Statistical Analysis of Organized Attacks in Football

Håkan Wahlström

Master's thesis
2022:E61

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Master's Theses in Mathematical Sciences 2022:E61
ISSN 1404-6342

LUTFMS-3457-2022

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/

Abstract

This report investigates what defensive and attacking strategies are most efficient in an organized attack situation in football. Defensive shape, defending player synchronizations, number of passes, types of passes and other variables are tested to see which methods statistically generate more and better goal scoring opportunities. Data from the Eredevise (highest Dutch football division) 2021/2022 season is investigated where organized attack attempts are found. The data is then filtered to find defensive and attacking team characteristics of each attack attempt respectively. Three regression methods are then used to statistically analyse what strategies are the best predictors for whether an attack is successful.

In the Eredevise 2021/2022 season, playing crosses was the best way to improve quality of chances created and a good way to create more chances. Keeping a high movement speed among the players playing in attacking positions also tends to generate more and better chances. From a defending team perspective, moving synchronously and spreading out the team over greater areas are good ways to prevent opponents from scoring.

Preface

Acknowledgements

This thesis couldn't have been done without the constant help of professor David Sumpter. From helping with connections in Ajax and Hammarby, to giving valuable insights and guidance when creating the project and organizing regular meetings, you are the whole reason this project exists. Thank you from the bottom of my heart David!

I'd also like to thank the data analytics team at Ajax for their consistent feedback, ideas and help in moving this project in the right direction. Big thanks to Mirjam, Vosse and Ya'gel for your enthusiasm and efforts in helping to create this thesis. I'd also like to thank Abel Lorincz for valuable feedback and insights especially during my visit to Årsta!

Thank you to Aleksander and Joakim for being my colleagues and friends throughout this project, you've been incredibly supportive especially in sharing code and the work you created. I hope to work with you again in the future!

Lastly, I'd like to thank Johan Lindström for supervising the project on Lund University's side, and Anna Lindgren for being the examinator.

1 Introduction

This report aims to investigate what strategies are most successful when attacking against an organized defense. In the world of football today, there are many tactics and strategies available on how to play the game. Different play styles have their advantages and disadvantages, and the game is constantly evolving. A game situation that occurs several times per match is that one team controls possession of the ball in the opponents half, and the defending team is well organized. These are often controlled situations where tactics and strategies play a big part, both offensively and defensively. Therefore, it's interesting for teams to understand what the most efficient strategies are both for attacking and defending in this type of situation.

In recent years, data analytics has gained increasing importance and usage within clubs as a way to better understand the game, and gain a competitive advantage on opponents. In this project, I will use data from the 2021/22 season as a foundation for making a statistical analysis on what methods are most successful when attacking against an organized defense.

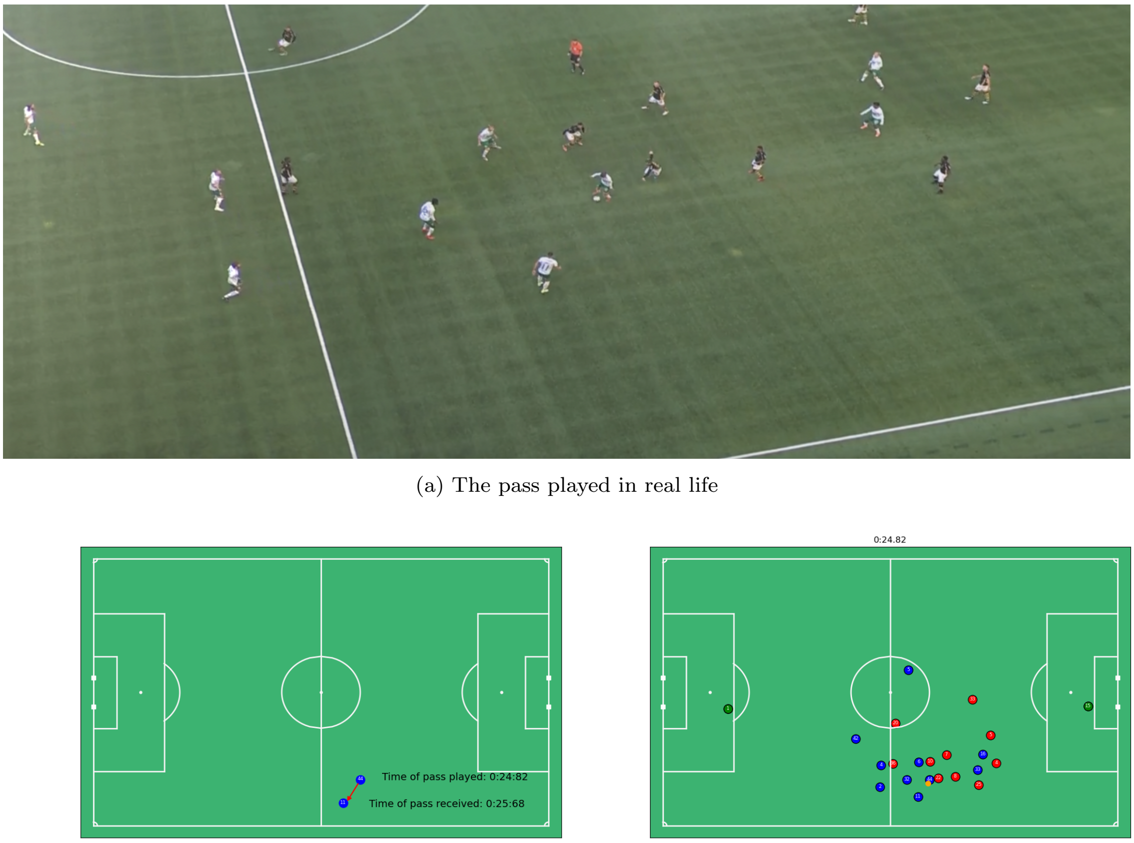
This project's method of finding successful attack strategies uses a combination of event and tracking data as input - raw information about events happening in games and player positioning on the field (discussed in chapter 2.1). The data is filtered so that only the instances in games where a team is defending in their own half is investigated - each such instance is called an attack attempt. Then, information on attack attempts is further processed through algorithms aimed to describe characteristics of an attack based on known football concepts. The football concepts are described in chapter 2.2 and the variables are described in chapter 4. For example, each attack is investigated in terms of number of passes played by the attacking team, average length of passes, average movement speed of attackers and so on. The product is a data set of attack attempts and their characteristics, which mostly consist of different attacking and defending strategies. To find which strategies are more or less successful, two different regression methods were used and which are discussed in chapter 3.6. First, a logistic regression with response variable being whether an attack ended up with a shot or not. Then, a generalized linear regression used only on attacks where a shot is taken, and with response variable being a transformation of expected goals — the probability that the shot ends up being a goal. For both of the regression techniques, some model selection methods are applied to remove characteristics which have a rather small impact on attack success, which is described in chapter 3.6 and further in chapter 4.8 and 4.9.

The results are presented in the results section, where specifically the size of coefficient estimations and statistical significance are interesting to look at. A discussion about the implications of the study and it's results is made, as well as some possible future work based on this project.

2 Data and Football Concepts

2.1 Raw Data

Access to good and reliable data is essential for this project. With the help of Swedish professional football club Hammarby IF and Dutch professional footbal club AFC Ajax, I have gained access to *event data* and *tracking data*. Event data contains information on events that happen with the ball in a match. For example, every pass is logged with location on the field (x,y -coordinates), time of the pass, player performing the pass, whether it was successful, what body part was used to make the pass as well as other characteristics of the event. This data only includes information about one or two players at a time. Therefore, tracking data is used to complement event data. Tracking data contains information about all the players' and ball's position on the field at 25 Hz. However, tracking data does not specify what is happening with the ball, who controls it etc. It just contains information about the position of the ball and players. The two types of data complement each other very well.



(a) The pass played in real life

(b) Visualization of event data

(c) Visualization of tracking data

Figure 1: Visualization of the data types available, tracking and event data

Event data was collected from wyscout [1], a well known data provider for top clubs. Tracking data was provided by Tracab [3]. In this project, both event data and tracking data were consistently and widely utilized.

2.2 Football Concepts

Some general football concepts and ideas are required in order to make a statistical analysis of how to attack against an organized defense. First of all, when defending the most common strategy is to do so in three "lines": a back-line, a midfield and an attacking line. See an illustration example in figure 2. The number of players in each line varies depending on team, as does the space they cover and how high up the field they stand. The common theme across almost all teams is that there are precisely three defensive lines.

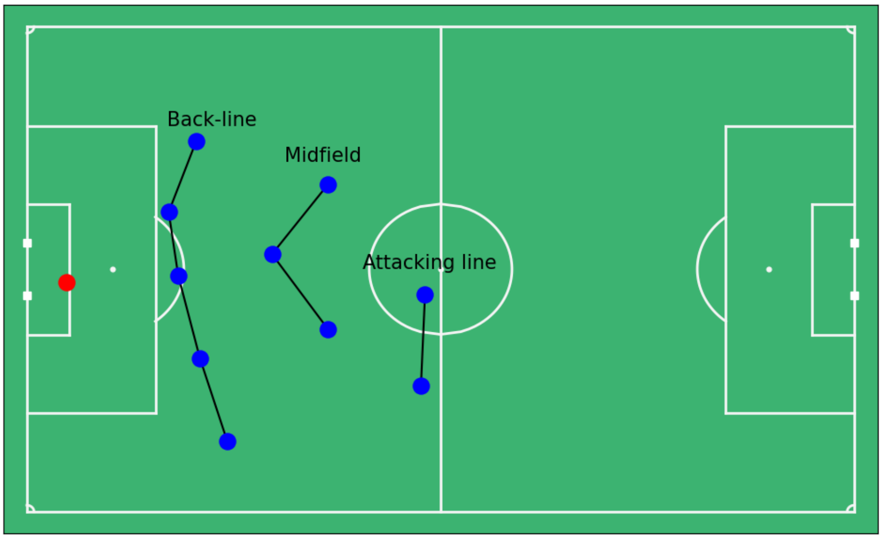


Figure 2: The three lines of defence illustrated

Second, the most common way of defending among professional teams today is what's called zonal-defending. It is based on the idea that there are positions on the field you as a defending team try to control, and each defending player has some area which he is responsible for. An important space to control is that covered by the back-line and midfield. Allowing the opponents to play freely in the space between the midfield and back-line is considered very dangerous. From the attacking team perspective, contesting this space and playing inside it could be a great way to create chances.

Today there are several concepts in football that are considered by various experts as keys to breaking down organized defenses. For example, Swedish Viasat expert Martin Åslund described when commentating on the game between Tottenham Hotspur and Burnley in the Premier League on the 15th of May 2022 how, in his opinion, Tottenham didn't play the ball fast enough and didn't switch sides enough to create goal scoring opportunities. So in his opinion, pass tempo and switching sides quickly are important factors when trying to attack against an organized defense. Other factors that are often said to affect how to successfully break down low block defenses are movement speed of attackers/making disruptive runs and how to position the attacking players among other things.

3 Statistical Methods and Theory

3.1 Pitch Control

One of the keys to understand defensive teams' organization is what parts of the pitch they control. To do so, this project utilizes Manchester City's lead data scientist Laurie Shaw's method of calculating pitch control [4]. The essence of this method, is that each point of the field is assigned a value based on which team would most likely control the ball if it travels directly to that point. Shaw's algorithm uses player current positions, player velocity and movement direction, reaction times, acceleration speeds, top speeds and time to control the ball as inputs in evaluating the probability of players from each team controlling the ball. The mathematics of Shaw's pitch control model is developed from Liverpool FC's head of data Science William Spearman, and was introduced in his article beyond expected goals in 2018 [5]. It makes the assumption that while a player is close to the ball, his ability to control the ball can be treated as a Possion point process. A visual illustration of the pitch control model can be seen in figure 3.

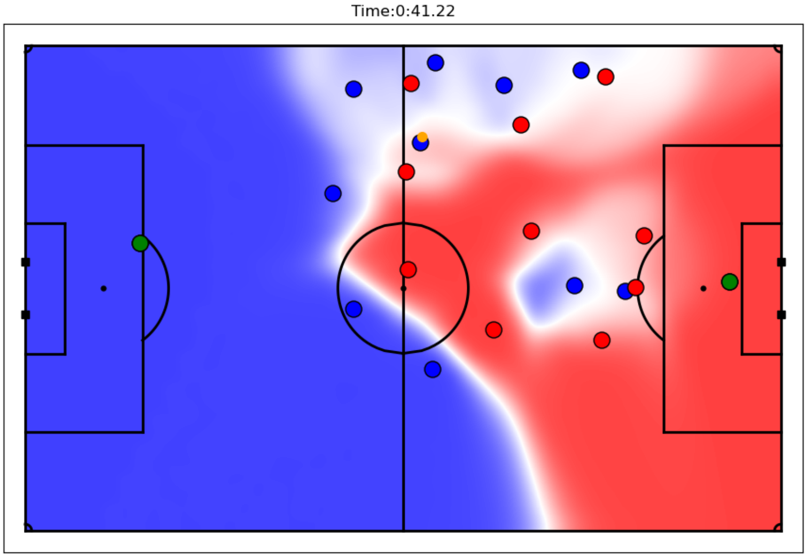
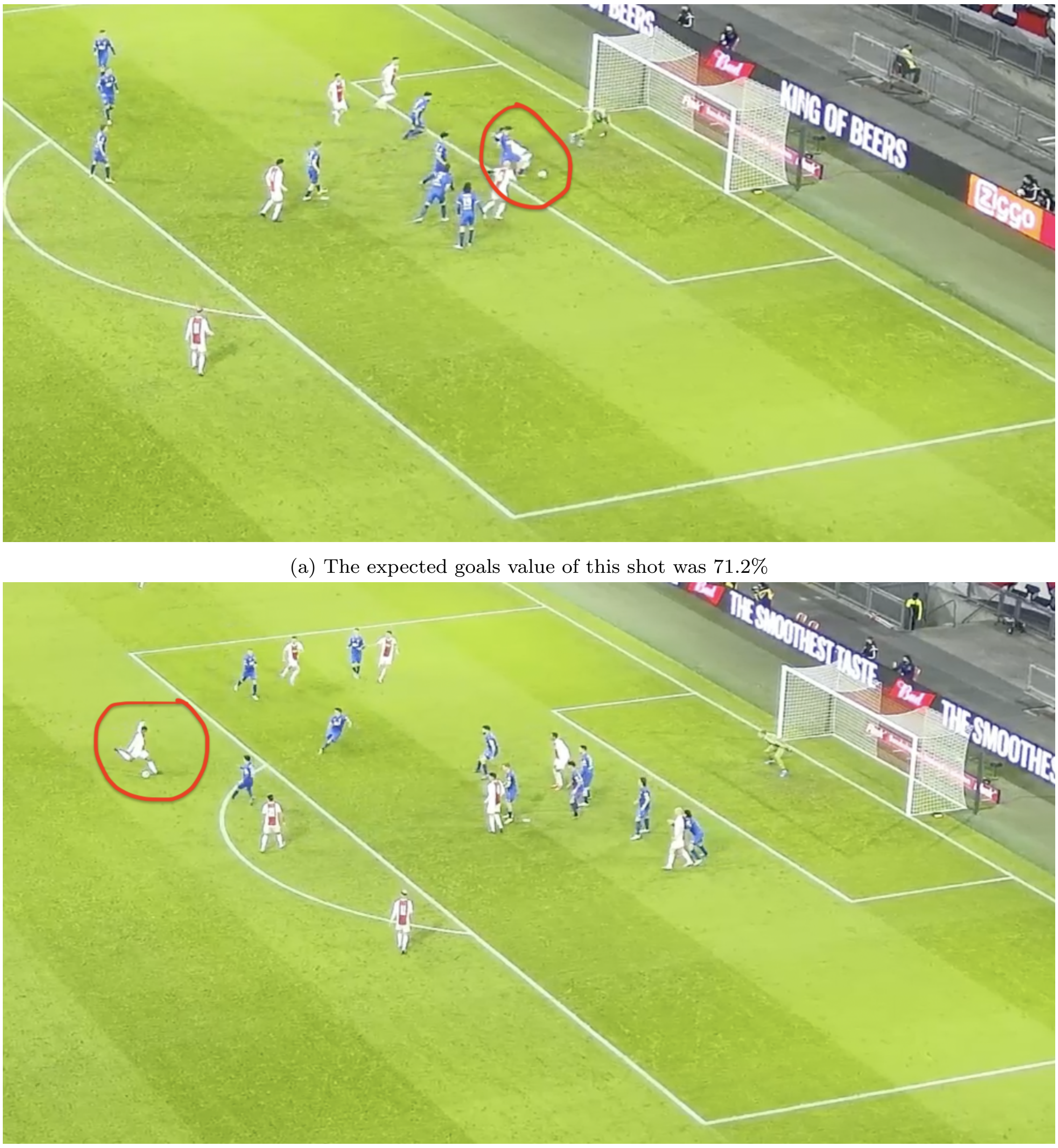


