-driven recommendations to teams. Existing analytics infrastructure allows scouts to shortlist 5 players in the world that they should be looking at. Our model narrows the analysis down to the question of “What exactly should we be looking for”, and subsequently identifying the top players that meet those charactertistics. Such targeted scouting is cost effective and can even reveal previously unnoticed players.

For the second business opportunity of training, team managers can make use of the model results to emphasise which ‘impact areas’ they can train the players on further. For instance, ‘reactions’ has been determined to be the most important predictor of a player’s overall rating by several predictive models. Acts like dribbling the ball demand tons of focus, control, and quick movements. Hence, by training to attain faster reaction speeds, it will help decrease the pressure felt in your mind during a match (Cruzcoaching, October 2020). The training regime can incorporate various techniques such as those mentioned by SoccerCoachWeekly, and Norwich City Football Club’s real life reactions training for goalkeepers. This will allow their players to become better rounded, both in the soft or ‘skill’ aspects of their game (Finishing, ball control) and the hard, or ‘physical’ aspects of their game (Sprinting speed, reactions). The latter may be especially useful when combined with some of the output from the match outcome dataset’s findings, and sprint speed is an important consideration for attacking players against teams who play the “offside trap”. According to external sources, faster sprinting speed is linked with a higher chance of ‘beating’ the offside trap, in which case the player usually ends up with a “big chance” (A chance worth 0.38 xG (Fantasy Football Scout), which essentially implies that historical data and simulations show that it would be scored 38% of the time).

Both sides can be used to build a more efficient and a more capable team, which would not only better their chances of fighting for and securing domestic and continental cup competitions, but also yield a higher figure in the form of transfer fee, if the player was to be acquired by an external club. A higher transfer fee could be reinvested in signing a replacement player, or in better training staff/facilities, starting the loop once more.

## **6.3 Dashboard**

As mentioned previously, this is a unique selling point of our solution. To draw useful insights from our analytics solution, a dashboard is used to track each player’s ability and performance and identify areas of improvement. Dashboards can summarise relevant complex information into a user-friendly interface with interactive features such that team managers and coaches can readily access this information and analyse accordingly. Given the limited application of data-driven approaches to real life scenarios such as scouting (Rein & Memmert, 2016), the dashboard also helps team managers to incorporate the model insights into their existing strategies.

Each dashboard is specific to each team and the team managers and coaches can track their current team’s and players’ performance. As shown by the predictive models, the most significant variables affecting player and team performance are highlighted on the dashboard.

On the Team page, the features are as follows ([Appendix V1](#V1)):

|  |  |
| --- | --- |
| Feature | Description |
| Latest player names table | Displays the team’s latest lineup in ascending order, with player number 1 being the goalkeeper. |
| Last 10 matches table | Classified into two tables of the team acting as a home team and away team. By default, the tables will display the last 10 matches that the team has played as a home and away team.  Season slicer: to get the last 10 matches for the season selected.  Opponent slicer: To get the last 10 matches against the opponent selected. |
| Formation graph | Displays the number of times that the team has used the respective formations in their matches. |
| Team rating | Shown at the top right of the dashboard is the team’s overall rating – summation of all the players’ average ratings |
| Team statistics | Line graphs of their scoring with regards to Build Up Play, Defense and chance creation matrics over the past 5 years. |

Table 6. Features of dashboard Team page

On the Player page, the features are as follows ([Appendix V2](#V2)):

|  |  |
| --- | --- |
| Feature | Description |
| Latest player names table | Similar to the Team’s page, it displays the team’s latest lineup in ascending order, with player number 1 being the goalkeeper. The user can use the table as a filter look at each player’s performance individually or select multiple players to compare the between each other. |
| Year slicer | The user can select a range of years which would filter the entire page accordingly. |
| Overall player rating graph | Displays each team player’s overall rating performance with respect to time. |
| Line graphs | Line graphs of the important attributes based on the player attributes prediction model (Reactions, ball control, standing tackle, finishing, sprint speed and heading accuracy). |

Table 7. Features of dashboard Player page

Sample analysis using the dashboard ([Appendix V1](#V1), [Appendix V2](#V2)):

Illustrated on the dashboard is team Arsenal. Based on team formations, we can note that the team used the formation 4-2-3-1 the most and achieved the most wins with that formation. Their last 10 matches indicates that Arsenal performed better as a home team. The team statistics show a decreasing trend in chance creation shooting and crossing. Coaches can look into improving those aspects of the team. As for their players, their overall rating has been consistent which is good for the team. As for the player Nacho Monreal, he might need more focus in his finishing as it is below 50, and his sprint speed as it has been decreasing.

The dashboard also allows coaches to integrate team and player analysis. Executing a 4-2-3-1 formation may require players with certain qualities such as better ball control or sprint speed. Players with these qualities can be identified through the Player page.

Implications:

With these visuals included in the dashboard, coaches can intuitively understand the influence of formation on their win rate and the team’s performance in build-up play, defence and chance creation. Coaches can further delve into each individual via the player’s page to track the current performance of each player and identify if he needs additional coaching, specifically in the more significant qualities as identified by our model. Dashboards are very flexible and further refinements and improvements can be made to reflect each team’s data so that it is tailored to each team.

## **6.4 Limitations**

The model is trained on publicly available historical data of matches, and hence does not consider internal factors such as the kind of training involved and quality of coaching, as such information is not public. This information would be useful to understand how the preparation phase affects match outcome. The model is also biased towards the data it has been trained on. For example, the train set only has match information on 9 football leagues, so the model performance may suffer if used on an unseen foreign league.

As for individual players, the Random Forest model is recommended to give team managers an estimate of the overall rating of each of their players, and the important factors they could focus on for training or recruitment. However, the dataset used for this analysis may not be reflective of real-life player performance as they are extracted from the FIFA video game. Concrete data is required to make decisions with millions on the line (in the form of prize money). Further, we do not expect each individual statistic to be divisible to the degree that it is in the dataset.

# **7.** **Conclusion**

Various analysis tools in R were used to meet our business objectives of predicting match outcome along with its determining factors and identifying important indicators of player performance. Our insights can enable team managers to reduce training costs, recruit more economically, and formulate better winning strategies. This is especially useful in light of COVID-19, where team’s finances are stretched to the limit with the lack of ticket revenue inability to conduct stadium tours.

While predictive models are useful for decision making in soccer, the models tested in this report should be used as a reference point rather than a solution. Given the match outcome’s model accuracy of 52.9% and other limitations, there is room to improve the models. While the individual players’ ratings may be high, the performance of players will also be affected by other variables during the match, such as team line-up and unforeseen conditions. Therefore, it should not replace traditional analysis but act as supplemental knowledge to streamline the analysis. The dashboard is especially useful to speed up the analysis process through quick visualisation.

We are confident that our predictive models and dashboard will help to make soccer analysis more accessible to team managers and supplement their current resources in formulating team strategies and planning programs to improve player performance.

# **8. Appendices**

Appendix A: Original Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Filename | Size | Description | Other information |
| league | 11 rows, 3 columns | League names |  |
| country | 11 rows, 2 columns | Participating countries |  |
| match | 25979 rows, 115 columns | Contains match details:   1. Team line up with squad formation (X, Y coordinates) 2. Betting odds from up to 10 providers 3. Detailed match events (goal types, possession, corner, cross, fouls, cards etc…) for +10,000 matches 4. Seasons 2008 to 2016 |  |
| team | 299 rows, 5 columns | Distinct team names with respective apis |  |
| player | 11060 rows, 7 columns | Distinct players’ names and details |  |
| team\_attributes | 1458 rows, 25 columns | Team attributes over different years | Players and Teams' attributes sourced from EA Sports' FIFA video game series, including the weekly updates |
| player\_attributes | 183,978 rows, 42 columns | Players’ attributes over different years |

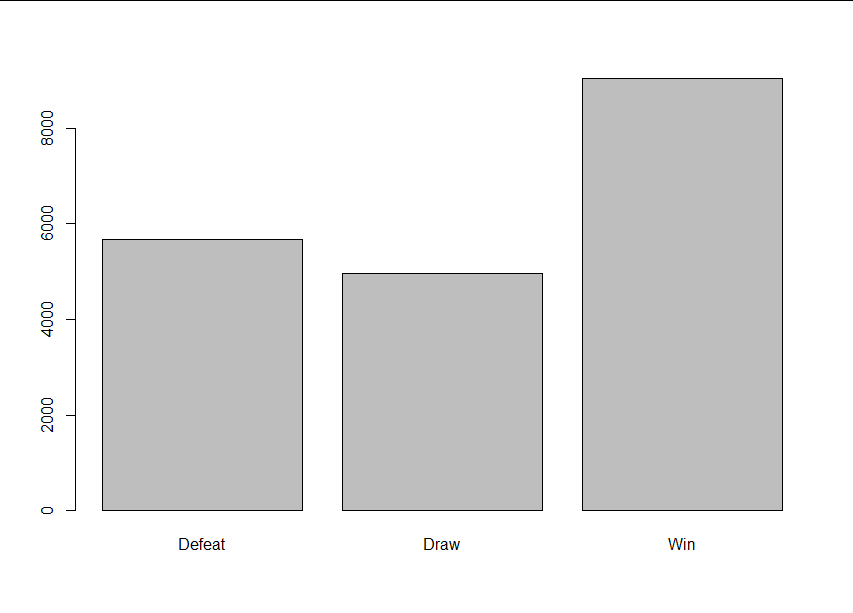
Appendix B1 – Columns in the final ‘Match’ Dataset

|  |  |
| --- | --- |
| Attribute | Description |
| Home team goals difference | Number of goals won minus number of goals conceded by home team in the last 10 matches |
| Away team goals difference | Number of goals won minus number of goals conceded by away team in the last 10 matches |
| Games won by home team | Number of game won by home team in the last 10 matches |
| Games won by away team | Number of game won by away team in the last 10 matches |
| Games against won | Number of game won by home team in the last 10 matches between the home and away team |
| Games against lost | Number of game won by away team in the last 10 matches between the home and away team |
| Home team overall rating | Overall performance rating of the home team |
| Away team overall rating | Overall performance rating of the away team |
| League name | Name of the league the match is in |
| Home team formation | Formation of the home team |
| Away team formation | Formation of the away team |
| Home team buildUpPlaySpeedClass | The pace, or the “tempo” with which the home team “builds up play”, i.e. moves the ball from their own goal towards their opponent’s - Taken from the team\_attributes dataset |
| Home team chanceCreationPassingClass | The nature of the home team’s passing when trying to create a goal-scoring chance (Risky passes, safe passes, usual etc.) - Taken from the team\_attributes dataset |
| Home team defenceDefenderLineClass | The nature of the home team’s defensive line (Do they cover the ball, or play an “offside trap”; a risky tactic to disrupt the opponent’s attacking play) - Taken from the team\_attributes dataset |
| Away team buildUpPlaySpeedClass | The pace, or the “tempo” with which the away team “builds up play”, i.e. moves the ball from their own goal towards their opponent’s - Taken from the team\_attributes dataset |
| Away team chanceCreationPassingClass | The nature of the away team’s passing when trying to create a goal-scoring chance (Risky passes, safe passes, usual etc.) - Taken from the team\_attributes dataset |
| Away team defenceDefenderLineClass | The nature of the away team’s defensive line (Do they cover the ball, or play an “offside trap”; a risky tactic to disrupt the opponent’s attacking play) - Taken from the team\_attributes dataset |
| B365 (Won) | Bet365 betting probability for home team winning |
| Label | Match outcome for the home team (Win/Lose/Draw) |

