KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree Project in Mathematical Statistics

Second Cycle 30 credits

Identifying Optimal Throw-in Strategy in Football Using Logistic Regression

STEPHAN NIETO

Stockholm, Sweden 2023

Abstract

Set-pieces such as free-kicks and corners have been thoroughly examined in studies related to football analytics in recent years. However, little focus has been put on the most frequently occurring set-piece: the throw-in. This project aims to investigate how football teams can optimize their throw-in tactics in order to improve the chance of taking a successful throw-in. Two different definitions of what constitutes a successful throw-in are considered, firstly if the ball is kept in possession and secondly if a goal chance is created after the throw-in. The analysis is conducted using logistic regression, as this model comes with high interpretability, making it easier for players and coaches to gain direct insights from the results. A substantial focus is put on the investigation of the logistic regression assumptions, with the greatest emphasis being put on the linearity assumption. The results suggest that long throws directed towards the opposition's goal are the most effective for creating goal-scoring opportunities from throw-ins taken in the attacking third of the pitch. However, if the throw-in is taken in the middle or defensive regions of the pitch, the results interestingly indicate that throwing the ball backwards leads to increased chance of scoring. When it comes to retaining the ball possession, the results suggest that throwing the ball backwards is an effective strategy regardless of the pitch position. Moreover, the project outlines how feature transformations can be used to improve the fitting of the logistic regression model. However, it turns out that the most significant improvement in accuracy of logistic regression occurs when incorporating additional relevant features into the model. In such case, the logistic regression model achieves a predictive power comparable to more advanced machine learning methods.

Keywords

Set-piece, throw-in, football analytics, optimal strategy, logistic regression, model assumptions, feature importance, feature transformations, gradient boosting.

Sammanfattning

Titel: Identifiering av optimal inkaststrategi i fotboll med logistisk regression

Fasta situationer såsom frisparkar och hörnor har varit välstuderade i studier rörande fotbollsanalys de senaste åren. Lite fokus har emellertid lagts på den vanligast förekommande fasta situationen: inkastet. Detta projekt syftar till att undersöka hur fotbollslag kan optimera sin inkasttaktik för att förbättra möjligheterna till att genomföra ett lyckat inkast. Två olika definitioner av vad som utgör ett lyckat inkast beaktas, dels om bollinnehavet behålls och dels om en målchans skapas efter inkastet. Analysen görs med logistisk regression eftersom denna modell har hög tolkningsbarhet, vilket gör det lättare för spelare och tränare att få direkta insikter från resultaten. Stort fokus läggs på undersökning av de logistiska regressionsantagandena, där störst vikt läggs på antagandet gällande linjäritet. Resultaten tyder på att långa inkast riktade mot motståndarnas mål är de mest gynnsamma för att skapa en målchans från inkast tagna i den offensiva tredjedelen av planen. Om inkastet istället tas från de mellersta eller defensiva delarna av planen tyder resultaten intressant nog på att inkast riktade bakåt leder till ökad chans till att göra mål. När det kommer till att behålla bollinnehavet visar resultaten att kast bakåt är en gynnsam strategi, oavsett var på planen inkasten tas ifrån. Vidare visar projektet hur variabeltransformationer kan användas för att förbättra modellanpassningen för logistisk regression. Det visar sig dock att den tydligaste förbättringen fås då fler relevanta variabler läggs till i modellen. I sådant fall, får logistisk regression en prediktiv förmåga som är jämförbar med mer avancerade maskininlärningsmetoder.

Nyckelord

Fasta situationer, inkast, fotbollsanalys, optimal strategi, logistisk regression, modellantaganden, variabelvikt, variabeltransformationer, gradient boosting.

Acknowledgements

I would like to give a special thanks to my supervisor and examiner Joakim Andén-Pantera for his supervision and being highly helpful throughout this project. I would also like to give my warmest thanks to David J.T. Sumpter for providing me with the opportunity to work on this project and for giving valuable advice during our weekly meetings.

Furthermore, I would like to express my appreciation to Zacharias Ljungström and Ágúst Pálmason Morthens, who also conducted their master theses at Twelve, for their meaningful inputs and and insightful discussions.

Stephan Nieto, May 2023

Author

Stephan Nieto, snieto@kth.se
M.Sc. Applied and Computational Mathematics, Mathematics of Data Science
KTH Royal Institute of Technology

Place for Project

Twelve Football AB
Stockholm, Sweden

Examiner

Joakim Andén-Pantera
Department of Mathematics
KTH Royal Institute of Technology

Supervisor

Joakim Andén-Pantera
Department of Mathematics
KTH Royal Institute of Technology

David J.T. Sumpter
Twelve Football AB
Stockholm, Sweden

Contents

Chapter 1

Introduction

1.1 Relevance

Set-pieces, such as free-kicks, corners, and penalties, play a crucial role in football games, offering valuable opportunities for teams to score. In fact, prior studies have demonstrated that set-pieces account for approximately 35% of all goals scored [26] and as a result teams have put great emphasis on preparation and practice of corners, free-kicks and penalties [20]. These set-pieces have also been the focus of recent research in this area. For instance, in a paper by Shaw and Gopaladesikan [19], offensive and defensive strategies used by football teams during corner kick situations were identified.

Less focus has been put on throw-ins, which is a set-piece awarded to a team when the opponents play the ball over the touchline, either on the ground or in the air [22]. In the MLS, the top professional football league in US and Canada, an average of 44 throw-ins per game occurred during the 2015–2018 seasons [11], making throw-ins more frequent than corner kicks, free kicks and goal kicks [20]. One of the reasons for the lack of research in the area of throw-ins could be explained by few goals originating directly from throw-ins. However, throw-ins could be seen as an opportunity for a team to increase possession of the ball, which in turn has shown to affect a team's chance of scoring and winning football games [9].

Stone et al. [20] conducted a recent study investigating the relationship between throw-ins and team performance, as well as the effect of the direction and length of a throw-in on possession retention and shot creation. The study found that 54% of the throw-ins led to retained possession during the 2018–2019 Premier League season. In another study, McKinley [11] created a model to predict possession retention using a

gradient-boosted ensemble of decision trees. Apart from these investigations, there is a lack of studies specifically focused on throw-ins, which justifies further exploration of this aspect of football.

1.2 Goals

With the absence of extensive research on throw-ins, this project aims to further investigate what factors contribute to a successful throw-in. In particular, the following research question will be addressed in the project:

> *How can players and teams optimize their throw-in tactics to improve their chances of success?*

To answer this question, two different ways of defining *success* will be explored. The first definition considers if the control of the ball, or ball possession, is kept by the team taking the throw-in, while the second definition takes into account whether the execution of the throw-in leads to the creation of a goal-scoring opportunity.