Figure 3: Pitch control model visualized. Color scale is made according to the probability of a player of a team (blue or red) to control the ball if it travels to that position directly from where it is now

3.2 Expected Goals

Expected goals (xG) is one of the most commonly used piece of football statistics. Expected goals is a metric that evaluates the quality of goal scoring opportunities generated by teams [6]. It looks at several factors such as position on the pitch from where a shot was taken, which body part was used, number of players between the ball and the goal, goalkeeper position, whether it's a set piece or in open play and possibly more [7]. Expected goals evaluates the quality of goal scoring opportunities by producing a probability of any shot to result in a goal. As it is a probability measurement, it assumes values between 0 and 1. In figure 4 we see two examples of shots with different expected goals values (calculated by wyscout). The first shot is taken from only a few meters away from goal, with no defenders between the ball and goal, and with the goalkeeper out of position. It has a high probability of resulting in a goal. The second one is a shot from a few meters outside the box and with a lot of defenders standing between the goal and the ball. This shot has a low probability of yielding a goal.



(a) The expected goals value of this shot was 71.2%

(b) The expected goals value of this shot was 1.2%

Figure 4: Two shots with very different probability of ending up in a goal - as measured by expected goals (xG)

3.3 Convex Hull

As mentioned in chapter 2.2, the space between the midfield and back-line is important to control from a defensive team perspective. This dangerous space is in this project defined as the area created by the convex hull between the defending and midfield line. The convex hull is defined as the minimum convex set enclosing all points (all defenders and midfielders), and will be further discussed in chapter 4 [8].

3.4 Synchronization

A few attempts have previously been made to categorize and evaluate teams' structure and overall organization. One recurring method of doing this is using a measure of synchronization of players' movements, as suggested by Duarte et al [9]. Duarte et al analyzed a small dataset, namely one match from the Premier League, and looked at player movements as phase, which can most easily be thought of as the running direction of players when viewing the field from above. First the overall average movement direction is found, followed by each player's individual phase in relation to this group phase, and finally group synchronization is calculated with this as basis [9]. Cluster phase is calculated through

$$r'(t_i) = \frac{1}{n}\sum_{k=1}^{n} e^{i\theta_k(t_i)} \tag{1}$$

$$r(t_i) = atan2(r'(t_i)) \tag{2}$$

where $r'(t_i)$ is the cluster phase at time t_i, n is the total number of players in the team (11 usually) and θ_k is the angle of player k. Then each player's relative phase is calculated as

$$\phi_k(t_i) = \theta_k(t_i) - r(t_i) \tag{3}$$

which is a measure of how well synchronized each individual player is with the whole team. The group synchronization measurement, or as it's called in Duarte et al's article the cluster amplitude, is then calclulated as

$$\rho_{group}(t_i) = \left| \frac{1}{n}\sum_{k=1}^{n} exp\{i(\phi_k(t_i) - \overline{\phi}_k)\} \right| \tag{4}$$

where $\overline{\phi}_k$ is the mean deviation of player k over the entire time interval investigated (in other words, over the period of the match that player was on the field).

3.5 Gaussian Mixture Model

Detecting the three defensive lines automatically is a difficult task, and requires a good choice of clustering method. Here, a one dimensional Gaussian Mixture Model is used. The method divides a set of data into subgroups, and is useful when knowing a priori how many subgroups there are [10]. Mathematically it's computed through

$$p(x) = \sum_{i=1}^{K} \phi_i \mathcal{N}(x|\mu_i, \sigma_i) \tag{5}$$

$$\sum_{i=1}^{K} \phi_i = 1 \tag{6}$$

where x is the population of data points, p is the joint distribution of x when summing over all clusters, K is the number of subgroups/clusters, ϕ_i is the component weight for subgroup i, μ_i is the mean value of subgroup i, and σ_i is the standard deviation of subgroup i [11].

Parameter estimation is done through Expectation Maximation (EM), which is a numerical technique for maximum likelihood estimation. It begins by randomly assigning group mean estimates to some samples in the data, setting all variance estimations to the sample variance, and component distributions to equal probabilities [11]. In other words, for the example of K = 3

$$\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_3 = x_{r1}, x_{r2}, x_{r3} \tag{7}$$

$$\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\sigma}_3^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{8}$$

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{9}$$

$$\hat{\phi}_1 = \hat{\phi}_2 = \hat{\phi}_3 = 1/3 \tag{10}$$

initializes the algorithm, where x_{ri} implies a randomly selected observation from the data set investigated. After this initialization, one iterates between calculating expectation and maximization. The expectation step is computed through

$$\hat{\gamma}_{ik} = \frac{\hat{\phi}_k \mathcal{N}(x_i|\hat{\mu}_k, \hat{\sigma}_k)}{\sum_{j=1}^{K} \hat{\phi}_j \mathcal{N}(x_i|\hat{\mu}_j, \hat{\sigma}_j)} \tag{11}$$

where $\hat{\gamma}_{ik}$ represents the probability that x_i belongs to cluster k [11]. This is calculated for all combinations of i and k, and is then used in the maximization step as

$$\hat{\phi}_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik}}{N} \tag{12}$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik} x_i}{\sum_{i=1}^{N} \hat{\gamma}_{ik}} \tag{13}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik}(x_i - \hat{\mu}_k)^2}{\sum_{i=1}^{N} \hat{\gamma}_{ik}} \tag{14}$$

which is calculated for each cluster k. This results in an updated set of parameters ϕ, μ, σ for which the expectation is calculated and so on. The iterations stop at convergence or after a predefined number of iterations[11].

3.6 Regression Methods

The statistical method used to evaluate association between different attacking strategies and creating good goal scoring opportunities is linear regression and generalized linear regression. By categorizing each attack attempt according to what attacking strategies were utilized, it's possible to view each attack attempt as an observation and the attacking strategies as characteristics of an attempt.

3.6.1 Linear Regression

Linear regression mathematically has the formula

$$E(Y) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{15}$$

where Y is the response variable (in this project some transformation of expected goals), β_i is the coefficient explaining relationship between the response variable and the explanatory variable i (in this project this would be some charactersitic of the attacking or defending team), and X_i is the value for explanatory variable i. Explanatory variables can be measured numerically, categorically or binary (yes or no). This means that X_i can essentially be either a binary (1 or 0) or numerical depending on it's characteristics. The act of fitting a linear regression model revolves around choosing explanatory variables i, gathering data on several observations for these, and fitting the coefficients β according to least square error between true observations and the model's predictions. In practice, this is done in statistical software R.

In addition to this, interaction terms were investigated and evaluated for some combinations of measurement variables. Interaction occurs when an independent variable has a different effect on the outcome depending on the value of another independent variable. For two measurement variables the formula would look as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \tag{16}$$

with $X_1 X_2$ being the interaction between the two variables.

3.6.2 Logistic Regression

When looking at a binary response variable, logistic regression is the preferred method. It works very similar to linear regression, with the difference being that the modeled response is the log odds of a ”positive” outcome of an observation. In this project, logistic regression will be used to investigate what attacking and defensive strategies tend to lead to a shot being taken (1 if a shot is taken, 0 if it is not). The response variable Y then follows a binomial distribution with probability of a shot p as

$$Y \sim Bin(1, p) \tag{17}$$

$$E(Y) = p \tag{18}$$

Rather than using p directly as response variable, it's preferable to use a logodds (or logit) transformation of p as it transforms the response variable across the whole real axis. The logit/logodds function looks as

$$logit(p) = \log\left(\frac{p}{1-p}\right) \tag{19}$$

where p is the probability of a positive outcome. Thus the regression model in it's entirety is

$$log(\frac{p}{1-p}) = X\beta = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \tag{20}$$

from which p can be extracted.

3.6.3 Logit-Linear Regression

Expected goals was used as response variable for the attack attempts where a shot was taken. For this, the logit-transformation was utilized again but for the response variable. As mentioned, this transfers a values between 0 and 1 to the entire real axis through

$$logit(xG) = \log\left(\frac{xG}{1-xG}\right) \tag{21}$$

which results in the regression formula

$$logit(xG) = X\beta = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \tag{22}$$

3.6.4 Gamma Regression

When the response variable of a regression follows a gamma distribution, the gamma regression technique is suitable. A variable y follows a gamma distribution if its probability density function is given by

$$f(y|\alpha, \lambda) = \frac{\lambda^{\alpha} y^{\alpha-1} e^{-\lambda y}}{\Gamma(\alpha)} I_{(0,\infty)}(y) \tag{23}$$

where α and λ are parameters explaining the shape of the gamma distribution, Γ is the gamma function, and I is the indicator function. In this project, expected goals is the used response variable, and it assumes values between 0 and 1. The gamma distribution is generally applicable for response variables that assume values over the entire positive axis, however as xG chances close to 1 are very rare, this will likely not cause statistical issues.

In this project a log-link is used in the gamma regression. Thus the relationship between response variable mean and parameters α, λ and β is

$$ln(E(xG)) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n. \tag{24}$$

3.6.5 Model selection

There are several ways to select the best possible linear regression model out of a set of potential ones. An easy one to implement when dealing with a manageable number of explanatory variables is the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These two methods simply put evaluate the explanatory power of different models in relation to each other. They are mathematically calculated as follows

$$AIC = -2 \cdot \frac{l}{n} + 2\frac{k}{n} \tag{25}$$

$$BIC = -2 \cdot \frac{l}{n} + 2\frac{k \cdot ln(n)}{n} \tag{26}$$

where l is the log-likelihood function of the observed values given the predictions made by the model, k is the number of estimated parameters, and n is the number of observations in the data. The best model is the one with the smallest AIC or BIC respectively. Note however that BIC and AIC values can not be compared to each other, and that neither of the values give any information about the absolute quality of fit of the different models, just which one is relatively speaking better. The BIC does, as can be seen mathematically in the equation above, punish additional parameters more than does the AIC. Thus there will always be fewer or the same number of variables in a model selection made with BIC than with AIC [12]. In practice, when deciding which variables to include in a linear regression model for best explanatory effect one iteratively adds and/or removes variables and compare AIC or BIC values. So for example, if we are to select the best model in terms of BIC, and we have a dataset with three available explanatory variables and a response variable, it might look like this.

1. Calculate BIC of the null model - in other words predicting that the outcome variable is always constant

2. Calculate BIC of all of the possible one variable models separately. If all of their BIC scores are higher than the one for the null model, we stop and conclude that the null model is the best. Otherwise, continue with the lowest score BIC model.

3. Calculate the BIC scores for all two variable models containing the variable selected in step 1. Compare BIC scores and choose the one with the lowest value.

4. Continue doing this until the model cannot be improved by either removing or adding another variable.

This will henceforth be referred to as AIC and BIC model selection methods.

The stepwise AIC and BIC methods are suitable when the number of variables is manageable. When there are larger amounts of explanatory variables, a better method for model selection is the Least Absolute Shrinkage and Selection Operator, or Lasso. The concept revolves around penalizing non-zero coefficients to remove variables with little significance from the model. Mathematically, a Lasso regression technique solves the equation

$$min_{\beta_0,\beta} \quad \frac{1}{N}\sum_{i=1}^{N} \omega_i l(y_i, \beta_0 + \beta^T x_i) + \lambda|\beta| \tag{27}$$

where β_0, β^T are the intercept and variable coefficients of the model respectively, N is the number of observations, l denotes the negative log-likelihood contribution of observation i, and λ is the penalty term. λ can be adjusted depending on how big of a penalty one wants, which in turn determines the number of coefficients.

The Lasso technique tends to pick one of the variables that are highly correlated and discard the rest. Sometimes it's more desirable to select groups of variables that are highly correlated but explain the response variable well, and discard entire groups of variables that have smaller significance on the response. For this, an elastic net could be used. It's mathematically similar to the Lasso regression, except for that it also incorporates a ridge regression penalty term as

$$min_{\beta_0,\beta} \quad \frac{1}{N}\sum_{i=1}^{N} \omega_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha|\beta|] \tag{28}$$

where α is a term that controls the elastic net, that is, how much of the penalty is to be determined by traditional Lasso and how much is to be determined by ridge. In this project, α will only be set to 0.5 for the elastic net.

