ISyE 6414 Regression: Team 16 Project Report

Team Members: Chinmay Palande, Patrick Murphy, Addison Rogers

I. Introduction

A. Background and Motivation

Soccer is one of the most popular sports in the world, with over 265 million players worldwide[1]. The English Premier League by itself generated around 3.2 billion pounds sterling total revenue in the 2015/2016 season.[2]

With so many people interested in watching (and spending money on) the sport of soccer, interest in predicting the outcomes of games has risen. Around 700 billion to 1 trillion dollars have been sunk in legal (and illegal) betting markets for predicting game outcomes per a BBC report.[3] Since the amount of money circulating around the results of matches is enormous, we consider the problem of how to best predict the outcome of matches.

The popular FIFA video game franchise attempts to capture the essence of soccer by deriving different player and team ratings and attributes using real life game data and human intuition in order to simulate matches. As soccer statistical research improves and video games continue to converge closer to real-life simulation, it is of interest to analyze the accuracy of video game ratings. This paper attempts to tackle the prediction problem by using these ratings and attributes to predict real-life game outcomes.

B. Summary of Contents

What follows is an exploration of the data set, exploratory analysis, multiple model selection approaches for choosing model features, fitting of a logistic regression model in R to determine the likelihood of a game being won based on FIFA attributes, and discussion of the results and potential explanations. The *a priori* assumption that the FIFA attributes would be decent predictors of real game world success will turn out to be supported. Discussion of the implications will follow, and ideas for future research will conclude the paper.

II. Methods

A. Data Source

The data source is a Kaggle soccer database including seven different datasets.[4] These datasets are a mixture of real-life soccer information, such as game statistics, and attributes from the FIFA video game franchise. Match data begins with the 2008/2009 season and continues through the 2015/2016 season. Because FIFA Team Attribute data begins with the 2010 season, match data from 2008 and 2009 was excluded from the analysis. The datasets utilized for this report are the following:

[1] FIFA Big Count Survey 2007

[2] Ahead of the Curve: Annual Review of Football Finance

[3] Football betting - the sport worth millions

[4] European Soccer Database

- Team Attributes - dataset with FIFA attributes for each soccer team. Attributes include a number of numerical and categorical variables such as Playing Speed, Team playing strategy, Defense Aggression and Pressure, etc. Although there are 18 numerical attributes and 24 categorical attributes, some of the numerical and categorical factors overlap. For example, chanceCreationPassing and chanceCreationPassingClass represent the same information but one is numerical and one is categorical.
- Match - dataset with with real life soccer game results. Dataset includes home team, away team, game score, season, data, country, and betting odds for each match.
- Player Attributes - dataset with FIFA attributes for each player. Attributes include individual player identification numbers, stamina, agility, preferred foot, overall ratings, position, etc.

B. Data Preprocessing

Several different approaches were developed during data preprocessing to try to find the best answer to the question if FIFA attributes are strong predictors for real-life soccer matches. The first approach was to combine all numerical factors from the Team Attributes and Match datasets above into a single set. A binary variable for whether or not the home team in a given game won was created. However, since a team can win, lose, or tie in soccer, the response was not actually binary. In order to create a binary dataset, draws were counted as losses, and the logistic regression calculated an estimate for the probability of winning.

The second approach taken calculated number of home wins, home losses, away wins, and away losses for each team in a given year and combined these columns with the Team Attributes data set. This newly created dataset contained one datapoint per team rather than one datapoint per game. It allowed transformation of the data using the Logit function to find the ratio of wins vs. total games played . This approach was removed from consideration because opponent strength would not be taken into account.

A third approach involved using only categorical variables with a binary win or lose (or draw) response variable so that logistic regression with replications could be used. This model failed Goodness of Fit deviance tests with values of approximately 0.

Counting draws as wins was considered, but the predictive power of this arrangement was estimated to be low. Therefore, the final approach was to merge the FIFA team attributes with the match dataset. This approach, while similar to the first with the same response, includes numerical variables in addition to categorical variables that were not already covered by the numerical variables. For example, buildUpPlayingSpeed shows 70, 50, and 30 where buildUpPlayingSpeedClass shows Fast, Balanced, and Slow respectively, so buildUpPlaySpeedClass was not added to the model.

Two columns in the final dataset contained a large number of NA values. Both buildUpPlayDribbling and buildUpPlayDribbling_away factors were missing over 60% of their data. These columns were removed from analysis as too much data was missing for imputation to be useful.

After preliminary experimentation with model selection, it was determined that further variables needed to be derived to improve model accuracy. It was observed that the independent variables with the highest individual predictive power were related to defense and goalkeeping. To capture this relationship, four goalkeeper ratings from the Player Attributes dataset were introduced: Home_GK_Rating, Home_GK_Potential, Away_GK_Rating, Away_GK_Potential. The *summarise* function was used to take the maximum rating recorded for each goalkeeper in a year as ratings are updated multiple times throughout a season. Additionally, interaction terms among predictors were included and tested in the model.

As all numerical data was already scaled from 0-100, it was determined that standardizing the data was was not necessary. Additionally, for the model selection methods used, the data is scaled within the respective *glmnet* library functions along with several categorical variables.