The results are aimed to be presented in such a way that football coaches and players can gain direct insights from the findings. The main goal is not to achieve the highest possible prediction accuracy but rather to understand which factors are the most important for the throw-in being successful. This is a key reason for why this project will primarily use logistic regression as this model comes with a high degree of interpretability. The analysis will be conducted using event data and a crucial part of the project will be to extract relevant features from the data set which have a high impact on the throw-in outcome based on football intuition.

In addition to seeking insights about the optimal throw-in strategy, this thesis also aims to explore the use of logistic regression and gain a more thorough understanding of how the underlying assumptions can be examined. In particular, great emphasis will be put on examining the linearity assumption before fitting a logistic regression model. Furthermore, another objective is to compare the logistic regression model with a more advanced machine learning method, with the purpose of gaining a deeper understanding of the strengths and limitations of logistic regression.

1.3 Data

The data will be taken from the 2022 season of the Swedish top tier Allsvenskan and will consist of event data from the data provider Wyscout. This data contains information from the 240 games played during the Allsvenskan season. In total, the data set

contains approximately 420 000 events, with an average of 1 750 events per game. Every event is categorized based on the type of action that an event corresponds to. The most frequent event type in the data set is the *pass*, with around 207 000 instances. Throw-ins are less common, with about 10 000 instances recorded throughout the season, resulting in an average of 42 throw-ins per game.

Apart from the event type, the Wyscout event data also includes a large number of other parameters which describe the events happening on the pitch, for example the match minute, start and end coordinates of events, which team is having possession of the ball and metrics which estimate the goal-scoring probability when a shot is taken. Using these parameters, in particular the start and end coordinates of a throw-in, it will be possible to create new features, such as the angle of the throw, which could yield significant value for the models.

1.4 Project outline

The report is structured as follows. Chapter 2 provides the necessary mathematical background for the project. It primarily focuses on the theory of logistic regression, but also provides an overview of gradient boosting and relevant football terminology. Chapter 3 outlines the project methodology, including the definition of features and targets. Another important part of this chapter is the description of how the linearity assumption of the logistic regression model is investigated. Chapter 4 presents the project's results, including the analysis of the linearity assumption, evaluation of the logistic regression models, results of gradient boosting and an analysis of the throw-in strategies among the teams in Allsvenskan 2022. The results are then analyzed and discussed in Chapter 5, followed by a summary of the main conclusions of the project in Chapter 6. Lastly, the report includes an appendix that contains additional results, which will be occasionally referred to throughout the report.

Chapter 2

Background

This chapter provides the mathematical foundation for the project. First, it describes the mathematics behind logistic regression, including the assumptions made in the model and how to interpret the model coefficients followed by a presentation of the evaluation metrics which will be used to assess the accuracy of the model. These sections aim to create a thorough understanding of the logistic regression model, enabling the reader to understand the model's potential and limitations. After this detailed review of logistic regression, gradient boosting is briefly introduced, as it will be used for comparison purposes later in the project. Finally, relevant football terminology is described to ensure that the reader is familiar with the key football concepts used throughout the thesis.

2.1 Logistic regression

Logistic regression is selected as the main model of this project because of its high degree of interpretability. This aspect is essential in this project as the aim is to provide teams, and more specifically coaches and players, with insights on how throw-ins could be taken in the most successful way. Understanding which aspects of the throw-in play the most important role and how these factors affect the throw-in outcome is of high importance and the high interpretability of logistic regression enables to identify these factors. Moreover, logistic regression has shown to be a good choice of model when dealing with small data sets [15], which makes it a suitable choice for this project.

2.1.1 The model

Logistic regression is a regression model which can be used for predicting a binary output, denoted by $y_i \in \{0, 1\}$. More specifically, given a number n of p dimensional data points $\mathbf{x} = (\mathbf{x}_i)_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$, one is interested in finding $P(y_i = 1|\mathbf{x}_i) = \pi_i$.

If the probability π_i is greater than a predefined threshold value, the data point $\mathbf{x}_i$ is assigned to class 1 and otherwise it is assigned to class 0. This *classification threshold value* is often set to 0.5.

In order to obtain a value between 0 and 1 for the probability π_i, one can make use of the logit function, also known as the *log odds*, which is defined as $\ln\frac{\pi_i}{1-\pi_i}$ in the current setting. Assuming that the logit transformation of π_i is linear with respect to the regression coefficients $\boldsymbol{\beta}$ the following model can be defined:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta} \tag{2.1}$$

Having defined this structure of the model, the probabilities $\boldsymbol{\pi} = (\pi_i)_{i=0}^n$ can be expressed as follows:

$$\boldsymbol{\pi} = \frac{\exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)} \tag{2.2}$$

The coefficients $\boldsymbol{\beta} = (\beta_i)_{i=0}^k$ are learned using maximum likelihood estimation (MLE). Taking the log-likelihood and assuming that the observations are independent yields:

$$\ln L(\mathbf{y},\boldsymbol{\beta}) = \sum_i \ln \pi_i^{y_i}\left(1-\pi_i\right)^{1-y_i} = \sum_i y_i \ln\frac{\pi_i}{1-\pi_i} + \sum_i \ln\left(1-\pi_i\right) \tag{2.3}$$

The MLE can then be obtained numerically via an optimization algorithm. In this project, the Python package *statsmodels* is used for this purpose. By default, the MLE in *statsmodels* is obtained through the iteratively reweighted least squares (IRLS) algorithm. This optimization algorithm can handle non-linearities in the model by iteratively reweighting a least squares problem based on the current estimates of the model coefficients [5].

One can show that the estimation of the model coefficients in logistic regression $\hat{\boldsymbol{\beta}}$ is an unbiased estimator, i.e. that it satisfies $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. In addition, it can be shown that variance of $\hat{\boldsymbol{\beta}}$ is given by

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \left(\boldsymbol{X^T V X}\right)^{-1} \tag{2.4}$$

where $\boldsymbol{X}$ contains the unique samples of the predictor variables and $\boldsymbol{V}$ is a diagonal matrix which contains the estimated variance of each observation, i.e. the i-th diagonal element of $\boldsymbol{V}$ is given by $V_{ii} = n_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)$ [13] where n_i is the total number of samples of the i-th observation and $\hat{\pi}_i$ is the estimated probability that the i-th observation

belongs to class 1.

2.1.2 Assumptions and how to verify

The inferences drawn from binary logistic regression rely on three main underlying assumptions, including $i)$ linearity in the logit for continuous variables, $ii)$ absence of perfect multicollinearity, and $iii)$ independence of observations [8]. If at least one of these assumptions is not met, the logistic regression model may produce misleading results. Apart from the three main assumptions, some literature also specifies that there should be no outliers, high leverage values or highly influential points [18]. This requirement will be checked in this project by investigating if there exist any extreme points in the data sets, which will be more discussed in the Method section. The paragraphs below will describe the three main assumptions in more detail and discuss tests that can be conducted to ensure that the assumptions are met.