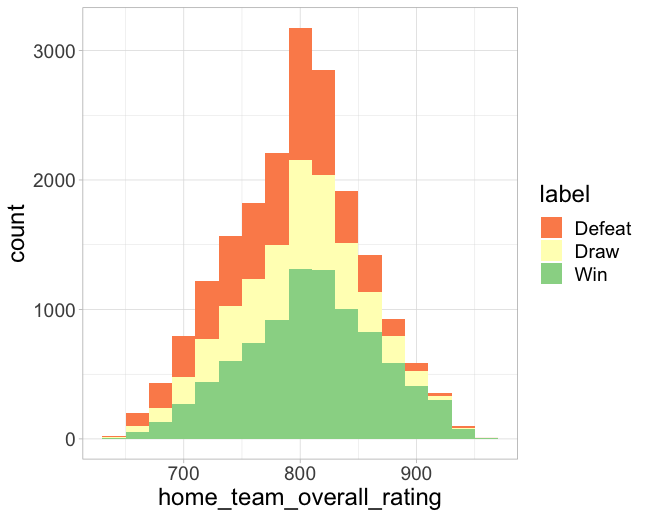
Appendix B2 – Columns in the final ‘Player Attributes’ Dataset

|  |  |
| --- | --- |
| Attribute | Description |
| height | The player’s height |
| weight | The player’s weight |
| age | The player’s age |
| overall\_rating | The player’s overall rating |
| preferred\_foot | The player’s stronger foot (Left, right) |
| crossing | The player’s ability to ‘cross’ the ball (Make a traditionally long-range horizontal pass in attack) |
| finishing | The player’s ability to score given a goal-scoring chance. A higher ‘finishing’ ability indicates ability to consistently convert mid-to-easy chances |
| heading\_accuracy | The player’s accuracy while attempting to score a goal with a header |
| short\_passing | The player’s ability to make a short-range pass |
| volleys | The player’s ability to ‘volley’ the ball (Hit it while it is in the air) |
| dribbling | The player’s dribbling ability |
| curve | The curve/swerve the player is able to achieve on his free-kicks |
| free\_kick\_accuracy | The accuracy of the player’s free-kicks |
| long\_passing | The player’s ability to make a long-range pass |
| ball\_control | The player’s ability to retain possession of the ball |
| acceleration | The increment of the player’s running speed on the pitch. |
| sprint\_speed | The speed rate of the player’s sprinting |
| agility | How quick and graceful the player is able to control the ball |
| reactions | The player’s reaction speed |
| balance | The player’s ability to not lose his balance |
| shot\_power | The power the player is able to achieve on his shots |
| jumping | The player’s jumping ability |
| stamina | The player’s stamina |
| strength | The player’s physical, upper body strength |
| long\_shots | The player’s ability to score a goal from a ‘long shot’ (Unofficially considered to be from further than 18 yards from goal) |
| aggression | The aggression level of the player on pushing, pulling and tackling |
| interceptions | The player’s ability to intercept a pass meant for an opposing player and win possession for his team |
| positioning | The player’s capability in performing the positioning on the field as an outfield player |
| vision | The player’s mental awareness about his teammates’ positioning, for passing the ball |
| penalties | The player’s ability to score a penalty kick |
| marking | The player’s ability to stick close to his counterpart and prevent him from being involved in the play (Same concept as a man-to-man mark in other sports like hockey) |
| standing\_tackle | The player’s ability to perform a standing tackle |
| sliding\_tackle | The player’s ability to perform a sliding tackle of a player in a match |
| gk\_diving | The goalkeeper’s diving ability (Or for an outfield player, his ability to dive as if he were a stand-in goalkeeper) |
| gk\_handling | The goalkeeper’s handling ability (Or for an outfield player, his ability to handle the ball as if he were a stand-in goalkeeper) |
| gk\_kicking | The goalkeeper’s kicking ability (Or for an outfield player, his ability to kick as if he were a stand-in goalkeeper) |
| gk\_positioning | The goalkeeper’s positioning ability (Or for an outfield player, his ability to position himself appropriately as if he were a stand-in goalkeeper) |
| gk\_reflexes | The goalkeeper’s reflexes (Or for an outfield player, his reflexes as if he were a stand-in goalkeeper) |

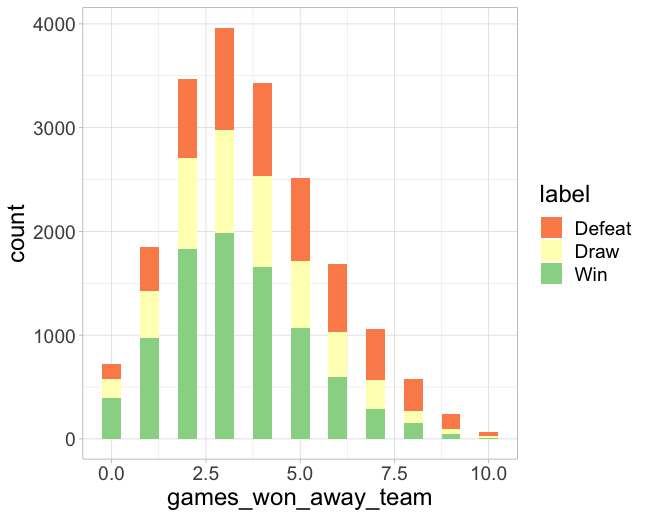
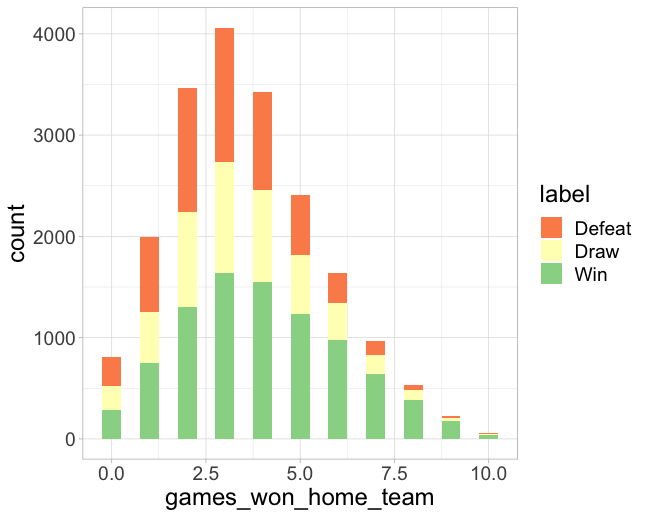
Appendix C – Graphical Distribution of ‘Label’ in Match



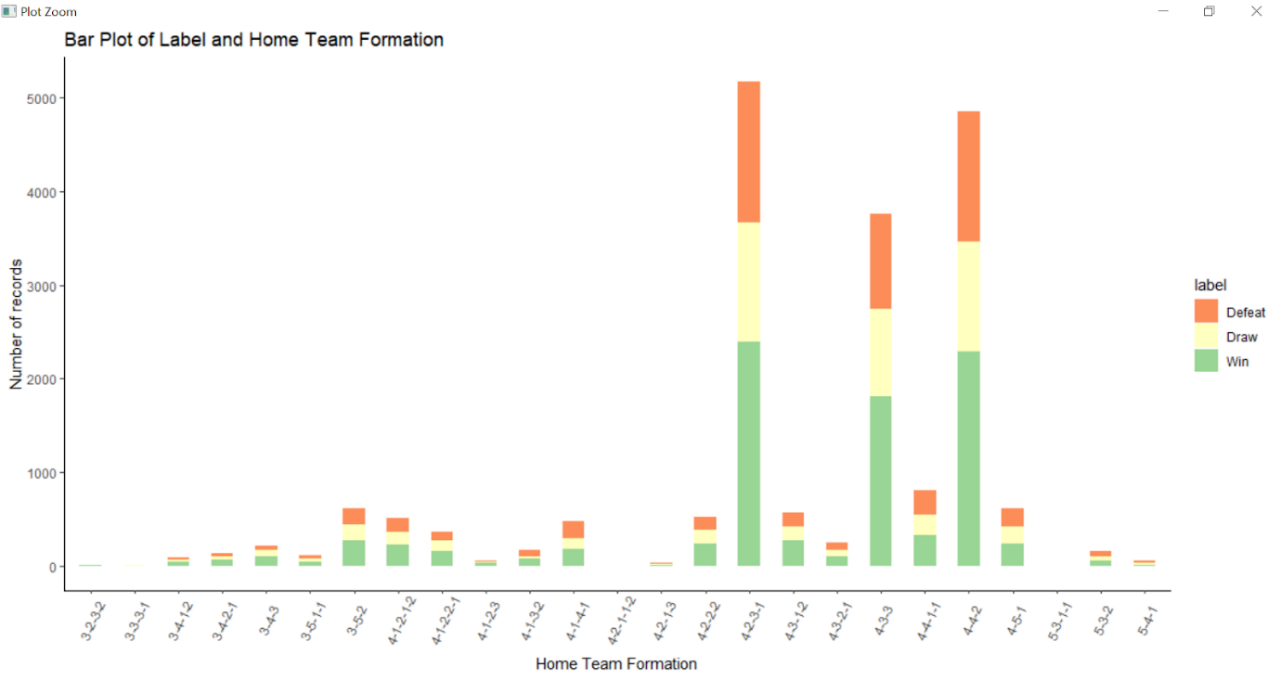
Appendices D1 and D2 – Distribution of Team Overall Rating Grouped by Match Label