C. Exploratory Analysis

Multiple scatterplot matrices were obtained using *scatterplotMatrix* in order to evaluate the linear relationships among the predictors in the dataset. Box plots evaluating different categories against the response were used to assess impact of different categories on the proportion of winning versus losing/draw. A correlation matrix was obtained using *corrplot* to evaluate multicollinearity among predictors and correlation coefficients between predictors and the response. Cook's distances were calculated to test for point outliers according to the following equation:

$$D_i = \frac{(\widehat{Y}_i - \widehat{Y})^T - (\widehat{Y}_i - \widehat{Y})}{q\sigma^2}$$

Where $\widehat{Y}_i$ are fitted model values without the ith observation and $\widehat{Y}$ are fitted values with all observations.

D. Model Selection

Three different methods for model selection were used for this data set individually. The first model selection method utilized was backwards selection using the AIC criterion to find a reduced model with the lowest AIC.

For the other two methods, the dataset was divided into training (80% of the data, randomly sampled) and testing (remaining 20%) sets. The R function "glmnet" was used to build a LASSO model using 10-fold cross validation to find lambda with the minimum binomial deviance (lambda.min) and lambda for the minimum binomial deviance one sigma away from the minimum lambda (lambda.1se). A reduced model was obtained for each. An elastic net using 10-fold cross validation was also fit to the data to find a reduced model.

A classification threshold was determined, using the *cv.glm* function, in order to minimize misclassification errors. To accomplish this, cost functions were defined at different levels and then used to compute the error at different thresholds. These errors were compared in order to choose the threshold with the smallest prediction error.The resulting threshold was used to compare the three model selection methods for maximum prediction accuracy using *predict.cv* from the *glmnet* library.

The accuracy of the classification model is defined as the ratio of correctly classified responses to total data points. If the response is selected as 1 for every single game, the accuracy of the results is 45.6%. The accuracy results for the various models were compared against this baseline value to determine the benefits of each model.

E. Final Model:

The final model can be selected based on either of the three methods described above. The stepwise regression based on AIC as the selection criterion is a greedy algorithm that considers addition of one variable at a time. LASSO and elastic net approaches consider all variables simultaneously. However, LASSO regression has certain disadvantages when the data has multicollinearity. As the model will be used primarily for prediction, elastic net regularization is a good choice as it combines the advantages of LASSO regression (variable selection) and ridge regression (predictive power/encourages grouping effect of predictors[5]).

III. Results

A. Exploratory Analysis Results

Scatterplot matrices of the predictors yielded little evidence of linear relationships among the predictors of the model (see Appendix A1-A6). After preliminary analysis it was determined that additional factors were needed to improve the predictive power of the model, leading to the inclusion of the four goalkeeper predictors to the model (see Section II.B). These predictors demonstrated much stronger linear relationships (see Appendix A7). All model predictors displayed independence as there was no evidence of clustering on any plots.

The boxplots (see Appendix A8-A15) showed most mostly equivalent proportions of wins across all categorical variables. This suggests low individual predictive power for the categorical variables.

Other than GK_Rating and GK_Potential for home and away teams, no factors have a correlation with home_win above 0.07. Minimal multicollinearity was observed in the correlation matrix (see Figure 1). Only four coefficients exceed 0.3: GK_Rating and GK_Potential for both the home and away teams were observed to have correlation coefficients of 0.87. Because of this multicollinearity, GK_Potential variables for home and away teams were removed from the model.

Cook’s distance calculations resulted in distances no greater than 0.0006 (see Appendix A16). No points in the data set appear to be obvious outliers.