The linearity assumption requires that every continuous regression variable is linear with respect to the log odds of the predicted probabilities of the model. This assumption is critical because the logistic regression model is based on a linear relationship between the log-odds and the regressor variables, as shown in Equation 2.1. To test this assumption, an investigation of the relationship between the target and every regressor has to be done before fitting a model, for example by plotting the log odds for every regressor and then visually inspecting the relationship. This procedure will be described in the Method section.

The absence of perfect multicollinearity implies that regressor variables are not perfectly correlated to one another. To be more precise, it means that there does not exist any linear relationship between the regressor variables. Including regressor variables with strong linear dependencies, or correlation, could lead to unstable results, meaning that the coefficients of the regression model could change substantially if small changes in the model or data are made [12]. This can be motivated by considering Equation 2.4 where the matrix $\boldsymbol{X^T V X}$ could become close to non-invertible if there is strong multicollinearity among the regressors, resulting in high variance for the estimated model coefficients $\hat{\boldsymbol{\beta}}$. Nevertheless, it is worth mentioning that presence of multicollinearity does not affect the overall predictive power of the model, but only the inferences made regarding individual regressor variables, such as the variable importance [12].

To test for multicollinearity a first approach can be to calculate the correlation between the features. However, this does not consider the correlation between a feature and a set of other features. In order to handle this, it is common to use variance inflation factor (VIF) analysis. The VIF is a measure that quantifies the increase in the variance of a coefficient estimate due to multicollinearity among the regressor variables in a

regression model. Typically, a VIF value greater than 10 is considered to indicate the presence of strong multicollinearity [12]. If the VIF of the i-th regression coefficient β_i is 10, it implies that the variance of β_i is 10 times higher than it would have been if the i-th regressor variable had been linearly independent of the other regressors in the model. However, as suggested in [14], caution should be made regarding defining such thresholds as it could vary in different contexts. For this reason, apart from considering the VIF values, the confidence intervals of the model coefficients will be examined to determine to what extent multicollinearity is present in a model.

Lastly, the assumption regarding independent observations is related to the maximum likelihood estimation of the coefficients, as could be seen from Equation 2.3 and in this project it will be assumed that the observations are independent. This is considered as reasonable as throw-ins from all teams and over the entire season are analysed.

2.1.3 Interpreting model coefficients

Interpreting the model coefficients of logistic regression is essential for using the model for inference. The interpretation of the coefficients can be motivated by considering the case when the model has one regressor. If one denotes the log odds as $\eta(x_i) = \ln \frac{\pi(x_i)}{1-\pi(x_i)}$ where $\pi(x_i)$ is the probability that the regressor value x_i belongs to class 1, the log odds after coefficient estimation can be expressed as $\hat{\eta}\left(x_i\right) = \hat{\beta}_0 + \hat{\beta}_1 x_i$ based on the linearity assumption according to Equation 2.1. Now, if the regressor variable value is increased with one unit to $x_i + 1$, the difference $\hat{\eta}\left(x_i + 1\right) - \hat{\eta}\left(x_i\right)$ can be expressed as :

$$\hat{\eta}\left(x_i + 1\right) - \hat{\eta}\left(x_i\right) = \hat{\beta}_1$$

Rewriting the above in terms of log odds gives

$$\hat{\eta}\left(x_i + 1\right) - \hat{\eta}\left(x_i\right) = \ln \frac{\pi(x_i + 1)}{1 - \pi(x_i + 1)} - \ln \frac{\pi(x_i)}{1 - \pi(x_i)} = \ln \left(\frac{\pi(x_i + 1)}{1 - \pi(x_i + 1)} \Big/ \frac{\pi(x_i)}{1 - \pi(x_i)} \right) = \hat{\beta}_1$$

which provides an expression for the *odds ratio*

$$\frac{\pi(x_i + 1)}{1 - \pi(x_i + 1)} \Big/ \frac{\pi(x_i)}{1 - \pi(x_i)} = e^{\hat{\beta}_1} \tag{2.5}$$

In this way, $e^{\hat{\beta}_1}$ represents the multiplicative change in odds of the regressor value belonging to class 1 for a one unit increase in the corresponding predictor variable. For example, if class 1 corresponds to a successful throw-in and if $\hat{\beta}_1$ is the estimated coefficient for the predictor variable representing throwing length, then $e^{\hat{\beta}_1}$ represents

the increase in odds of a successful throw-in for a one-meter increase in throwing length, assuming that the length is the only regressor variable in the model.

For the case when having multiple regressors, i.e. for multiple logistic regression, the interpretation of each coefficients is the same as for the case of one regressor, assuming that the rest of the regressors are held constant [13].

2.1.4 Evaluating logistic regression

Evaluation Metrics in Binary Classification

Given that logistic regression is used for binary classification, there are four possible outcomes of the model given a classification threshold: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). For example, if predicting whether a throw-in is successful, a positive outcome indicates that the throw-in is indeed successful given a classification threshold, while a negative outcome corresponds to an unsuccessful throw-in. A true outcome means that the model has correctly classified the outcome of a throw-in, while a false outcome suggests that the classification is incorrect.

With these four outcomes in mind, the following terms can be defined. The true positive rate (TPR), or sensitivity, is the proportion of true positives out of all positives. The true negative rate (TNR), or specificity, is the proportion of true negatives out of all negatives. One can also define the false positive rate (FPR) which is the proportion of false positives out of all negatives and the false negative rate (FNR), i.e. the proportion of false negatives out of all positives. Together, TPR, TNR, FPR, and FNR can provide a comprehensive evaluation of the performance of a classification model. Note that these metrics can be of different importance in different situations, as for example minimizing the FNR could be priority in medical applications. The metrics are summarized with the following equations:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad \text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \qquad \text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}}$$

ROC curve

One way to evaluate the accuracy of a logistic regression model is by using a receiver operating characteristic (ROC) curve. A ROC curve is created by plotting the TPR on the y axis against the FPR on the x axis for a range of different classification thresholds after which the points are connected to form a curve. By examining the ROC curve, it is possible to obtain a measure of how well the model performs. A perfect model with a perfectly defined classification threshold, i.e. resulting in a TPR and TNR of 100%,

would generate a point in the upper left corner $(x = 0, y = 1)$. Based on this, ROC curves can be used to find the optimal classification threshold in terms of maximizing the TPR and TNR, as a threshold of 0.5 is not always the most suitable choice in a model. The optimal threshold corresponds to the point on the ROC curve that is the closest to the upper-left corner.

ROC curves also include a diagonal line, ranging from $(x = 0, y = 0)$ to $(x = 1, y = 1)$ which is called the line of equality or the random chance line. This line represents a model where the classifications are made randomly and thus the closer a ROC curve is to this line, the less accurate is the model. To quantify the goodness of a ROC curve, one can use the area under curve (AUC), also known as the c-statistic. A perfect classification would result in an AUC of 1 while a value 0.5 would suggest that the model is no better than a random prediction model [3].