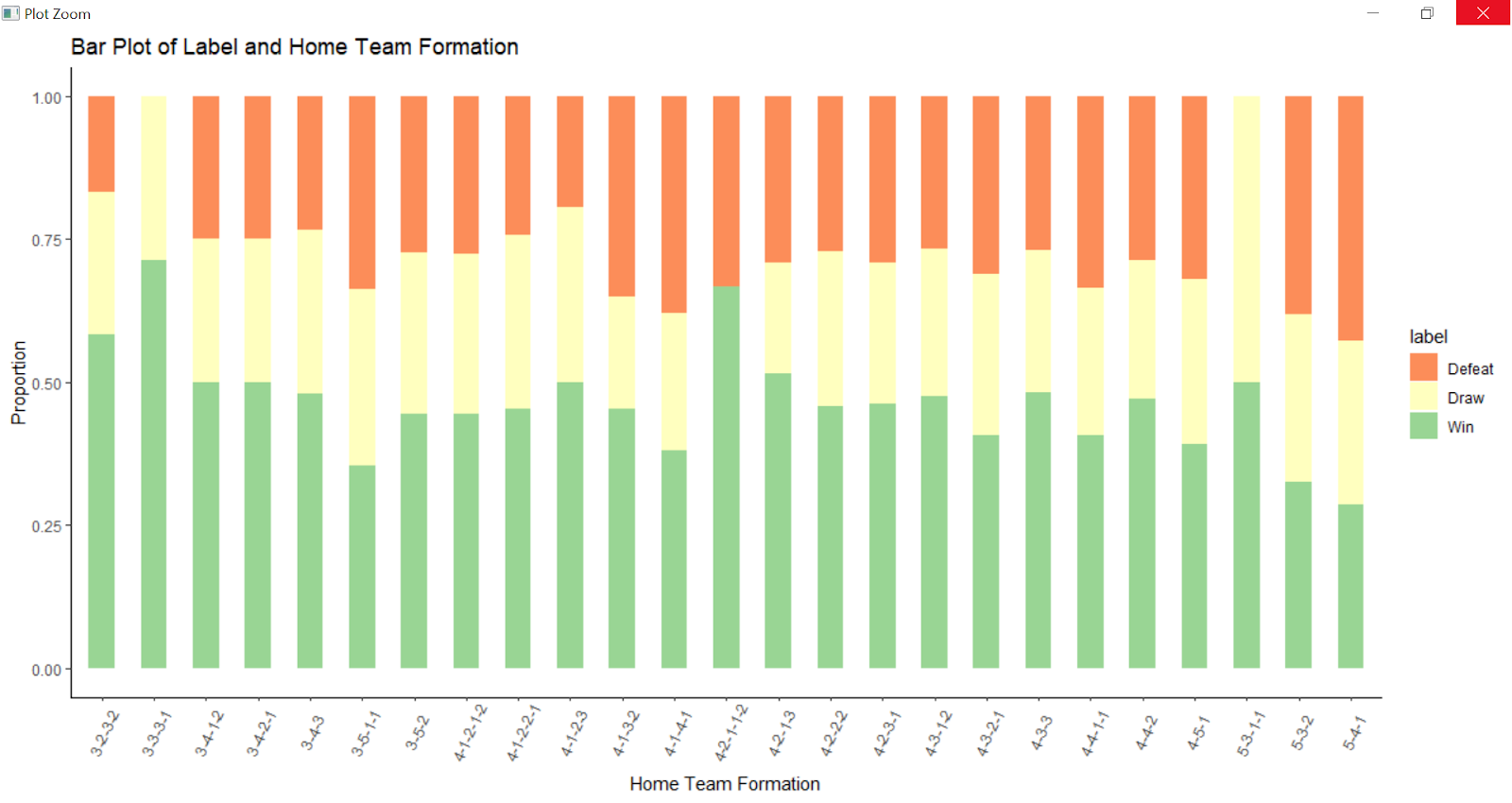
Appendices E1 and E2 – Histogram of Games Won by Home and Away Teams grouped by Label



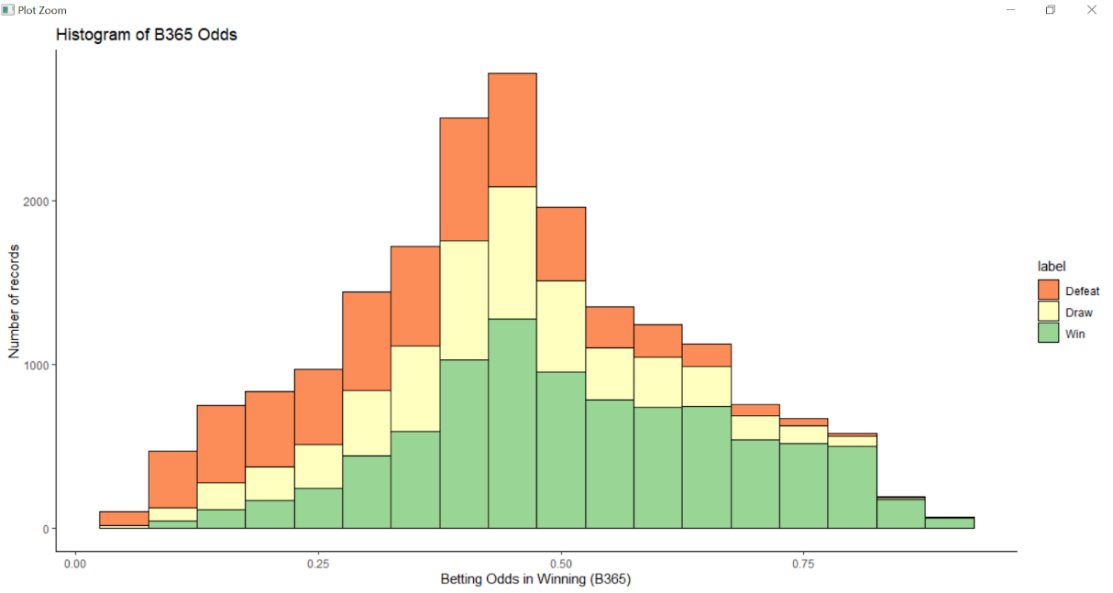
Appendix F1 – Bar plot showing the frequency of home team formations used



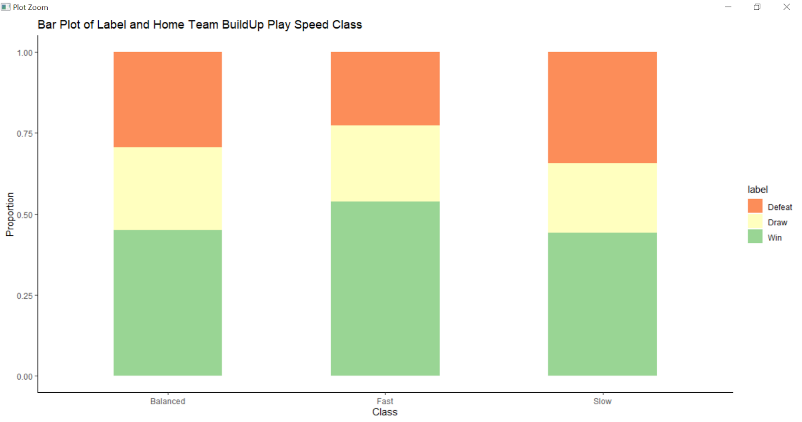
Appendix F2 – Bar plot showing home team formations used with label (As a proportion)



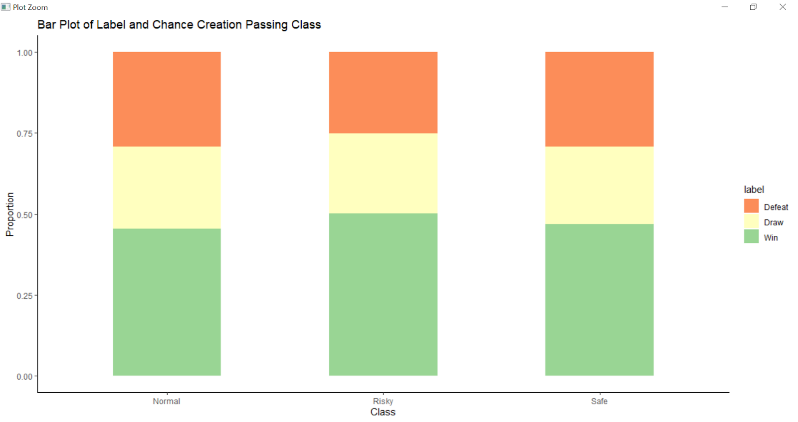
Appendix G: Histogram showing distribution of betting odds of winning with the match label



Appendix H1 – Bar plot showing distribution of label for each build-up play speed class



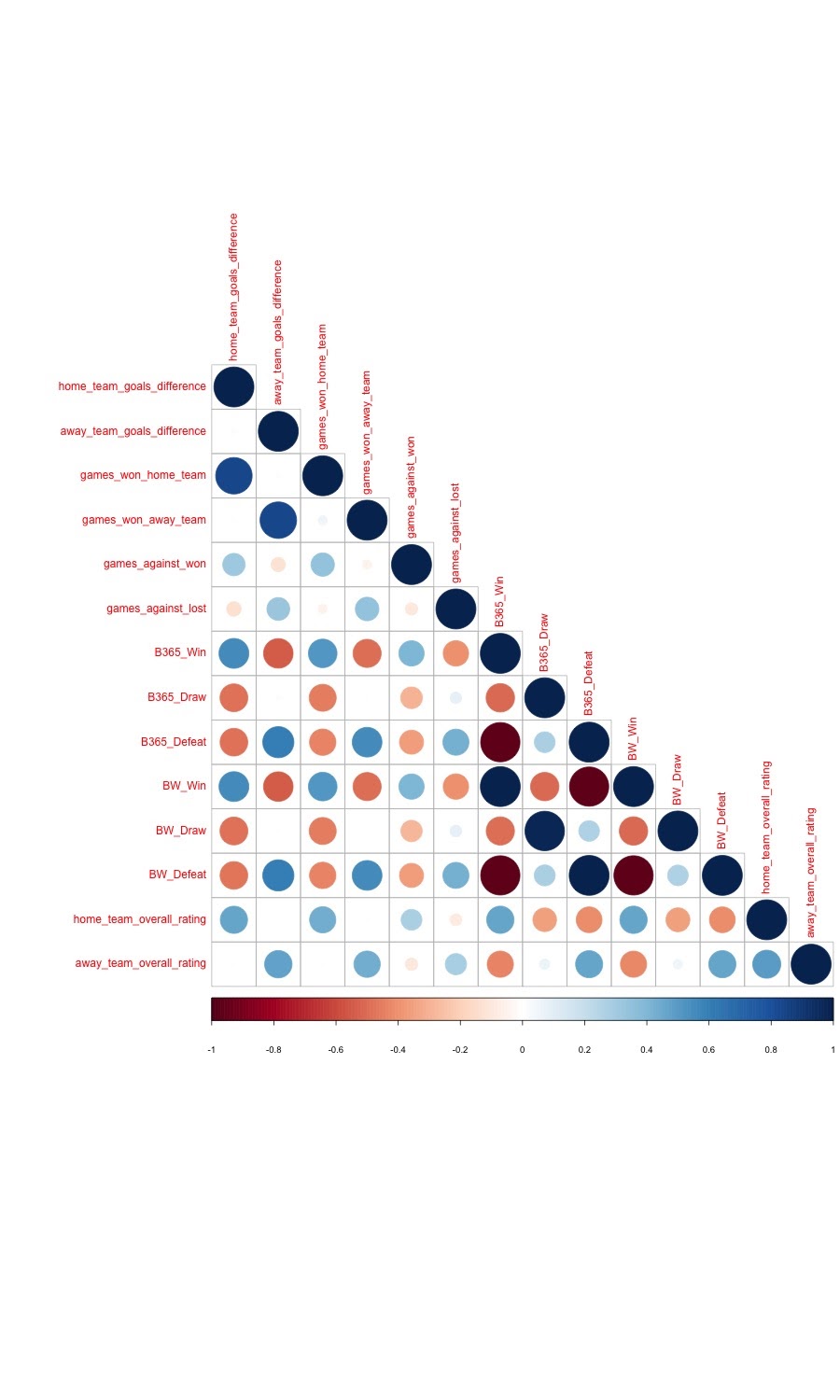
Appendix H2 – Bar plot showing distribution of label for each chance creation passing class



Appendix H3 – Bar plot showing distribution of label for each defence defender line class