[5] Regularization and Variable Selection via the Elastic Net

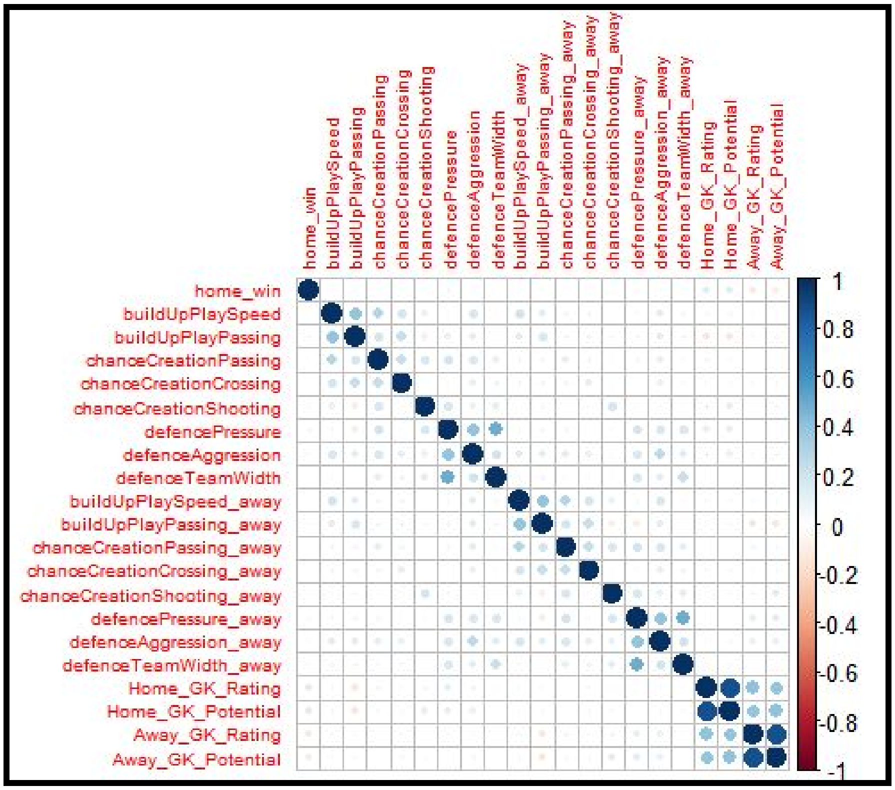

Figure 1: Correlation Matrix Plot

B. *Model Selection Results*

As previously discussed in Section II.D, the accuracy baseline for comparison is 45.6%. What follows is a discussion of the results of each individual model selection method.

stepAIC:

Stepwise regression using the AIC criterion selects 18 variables with an AIC of 22154 reduced from 23235. The threshold for probability to classify the response as 1 or 0 is taken to be 50% as it gives the minimum misclassification error. Using this threshold, the model selected using stepAIC gives an accuracy of 60.85%. Figure 2 shows a plot of the cross validation error thresholds by classification error. This threshold is shared among all models.

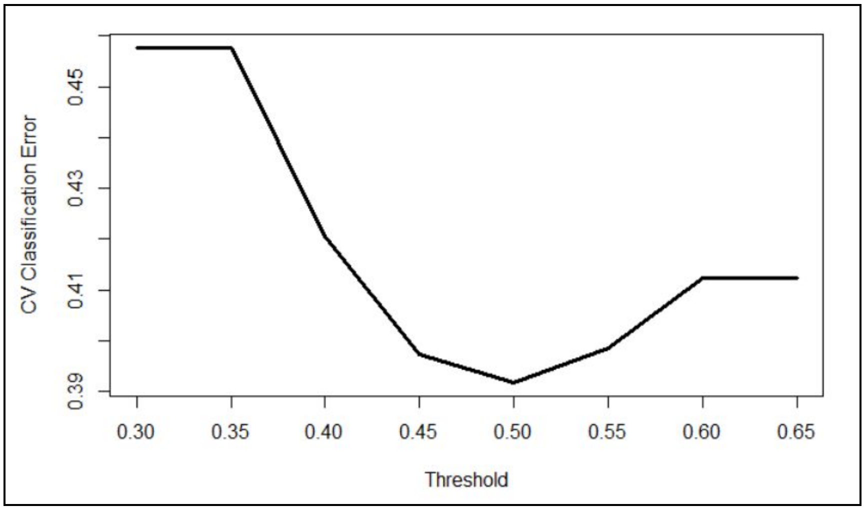

Figure 2: Error Threshold Plot

LASSO:

As described in Section II.D, two different lambda parameters from *cv.glmnet* were tested for model selection. The first lambda for minimum binomial deviance (*lambda.min = 0.001710454*) gives a model with 24 variables. The second lambda (one sigma away from the minimum binomial deviance lambda.1se = 0.009128181) gives a model of 13 variables. Both models give an accuracy of 61.6% on the test dataset. The simpler model with fewer variables was selected because of the equivalent accuracy. See Appendix B1 for a plot of the Binomial Deviance by log(lambda) using LASSO.

Elastic net:

The elastic net model gave similar results to the LASSO model. For a weight coefficient alpha of 0.5, *lambda.min (=0.003117005)* and *lambda.1se (=0.01663452)* select 14 and 24 variables respectively. The accuracy for both parameters is 61.7%. The simpler model with fewer variables was selected because of the equivalent accuracy. See Appendix B2 for a plot of the Binomial Deviance by log(lambda) using Elastic Net.

It should be noted that all potential interaction terms were considered in order to improve the prediction accuracy of the models. However, none of these terms were selected by any of the methods mentioned above.

C. Regression Results

The final logistic regression model gives the probability of the home team winning based on the variables below. The model is built using Elastic net regularization with equal weight for ridge regression and LASSO components.

Model: $log(p/(1-p)) = \beta_0 + \Sigma_j \ \beta_j X_j \ ; j \ \in \{1, 2, ... , 14\}$

Assumptions:

1. Linearity assumption: $log(p/(1-p)) = \beta_0 + \Sigma_j \ \beta_j X_j$

2. All the response variables are independent random variables
3. Link function is given by $g(p) = log(p/(1-p))$

The estimates for regression coefficients corresponding to the variables were calculated as follows:

Variable	**Coefficient Value**
(Intercept)	-0.326726139
buildUpPlayPassing	-0.001732058
chanceCreationShooting	0.000311557
defencePressure	0.005833904
defenceAggression	0.001140971
buildUpPlayPassing_away	0.002331441
chanceCreationCrossing_away	-0.001385744
defencePressure_away	-0.002671476
defenceAggression_away	-0.002850468
Home_GK_Rating	0.045485974
Away_GK_Rating	-0.042499476
buildUpPlayPositioningClassOrganised	-0.226330632
chanceCreationPositioningClassOrganised	-0.248275397
buildUpPlayPositioningClass_awayOrganised	0.140126723
chanceCreationPositioningClass_awayOrganised	0.222823329

Table 3: Regression Results

This logistic model does not contain replications and as such it does not have valid residual values. For this reason it is not feasible to conduct any goodness of fit tests.