3.6.6 Model evaluation

Three methods will be utilized to evaluate the quality of the models. The already mentioned AIC and BIC metrics will be used for the gamma regression. For the linear regression with logit-transformation, R^2 will be used. R^2 evaluates how much of the variability in the response variable that can be explained by the model. Mathematically it's computed by

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{29}$$

where $\hat{Y}$ is the model's prediction of the outcome, $\overline{Y}$ is the mean value of the outcome and Y is the actual observed outcome. The higher the value of R^2 the more of the variability the model is able to explain.

For the logistic regression, the confusion matrix will be used along with the ROC curve and AUC. The confusion matrix is a two-by-two matrix that shows the distribution of predicted values against true values of observations. It looks as in table 1.

	Predicted Positive	Predicted Negative
Actual Positive	Correct Positive	False Negative
Actual Negative	False Positive	Correct Negative

Table 1: Shape of confusion matrix

There are four interesting measurements to consider:

1. Accuracy, the overall correctness of the model: (True positive + True Negative) / Number of observations

2. Precision, the share of predicted positives which are correct: True Positive / (True Positive + False Positive)

3. Sensitivity, the share of correctly predicted positives out of all actual positives: True Positive/(True Positive + False Negative)

4. Specificity, the share of correctly predicted negatives out of all actual negatives: True Negative/(True Negative + False Positive)

Out of these four, sensitivity and specificity are the two that will be mostly considered.

In this report a positive represents an attack ending with a shot, and negative represents an attack not ending with a shot. The logistic regression response variable is log-odds of observations with certain parameter values being positive. This can be translated to probability by solving for p in equation (18). Naturally, if one is to label each of these predicted probabilities as either positives or negative binarily, a threshold is required. The most obvious threshold is 0.5, meaning that observations with a probability higher than 0.5 are predicted as positives, and the rest as negatives. However, when the overall frequency in the training data of a positive is very small or very big, a 0.5 threshold results in all or almost all observations being predicted by the most common case, which isn't very useful. In these cases, the specificity and sensitivity will be vastly different to each other. It's often desirable to provide a more balanced error between specificity and sensitivity, which can be done by changing the threshold value for p. An often used choice of threshold for p is one where specificity and sensitivity are the same. This threshold value will be used in this report.

By testing specificity and sensitivity continuously for a lot of threshold values between 0 and 1, we can create a receiver operating characteristic (ROC) curve. This illustrates how the specificity and sensitivity varies as the threshold changes. AUC, or area under curve, measures how well the model performs over this whole range of thresholds - which is illustrated in figure 5.

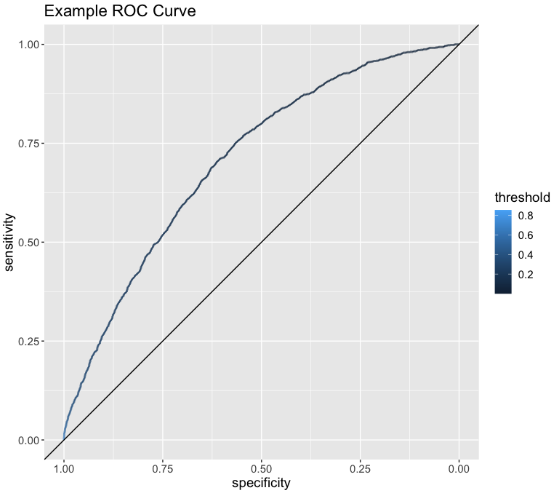


Figure 5: Example of ROC curve. The diagonal black line is the performance of the null model, i.e. just guessing.

4 Method

This chapter describes the novel work performed in this project, which is based on the concepts presented in chapter 2 and 3.

4.1 Finding Longer Attacks

In order to analyze how to beat an organized defense, the first task is to identify the situations where there is an organized attack. A football match consists of duels in the midfield, set piece plays, counter attacks, long balls from the defense towards the strikers etc. In other words, some filtering of the raw data is necessary to identify the times there is an organized attack against a team defending on their own half. Here, the main criteria for defining organized attack is that the attacking team held possession for at least 20 seconds, out of which at least 10 consequtive seconds are on the attacking 3/5ths of the field (see figure 6 for an illustration of what this means). Furthermore, the first ten seconds from when a corner or longer free kick (longer than 15 meters) was played were excluded as set piece plays are a different game scenario. Running the data through these criteria produced a set of organized attack attempts.

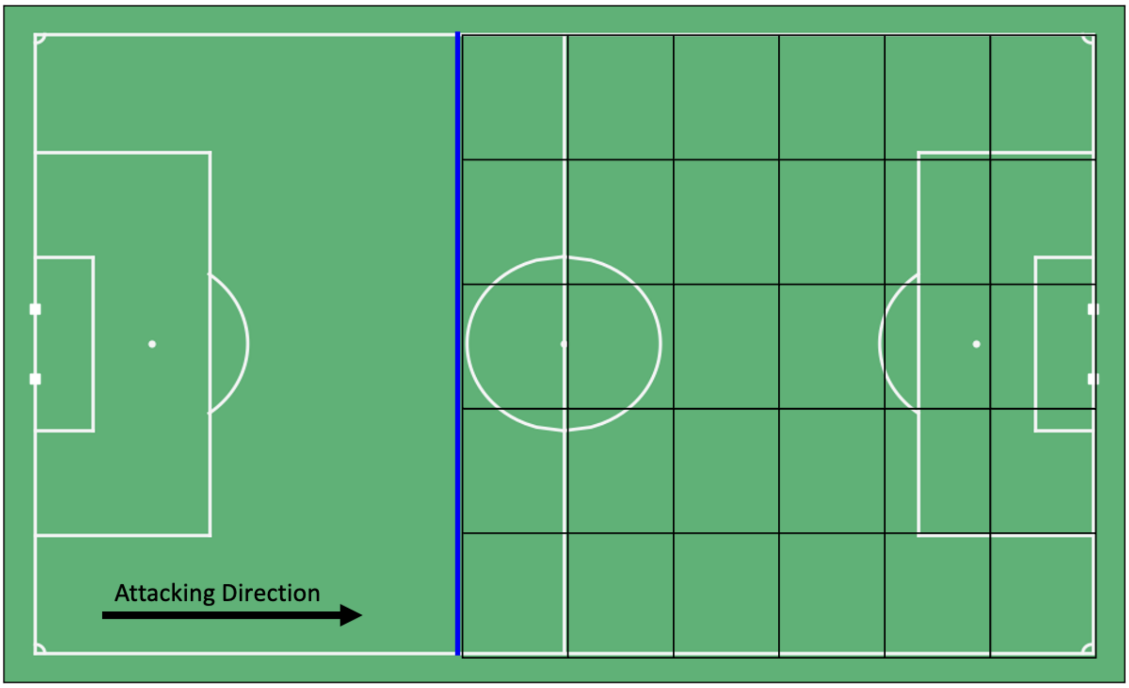


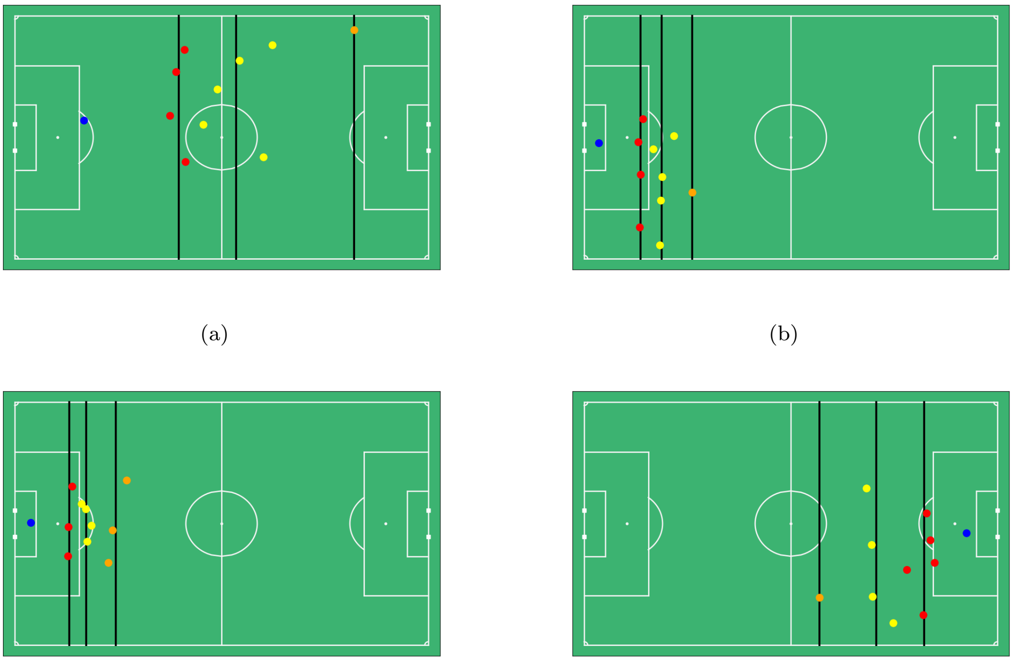
Figure 6: The area with a grid is the attacking 3/5ths of the field

4.2 Finding Lines of Defense

As mentioned in chapter 2.2, teams generally tend to defend in three clusters or lines. There is the attacking line, which is the first line of defense closest to the opponents goal. The midfield, which is the second line of defense, and the back-line, or defense as it's also called, which is the last line of defense in front of the goalkeeper.

In the situation shown in figure 2, it would be easy to implement an algorithm that automatically detects and categorizes the players into one of the three lines. However, in order to continuously detect these lines in various examples and scenarios a good clustering algorithm is needed. In this project, several methods were tried and tested with varying success. The best method proved to be the Gaussian mixture model (GMM) applied one-dimensionally on only the horizontal coordinates.

This method requires a specification of how many clusters there are (three in this case), and is generally applicable to two-or-more dimensional problems. The problem of finding three defensive lines is at it's core a two dimensional problem - there are two dimensions of player positions on the field, horizontal and vertical. However, in the context of finding defensive lines the horizontal axis is far more important than the vertical one which is why the Gaussian Mixture model has been implemented one-dimensionally. Figure 7 shows four examples of the Gaussian mixture model finding lines of defense.



(a)

(b)

(c)

(d)

Figure 7: Four examples of lines of defense found with Gaussian mixture model. Red players are in the defensive line, yellow are midfielders and orange are in the attacking line

4.3 Defining Dangerous Area

As explained in chapter 2.2 the space between defensive line and midfield is of interest. This area is here defined as the convex hull made up by the midfielders and back-line. Figure 8 illustrates the definition of this space for a generic example.

The dangerous space will be part of several features of attack attempts. For example, pitch control as described in chapter 3.1 is calculated for and valued specifically in this area. To recap, the pitch control for a specific location on the pitch describes the likelihood that the ball will be controlled by the attacking/defending team should the ball travel directly there. Since this is space that the defending side should aim to control, taking the average pitch control of the defending team on each of these location gives the likelihood that the defending team will control the ball if it travels to an arbitrary point within the area and thus measures how well the defending team is controlling that space.

Furthermore, a few metrics regarding passing strategies inside this space shall be used, and will be described later. Additionally, the structure of the convex hull itself was evaluated as a way of understanding how the defense positioned themselves. In other words, the distance between the edge players in horizontal and vertical direction respectively was found, as well as the area of this convex hull (illustrated in figure 13).

The convex hull area created between the defensive team's defensive line and midfield will henceforth interchangeably be called *convex hull* and space *between lines*.

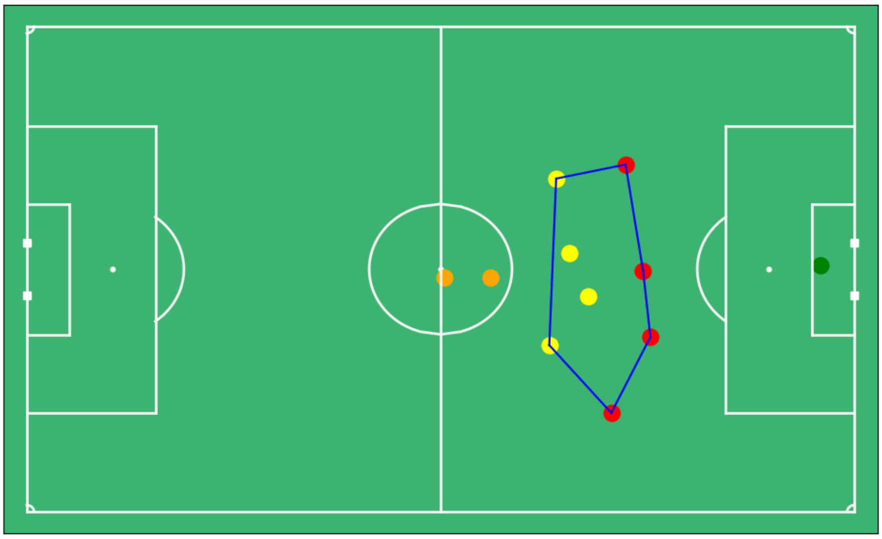


Figure 8: Illustration of the dangerous space created by the convex hull with midfielders (yellow and defenders (red).

4.4 Synchronization

Movement synchronization between the defending team has been investigated according to chapter 3.4. The only adjustment to the model presented by Duarte et al is that the goalkeeper is excluded when calculating cluster amplitude [9]. The reason for this is that the goalkeeper has a separate role and he's not participating in the pressing and defending in a football match in the same way that the outfield players do, and thus including him would make the results less interpretable in a football context.

Synchronization was also calculated in the vertical and horizontal direction separately so as to provide a more detailed interpretation of the defending team's organization regarding movement direction. Mathematically this was done with a different approach compared to cluster amplitude as there is no angle/phase to utilize in a one dimensional problem. Rather, the decision was made to use movement speed according to

$$\rho_i = 1 - \frac{x_i - \overline{x}}{max_i|x_i|} \tag{30}$$

$$\rho = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \overline{x}}{max_i|x_i|} \tag{31}$$

where ρ is the synchronization (ranging from 0 to 1), x_i is the horizontal movement speed of player i (negative if movement is to the left and positive if the movement is to the right), $\overline{x}$ is the average horizontal movement speed of all players, and $max|x|$ is the largest absolute value movement speed of any player.

In figure 9 we see an example with only two players. The two players are completely sychronized in the vertical direction ($\rho = 1$), but also completely asynchronized in the horizontal direction ($\rho = 0$).