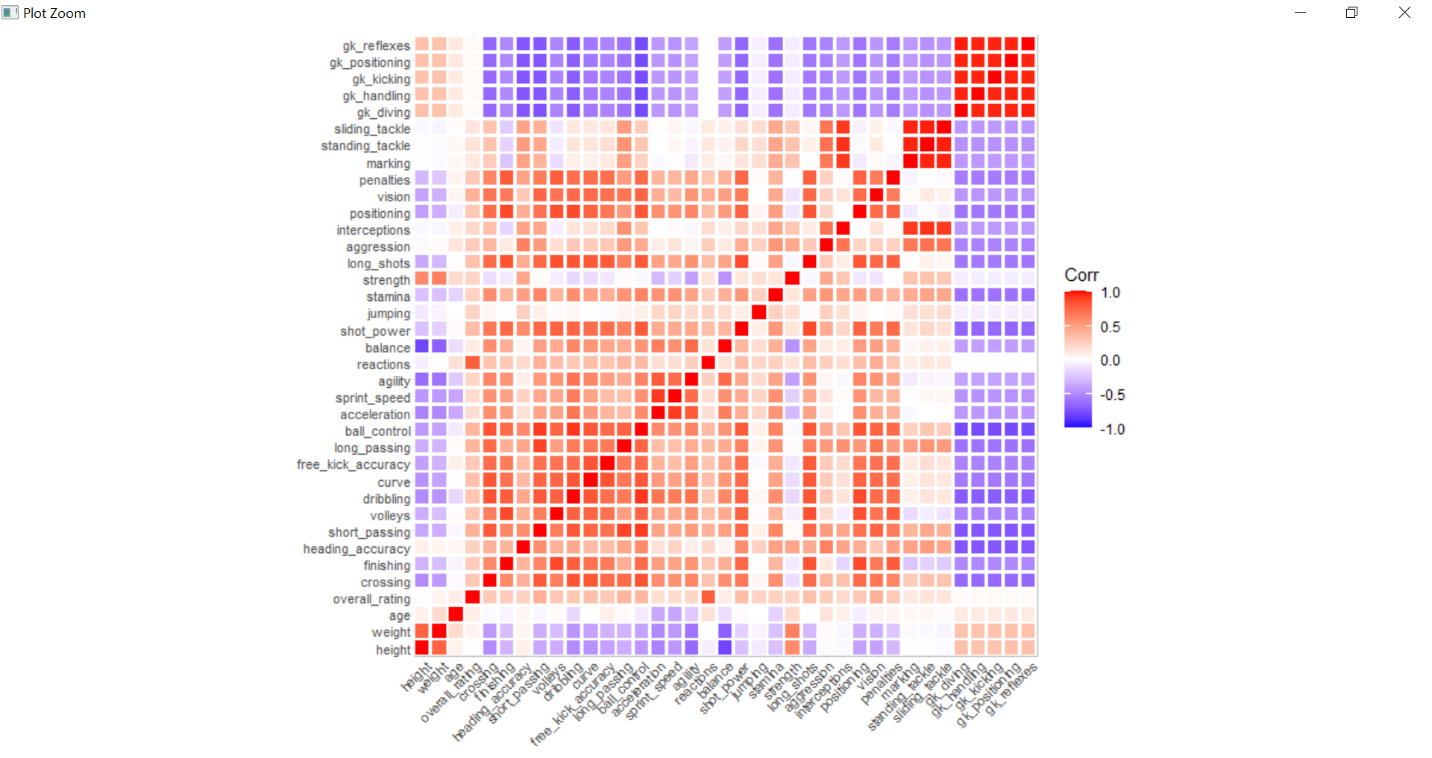
Appendix I – Correlation plot of numerical variables in match dataset



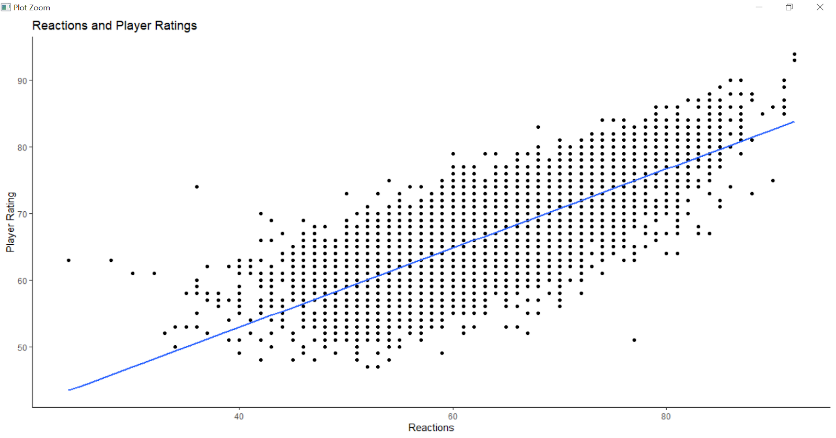
Appendix J – Histogram showing distribution of Player Rating and Preferred Foot



Appendix K – Correlogram of variables



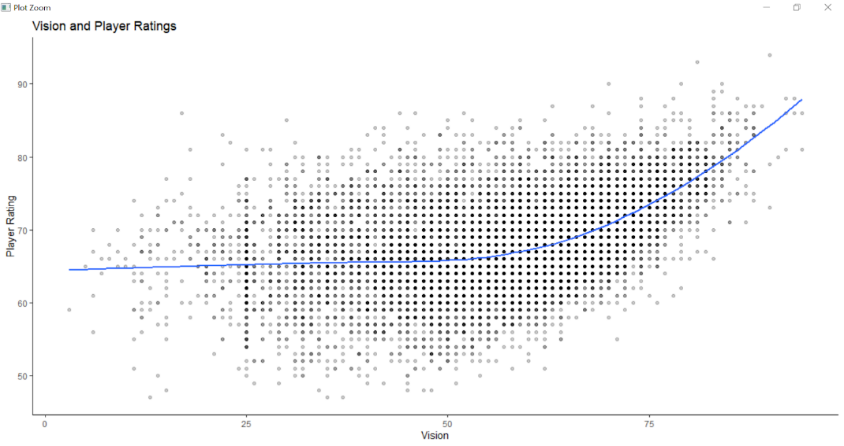
Appendix L1 – Scatterplot between Reactions and Overall\_Rating



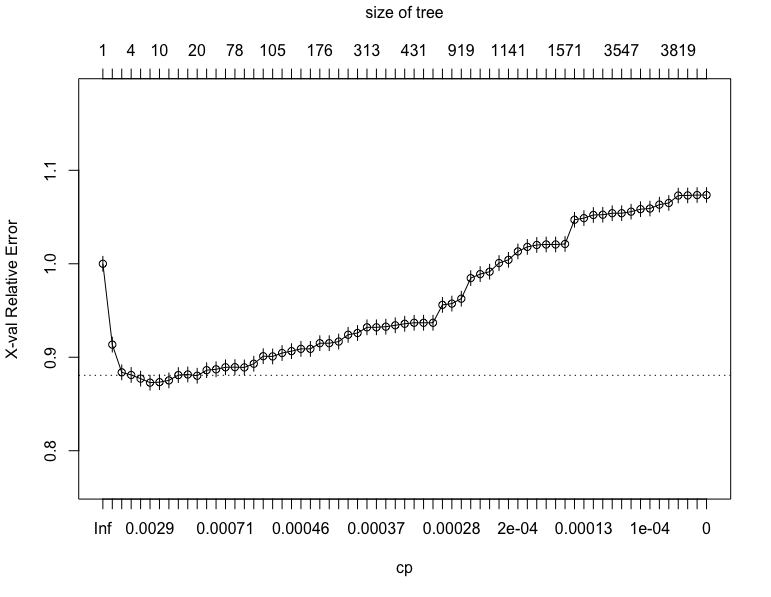
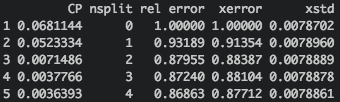
Appendix L2 – Scatterplot between Long Passing Score and Overall\_Rating



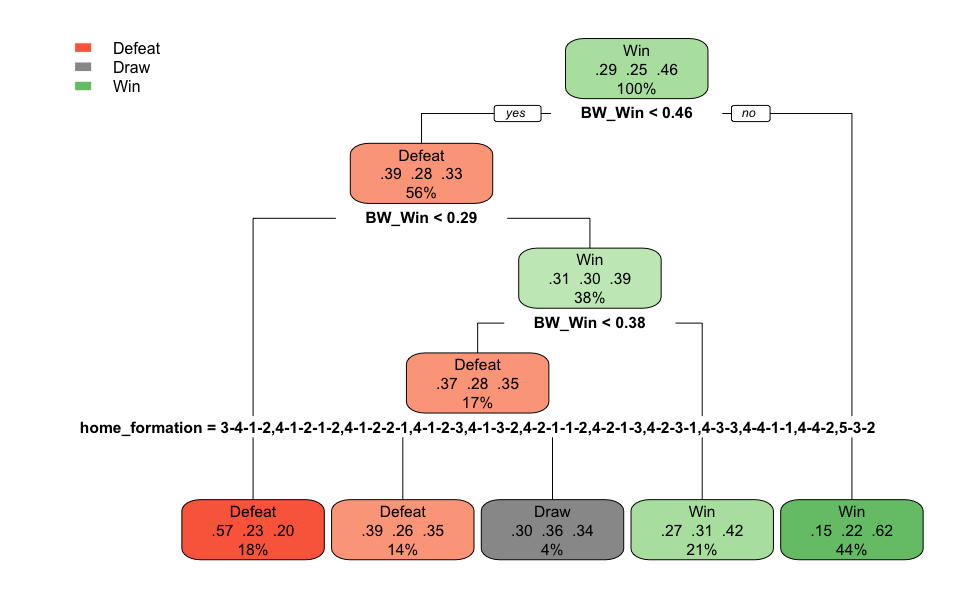
Appendix L3 – Scatterplot between Vision and Overall\_Rating



Appendix M1 – CART: Growing to the Maximum and CP Table



Appendix M2 – CART: Optimal Classification Tree



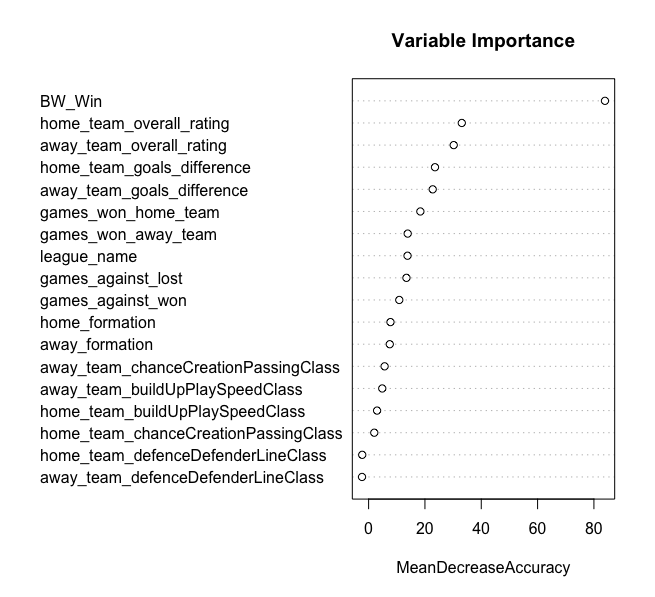
Appendix M3 – CART: Testset Performance Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CART** | | Ground truth | | |
| Defeat | Draw | Win |
| Prediction | Defeat | 906 | 499 | 496 |
| Draw | 82 | 65 | 66 |
| Win | 707 | 918 | 2136 |