IV. Conclusion

A. Regression Interpretation

Given a prediction accuracy of 45.6% if the home team is picked to win every game, a prediction accuracy of 61.7% is adequate. Because of having three different game results (win, lose, tie), soccer predictions are a lot different than other sports like American football, basketball, and baseball. Therefore, it is difficult to evaluate the value in the 61.7% prediction accuracy. It is safe to state this level of accuracy in sports betting against the spread would be very impressive, but the model does not factor into account the spread of games or the margins of victories. Therefore this model could be appropriate for moneyline betting (win vs. lose) but is not suitable for point spread betting or over/under betting, and it is somewhat unclear if 61.7% is impressive for moneyline betting with three results.

Several features of the model are worth examining in some detail:

- As the coefficients have been calculated based on elastic net regularization, there are no standard errors or appropriate confidence intervals for the regression coefficient estimates. This leads to difficulty in evaluating the statistical significance of specific coefficients. The validity of the coefficients can only be evaluated based on the accuracy of predictions.
- Each coefficient can be interpreted in terms of the log odds ratio. For example, one unit increase in the Home_GK_Rating increases the log odds of home team winning the game by 0.045 given all other variables are held constant. A similar interpretation can be used for all the selected variables in the model.
- If a variable is included for home team it is also included for the away team. It is also evident that if a home attribute has a positive or negative association with log odds of home team winning, the equivalent attribute for the away team has the opposite association.
- Given all other predictors, home and away goalkeeper rating have a the highest impact on the log odds ratio given all other variables in the model. These are closely followed by the categorical variables buildup play positioning and chance creation positioning for both the home and away teams. It is slightly counter-intuitive to observe that given all other variables, home team that are “Organized” in these categorical variables have lower log odds of winning. This is most likely due to scaling since these variables only take on values of 0 or 1.

B. Findings and Further Investigation

The purpose of this analysis was to analyze the predictive power of FIFA attributes on real-life soccer results by attempting to predict match outcomes. It is clear from the results that FIFA team attributes alone have limited predictive power for real-life matches. Model predictive power improved from 58% to 62% when including GoalKeeper attributes. Although there is insufficient evidence to support the notion that FIFA team attributes do not have predictive power on real matches, it was shown that these attributes by themselves are limited.

Although FIFA Team Attributes have limited predictive power, it does not mean FIFA cannot improve accuracy for predicting the outcome of matches. The addition of player attributes and betting odds may lead to improved predictive power. Specifically, the Match dataset has team lineup strategy and the Player Attributes dataset has specific player ratings. Comparing offensive players on one team vs. defensive players on the other team may allow for better predictions. Implementing this strategy would take advanced knowledge of soccer tactics, formation, and play, but the potential gain is likely valuable. As a result, the answer to the problem statement of whether FIFA video game ratings can predict real-life outcomes is somewhat inconclusive.

A potential implication from this analysis is need for a thorough investigation of the methods taken by FIFA to rate players. Currently, FIFA collects game statistics, guesses a starting attribute for each player's skills, then reviews ratings with "coaches, professional scouts, and a lot of season ticket holders."[6] All the data is put into a formula and a rating is calculated.

Calculating video game ratings from real-life soccer statistics is difficult but there are ways to potentially improve the methods. An improvement could be derived from incorporating teamwork abilities and player combination statistics. For example, if a midfielder on the team is really strong at cross passes, the attacker on the team will perform better. If the midfielder is not as strong at crossing, that attacker's rating will not be as important because he likely will not receive the ball in high-probability scoring situations. Rather than just looking a specific player statistics and human inputs to calculate player and team ratings, FIFA could incorporate a teamwork model that improves ratings when certain people are on the field together.

In addition to accounting for player positioning, three other major areas present themselves for further investigation:

1. **Events within the game**
   This paper attempts to predict the outcome before the game using the ratings and attributes of the teams. However, there are many real time events during matches, such as a player receiving a redcard or injuries/substitutions, that may significantly affect game outcomes. Incorporating this type of real time event data in a future investigation could potentially help in explaining the large number of misclassifications.

2. **Multinomial logistic regression**
   As match results are not binary (win/lose) but instead have three classes(win/lose/draw), a multinomial logistic regression approach may be more appropriate and may lead to more accurate results.

3. **Comparing these results with betting odds**
   The betting industry has been trying to predict the outcomes prior to the game and may have a lot of conventional wisdom when it comes to these predictions. A next step to the study could be to compare the model results with bettings odds offered by various agencies.