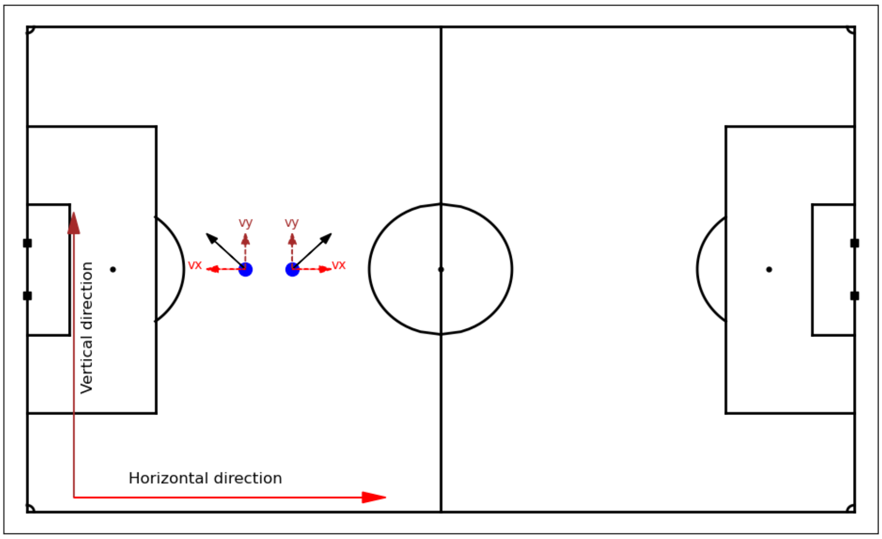


Figure 9: Illustration of horizontal and vertical direction. The black arrows represent the movement direction of the player, and the brown and red ones the horizontal and vertical components.

4.5 Defining Attacking Positioning

The way the attacking team positions themselves and how the players move was interpreted from a few different perspectives. First of all, how spread out the players are was evaluated as the distance in horizontal and vertical direction respectively between the edge players. See figure 10 for an illustration. Second, in order to make some interpretation of the space covered by the team, the concept of convex hull was reused. This time however, all the outfield players were included, as the purpose is to get an idea of how the whole attacking team is positioned. This is also illustrated in figure 10.

For movement, the footballing argument is that movement creates gaps in defence that can be exploited by the attacking team. Here, movement in the vertical and horizontal direction were investigated individually as follows

$$Vx = \frac{\sum_{i=1}^{n} |v_{x,i}|}{n} \tag{32}$$

$$Vy = \frac{\sum_{i=1}^{n} |v_{y,i}|}{n} \tag{33}$$

where n is the number of outfield players, and $v_{x,i}, v_{y,i}$ represent horizontal and vertical movement direction of player i respectively.

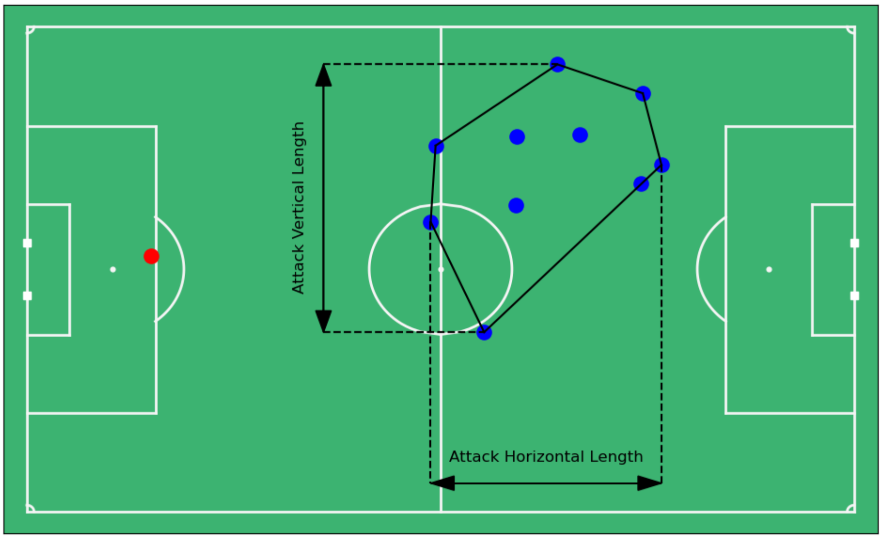


Figure 10: The convex hull, horizontal length and vertical length metrics of attacking team visualized

4.6 Defining Generic Attack Characteristics

A few basic definitions of attack characteristics were made. These include duration of attack (time between the first touch of the ball until losing it to the opposition, taking a shot, or the ball going out of play), number of passes played during the attack, average pass length and whether a cross was played.

4.7 Creating Variables to Investigate Attacks

Attack attempts were found according to chapter 4.1, and for each of those the metrics described so far were used to describe the individual attempt's characteristics. The metrics look as follows

4.7.1 Passing Strategies

1. Duration of the attack measured in seconds from when the ball was won until either a shot was taken, the ball was lost to the opponents, or the ball went out of play.

2. Number of passes played during the attack.

3. Average pass distance.

4. Number of times the team switches side. Switching side in football means attacking with the ball on one wing, and then play the ball across the field to the other wing. See figure 11d for two examples.

5. Is there a cross played? A cross is when a longer ball is played towards the penalty area. See figure 11a for two examples.

6. How many longer passes are played from in front of the defensive team's midfield (see chapter 4.2 for reference) to behind the back-line. See figure 11b for an illustration.

See figure 11c for an illustration of the following metrics.

7. How many passes are played from between defensive lines (see chapter 4.3 for definition).
8. How many passes are played from between defensive lines back to in front of the defensive side's midfield.
9. How many passes are played from between defensive lines out to either of the wings.
10. How many passes are played from between defensive lines to behind the opponent's defence.
11. How many passes are played from between defensive lines to another player also between defensive lines.
12. How many dribbles are made between defensive lines.

4.7.2 Attacking Positioning and Movement Variables

1. Vx of all outfield players as described in chapter 4.5
2. Vy of all outfield players as described in chapter 4.5
3. Vx calculated only including players that are positioned deeper than the furthest forward standing defensive midfielder. See figure 12 for reference
4. Vy calculated only including players that are positioned deeper than the furthest forward standing defensive midfielder. See figure 12 for reference.
5. Area of convex hull of attacking team - see figure 10.
6. Horizontal distance between edge players in attacking team
7. Vertical distance between edge players in attacking team - see figure 10

Football is a multifaceted sport. There are a plethora of ways you can play the game. This list of variables is not exhaustive, but it does cover some of the characteristics that make up attacking strategies. It's possible for the attacking team to affect all of the mentioned variables, and they can be extracted from the available data.

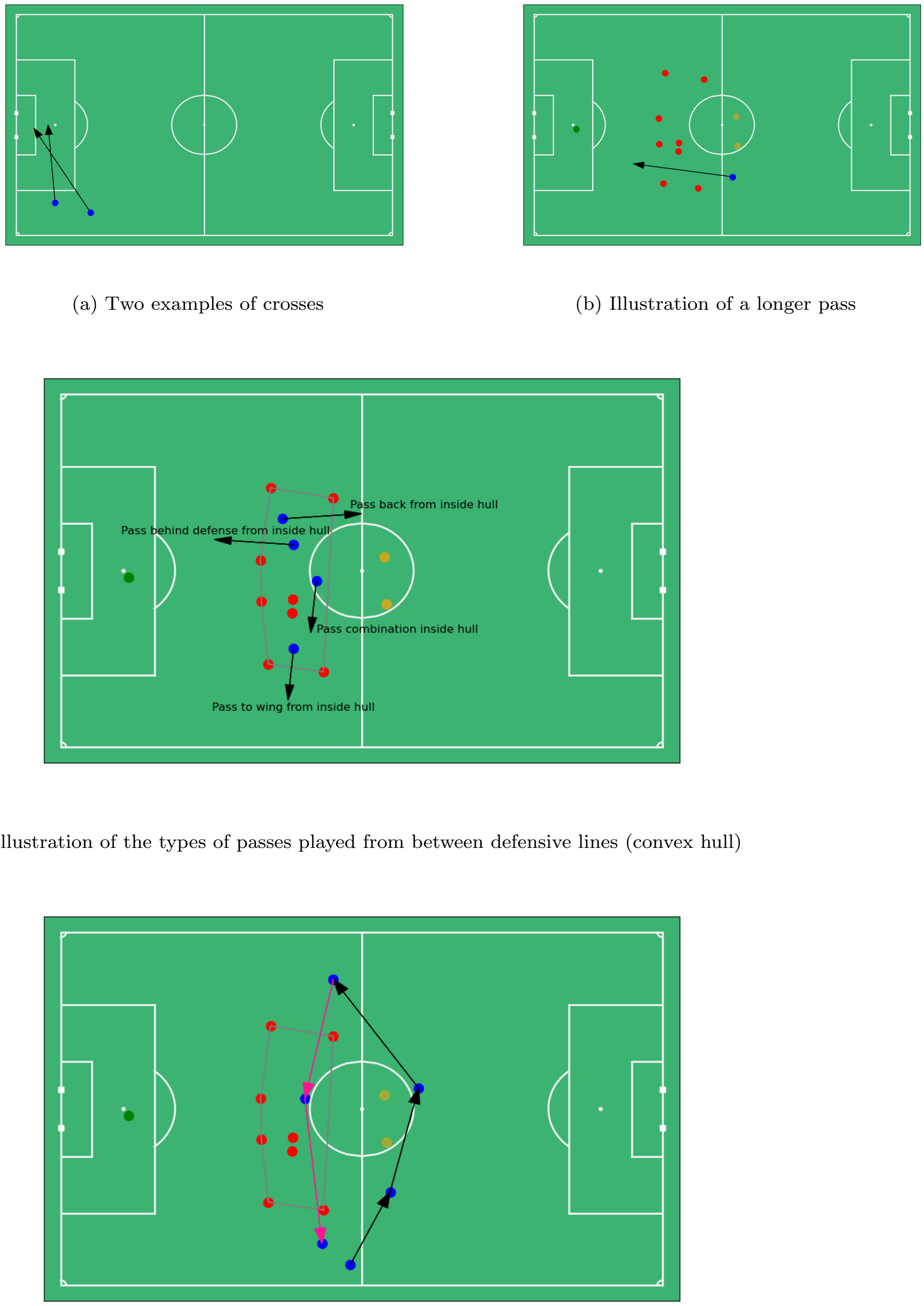
4.7.3 Defensive variables

In contrast, some variables are included that aim to capture the structure and organization of the defending team. These include the following

1. Cluster amplitude of defending team's movement as described in chapter 4.4.
2. Synchronization in the vertical direction of the defending team as described in chapter 4.4.
3. Synchronization in the horizontal direction of the defending team as described in chapter 4.4.

   See figure 13 for illustration of the following metrics

4. Vertical length of the convex hull between defensive lines
5. Horizontal length of the convex hull between defensive lines.
6. Area of the convex hull between defensive lines
7. Number of players that make up the vertices or are inside the convex hull.


Pass back from inside hull
Pass behind defense from inside hull
Pass combination inside hull
Pass to wing from inside hull

(a) Two examples of crosses

(b) Illustration of a longer pass

(c) Illustration of the types of passes played from between defensive lines (convex hull)

(d) Two different ways to switch side, one shown in black and another in pink

Figure 11: Visualization of some of the metrics/variables used. Attacking direction to the left. Red dots are defending team midfielders/defenders. Orange (barely visible) dots is the defending team's attacking line. Blue dots are players on the attacking team. Arrows indicate pass travel path.

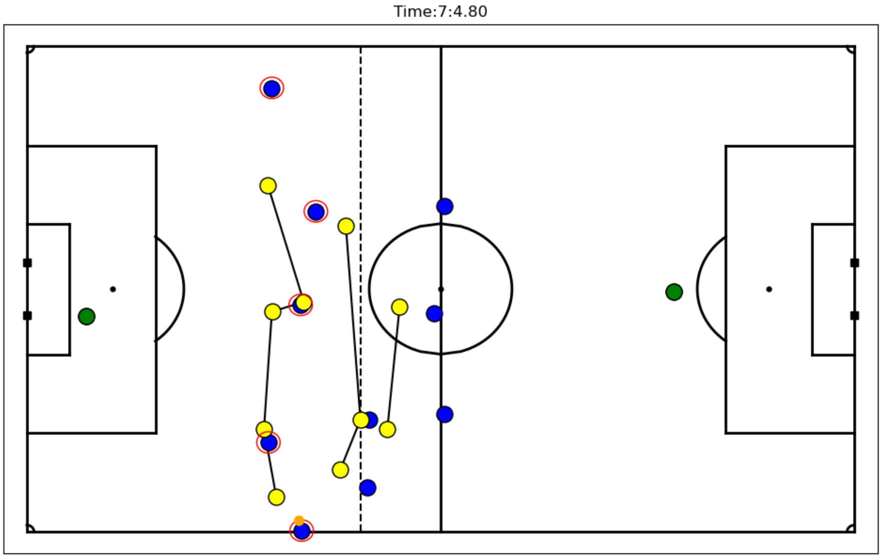


Figure 12: Yellow team is defending and blue is attacking. The black lines between yellow players indicate the three lines of defense. Red circled players in the blue team are positioned deeper than the furthest forward standing defensive midfielder.

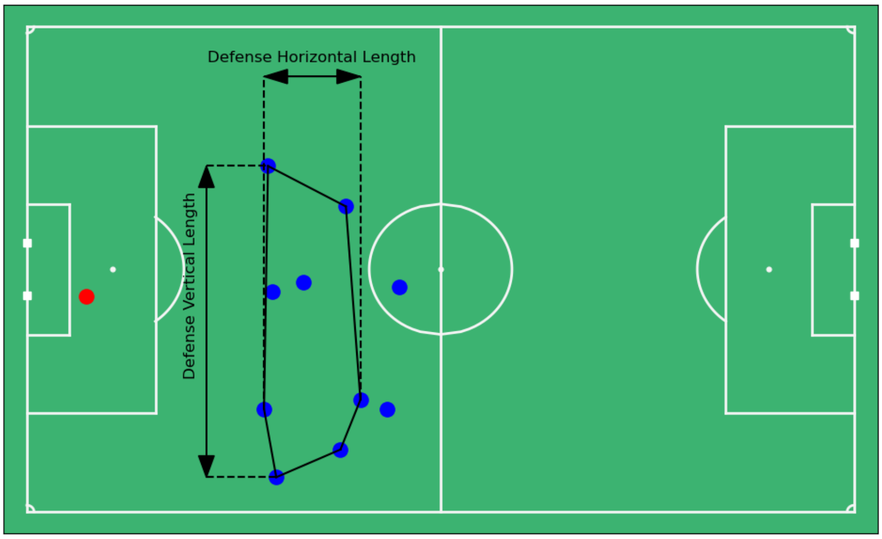


Figure 13: Visualization of convex hull, vertical length and horizontal length of defensive team

4.8 Logistic Regression Models

The regression techniques used is separated in two. First, there is the logistic regression with response variable being whether or not a shot was taken at the end of an attack attempt. Remember, the purpose of this study is not to produce a model that accurately predicts whether or not a shot is taken in an attack based on some characteristics. Rather, the purpose is to explain which attacking strategies and defensive shapes/characteristics are most and least effective in organized play. To start with, the full logistic model was created. This model includes all the variables in chapter 4.7. Then, in order to find which variables best predict the outcome of an attack, the stepwise AIC method was used with the full model as the limit. Also, to find which variables best explain the relationship between shot attempts being taken and attack characteristics, the BIC methods was used with the full model as the limit.