Appendix N1 – Random Forest: Testset Performance Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Random Forest** | | Ground truth | | |
| Defeat | Draw | Win |
| Predicted | Defeat | 762 | 424 | 399 |
| Draw | 119 | 91 | 145 |
| Win | 814 | 967 | 2154 |

Appendix N2 – Random Forest: Plot of variable importance for Random Forest

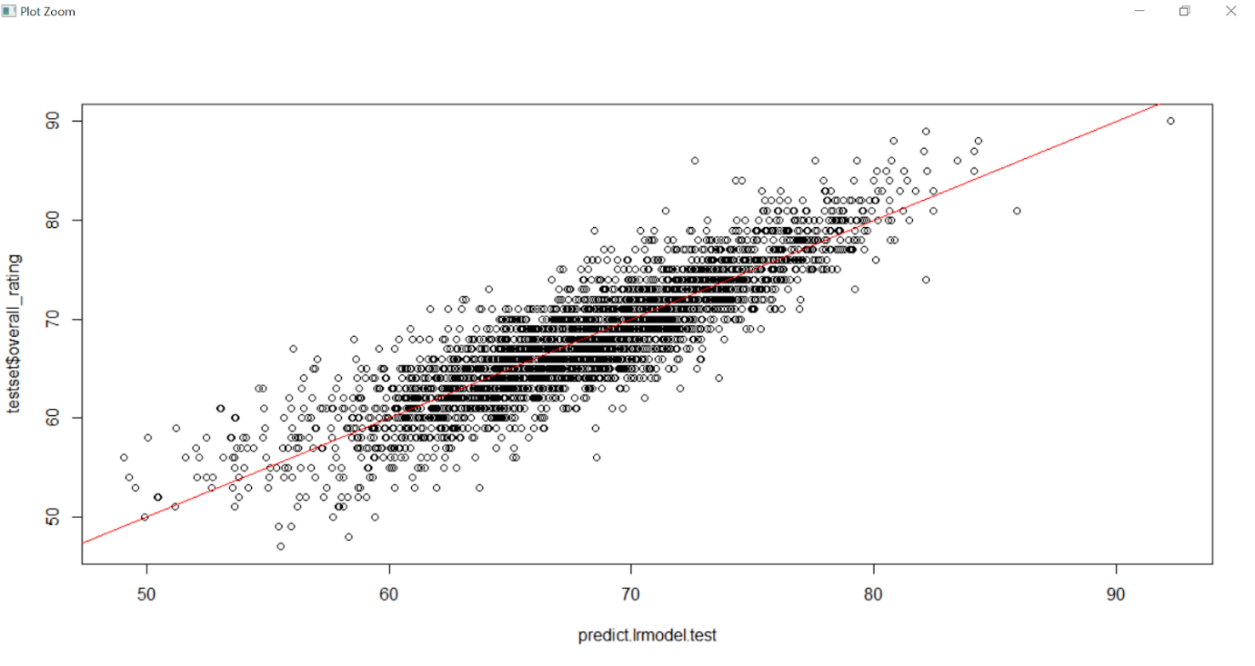


Appendix O – Confusion Matrix from Multinomial Regression Model on Test Set

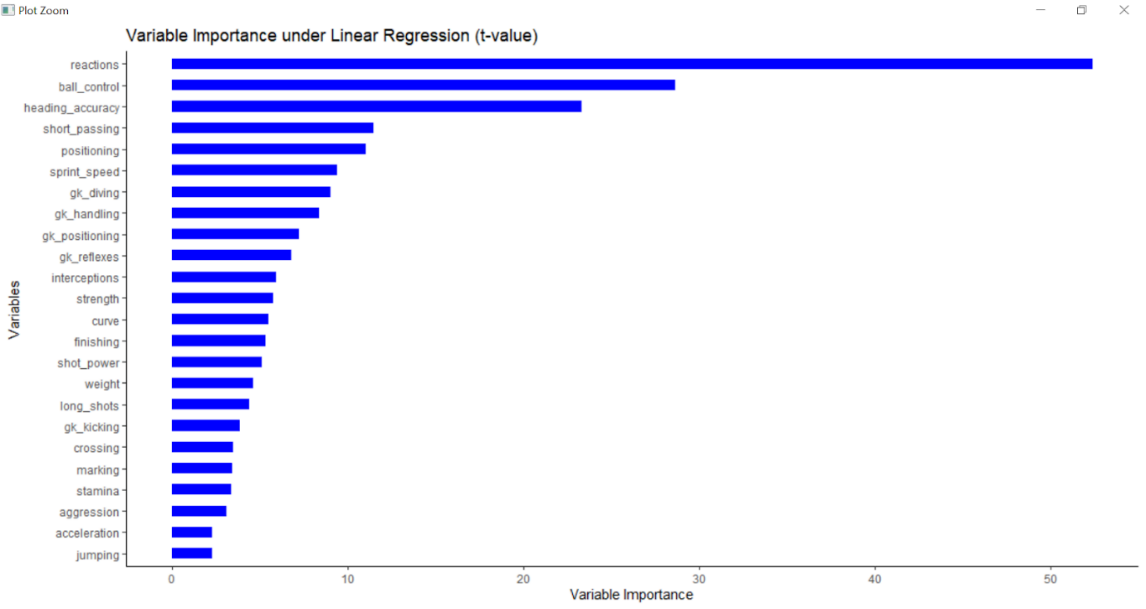
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Multinomial Regression** | | Ground truth | | |
| Defeat | Draw | Win |
| Prediction | Defeat | 866 | 468 | 445 |
| Draw | 45 | 26 | 38 |
| Win | 784 | 988 | 2215 |

Appendix P1 – Scatterplot showing the distribution of residuals from the linear regression model

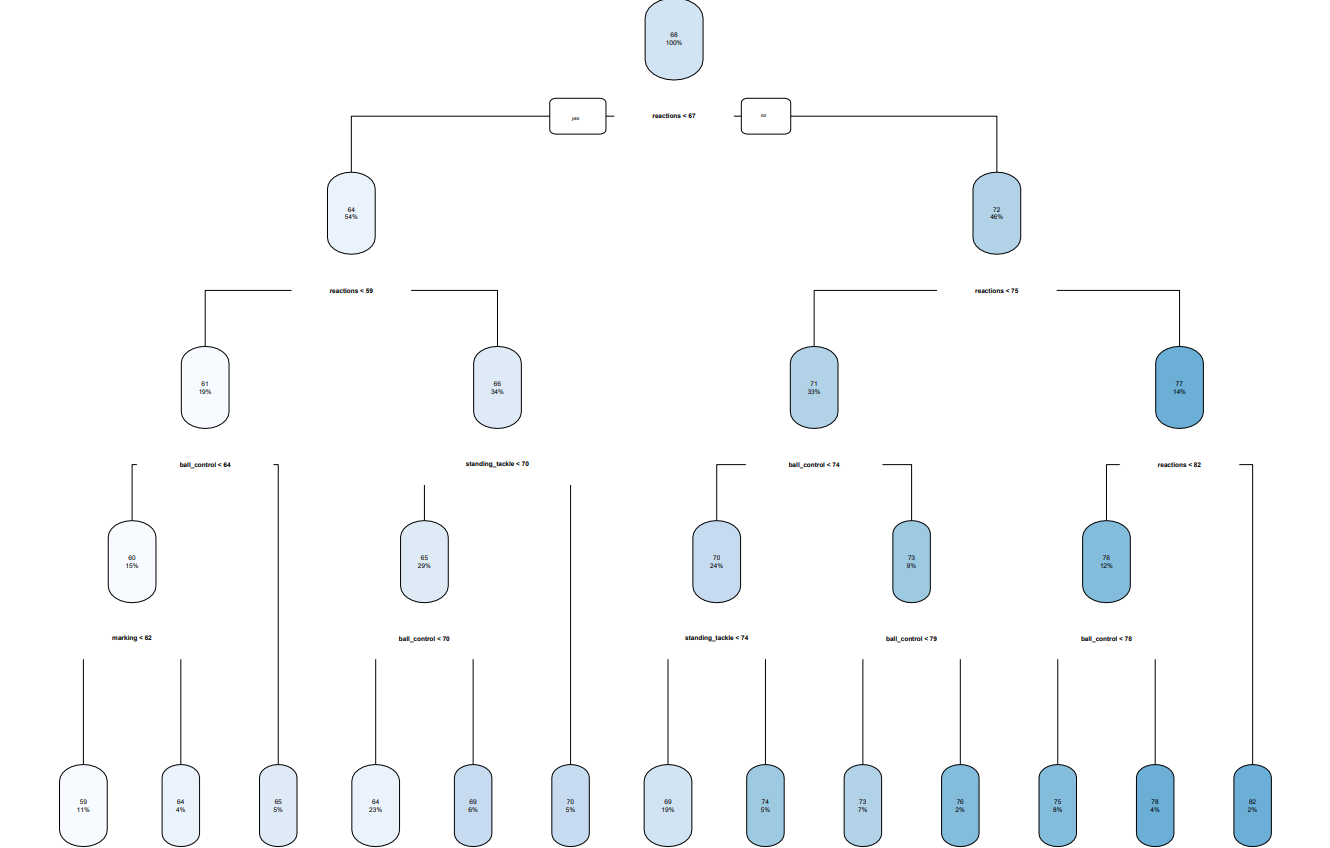
The red line represents the linear regression model if it were to predict all values accurately, and the distribution of residuals from the line.