[6] "EA explains how FIFA player ratings are calculated"

Appendix

A: Exploratory Analysis Results

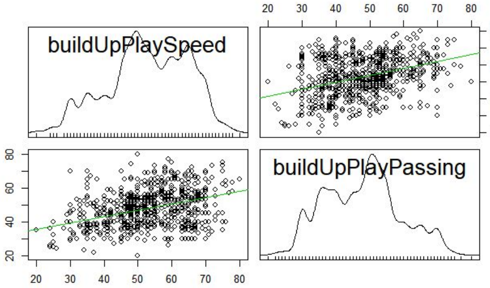

Figure A1: Home Team buildUpPlay

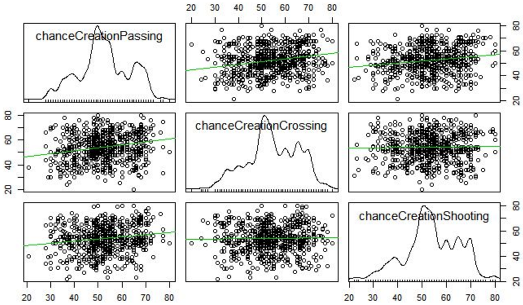

Figure A2: Home Team chanceCreation

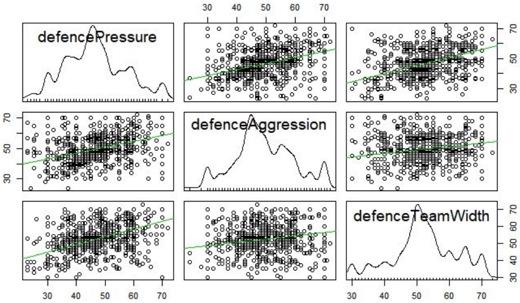

Figure A3: Home Team defenceRating

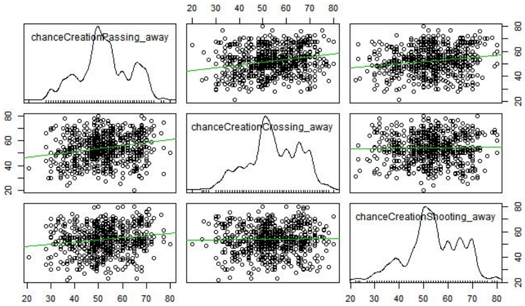

Figure A4: Away Team chanceCreation