Akaike information criterion (AIC)

The AIC is a statistical metric that allows for comparison between models and serves as a suitable criterion for model selection. It takes into account both how well the model fits to the data, as well as the model complexity measured in terms of the number of used parameters and by doing so AIC punishes overly complex models. The objective is to choose a model with the lowest AIC value, and in that way balancing the model's complexity and accuracy. Note that AIC is primarily used for models which have a defined likelihood function and thus it may not be appropriate for many machine learning models. The AIC is defined in the following way:

$$\text{AIC} = 2k - 2\ln(L)$$

where k is the number of parameters in the model and L is the maximum likelihood of the model [4].

Log loss

Next, the idea behind the log loss is presented. Log loss, or cross entropy loss, is a loss function which measures the difference between the predicted probabilities and the true outcomes in a binary classification problem. A lower log loss score indicates better model performance, with a value of 0 indicating perfect prediction accuracy, while increasing values indicate increasingly worse performance. The log loss function is defined as follows:

$$\text{log loss} = -\left[y \log(p) + (1 - y) \log(1 - p)\right]$$

where y is the true binary outcome, 1 or 0, while p is the predicted probability of class 1. With this, the average log loss among all data samples can be expressed as:

$$\text{average log loss} = \frac{1}{n}\sum_{i=1}^{n} -\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]$$

The log loss can be sensitive to outliers and class imbalance. However, unlike AIC, log loss is not restricted to a specific class of models and can be used to compare the performance of various families of models.

2.2 Gradient boosting

This section of the Background gives a brief explanation of the gradient boosting algorithm, as this model will later be used for comparison purposes. Gradient boosting is a powerful machine learning algorithm that can be used for both regression and classification problems. The basic idea behind gradient boosting is to iteratively combine many weak models, typically smaller decision trees, into a single strong model, by learning from the incorrect predictions of the previous weak learner.

More specifically, the gradient boosting algorithm takes as inputs a training set $\{(x_i, y_i)\}_{i=1}^{n}$ where n is the total number of data samples and a differentiable loss function $L(y, F(x))$ where $F(x)$ is the prediction of a weak learner and $x = (x_i)_{i=0}^{n}$. Additionally, the number M of weak learners used is also specified. The first step of gradient boosting is to initialize the predictions $F_0(x)$ with a constant value. This is done by choosing the constant γ which minimizes the sum of the loss functions across all target values y_i.

$$F_0(x) = \underset{\gamma}{\text{argmin}} \sum_{i=1}^{n} L\left(y_i, \gamma\right)$$

Having obtained an initial prediction $F_0(x)$, the second step of the algorithm is to calculate the so-called *pseudo residuals* r_{im}, where m is the index of the current weak learner, in the following way.

$$r_{im} = -\left[\frac{\partial L\left(y_i, F\left(x_i\right)\right)}{\partial F\left(x_i\right)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n$$

After calculating the pseudo residuals, which correspond to the negative gradient of the loss function, a weak learner is trained on the data set $\{(x_i, r_{im})\}_{i=1}^{n}$, producing predictions denoted as $h_m(x)$. A natural step would be to now add these predictions of the residuals to the previous prediction, which in the first step of the algorithm ($m = 1$)

is $F_0(x)$. However, the predictions $h_m(x)$ are first multiplied by a multiplier γ_m which is found by solving the following optimization problem, constituting step 3:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}\left(x_i\right) + \gamma h_m\left(x_i\right)\right)$$

Finding γ_m now makes it possible to carry out step 4, which is to update the predictions of the gradient boosting model, based on the previous prediction in the following manner:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Having obtained the new predictions $F_m(x)$, steps 2–4 of the algorithm are repeated until the total number M of the pre-specified weak learners have been trained, which yields the final prediction $F_M(x)$.

One potential issue with the gradient boosting algorithm is overfitting, which can be handled through several measures. A common approach is to use shrinkage, which involves reducing the step size in the updating step by introducing a *learning rate* ν in the following way:

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x)$$

where ν takes a value between 0 and 1. The downside of introducing the learning rate is that decreasing the learning rate requires more iterations to reach the optimum solution, which increases the computational time of the gradient boosting algorithm. Another way to reduce overfitting is to decrease the complexity of the weak learners, which in the case of decision trees could be to reduce the maximum depth of each tree or increasing the minimum number of data points required to split a node [1].

2.3 Football terminology

The following section provides a brief overview of the key football-related terms used throughout the report.

Possession chain

Throughout this report, the term *possession chain* will be frequently used. In football, a possession chain refers to a sequence of events where a team has control of the ball, starting from the moment the team gains control until the ball is lost. If the team

loses the ball, but recovers it immediately, the possession chain is not considered to be broken [23].

Expected goals

In football, the expected goal metric (xG) is a commonly used measure to assess the quality of a goal-scoring opportunity. This metric represents the probability that a shot will result in a goal and it is calculated based on machine learning models using historical data. For instance, if a shot has an xG value of 0.2, it means that a similar shot historically resulted in a goal 20 % of the times.

The xG for all shots is provided in the Wyscout event data. Although the specific machine learning model and the full description of the used features are not publicly available, some of the parameters considered in the calculations of xG by Wyscout are the shot location, shot type (foot or head), from where the ball was passed before the shot was taken and whether the shot came from a set-piece or not [25]. To obtain a better understanding of how xG varies across the pitch, an example of a heat map representing the xG is visualized in Figure 2.3.1. The heat map is based on a basic xG model developed by Sumpter et al. in the course Mathematical Modelling of Football at Uppsala University [21]. It shows how the xG varies for different shot locations on the pitch.

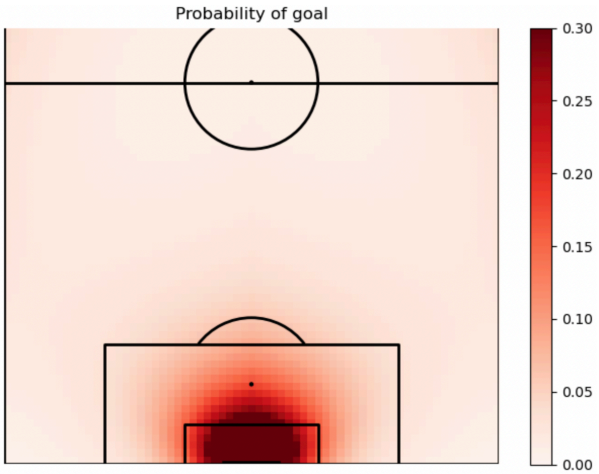


Figure 2.3.1: Heat map representing the xG for different shot locations on the pitch.