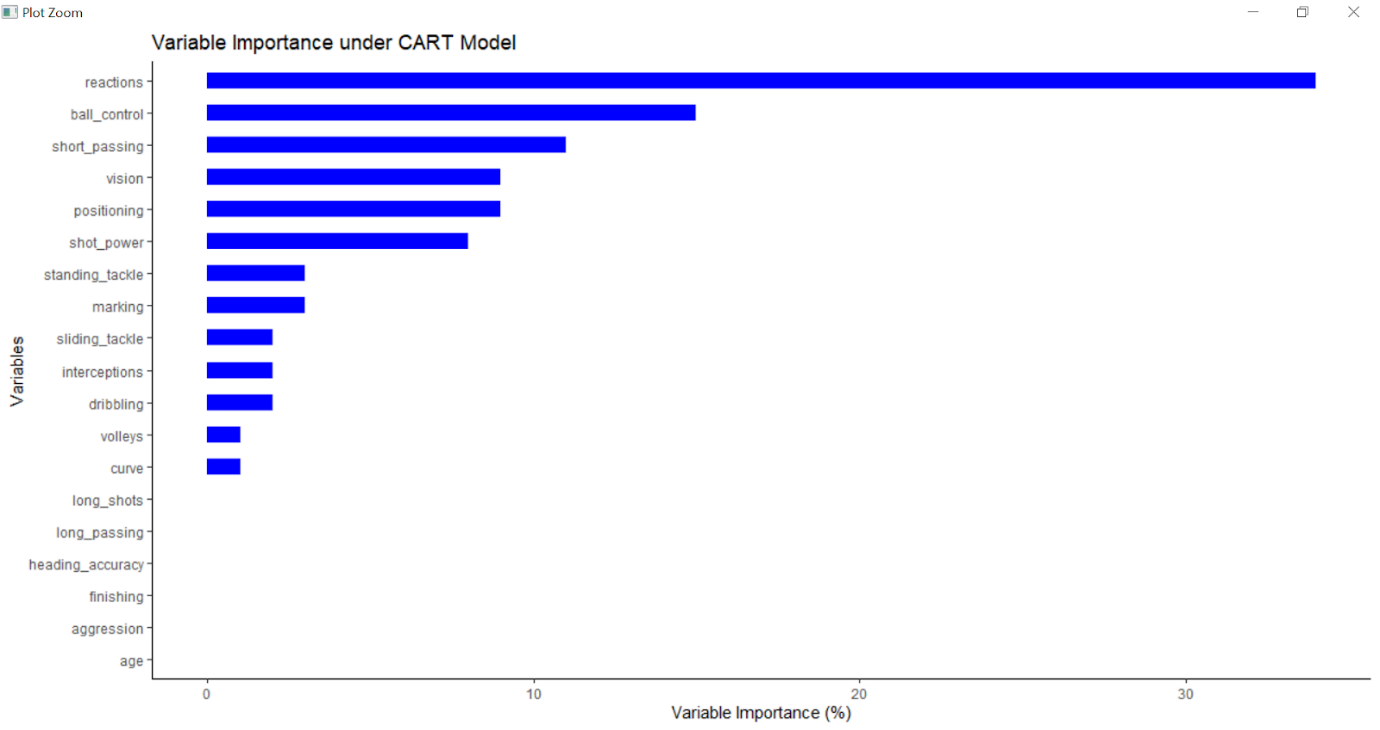
Appendix P2 – Bar graph in decreasing order of variable importance from the linear regression model



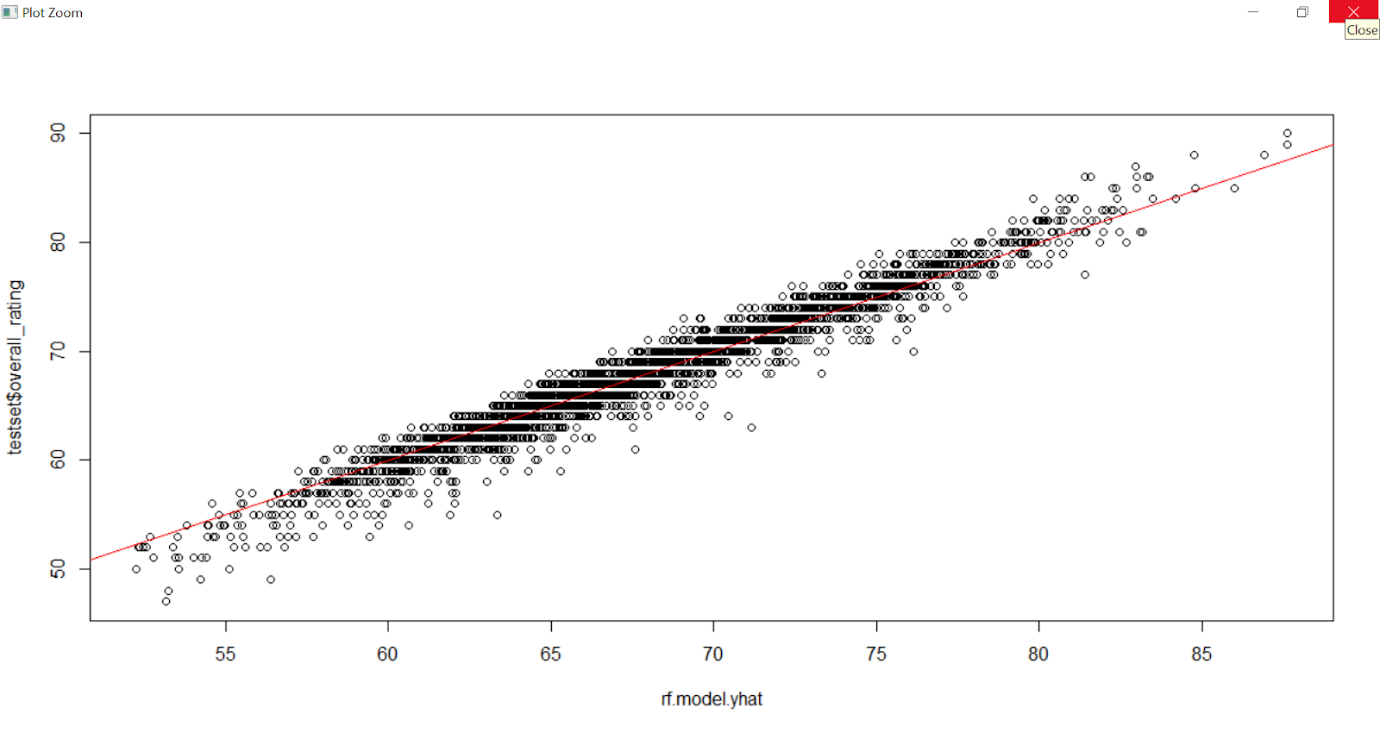
Appendix Q1 – Pruned CART Model



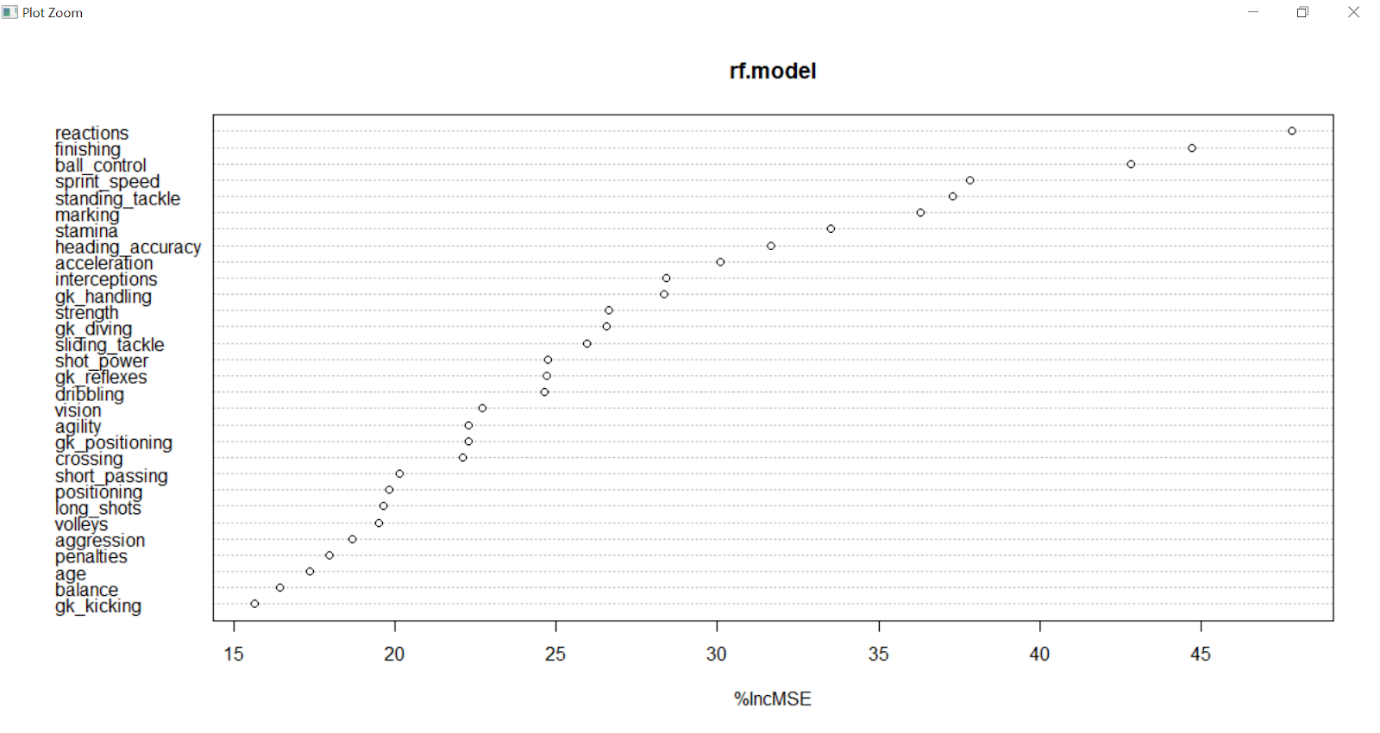
Appendix Q2 – Bar graph in decreasing order of scaled variable importance from the CART model



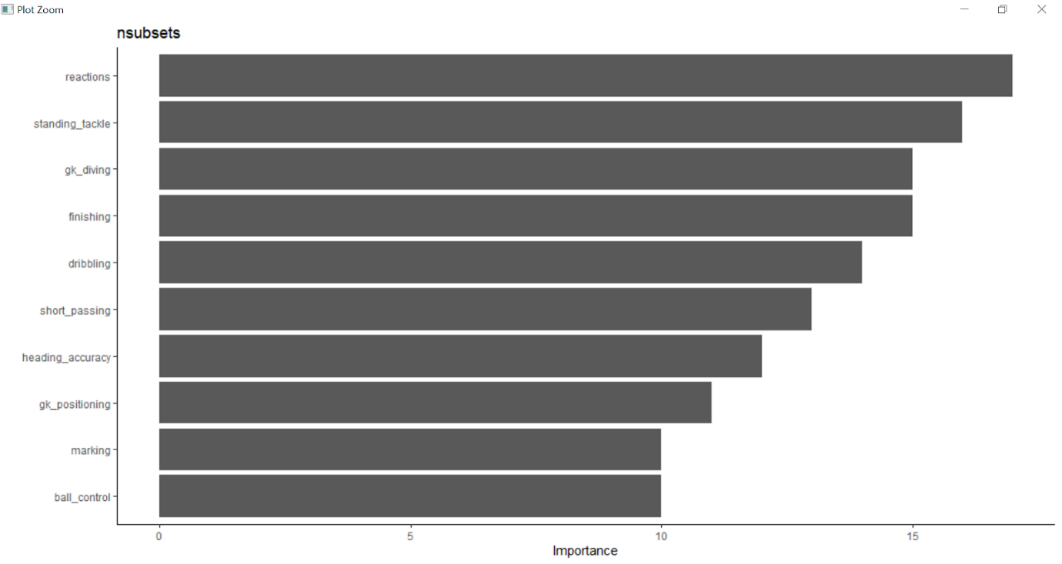
Appendix R1 – Scatterplot showing the distribution of residuals from the Random Forest model