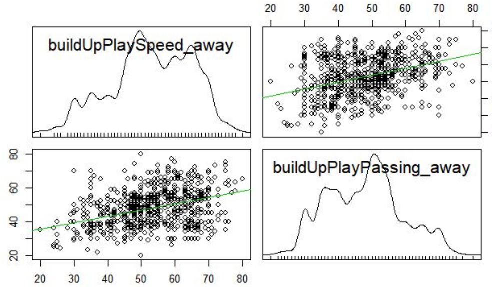

Figure A5: Away Team buildUpPlay

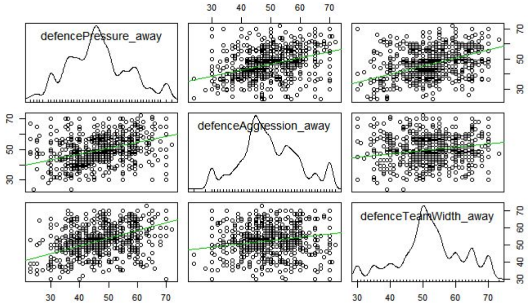

Figure A6: Away Team defenceRating

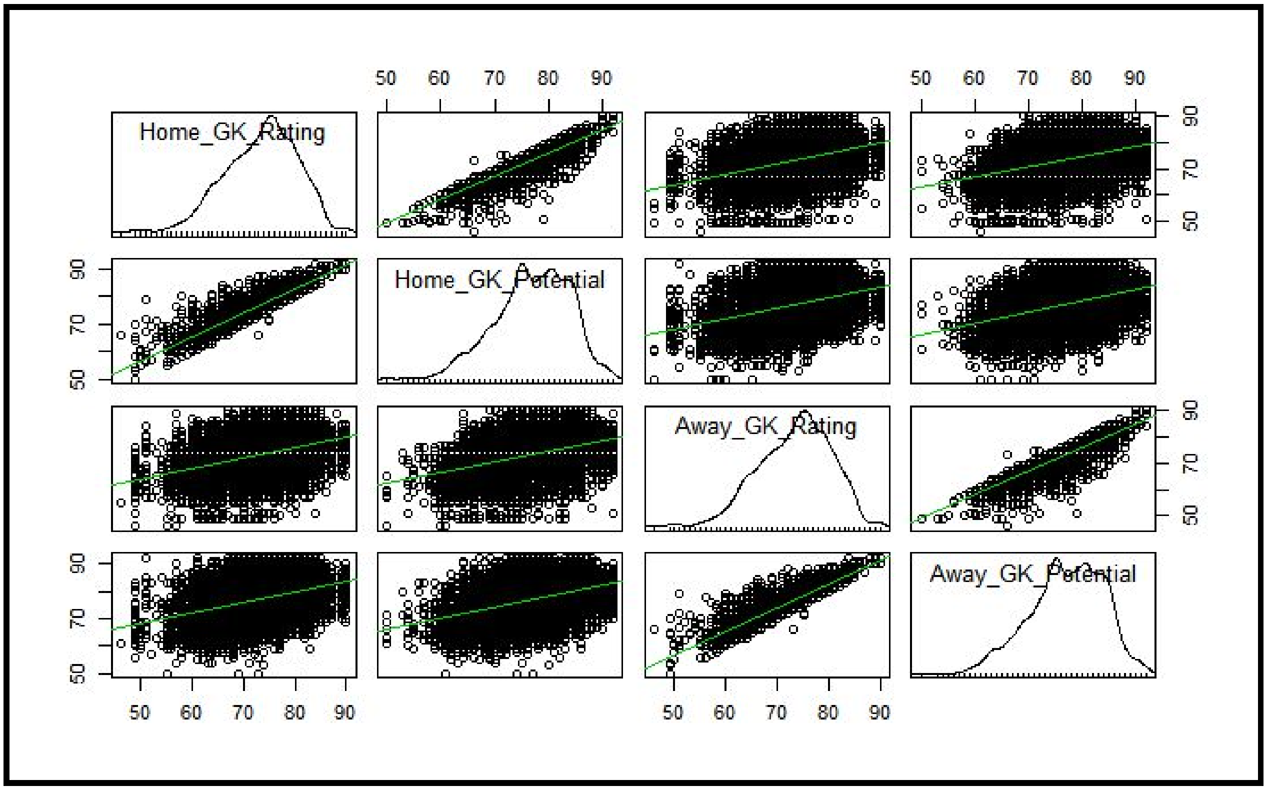

Figure B7: Home and Away GK Scatterplots

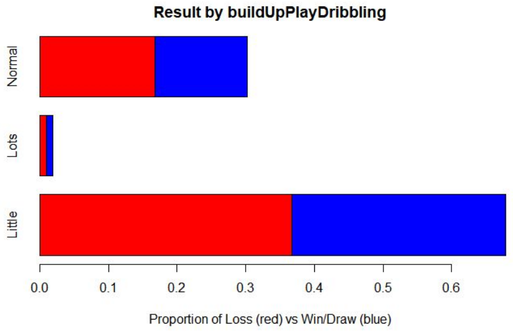

Figure A8: buildUpPlayDribbling Boxplot

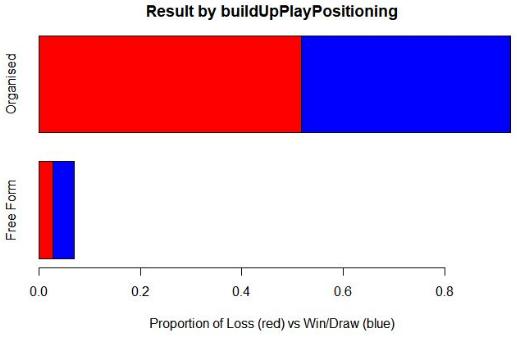

Figure A9: buildUpPlayPositioning Boxplot