This gives a total of three models, where none of them include any interaction variables. Therefore, the Lasso method was utilized for model selection in a model including all pairwise interactions because it provides the opportunity to efficiently select variables from a larger quantity of available ones. Choosing a model is done through estimating the prediction error with cross-validation for different values of λ. Then, an appropriate value for λ, and thus the number of coefficients in the model, is selected through finding the λ one standard deviation from the minimum error. This value resulted in a reasonable number of non-zero coefficients, where the minimum value itself gave an infeasibly large amount. To use a variable selection method with different features, an elastic net regression, that is a combination between ridge and Lasso regression was made with α at 0.5.

4.9 Gamma and Logit-transformation Linear Regression

This study investigates not only what characteristics of an attack explain whether a shot was taken at the end of it, but also the quality of goal scoring opportunities created. For this, the data-set was reduced to only include the attack attempts where a shot was taken. Then for response variable two different transformations of expected goals were used. The first one being the logit-transformation and then linear regression, and the second being gamma regression. A residual analysis was made to determine which of these methods best represent the distribution of expected goals, but it was inconclusive as neither of them had any significant issues in their residuals. Therefore, both of these transformations of expected goals will be used and shown here.

In terms of model selection and variations, the model choices and reasoning are the same as for the logistic regression. This means that for each of gamma- and logit-linear-regression, there are five models. The full model including all variables in chapter 4.7, the AIC selected model, the BIC model, the Lasso model and the Lasso/ridge combination model. Only the latter two include interaction variables.

5 Results

This chapter presents the results of the models. Discussion about model coefficients, comparison between models and footballing implications will then be made in chapter 6.

5.1 Correlation Between Variables

As one evaluates the success of different attack strategies it's interesting to consider the correlation between them. Therefore, figure 14 presents, with color coding, the correlations between all the explanatory variables in the project.

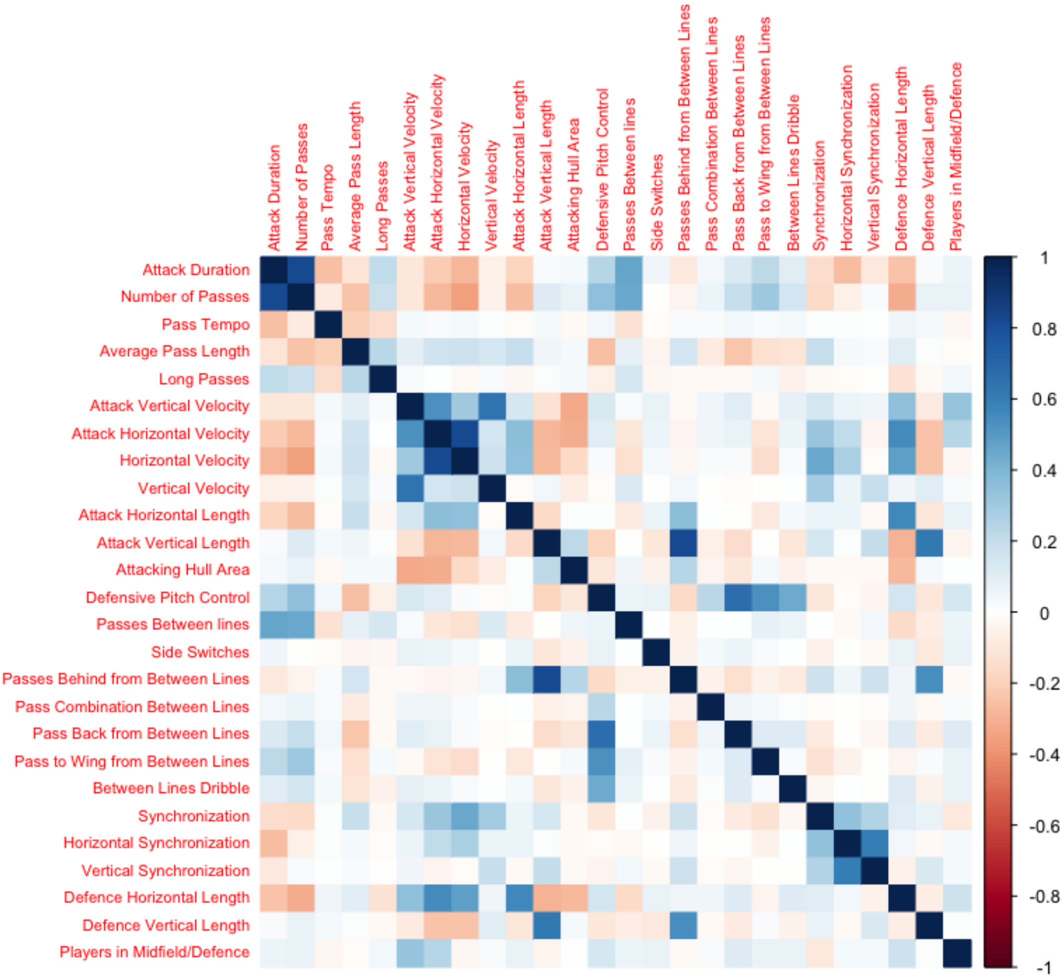


Figure 14: Correlation between all explanatory variables

5.2 Plots

In this section, only the statistically significant variables using a two sided 95% confidence interval will be displayed, to see all the model parameter estimates, standard deviation etc see appendix A. Summaries of all regression coefficients are presented in figure 15-18

It's important to notice that the coefficients shown are not the β-values from the models, but those values scaled by their standard deviation. In other words, what is plotted here is $\beta_i^* = \beta_i \frac{\sigma_{xi}}{\sigma_Y}$ for each coefficient β in each model where σ_{xi} is the standard deviation of explanatory variable x_i and σ_Y is the standard deviation of the response variable. This standardization means that the size of β_i^* for different coefficients i are directly comparable, at least within the same regression type (gamma/logit-linear). In other words, if a coefficient for one explanatory variable is bigger than that of another, then the effect of that variable on the response variable is bigger.

The following plots are separated between binary response variables (whether a shot was taken) and continuous response variables (expected goals), as well as between positive and negative coefficients. The separation between positive and negative coefficients is for visual ease. The main implications and key insights to be made from these plots will be discussed in chapter 6.

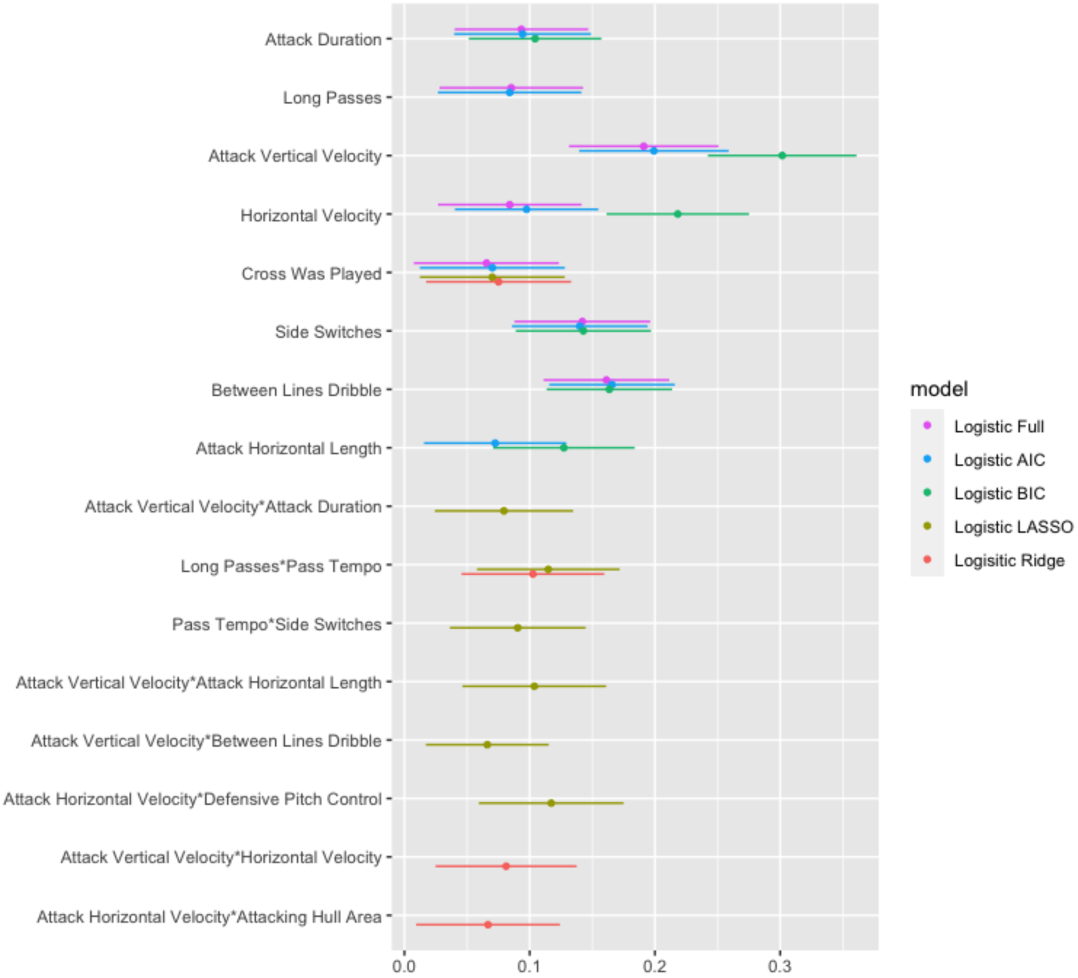


Figure 15: Logistic regression positive standardized coefficients. The likelihood of attacks ending with a shot increases with these variables, and the average effect of them on the outcome of attacks is comparable by the size of the coefficients (horizontal axis).

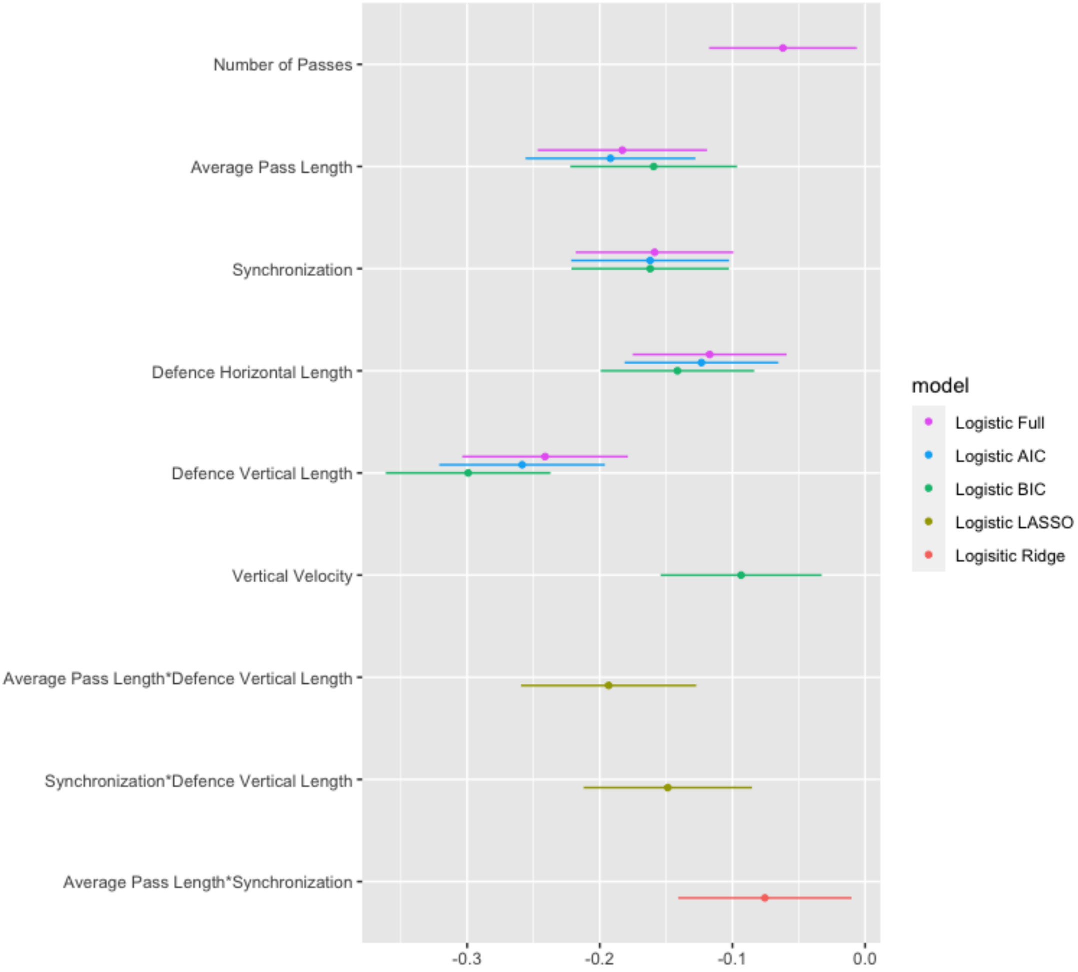


Figure 16: Logistic regression negative standardized coefficients. The likelihood of attacks ending with a shot decreases when these variables increase, and the average effect of them on the outcome of attacks is comparable by the size of the coefficients (horizontal axis). The bigger the absolute value of the coefficient (the further to the left in this case), the more of an effect that variable has on the probability of an attack ending with a shot.