Appendix R2 – Variable Importance Plot of X-variables in Random Forest model



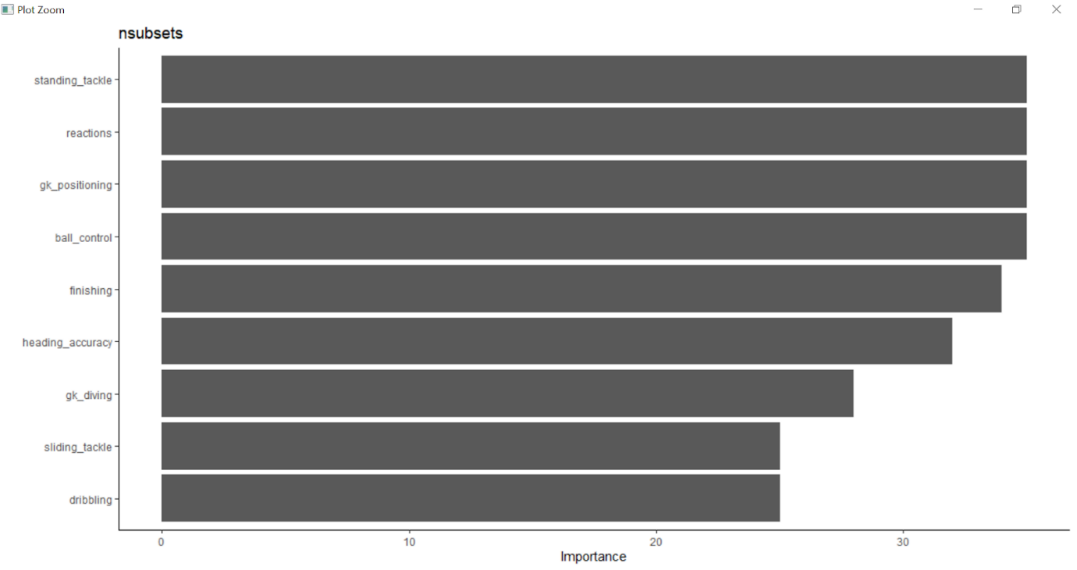
Appendix S1 – Variable Importance Plot of X-variables in MARS model (Degree = 1) by nsubsets



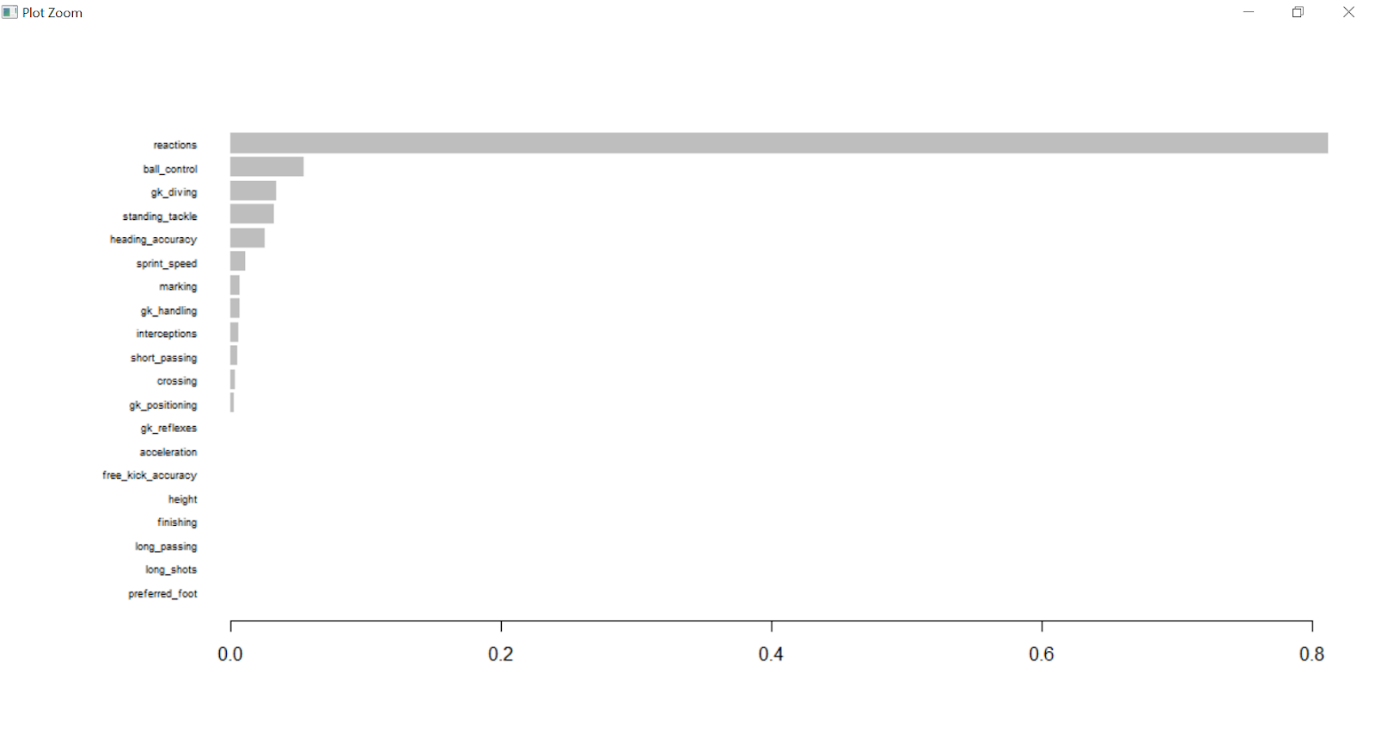
Appendix S2 – Scatterplot showing the distribution of residuals of the 2 MARS models



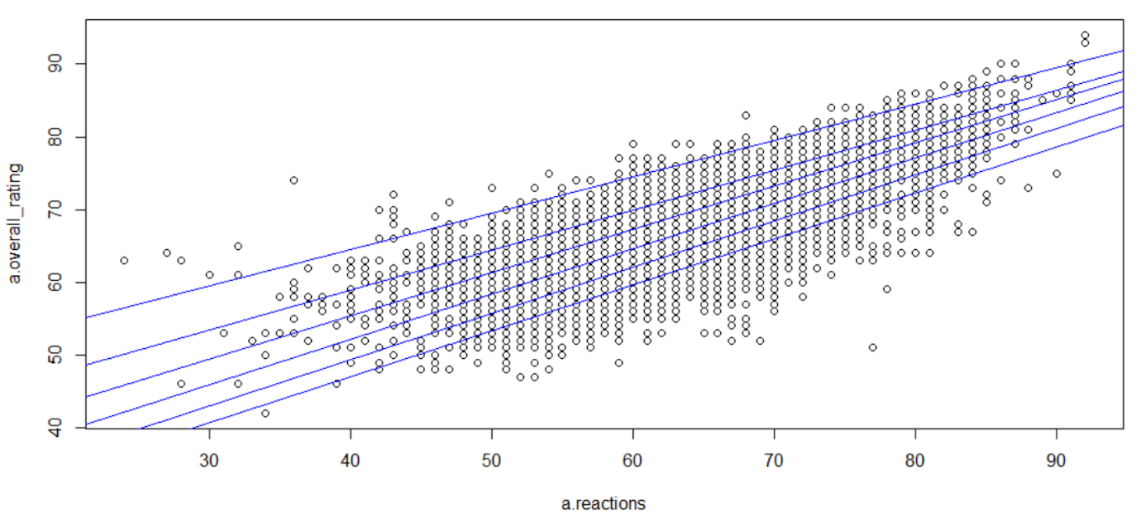
Appendix S3 – Variable Importance Plot of X-variables in MARS model (Degree = 2) by nsubsets