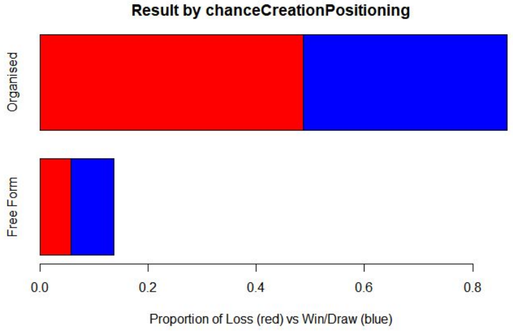

Figure A10: chanceCreationPositioning Boxplot

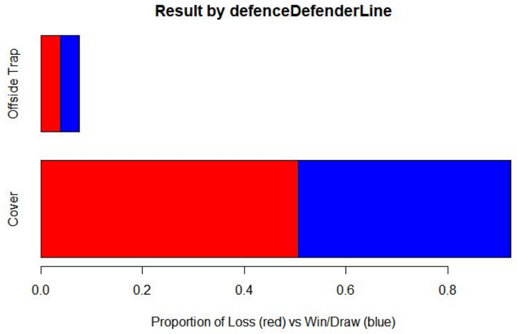

Figure A11: defenceDefenderLine Boxplot

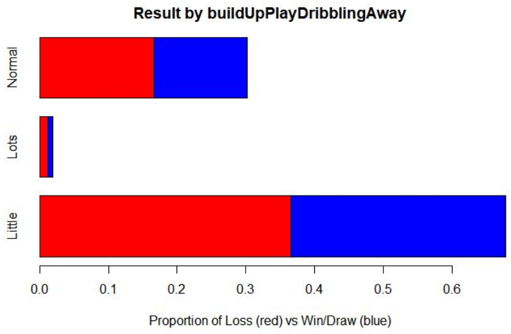

Figure A12: buildUpPlayDribblingAway

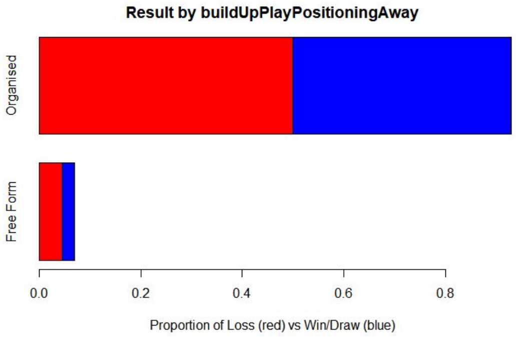

Figure A13: buildUpPlayPositioningAway

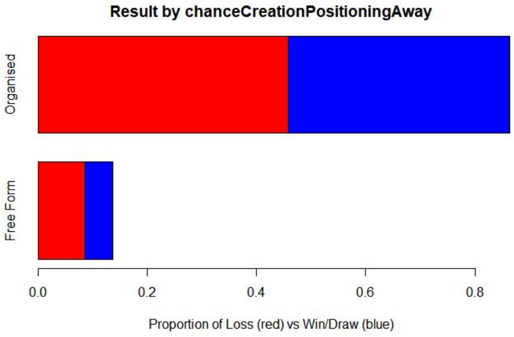