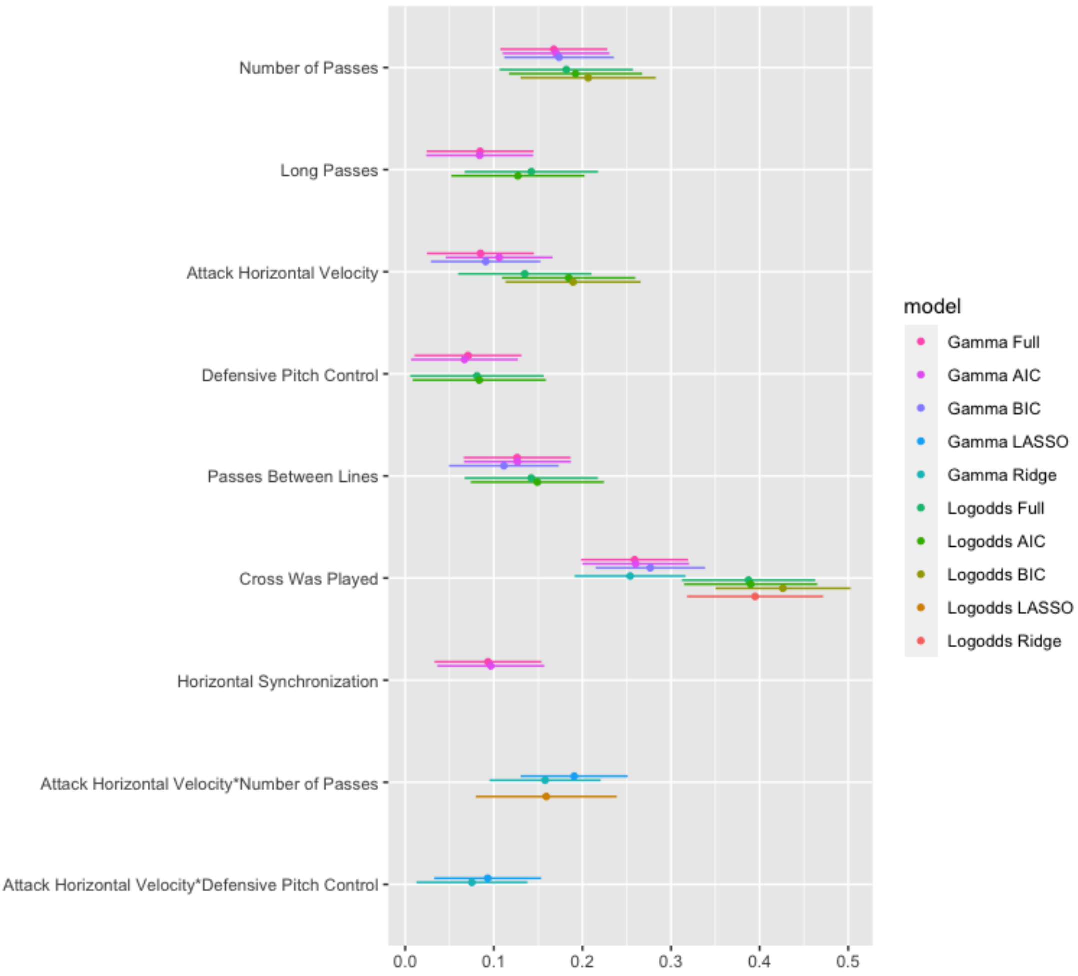


Figure 17: Gamma/Logit-linear regression positive standardized coefficients. The quality of chances created in attacks where a shot is taken increases with these variables, and the average effect of each variable is comparable by the size of the coefficient.

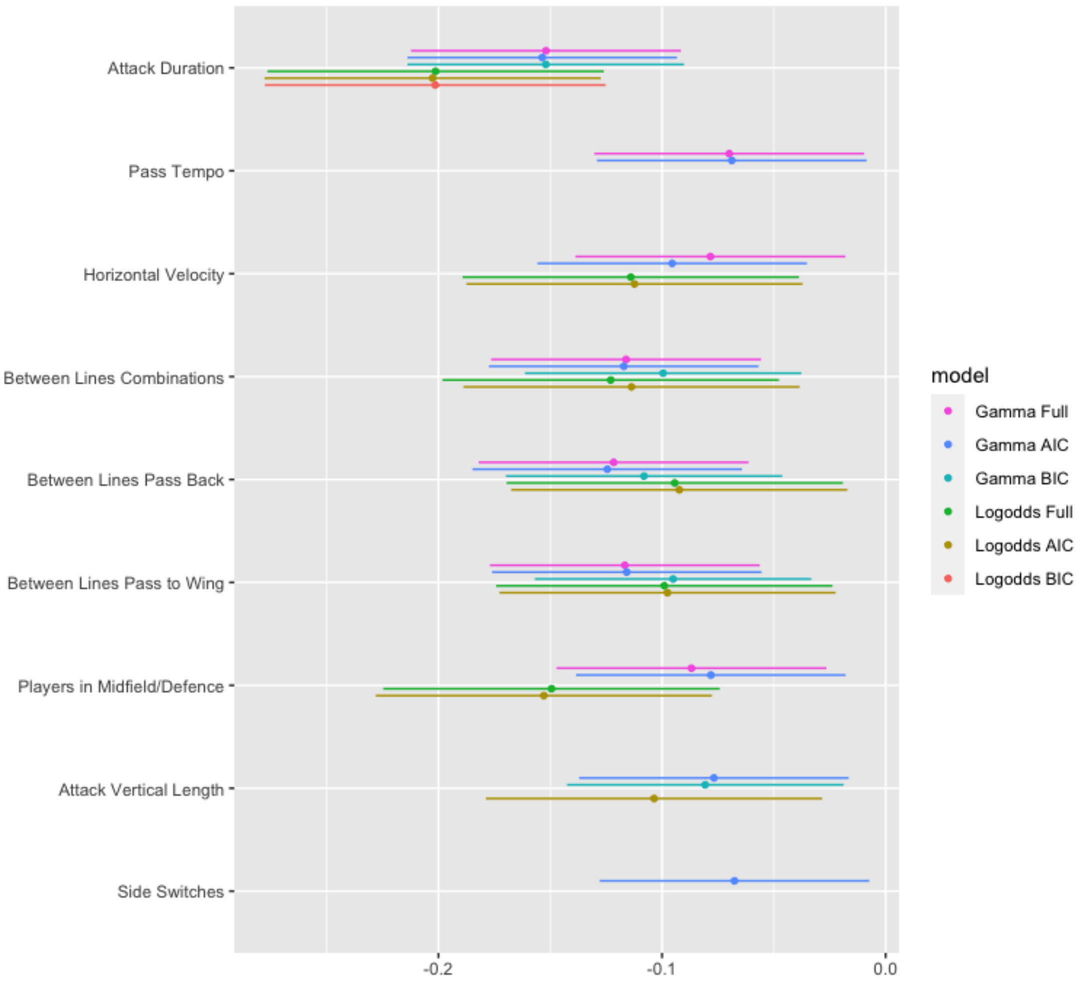


Figure 18: Gamma/Logit-linear regression negative standardized coefficients. The quality of chances created in attacks where a shot is taken decreases when these variables increase in size. The average effect of each variable is comparable by the size of the coefficient, the bigger the absolute value of the coefficient (the further to the left in this case), the more this variable affects the outcome of attacks ending with shots.

5.3 Model Evaluation

5.3.1 Logisitc Regression Evaluation

The logistic regression models are evaluated in terms of their ROC curves and values for area under curve. The ROC curves for all five models are given in figure 19:

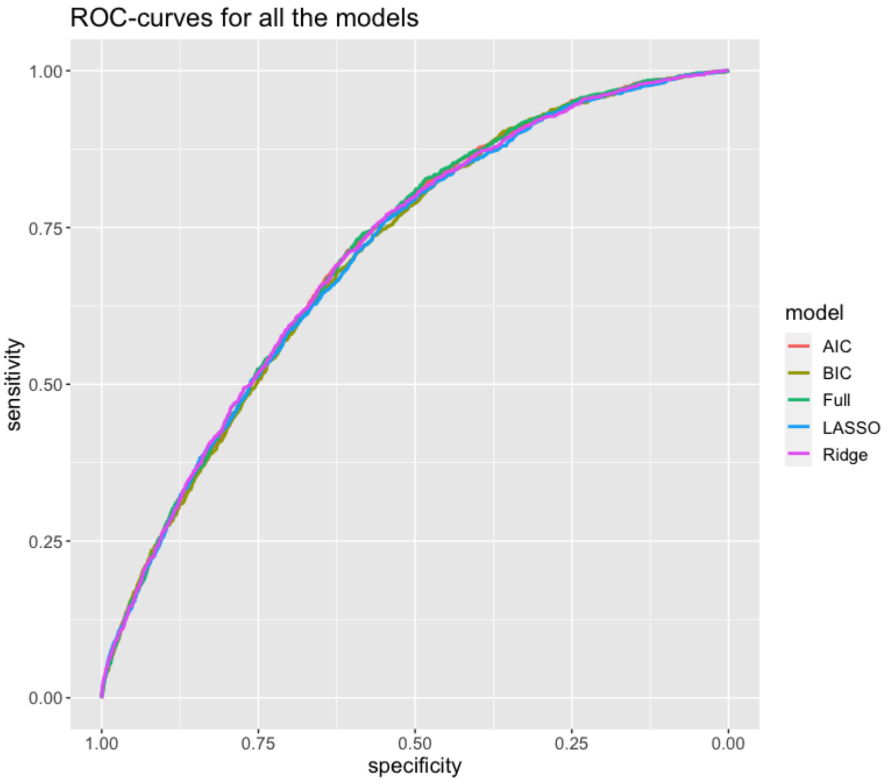


Figure 19: ROC curves for the logistic regression models predicting probability of there being a shot at the end of an attack attempt

Table 2 shows the area under curves for the five models, along with the corresponding 95% confidence interval.

Model	AUC estimate	Lower Limit	Upper Limit
Full	0.710	0.696	0.724
AIC	0.709	0.695	0.723
BIC	0.704	0.690	0.718
Lasso	0.704	0.690	0.718
Ridge	0.709	0.695	0.723

Table 2: AUC values for the five different logistic regression models. As can be seen both in figure 19 and in this table, the models are very similar in performance.

The optimal threshold value for p was derived by finding the value where specificity equals sensitivity. Here are confusion matrices for the five models given this respective threshold.

(a) Full model confusion matrix.
Specificity = Sensitivity = 0.649
Threshold = 0.162

	Predicted no shot	Predicted shot
True no shot	4616	2492
True shot	483	894

(b) AIC model confusion matrix
Specificity = Sensitivity = 0.653
Threshold = 0.164

	Predicted no shot	Predicted shot
True no shot	4638	2470
True shot	479	898

(c) BIC model confusion matrix.
Specificity = Sensitivity = 0.646
Threshold = 0.162

	Predicted no shot	Predicted shot
True no shot	4595	2513
True shot	487	890

(d) Lasso model confusion matrix
Specificity = Sensitivity = 0.646
Threshold = 0.164

	Predicted no shot	Predicted shot
True no shot	4589	2519
True shot	488	889

(e) Lasso/Ridge elastic net model confusion matrix
Specificity = Sensitivity = 0.653
Threshold = 0.166

	Predicted no shot	Predicted shot
True no shot	4641	2467
True shot	478	899

Table 3: Confusion matrices for the five logistic regression models

5.3.2 Gamma Regression Evaluation

For gamma regression, the models are compared relatively to each other with BIC and AIC scores. The following plot shows the ranking of the five different models with respect to these two metrics.

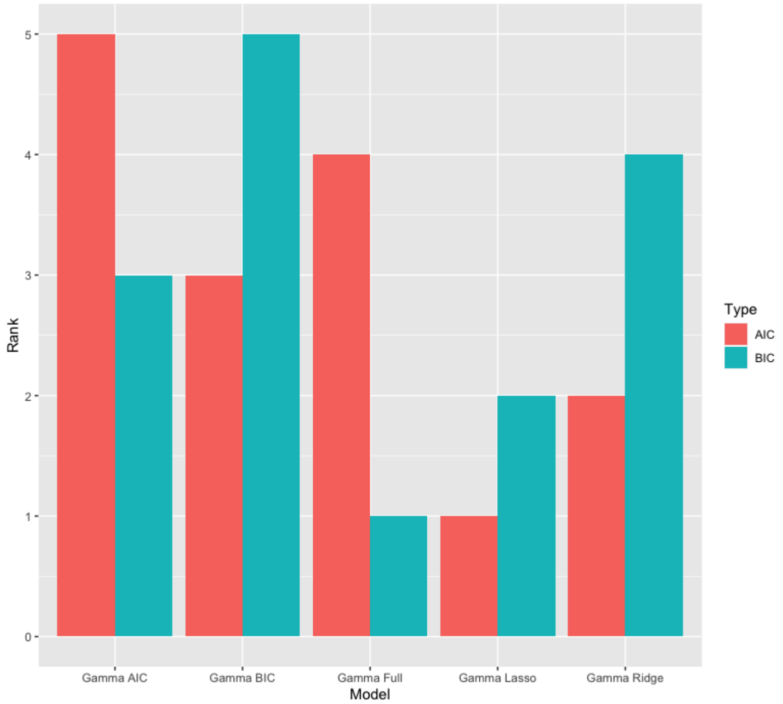


Figure 20: AIC and BIC rankings of the different gamma regression models. A higher ranking number implies a better model. Thus 5 is the best and 1 is the worst.

5.3.3 Logit-Linear Regression

These models are evaluated according to their R^2 and adjusted R^2 values, which are presented in the following plot.

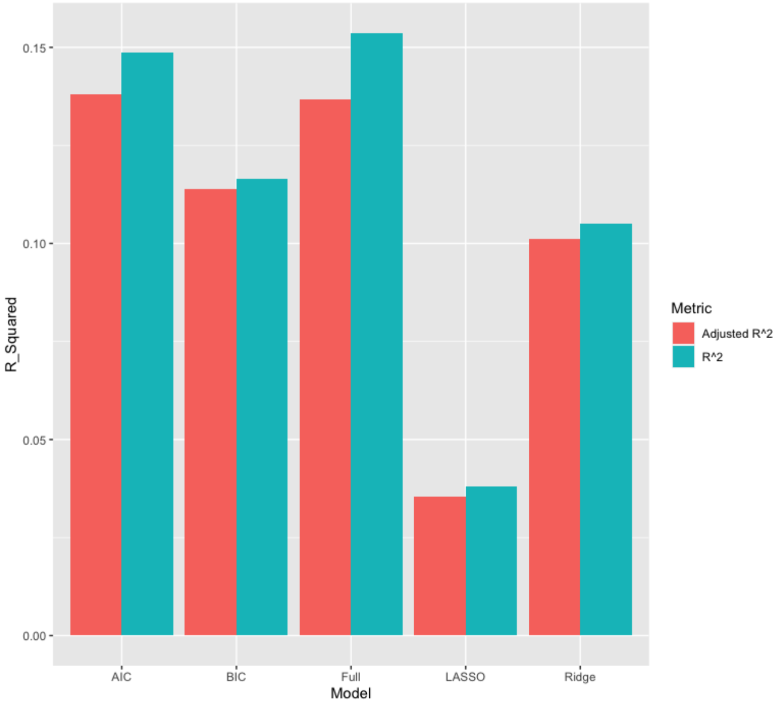


Figure 21: Logit-linear regression models and their respective R^2 and adjusted R^2 values

6 Discussion

6.1 Correlation discussion

Regarding correlation, an interesting thing to note is that there is a substantial negative correlation between nubmer of passes in an attack and the average pass distance. This means that the more passes a team plays, the shorter they tend to be on average. Intuitively, this can be explained with that if a team tries to attack quickly they likely do it by playing longer balls forward, and otherwise they play shorter passes in a more methodical and composed manner.