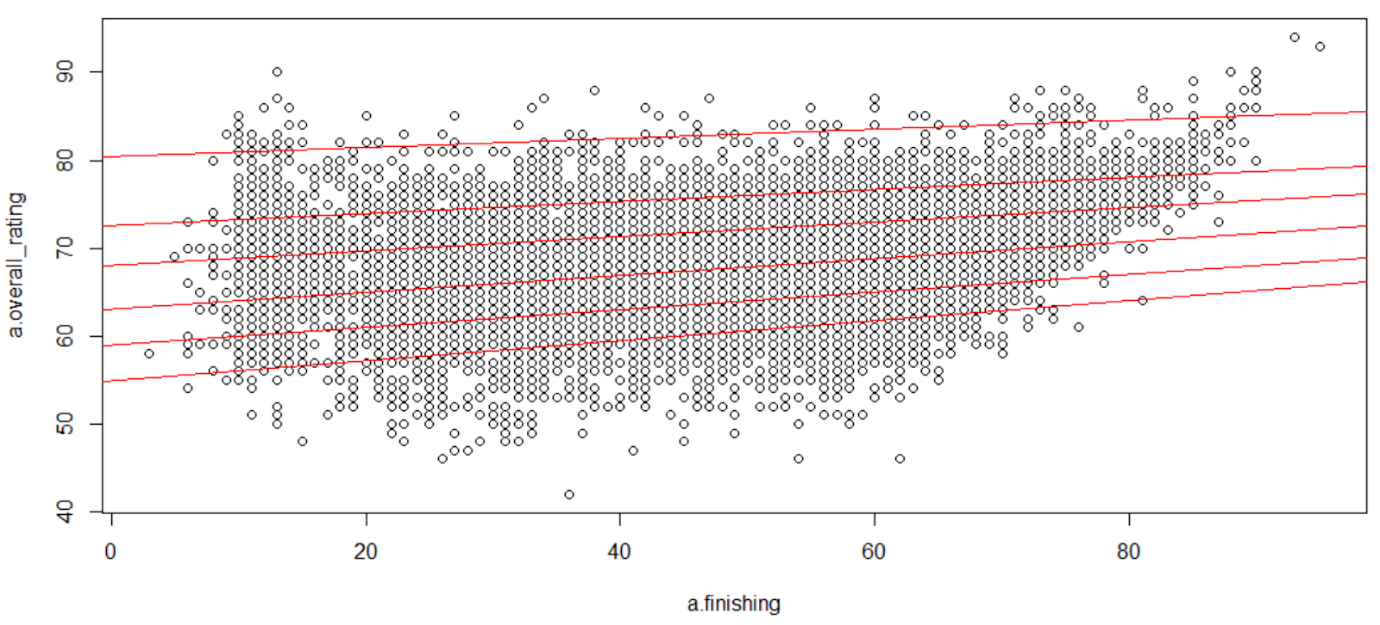


Appendix T – Variable Importance Plot of X-variables in XGBoost model by Gain

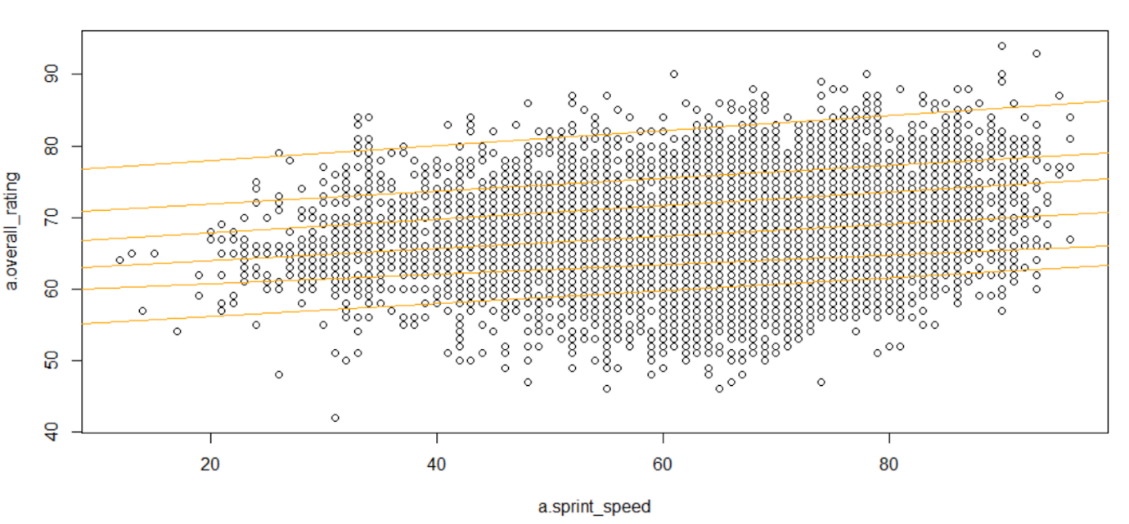


Appendix U1 – Quantile Regression: Reactions

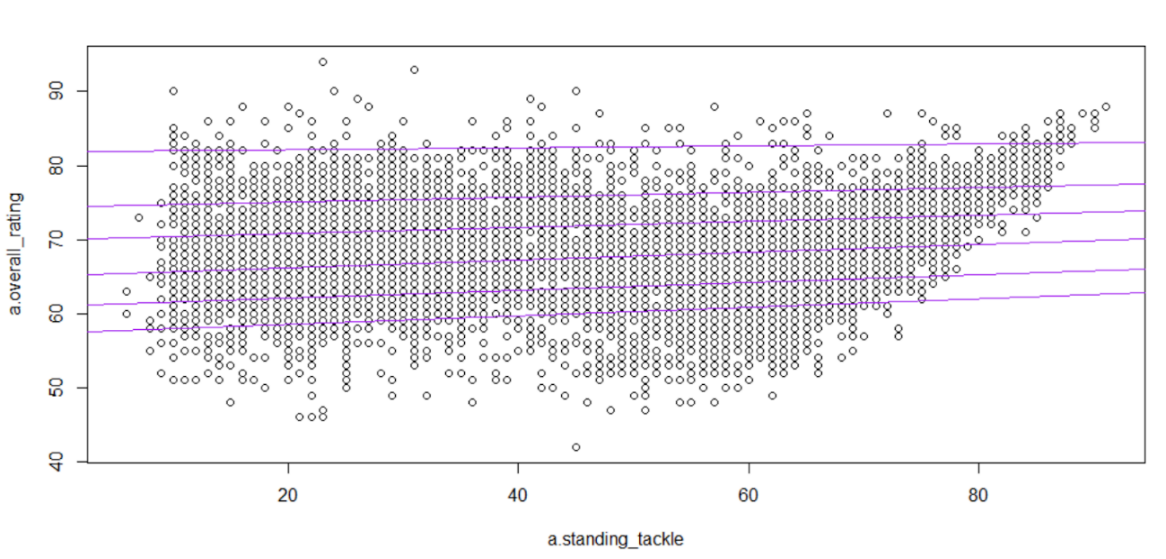


Appendix U2 – Quantile Regression: FinishingAppendix U3 – Quantile Regression: Ball Control

Appendix U4 – Quantile Regression: Sprint Speed



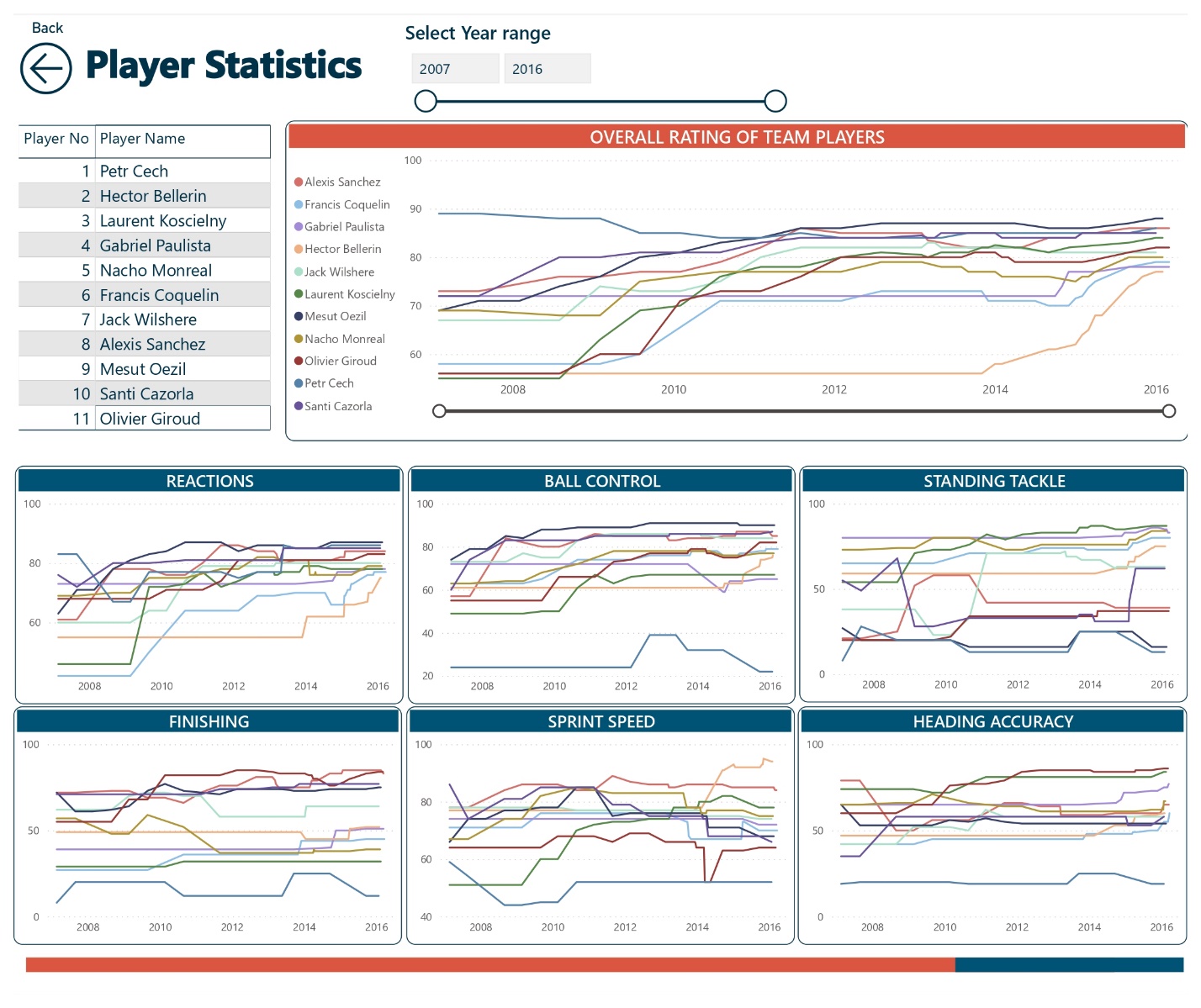
Appendix U5 – Quantile Regression: Standing Tackle



Appendix V1 – Dashboard Team Page



Appendix V2 – Dashboard Player Page



References

Panch, R. (2018). We Didn't Want to Lose to Man City in the Way Arsenal Did, Claims Antonio Conte. Retrieved 30 March 2021, from <https://www.90min.com/in/posts/5994723-we-didn-t-want-to-lose-to-man-city-in-the-way-arsenal-did-claims-antonio-conte>

FIFPlay (n.d.). FIFA Encyclopedia. Retrieved 30 March 2021, from <https://www.fifplay.com/encyclopedia/>

Saunders, T. (2019, September 23). FIFA 20 career mode changes | Player potential and more. GameRevolution. Retrieved 30 March 2021, from [https://www.gamerevolution.com/guides/599481-fifa-20-career-mode-changes-player-potential-manager-interactions](https://www.gamerevolution.com/guides/599481-fifa-20-career-mode-changes-player-potential-manager-interactions#:~:text=For%20as%20long%20as%20we,from%20their%20initial%20OVR%20rating)

Frost. (2021, March 24). Identifying the most important independent variables in regression models. Statistics by Jim. Retrieved 30 March 2021, from <https://statisticsbyjim.com/regression/identifying-important-independent-variables/>

Das, S. (2017, February 10). CART and Random Forests for practitioners. Retrieved 30 March 2021, from <https://dzone.com/articles/cart-and-random-forests>

Brownlee, J. (2016, August 17). A gentle introduction to XGBoost for applied machine learning. Machine Learning Mastery. Retrieved 30 March 2021, from <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>

MLS Soccer (2020). Analytics Guide: 3 Real-Life Examples That Show How Analysts Impact Their Clubs. Retrieved 30March 2021, from: <https://www.mlssoccer.com/soccer-analytics-guide/2020/three-real-life-examples>

Norwich City Football Club (2016). The Simple Training Drill That Went Viral – Here’s THAT Spikeball Clip. Retrieved 30 March 2021, from: <https://www.youtube.com/watch?v=UwpR5vc6KQU>

TopGoalkeeping (2019). How Can Goalkeepers Improve Their Reactions and Reflexes. Retrieved 30 March 2021, from: <https://topgoalkeeping.com/goalkeeper-improve-reaction-reflex-saves/>

SoccerCoachWeekly (2018). Turning Drill To Improve Strikers’ Reaction Time. Retrieved 30 March 2021, from: <https://www.soccercoachweekly.net/soccer-drills-and-skills/soccer-drill-to-improve-players-reaction-time/>

Taylor, Mark (2017). Why Bettors Should Look at Expected Goals in More Detail. Retrieved 30 March 2021, from: <https://www.pinnacle.com/en/betting-articles/Soccer/expected-goals-and-big-chances/QFU2JE3AQPK8VQV7>

Fantasy Football Scout (2020). Expected Goals Part One: What is xG and How is it Measured. Retrieved 30 March 2021, from: <https://www.fantasyfootballscout.co.uk/2019/06/14/expected-goals-part-one-what-is-xg-and-how-is-it-measured/>

BBC (2019). Financial fair play: All You Need to Know About How It Works. Retrieved 30 March 2021, from: <https://www.bbc.com/sport/football/29361839>

Lee, G. J., Jung, J. J., & Camacho, D. (2021). Exploiting weighted association rule mining for indicating synergic formation tactics in soccer teams. *Concurrency and Computation: Practice and Experience*, 1–20. <https://doi.org/10.1002/cpe.6221>

Bradley, P. S., Carling, C., Archer, D., Roberts, J., Dodds, A., Di Mascio, M., Paul, D., Gomez Diaz, A., Peart, D., & Krustrup, P. (2011). The effect of playing formation on high-intensity running and technical profiles in English FA Premier League soccer matches. *Journal of Sports Sciences*, *29*(8), 821–830. <https://doi.org/10.1080/02640414.2011.561868>