Figure A14: chanceCreationPositioningAway

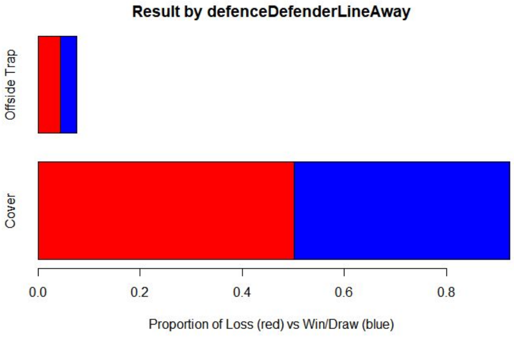

Figure A15: defenceDefenderLineAway

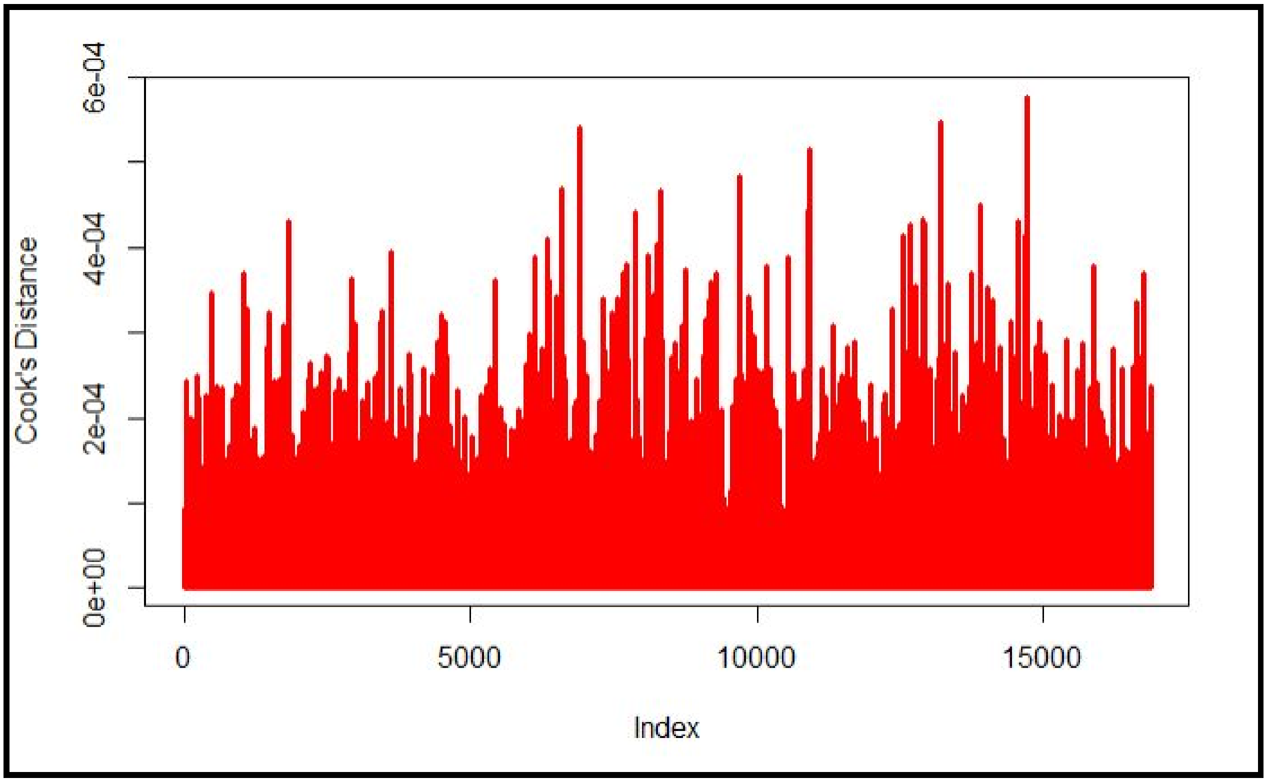

Figure A16: Cook's Distance

B: Model Selection Results

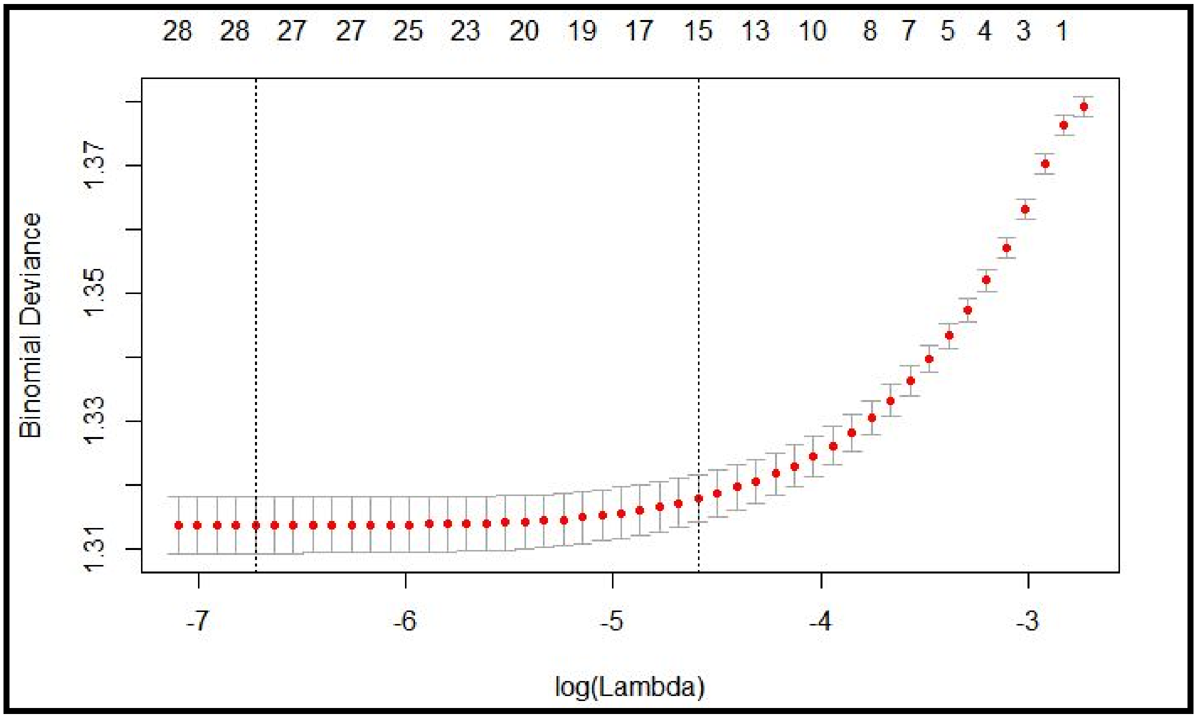

Figure B1: Binomial Deviance by log(Lambda) using LASSO

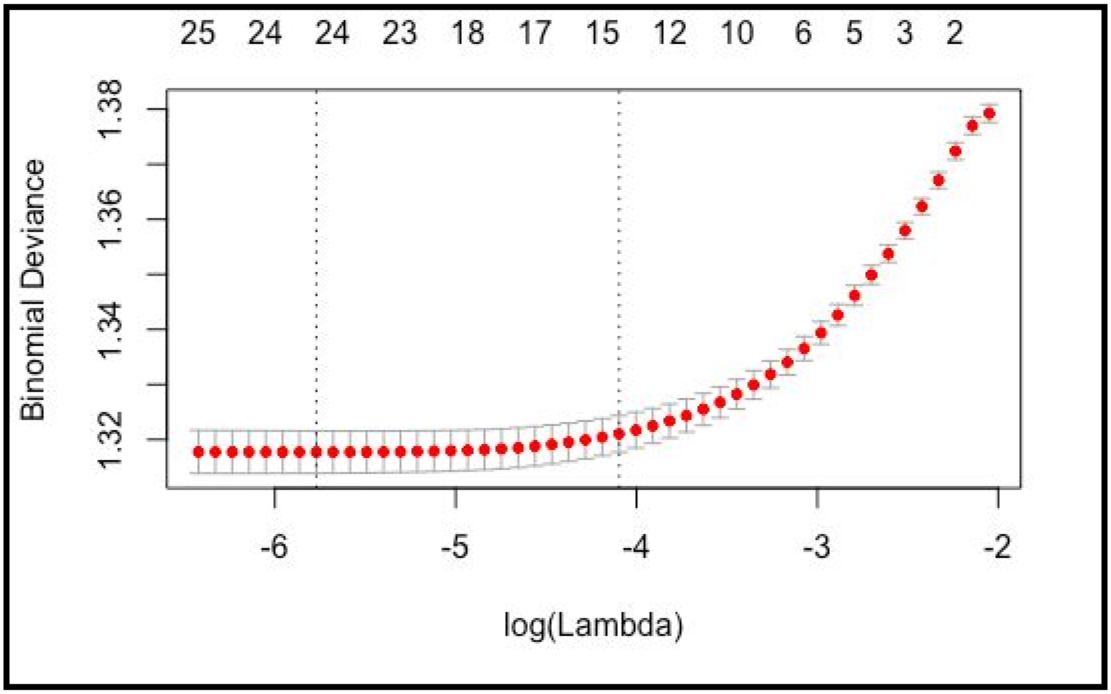

Figure B2: Binomial Deviance by log(Lambda) using Elastic Net