Pitch coordinates

The dimensions of a football pitch vary across different pitches, however the shape has to be rectangular. The longer sides are named *touchlines*, while the shorter sides are called goal lines and the recommended pitch dimensions by FIFA are 105 metres x 68 metres [6]. These pitch dimensions will be referred to as the *standard pitch dimensions* in this report. However, football data providers present pitch coordinates

based on other systems. For example, Wyscout uses a coordinate system which depends on the team currently having the possession of the ball. In this system, both the x and y axis range from 0 to 100 and each coordinate is expressed in % of 100. This way of defining the coordinates makes it easier to interpret a coordinate (x, y) relative to the team having ball possession, as the goal of the team with ball possession is always located at $x = 0\%$ while the opposition's goal is always at $x = 100\%$ [24]. In order to perform interpretable calculations based on Wyscout pitch data, it is however necessary to transform the coordinates. For example, to calculate the length of a pass based on the start and end coordinates, the Wyscout pitch coordinates have to be transformed to coordinates based on the standard pitch dimensions in order to obtain an approximation of the pass length.

Chapter 3

Method

This chapter presents an overview of the method used in this project. First, the targets and features are defined and explained. After this, a number of aspects of the data pre-processing is presented. In the third section, the procedure of investigating the linearity assumption is described. Lastly, the implementation of the gradient boosting model is briefly explained.

3.1 Defining features and targets

As mentioned in the Introduction, the success of a throw-in depends on how *success* is defined and in this project, two different definitions are considered. First, a throw-in is considered as successful if the throwing team manages to maintain control, i.e. possession, of the ball after the throw-in. To determine if the throwing team has maintained possession, it is necessary to define for how long time the team must have had control of the ball. This time is set to 7 seconds in this project, in the same way as done by Stone et al. in [20]. Thus, if the throwing team retains possession for at least 7 seconds, the throw-in is considered as successful according to this first definition.

The second definition of a successful throw-in considers if the throw-in leads to a goal chance opportunity. To determine this, the possession chain that follows the throw-in is considered. If the throwing team creates a chance with an expected goal (xG) of at least 3%, then the throw-in is considered as successful. This relatively low threshold is chosen in order to capture more instances of successful throw-ins, as this definition of success is practically relatively difficult to achieve.

With these two definitions of a successful throw-in, two corresponding target variables are defined. The target *retained* is a binary variable that indicates whether possession of the ball is retained 7 seconds after the throw-in. A value of 0 indicates that the possession is lost, while a value of 1 indicates that the possession is kept. The second

target *chance_created* is also a binary variable which indicates whether a goal-scoring opportunity is created during the possession chain following a throw-in. A value of 0 indicates that no goal-scoring opportunity is created, while a value of 1 indicates that a goal-scoring opportunity is created. The two targets are summarized in Table 3.1.1.

Table 3.1.1: Description of targets.

Target	**Description**
retained	Binary variable that indicates whether possession of the ball was retained after the throw-in.
chance_created	Binary variable that indicates whether a goal-scoring opportunity was created after the throw-in.

In order to obtain a better understanding of how common it is to carry out a successful throw-in according to the two definitions, the frequency of both success types is presented in Table 3.1.2. From here, it is seen that it is relatively rare to create a goal-scoring opportunity from a throw-in. Due to the lack of data of the positive instances (1) of this target, it could be more challenging for the logistic regression model to accurately predict the success of a throw-in based on this definition.

Table 3.1.2: The frequency of each definition of a successful throw-in out of 9861 throw-ins. Note that these numbers are obtained after conducting data pre-processing, which is described in Section 3.2.

Success definition	**Number of samples**	**Share of total**
Possession retained after 7 seconds	6502	65.9%
xG > 0.03 created in same possession chain	462	4.7%

Having provided the two definitions of a successful throw-in and the targets used in this project, the focus next turns to defining the features used in the models of this project. Four of these features can be seen as fundamental and these are named *start_x_adj*, *angle*, *length* and *time_since_last*. The feature *start_x_adj* is the start x position of the throw-in along the touchline, based on the standard pitch dimensions, as displayed in Figure 3.1.1. A value of 0 indicates the start of the pitch relative to the attacking team, while a value of 105 refers to the end of the pitch relative to the attacking team. The ending "adj " refers to the fact that the pitch length has been transformed from the original length of 100 in the Wyscout coordinate system to 105 meters.

Next, the feature *angle* is defined as the angle in radians between the throwing direction and the attacking direction and is denoted as α in Figure 3.1.1. An angle of 0

represent a throw straight up the pitch towards the opposition's zone, while an angle of π represent a throw straight down the pitch towards the player's own zone. Note that this angle is calculated based on the transformed start and end coordinates of the throw-in. The length of the throw-in in meters is simply named *length* and it is also calculated based on the start and end location of the throw-in, using the transformed coordinates. Finally, *time_since_last* reflects how fast the throw-in is being taken, i.e. the time in seconds since the ball went out of the pitch before a throw-in. It is calculated as the time difference between the game interruption caused by the ball going out of the touchline and the moment when the ball is thrown.

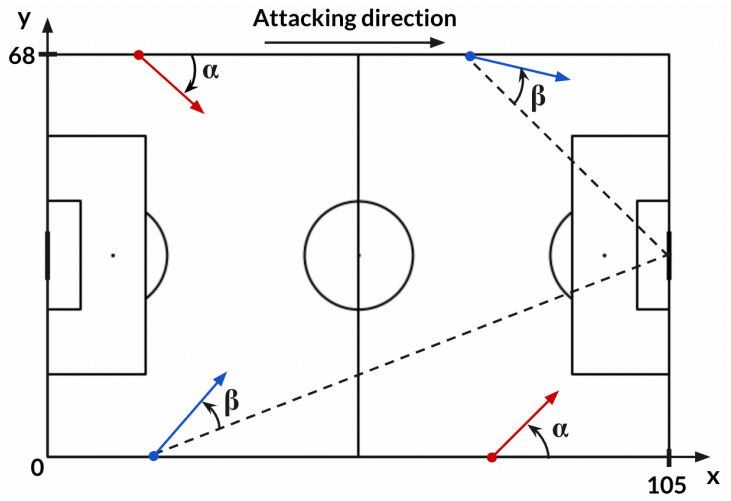


Figure 3.1.1: Scheme showing coordinate system of pitch with standard dimensions and definition of two different throwing angles.

In order to obtain a better understanding of these four fundamental features, the frequencies are plotted in histograms as shown in Figure 3.1.2.

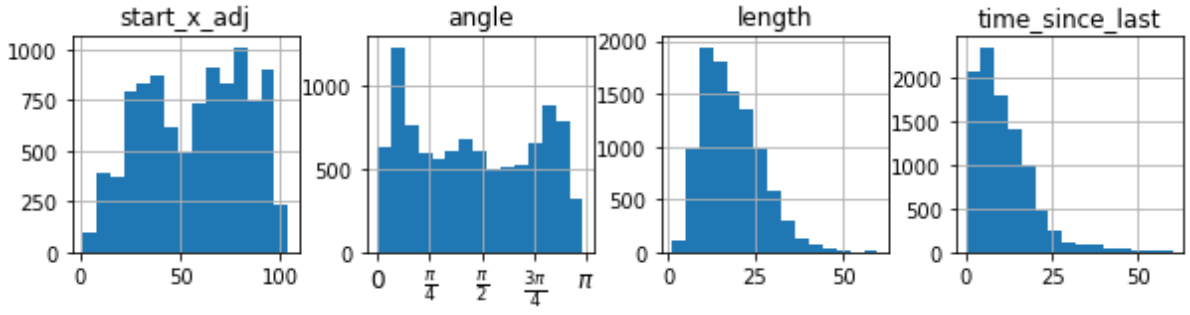


Figure 3.1.2: Histogram showing distribution of four event data features. The units of the four subplots are: meters, radians, meters and seconds respectively. Note that these histograms are obtained after conducting data pre-processing, which is described in Section 3.2.