There is an expected positive correlation between the synchronization measurements. There is also an expected positive correlations between the different features describing size and spread of attacking and defending team respectively. Except for between defence vertical length and defence horizontal length (look at figure 10 to see what this means). This is slightly surprising as I thought the teams were either spread out in their defensive shape or kept tight together. This correlation statistic implies that when a defending team is more spread out in the horizontal direction, they keep more together in the vertical direction and vice versa.

6.2 Model parameter discussion

Here I'll discuss the most important interpretations from parameters in the different models.

6.2.1 Efficiency of Crosses and Long Passes

First thing that catches the eye in the plots in chapter 5.2 is that crosses are incredibly effective passes when attempting to score goals. Playing a cross generally means a higher probability of getting a goal scoring opportunity, as can be seen by the positive coefficient in figure 15. Four of the five logistic regression models have this variable as statistically significant. What's more, playing crosses is by far the variable measured in this project that has the biggest effect on chance quality when a shot is taken. Looking at figure 17, we see that the standardized coefficients for this variable is head and shoulders larger than any other variable - for all but the lasso regression models. There is a slight risk here that this variable is biased as when a team is able to make a cross into the box, they have already made substantial attacking progress. When compared to other attacks, it's possible that what this variable/coefficient measures the progress already made when the cross is played. This seems plausible for the logistic regression case, but not for the case where only attacks resulting in a shot are included. Crosses are most often taken from close to the sideline, far from the goal, which is hardly an improved attacking position compared to the average position from which shots are taken. Rather, the conclusion to draw here is that crosses are very effective ways of scoring goals when attacking against an organized defense.

Similarly, what's called long passes in this project seems to be effective both in creating more and better goal scoring opportunities. The effectiveness of long passes is increased when the pass tempo is higher, as can be seen by the positive interaction term in figure 15.

6.2.2 Passes Between Lines

One might look at figure 17 and 18 and conclude that making passes between the lines of the opponents defence is effective in creating better goal scoring opportunities. However, only certain types of these passes have a net positive effect. Notice that the variables between lines pass back, between lines pass to wing and between lines combinations all have statistically significant negative coefficients that are of roughly the same size as that of passes between lines. The relationship between these variables is a bit special, for example when a pass is played to the wing from between the lines, both of the variables passes between lines and between lines pass to wing are increased by one. Therefore, what we actually see here is the variables cancelling each other out. Looking at one model in specific, for example the gamma full which can be found in figure 9 in appendix A, we can see the value of the unstandardized coefficients. For this model, the net result of any pass from between the lines to the wing, back to defense, or to another player between the lines has a slightly negative net impact on chance quality. The only related variable not included with a negative coefficient is that representing passes behind the defence from between the lines. Such passes thus have a positive effect on chance quality.

6.2.3 Higher Movement Speed of Attackers Improves Chance Creation

Moving fast and a lot when attacking is by many believed to be a key in breaking down an organized defense. By making disruptive runs the attacking team can open up spaces and advance towards the opponents goal. Looking at figure 15 we see that the most important positive feature of an attack to determine whether there is a shot taken is the attack vertical velocity. It can be considered a bit surprising that it's not horizontal velocity that's the most important one (although that also improves chance creation), but vertical i.e. moving from side to side. The effect of moving quickly is further increased with attack duration, which I interpret as that if attackers consistently move into new spaces over a longer time, eventually the defense will lose marking and an opportunity to advance will present itself. Notice that, much like the previous argument with passes between lines, although the coefficient for *vertical velocity* is negative in figure 16, it's effect is much smaller than that of *attack vertical velocity.* In other words, players that are positioned deeper than the furthest forward standing defending midfielder are the ones whose movement matters most.

Interestingly, the same relationship can be seen in the chance quality plots, where *attack horizontal velocity* and *horizontal velocity* have opposite signs. Also, vertical movement has bigger effect on chance creation in general, but horizontal movement has bigger effect on chance quality. This effect is also increased with number of passes. The total conclusion to draw from the velocity measurements is regardlessly that the players playing in attacking positions can increase chance creation and chance quality by moving more and faster.

6.2.4 Switching Sides Generates More Shots

Switching sides seems to almost unanimously result in more chance creation, but it has no noticeable effect on chance quality. If anything, one of the models indicate that side switching is negatively correlated with chance quality, although it's only one model and the coefficient is barely significant.

6.2.5 Effective Defensive Strategies

Looking at the results from a defending team perspective, high synchronization, or cluster amplitude, seems to be effective in stopping the opponents from taking shots as can be seen in figure 16. However, higher horizontal synchronization in the defending team for some reason generally tends to lead to better goal scoring opportunities.

Besides synchronization, defensively it seems as though defending more spread out is preferable. Both defence horizontal and vertical length have negative coefficients in logistic regression, figure 16. Also, having more players in midfield/defence seems to help make the opponents take shots from worse positions as seen in figure 18. In relation to this, there is also the most surprising result so far, which is that defensive pitch control has a positive coefficient for four of the models measuring chance quality. That means that the more the defensive team controls the space between their defensive and midfield lines, the higher quality chances the opponents tend to create. This completely contradicts the assumption made in this project of that defending teams should aim to control this space specifically and allow opponents possession in other parts of the pitch. I have no good attempt at explaining this result, but a wild guess is that the more the defenders stand close together and control the space between themselves, the more they open up space around them which can be used by the attacking team. This hypothesis further strengthens the argument that defending more spread out rather than tight together is preferable. Additionally, it ties well together with crosses being an effective attacking method. The more vertically stretched out the defending team is, the more likely they are to have players pressing the attacking teams' wingers and thus preventing them from making dangerous crosses.

6.3 Model Evaluation Discussion

The five logistic models have very similar ROC curves and AUC values, as can be seen in chapter 5.5.1. An AUC value of 0.7, which is the level at which these models perform, is neither great nor terrible. Football is a complex sport with a lot of stochastic elements, and finding models that accurately predicts the outcome of attacks based on variables at the level done here is a difficult task. For the gamma regression model evaluation, it's unsurprising that the AIC and BIC models perform well. What is surprising however is that the Lasso model performs very poorly in terms of both of these metrics. The Lasso model makes variable selection from all the variables in chapter 4.7, as well as all pairwise interaction terms between those. The AIC and BIC models had just the single variables available for selection, but still outperform the Lasso model.

Looking at figure 21, we again see the Lasso regression model performing very poorly. On the other hand, the AIC and full model both perform at a decent level even in absolute terms, with adjusted R^2 values of just under 0.14. To put this number in a football data science perspective, I'd like to compare it to (McFadden's pseudo) R^2 values of expected goals models. As touched upon in chapter 3.2, many data scientists and companies have attempted creating logistic regression models for expected goals. I consider shot situations less complex and more alike than organized attacks, meaning they should be easier to model. Still, the models for expected goals rarely tend to have (McFadden's pseudo) R^2 values higher than 0.2.

In this project, the full model, measuring probability of an attack resulting in a shot, has an R^2 value over 0.15. In the context of complex football logistic regression models, this can be considered a success.

6.4 Future work

The aim of this project was to identify which attacking and defending strategies/characteristics are most effective. A number of things have been tried, but football is, as mentioned, already a very complex sport with multiple tactical approaches. Therefore, there is certainly an opportunity to expand on the explanatory variables and include ones that might be of specific interest to certain teams or coaches. The explanatory variables could really be anything from complex and specific patterns of play, to formation of the defending team or even team variables specifying who's playing. The concept of filtering raw data and organizing it in attack attempts with explanatory variables and attack success as response variable can be widely utilized to investigate efficiency of many things.

A lot of the variables in this project are high level, they give the general idea of what has happened in an attack, but lack more detailed information. For example, the variable *cross was played* is the same for a low 30 meter cross into the box from close to the touchline as it is for a 60 meter punt into the box from close to the halfway line. It'd give more insightful results to see precisely which types of crosses are most efficient rather than just any cross. The same idea applies to many of the other variables investigated here. More detail means more insight, but also more advanced algorithms for processing the data.

Having expected goals and whether there was a shot as response variable seems intuitively like a good idea. However, there would be extra value added if rather than just having the binary response variable shot or no shot, you'd add other outcomes of an attack attempt. Clearly it's more valuable to get a free kick in a dangerous position than it is to lose the ball to the opponents attackers in a position where they can start a counter attack. The logistic regression in this project does not separate between these two outcomes, but rather says that both of them are failures. If on the other hand, one would find an appropriate way of measuring the value of the ”ending state” of an attack, then that would likely be preferable to what's been used here. There is for example a recently developed tool called expected threat which evaluates possessions on different parts of the field [13]. This could help improve the response variable.

There are many different ways of defending in football, and often what works when attacking against one team doesn't work when attacking against another. Therefore, it'd be interesting to find common themes in the organization of a defense and investigate separately what is efficient against each of those. For example, when a team is defending tight together and is moving with high synchronization levels, what attacking strategies are efficient might differ from when a team is more spread out and moving less synchronously. Using some sort of clustering algorithm to separate the data set of attack attempts in different defending structures would nuance the interpretation of what attacking strategies are most efficient.

Lastly, it shall be said that this project uses data from the 2021/2022 Eredevise season and nothing else. What strategies are effective might very well depend on how teams tend to defend, quality of football, style of play of teams in different leagues etc. Therefore, it'd be interesting to see a comparison between the Dutch and other leagues.

7 Conclusion

With tracking data and event data as input, the full Eredevise (highest Dutch division) 2021/2022 season was analysed and attack attempts against organized defenses were identified. The data was filtered and a set of variables explaining characteristics of each attack attempt was made. Further, three regression methods were used to find relationship between attack characteristics and whether a shot was made as well as the quality of chances created.

The best attacking strategies proved to be making crosses, playing longer passes, having high movement speed especially among the forwards, and playing lots of passes. Defensively moving with high overall synchronization among players, and spreading out to cover more space rather than keep a tight formation are the best ways to prevent the opponents from scoring.

The models perform at a satisfactory level considering the complexity of the sport being analyzed. However, these models will not have good accuracy in predicting the future outcome of an attack based on it's characteristics. Rather, the methods deployed here and the models being generated should be used to gain insight on what strategies are historically successful and are desirable for teams to use.

8 References

References

[1] Wyscout event data. 2020-2022. Wyscout

https://footballdata.wyscout.com/packages/

[2] Signality tracking data. 2021-2022. Signality

https://www.signality.com/product

[3] TRACAB tracking data. 2020-2022. TRACAB

[4] Shaw Laurie. 2020. Friends-of-Tracking-Data-FoTD/LaurieOnTracking. GitHub Repository.

https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking

[5] Spearman William, March 2018, Beyond expected goals, MIT Sloan Sports Analytics Conference, (2022-02-05)

https://www.researchgate.net/publication/327139841_Beyond_Expected_Goals

[6] Arastey, G. M. (2019, November 22). What are expected goals (XG)? Sport Performance Analysis. (February 15, 2022).

https://www.sportperformanceanalysis.com/article/what-are-expected-goals-xg

[7] Kelly Ryan. 2019-10-27. What is XG in Football, How is the statistic calculated? Goal.com. (2022-02-22).

https://www.goal.com/en/news/what-is-xg-football-how-statistic-calculated/h42z0iiv8mdg1ub10iisg1dju

[8] Baíllo, A. and Chacón, J. E. (2021). Statistical outline of animal home ranges: An application of set estimation. Handbook of Statistics, 44. Chapter 2.1.1.1

https://doi.org/10.1016/bs.host.2020.10.002

[9] Duarte Ricardo, Araújo Duarte, Correia Vanda, Davids Keith, Marques Pedro, Richardson Michael. 2013-07-03. Competing together: Assessing the dynamics of team-team and player-team synchrony in professional association football. Human Movement Science 32 (2013), p.555-566.

[10] Reynolds Douglas. (2015) Gaussian Mixture Models. In: Li S.Z., Jain A.K. (eds) Encyclopedia of Biometrics. Springer, Boston, MA.
https://doi.org/10.1007/978-1-4899-7488-4_196

[11] Brilliant https://brilliant.org/wiki/gaussian-mixture-model/

[12] Lord, D., Qin, X., Geedipally, S. R. (2021). Fundamentals and data collection. Highway Safety Analytics and Modeling, 2, 45–49.

https://doi.org/10.1016/b978-0-12-816818-9.00010-x

[13] Singh Karun. (2018). Introducing Expected Threat (xT).

https://karun.in/blog/expected-threat.html

Appendix A - Model Parameters and Statistical Signifance Level

Here are tables presenting all the 15 models created and discussed in this project.

Table 4: Logistic Full Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-0.6053	0.474	
Attack Duration	0.0162	0.0006	***
Pass Amount	-0.0342	0.029	*
Pass Tempo	0.3920	0.060	.
Average Pass Length	-0.0384	1.95e-08	***
Long Passes	0.1592	0.0035	**
Attack Vertical Velocity	2.2800	3.50e-10	***
Attack Horizontal Velocity	0.4000	0.129	
Horizontal Velocity	0.4331	0.0041	**
Vertical Velocity	-0.3443	0.126	
Attack Horizontal Length	0.0197	0.249	
Attack Vertical Length	-0.0073	0.657	
Attacking Hull Area	0.0098	0.335	
Defensive Pitch Control	0.5328	0.468	
Passes Between Lines	0.0709	0.225	
Cross Was Played	0.1572	0.026	*
Side Switches	0.2285	2.78e-07	***
Between Lines Pass Behind Defense	-0.0694	0.667	
Between Lines Combinations	-0.0146	0.849	
Between Lines Pass Back	-0.0816	0.300	
Between Lines Pass to Wing	-0.1426	0.107	
Between Lines Dribble	0.4675	2.93e-10	***
Synchronization	-2.4019	1.71e-07	***
Horizontal Synchronization	0.6391	0.364	
Vertical Synchronization	-0.3387	0.632	
Defense Horizontal Length	-0.0524	7.50e-05	***
Defense Vertical Length	-0.0828	3.56e-14	***
Players in Midfield/Defense	-0.0015	0.981	

Table 5: Logistic AIC Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-0.069	0.8822	
Attack Duration	0.014	0.0007	***
Pass Amount	-0.026	0.0547	.
Pass Tempo	0.353	0.0888	.
Average Pass Length	-0.039	4.17e-09	***
Long Passes	0.154	0.0040	**
Attack Vertical Velocity	2.265	5.97e-11	***
Attack Horizontal Velocity	0.394	0.1125	
Horizontal Velocity	0.470	0.000844	***
Vertical Velocity	-0.362	0.0942	.
Attack Horizontal Length	0.026	0.0124	*
Attacking Hull Area	0.006	0.1158	
Cross Was Played	0.167	0.0174	*
Side Switches	0.223	4.05e-07	***
Between Lines Dribble	0.477	9.17e-11	***
Synchronization	-2.351	9.34e-08	***
Defense Horizontal Length	-0.053	3.01e-05	***
Defense Vertical Length	-0.085	4.73e-16	***

Table 6: Logistic BIC Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	0.089	0.8400	
Attack Duration	0.009	0.0001	***
Average Pass Length	-0.028	6.71e-07	***
Attack Vertical Velocity	2.707	¡ 2e-16	***
Horizontal Velocity	0.685	5.40e-14	***
Vertical Velocity	-0.583	0.0026	**
Attack Horizontal Length	0.038	1.00e-05	***
Side Switches	0.219	2.17e-07	***
Between Lines Dribble	0.469	1.53e-10	***
Synchronization	-2.310	8.87e-08	***
Defense Horizontal Length	-0.058	1.66e-06	***
Defense Vertical Length	-0.077	¡ 2e-16	***

Table 7: Logistic Lasso Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	0.0294517226	0.9839970849	
Cross Was Played	0.1785087668	0.0108872133	*
Attack Horizontal Velocity	-1.8349054658	0.2396655241	
Defence Vertical Length	0.0322255512	0.7309677311	
Attack Duration*Pass Tempo	0.0129852559	0.4760897077	
Attack Vertical Velocity*Attack Duration	0.0079166931	0.4348389942	
Long Passes*Pass Tempo	0.5118672985	0.0004195313	***
Attack Horizontal Velocity*Pass Tempo	0.0114693062	0.9717333419	
Pass Tempo*Side Switches	0.6016953645	0.1619321427	
Between Lines Dribble*Pass Tempo	0.3102341022	0.3745046540	
Average Pass Length*Synchronization	-0.0938391132	0.0232942254	*
Average Pass Length*Defence Vertical Length	0.0009299609	0.3125100336	
Attack Horizontal Velocity*Attack Vertical Velocity	-1.0288065807	0.0678551972	.
Attack Vertical Velocity*Horizontal Velocity	0.5533115450	0.0047490827	**
Attack Vertical Velocity*Attack Horizontal Length	0.0386570397	0.2333318334	
Attack Vertical Velocity*Defensive Pitch Control	-0.0187534810	0.9917453043	
Attack Vertical Velocity*Between Lines Dribble	0.5654199803	0.2254909989	
Attack Horizontal Velocity*Attack Horizontal Length	-0.0166164545	0.4947089040	
Attack Horizontal Velocity*Attacking Hull Area	0.0099131253	0.0226638846	*
Attack Horizontal Velocity*Defensive Pitch Control	3.7665948604	0.0503538623	.
Attack Horizontal Velocity*Side Switches	-0.1428402548	0.3518350200	
Between Lines Dribble*Horizontal Velocity	-0.0187135613	0.9177508362	
Attack Horizontal Length*Between Lines Dribble	0.0021073755	0.8771092422	
Defensive Pitch Control*Defense Vertical Length	-0.0927412489	0.0900058959	.
Players in Midfield/Defense*Side Switches	0.0125134389	0.6026573942	
Between Lines Dribble*Defense Horizontal Length	-0.0034348778	0.9020138246	
Synchronization*Defence Vertical Length	-0.0018552954	0.9459562636	
xSynchronization*Vertical Synchronization	-2.6333680064	0.0664612840	.
Defense Vertical Length*Vertical Synchronization	0.0102684669	0.7977201602	
Defense Vertical Length*Defense Vertical Length	-0.0013952321	0.2634666906	

Table 8: Logistic Elastic Net Ridge/Lasso Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	0.0294	0.983	
Cross Was Played	0.1785	0.010	*
Attack Horizontal Velocity	-1.8349	0.239	
Defence Vertical Length	0.0322	0.730	
Attack Duration*Pass Tempo	0.0129	0.476	
Attack Vertical Velocity*Attack Duration	0.0079	0.434	
Long Passes*Pass Tempo	0.5118	0.0004	***
Attack Horizontal Velocity*Pass Tempo	0.0114	0.971	
Pass Tempo*Side Switches	0.6016	0.161	
Between Lines Dribble*Pass Tempo	0.3102	0.374	
Average Pass Length*Synchronization	-0.0938	0.023	*
Average Pass Length*Defence Vertical Length	0.0009	0.312	
Attack Horizontal Velocity*Attack Vertical Velocity	-1.0288	0.067	.
Attack Vertical Velocity*Horizontal Velocity	0.5533	0.004	**
Attack Vertical Velocity*Attack Horizontal Length	0.0386	0.233	
Attack Vertical Velocity*Defensive Pitch Control	-0.0187	0.991	
Attack Vertical Velocity*Between Lines Dribble	0.5654	0.225	
Attack Horizontal Velocity*Attack Horizontal Length	-0.0166	0.494	
Attack Horizontal Velocity*Attacking Hull Area	0.0099	0.022	*
Attack Horizontal Velocity*Defensive Pitch Control	3.7665	0.050	.
Attack Horizontal Velocity*Side Switches	-0.1428	0.351	
Between Lines Dribble*Horizontal Velocity	-0.0187	0.917	
Attack Horizontal Length*Between Lines Dribble	0.0021	0.877	
Defensive Pitch Control*Defense Vertical Length	-0.0927	0.090	.
Players in Midfield/Defense*Side Switches	0.0125	0.602	
Between Lines Dribble*Defense Horizontal Length	-0.0034	0.902	
Synchronization*Defence Vertical Length	-0.0018	0.945	
xSynchronization*Vertical Synchronization	-2.6333	0.066	.
Defense Vertical Length*Vertical Synchronization	0.0102	0.797	
Defense Vertical Length*Defense Vertical Length	-0.0013	0.263	

Table 9: Gamma Full Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-2.7805	0.00554	**
Attack Duration	-0.0321	0.00000..	***
Pass Amount	0.1073	0.00000..	***
Pass Tempo	-0.5192	0.02323	*
Average Pass Length	-0.0135	0.05784	.
Long Passes	0.1523	0.00612	**
Attack Vertical Velocity	0.2934	0.44600	
Attack Horizontal Velocity	0.7510	0.00587	**
Horizontal Velocity	-0.4074	0.01102	*
Vertical Velocity	-0.1373	0.56869	
Attack Horizontal Length	0.0031	0.86247	
Attack Vertical Length	-0.0183	0.27372	
Attacking Hull Area	-0.0001	0.99252	
Defensive Pitch Control	1.7357	0.02151	*
Passes Between Lines	0.2363	0.00004	***
Cross Was Played	0.5908	0.00000..	***
Side Switches	-0.0779	0.06354	.
Between Lines Pass Behind Defense	-0.2580	0.11276	
Between Lines Combinations	-0.2852	0.00017	***
Between Lines Pass Back	-0.3155	0.00008	***
Between Lines Pass to Wing	-0.3292	0.00015	***
Between Lines Dribble	-0.0051	0.93986	
Synchronization	-0.2072	0.69063	
Horizontal Synchronization	3.4441	0.00246	**
Vertical Synchronization	-1.5189	0.08365	.
Defence Horizontal Length	0.0126	0.33387	
Defence Vertical Length	0.0130	0.22231	
Players in Midfield/Defence	-0.2035	0.00486	**

Table 10: Gamma AIC Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-2.575	0.97942	
Attack Duration	-0.031	0.00636	**
Pass Amount	0.107	0.01948	*
Pass Tempo	-0.508	0.22716	
Average Pass Length	-0.012	0.00703	**
Long Passes	0.151	0.05528	.
Attack Horizontal Velocity	0.809	0.23465	
Horizontal Velocity	-0.453	0.14609	
Attack Vertical Length	-0.014	0.00575	**
Defensive Pitch Control	1.478	0.67904	
Passes Between Lines	0.236	0.05733	.
Cross Was Played	0.586	0.06930	.
Side Switches	-0.089	0.04087	*
Between Lines Pass Behind Defense	-0.258	0.16197	
Between Lines Combinations	-0.286	0.07522	.
Between Lines Pass Back	-0.320	0.07915	.
Between Lines Pass to Wing	-0.325	0.08654	.
Horizontal Synchronization	3.302	1.05071	
Vertical Synchronization	-1.599	0.84593	
Defence Horizontal Length	0.018	0.01085	*
Players in Midfield/Defence	-0.173	0.06821	.

Table 11: Gamma BIC Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-2.321	0.3033	
Attack Duration	-0.028	0.0058	**
Pass Amount	0.099	0.0181	*
Attack Horizontal Velocity	0.349	0.1214	
Attack Vertical Length	-0.014	0.0054	**
Passes Between Lines	0.185	0.0526	.
Cross Was Played	0.576	0.0657	.
Between Lines Combinations	-0.229	0.0726	.
Between Lines Pass Back	-0.265	0.0773	.
Between Lines Pass to Wing	-0.255	0.0848	.

Table 12: Gamma Lasso Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-2.1719	0.256	
Attack Horizontal Velocity*Pass Amount	0.0695	0.011	*
Pass Tempo*Side Switches	-0.4436	0.326	
Attack Horizontal Velocity*Defensive Pitch Control	0.5436	0.179	
Attack Vertical Length*Vertical Synchronization	-0.0140	0.013	*
Attack Vertical Length*Players in Midfield/Defense	-0.0010	0.001	**
Side Switches*Defense Vertical Length	0.0010	0.003	**

Table 13: Gamma Elastic Net Ridge/Lasso Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-2.3927	0.267	
Cross Was Played	0.5303	0.066	.
Attack Horizontal Velocity*Pass Amount	0.0580	0.011	*
Pass Tempo*Side Switches	-0.2544	0.338	
Attack Horizontal Velocity*Defensive Pitch Control	0.4406	0.186	
Attack Vertical Length*Vertical Synchronization	0.00007	0.013	*
Attack Vertical Length*Players in Midfield/Defense	-0.0019	0.001	**
Side Switches*Defense Vertical Length	-0.0022	0.003	**

Table 14: Logit Linear Full Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-2.727	1.247	
Attack Duration	-0.042	0.008	**
Pass Amount	0.116	0.024	*
Pass Tempo	-0.452	0.284	
Average Pass Length	-0.009	0.008	**
Long Passes	0.256	0.069	.
Attack Vertical Velocity	0.097	0.479	
Attack Horizontal Velocity	1.192	0.339	
Horizontal Velocity	-0.592	0.199	
Vertical Velocity	0.031	0.300	
Attack Horizontal Length	-0.011	0.022	*
Attack Vertical Length	-0.022	0.020	*
Attacking Hull Area	0.001	0.012	*
Defensive Pitch Control	1.983	0.940	
Passes Between Lines	0.266	0.071	.
Cross Was Played	0.884	0.087	.
Side Switches	-0.086	0.052	.
Between Lines Pass Behind Defense	-0.077	0.202	
Between Lines Combinations	-0.302	0.094	.
Between Lines Pass Back	-0.244	0.099	.
Between Lines Pass to Wing	-0.279	0.108	
Between Lines Dribble	0.111	0.085	.
Synchronization	0.643	0.649	
Horizontal Synchronization	2.722	1.415	
Vertical Synchronization	-1.341	1.093	
Defence Horizontal Length	0.023	0.016	*
Defence Vertical Length	0.018	0.013	*
Players in Midfield/Defence	-0.350	0.089	.

Table 15: Logit Linear AIC Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-3.061	1.150	
Attack Duration	-0.041	0.007	**
Pass Amount	0.116	0.023	*
Pass Tempo	-0.396	0.276	
Long Passes	0.220	0.066	.
Attack Horizontal Velocity	1.352	0.281	
Horizontal Velocity	-0.532	0.181	
Attack Vertical Length	-0.022	0.008	**
Defensive Pitch Control	1.841	0.845	
Passes Between Lines	0.255	0.065	.
Cross Was Played	0.841	0.082	.
Side Switches	-0.090	0.051	.
Between Lines Combinations	-0.264	0.089	.
Between Lines Pass Back	-0.229	0.095	.
Between Lines Pass to Wing	-0.264	0.103	
Horizontal Synchronization	2.299	1.175	
Defence Vertical Length	0.020	0.012	*
Players in Midfield/Defence	-0.337	0.084	.

Table 16: Logit Linear BIC Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-3.773	0.1853	
Attack Duration	-0.036	0.0071	**
Pass Amount	0.110	0.0208	*
Attack Horizontal Velocity	0.698	0.1433	
Cross Was Played	0.880	0.0803	.

Table 17: Logit Linear Lasso Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-3.981	0.163	
Attack Horizontal Velocity*Pass Amount	0.057	0.014	*
Attack Horizontal Velocity*Defensive Pitch Control	1.268	0.729	
Attack Horizontal Velocity*Passes Between Lines	0.011	0.037	*
Attack Horizontal Velocity*Horizontal Synchronization	0.020	0.564	

Table 18: Logit Linear Elastic Net Ridge/Lasso Model. Estimates can be interpreted as the change in response variable when the explanatory variable changes with one unit. The third column gives the level of significance, i.e. the probability that the coefficient has no real effect on the response variable when all other variables in the model are included. Confidence level column is . for 90%, * for 95%, ** for 99% and *** for 99.9%

Variable	Estimate(β)	$Pr(>\|t\|)$	Confidence level
(Intercept)	-4.019	0.163	
Cross Was Played	0.820	0.081	.
Attack Horizontal Velocity	-0.970	1.063	
Attack Horizontal Velocity*Pass Amount	0.025	0.014	*
Attack Horizontal Velocity*Defensive Pitch Control	1.035	0.782	
Attack Horizontal Velocity*Passes Between Lines	0.061	0.036	*
Attack Horizontal Velocity*Horizontal Synchronization	1.289	1.234	