Based on these fundamental features, together with the start y coordinate and end x

and y coordinates of the throw-in, most of the other features are defined. For example, *distance_to_goal* measures the distance from the end location of the throw-in to the center of the opposition's goal. Apart from this distance, it can also be relevant to measure how much closer to the opposition's goal the ball is after the throw-in. This difference in distance is represented by *distance_to_goal_diff* and it is calculated by taking the difference between the distance to the center of the opposition's goal at the end and start location of the throw-in. A positive value of this variable indicates that the ball has been thrown towards the opposition's goal, while a negative value indicates that the ball has been thrown away from the opposition's goal.

A similar feature is *x_diff* which measures the relative change in x coordinate between the end and start position of the throw-in. A positive value means that the ball has been thrown in the attacking direction, while a negative value indicates that the ball has been thrown in the direction of the throwing team's own goal. Another feature related to distance is *distance_to_middle* which measures the distance from the start x position of the throw-in to the middle x coordinate of the pitch.

Furthermore, another feature that is investigated is named *angle_throw_goal* and it is the angle in radians between the throwing direction and the line from the start location of the throw-in towards the center of the opposition's goal. An angle of 0 represents a throw straight towards the center of the opposition's goal. This angle is referred as β in order to differentiate from *angle* which is denoted as α and is also illustrated in Figure 3.1.1.

All of the above described features are summarized in Table 3.1.3. Using these features, it is then possible to create feature interactions and feature transformations. Feature interactions are created by for example multiplying two feature with each other, while feature transformations involve applying a function on a feature, such as taking a feature to the power of two. Relevant feature interactions and feature transformations will be presented in the Results section of this report.

3.2 Target exceptions and removal of data points

The data from Wyscout includes the duration of each possession chain, which is highly relevant when defining one of the target variables, namely the one which indicates if the possession was kept after a throw-in. As previously mentioned, the possession will be considered as kept if the throwing team still has possession of the ball 7 seconds after the throw-in. However, there could be situations when it is natural to classify the possession as being kept, even though the length of the possession chain is less than 7 seconds. Thus, a couple of exceptions related to the definition of possession

Table 3.1.3: Explanation of features described in the Method section.

Feature	**Description**
angle (α)	angle in radians between the throwing direction and the attacking direction.
angle_throw_goal (β)	angle in radians between the throwing direction and the line towards the center of the opposition's goal.
distance_to_goal	distance in meters from the end location of the throw to the center of the opposition's goal.
distance_to_goal_diff	indicates how much closer or further the ball is from the center of the opposition's goal after a throw-in.
distance_to_middle	the distance between the start x position of the throw-in and the middle x coordinate of the pitch.
length	length of the throw-in in meters.
start_x_adj	the start x position of the throw-in along the length of the pitch.
time_since_last	time in seconds since the ball went out of game before a throw-in.
x_diff	difference in x coordinates of the end location of the throw-in and the start position of the throw-in.

retention are specified. If a possession chain is shorter than 7 seconds, but ends by an infraction, i.e. a violation of the game rules, by the opposition team, the ball possession after the throw-in is considered as kept by the team taking the throw-in. Another introduced exception is if a possession chain is shorter than 7 seconds and ends with an event tagged as *interruption* but the next possession chain starts with the throwing team having possession, the throw-in is regarded as successful in regards to keeping possession. Also, if the team taking the throw-in scores within 7 seconds after the throw-in, the possession is considered as retained as well.

Continuing with the outlier removal, the following actions are taken. For the feature *length*, all samples with a length outside of the interval 0.1–60 m are removed. This is motivated by throw-ins being longer than 60 m are unrealistic since this is the world record for the longest throw-in in football [7]. Throw-ins shorter than 0.1 m are also considered as unrealistic and by removing these samples singularity issues are avoided.

For the variable *time_since_last*, samples having a value larger than 60 seconds are removed. This is since this time would most probably be due to an interruption of the game such as an injury, which in turn could then affect the execution of the succeeding throw-in. The starting x positions and the angles are all between 0–105 and 0–π respectively and thus no outliers are identified within these features. Also, throw-ins which have the event *fair_play* in the same possession chain are removed. For these throw-ins, it is probable that the throwing team hands the ball over to the opposition

after the throw-in as a gesture of fair-play and these throw-ins are not considered in the analysis.

3.3 Investigating feature target relationship

In order to check whether the linearity assumption of logistic regression is fullfilled, it is highly helpful to investigate the relationship between every feature and the target before fitting a model. If the relationship in the logit scale turns out to be non-linear, relevant feature transformations can be explored. The analysis of the regressor feature dependence is conducted using a procedure presented by Marin [10]. This procedure is visualized in Figure 3.3.1 and is described below.

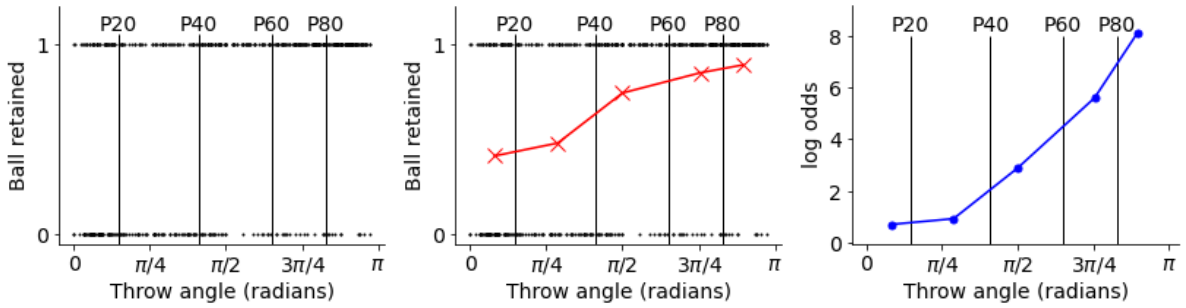


Figure 3.3.1: Scheme of procedure for investigating feature and target relationship. In (a) the binary target values are plotted for all regressor values and the regressor values are divided into five bins based on percentiles. In (b) the share of successful throw-ins (class 1) is plotted for every bin. Finally, in (c) the log odds are plotted and the obtained relationship approximates to what extent the linearity assumption is fulfilled.

In this example, which has the purpose of explaining the procedure of this method, the effect of the throwing angle α has been analysed in relation to the ball retention based on 500 throw-ins. When choosing a feature for a binary class problem, one wants the two classes (0 and 1) to be clearly distinguishable by the feature. Looking at Figure 3.3.1a, it is seen that high angle values imply a higher rate of successful throw-ins (having class 1), while smaller throwing angles tend to lead to more unsuccessful throw-ins (having class 0). This indicates that the angle could be a useful feature in a logistic regression model, however it is at this point hard to say if the relationship is linear in the logit scale. The first step of the feature investigation is thus to divide the feature values into five bins, determined by the 20th, 40th, 60th and 80th percentile, as seen in subfigure (a). Then, for the data points in every bin, the fraction, or probability p, of points classified with target 1 is calculated. In subfigure (b), these probabilities have been plotted against the median feature value of every bin. The red curve in (b) intends to illustrate that the median values and the retention probabilities could have a

relationship that reminds of the shape of a sigmoid. Now, if the data is in fact linear in the logit scale, applying the logit function to the retention probabilities, i.e. $\log\frac{p}{1-p}$, should result in a linear relationship. This is investigated by plotting the log odds against the median values of every bin, as shown in subfigure (c). From this figure, it is seen that the relationship is not completely linear. Thus this suggests that it could be suitable to apply a transformation on the angle, in order to better fulfill the linearity assumption.

Thus this procedure serves as an approximate method to investigate if the features are linear to the target in the logit scale, which is one of the assumptions of logistic regression as previously stated. The plots corresponding to subfigure (c) will hereafter be referred to as *logit plots* and such plots will be presented in the Results section. If the logit plots show a non-linear relationship, feature transformations can be investigated in order to create a linear relationship. Using this method, relevant features and feature transformations are found.

3.4 Training a gradient boosting model

In order to compare the predictive power of the logistic regression model to a more advanced method, a gradient boosting classifier is trained. Separate hyperparameter optimization is conducted for each of the targets, *retained* and *chance_created*. The hyperparameter optimization is based on the four fundamental features shown in Figure 3.1.2. The log loss is chosen as the loss function in the gradient boosting algorithm and different parameter options are compared in regards to AUC using grid search [17]. The considered parameters are the number of learners (*n_estimators*), the learning rate (*learning_rate*), the minimum number of data samples in the leaf nodes (*min_samples_leaf*) and the maximum depth of every tree (*max_depth*).

The optimization is started by finding a suitable combination of *n_estimators* and *learning_rate* while keeping the parameters *min_sample_leaf* and *max_depth* constant at 10 and 3 respectively. The learning rate is set to 0.05 after which *n_estimators* is varied, with values ranging from 20 to 100 with increments of 10. Having obtained a first rough optimum of *n_estimators*, the grid is refined around the optimum with increments of 5 in order to further improve the optimization.

After finding an optimized combination of *n_estimators* and *learning_rate*, the tree-specific hyperparameters are optimized by considering *min_samples_leaf* in the range 10 to 60, with increments of 10 and *max_depth* in the range 2 to 6 with increments of 1. All combinations of these two parameters are evaluated using grid search. After finding a rough optimum, the grid for *min_samples_leaf* is further refined with increments of 5.

Using this procedure, the hyperparameters for the possession retention model are set to: *learning_rate* $= 0.05$, *n_estimators* $= 70$, *min_samples_leaf* $= 55$ and *max_depth* $= 4$. For the chance creation model, the following choices are made: *learning_rate* $= 0.05$, *n_estimators* $= 40$, *min_samples_leaf* $= 40$ and *max_depth* $= 2$. The rest of the hyperparameters are kept to the default values, according *scikit learn's* implementation (version 1.2.0) of the gradient boosting classifier [16].

Chapter 4

Results

This section presents the results of the project in five parts. Firstly, relevant features are investigated in terms of their variability, linearity and different feature transformation are explored. Secondly, the results of the models used to predict possession retention are presented, which includes a comparison of different feature collections, visualization of the results, feature importance analysis and an investigation of the multicollinearity assumption. Thirdly, the corresponding results are presented for the model used to predict goal chance creation. In the fourth part, the results of the gradient boosting method are presented and lastly, an investigation regarding the throw-in strategies of the Allsvenskan teams is presented.

4.1 Feature and feature transformation analysis

In this section, an investigation of how the target variables depend on the features is presented and relevant transformations are outlined. The investigation is first based on the target *retained*, and then on *chance_created*. Note that only a selection of the regressors and transformations is presented. A supplementary presentation of the feature analysis is given by Figures B.0.1 and B.0.2 in the Appendix.

Before presenting the logit plots of the different features, the considered feature transformations and interactions are presented and the naming convention is explained. Starting with the feature transformations, these mostly include squaring, cubing, taking exponential with various bases and the natural logarithm of a feature. The naming convention is made in the following way: if a feature is squared or cubed, the endings *_squared* and *_cubed* respectively are added to the feature name. When applying an exponential function on a feature, the ending *_exp* is added followed by the value of the base of the exponential unless the base is e in which case only *_exp* is

added. For example, if applying an exponential transformation with base 0.2 on the angle α, i.e. 0.2^{α}, the transformed feature is named *angle_exp_0_2*. Moreover, when applying the natural logarithm on a feature, the ending *_log* is added.

The feature interactions are created by multiplying two different features. In this case, both names are added to the new feature name, separated by an underscore. For example, the feature which is obtained by multiplying the feature *angle* with *length* is given the name *angle_length*.

With these naming conversions in mind, it should be possible for the reader to understand the names of the features. However, a complete description of the feature transformations and interactions is given by Table A.0.1 in the Appendix.

4.1.1 Possession retention model

Firstly, an investigation of the dependence between the possession retention and the angle α of the throw-in, as well as various transformations, is presented as shown in Figure 4.1.1.

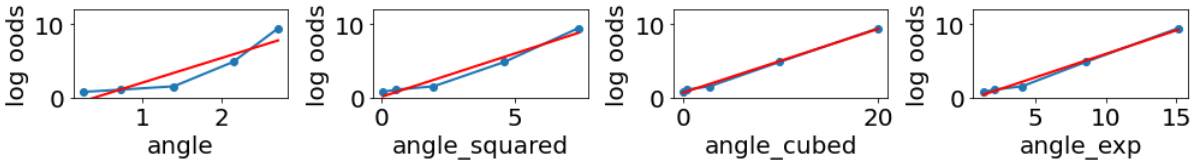


Figure 4.1.1: Dependence between the log odds of the retention rate and the angle α, together with relevant transformations. The red line is a linear fit of the five log odds values.

The leftmost subfigure of Figure 4.1.1 shows that the log odds of retaining the ball after a throw-in increase noticeably when the throwing angle α is increased, indicating that throwing the ball more backwards leads to higher retention rates. The increase appears to be gradual for smaller angles, but becomes more prominent as the angle increases. This suggests that the relationship between the throwing angle and the retention rate is not linear, as one can see from the difference between the blue and the red line in the leftmost subfigure, where the red line is the linear fit of the points. To explore this further, three additional transformations, namely the angle raised to the power of two and three as well as the the exponential of the angle are presented in Figure 4.1.1 as well. These transformations result in a more linear relationship indicating that they could be useful options when selecting features for the model. In particular, the features *angle_cubed* and *angle_exp* result in the most linear dependencies, which make these transformations particularly interesting.

The feature analysis is continued by investigating how the start x position of the throw-in affects the retention rate, as presented in Figure 4.1.2.

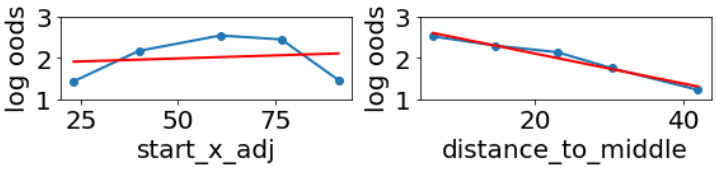


Figure 4.1.2: Dependence between the log odds of the retention rate and *start_x_adj* as well as *distance_to_middle*.

The left subfigure of Figure 4.1.2 shows that the start x position of the throw-in seems to have an impact on the retention rate, even though this impact is smaller compared to the impact of the angle, as shown in Figure 4.1.1. This is seen as the variation in the log odds is smaller for *start_x_adj* compared to *angle*. Furthermore, the retention rates are the highest when the ball is thrown from the middle region of the pitch, and decreases as the throw-in is taken closer to one of the goals. This suggests that the relationship between the start x position of the throw-in and the retention rate is not linear. Based on this, *start_x_adj* is transformed such that the distance from the middle of the pitch is measured instead, represented by the feature *distance_middle*, resulting in a more linear relationship, as seen in the right subfigure of Figure 4.1.2.

The feature analysis is continued by examining the effect of the feature *distance_ to_ goal_diff*, as shown in Figure 4.1.3.

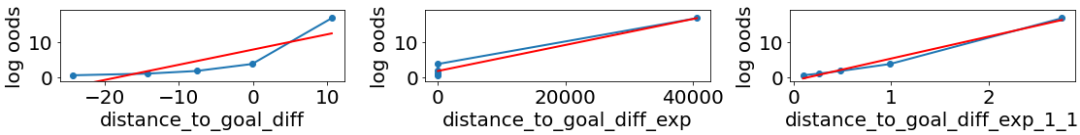


Figure 4.1.3: Dependence between the log odds of the retention rate and *distance_to_goal_diff* and relevant transformations.

Figure 4.1.3 shows that that *distance_to_goal_diff* seems to be a feature with a high impact on the retention rate, as the variation in log odds is relatively large between low and high values of the regressor. One can see that the variation in log odds is higher for *distance_to_goal_diff* than for *angle*, shown in Figure 4.1.1. Taking a closer look at the dependence it can be noticed that the retention rate is relatively low for small values of *distance_to_goal_diff*, and then increases substantially for higher values. This indicates that it is easier to retain the ball when throwing away from the opposition's goal. The relationship between the log odds of retaining the ball and *distance_to_goal_diff* appears to be exponential and thus exponential transformations are considered. First, an exponential transformation with base e is considered, denoted by the feature *distance_to_goal_diff_exp* which results in the logit plot shown in the middle subfigure of Figure 4.1.3. With this transformation

a higher degree of linearity seems to be achieved. However, it is noticed that this transformation could be too aggressive, as the first four points appear to end up in the same x value. Thus, an exponential with a slower growth rate is considered as well, by decreasing the value of the base of the exponential from e to 1.1. Applying the transform $1.1^{distance\_to\_goal\_diff}$ results in the rightmost logit plot. This transformation seem to result in a linear dependence without squeezing points too much.

As mentioned above, a supplementary overview of the dependence between the retention rate and the considered features and feature transformations is given in the Appendix.

4.1.2 Goal chance creation model

Next, the relationship between a number of features and the log odds the goal opportunity creation rate is presented. Note that since a goal-scoring opportunity is a rarely occurring event, the log odds values are lower compared to the log odds of the possession retention.

To start with, the effect of the throwing angle α is considered.

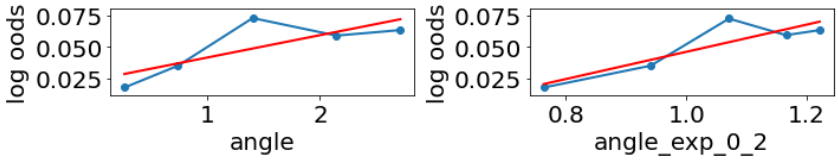


Figure 4.1.4: Dependence between the log odds of the goal chance creation rate and the angle α as well as an exponential transform.

Figure 4.1.4 shows that the goal-scoring opportunity rate seems to be highly dependent on the throwing angle α, relative to other features, presented in Appendix B. The increase is significant for smaller values of the angle and then attenuates for larger values. For this reason, a function which also attenuates for increased values of the independent variable could be relevant to use for the feature transformation in this situation. A reasonable transform to consider would be taking the logarithm of the angle since the log function satisfies the desired behaviour, however since a number of angles take a value of 0, this transformation is not feasible. Another class of functions with the desired property are exponential functions with a base smaller than 1. After having investigated different bases, a base value of 0.2 is considered as the exponential 0.2^{α} has a suitable decline rate. More specifically, since the values of the angle span from 0 to 3.14 (π), one wish to choose a base of the exponential for which the attenuation is reached for these values and this is rather satisfied by the base 0.2. The result of the transform is presented in the right subfigure of Figure 4.1.4 and shows that the relationship is more linear.

Next, the effect of the start x position of the throw-in in relation to goal chance creation is investigated, as shown in Figure 4.1.5.

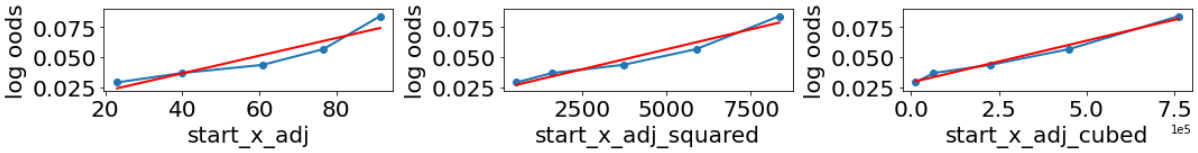


Figure 4.1.5: Dependence between the log odds of the goal chance creation rate and the start x position of the throw-in and a quadratic and cubic transformation.

Figure 4.1.5 illustrates that the log odds of creating a goal chance increase when the start x position is increased. One can note that the increase is more prominent for larger values of the regressor. Thus a transformation, taking *start_x_adj* to the power of 2 and 3 is considered. After the transformation the dependence appears to be more more linear, especially for *start_x_adj_cubed* as seen in the right subfigure of Figure 4.1.5.

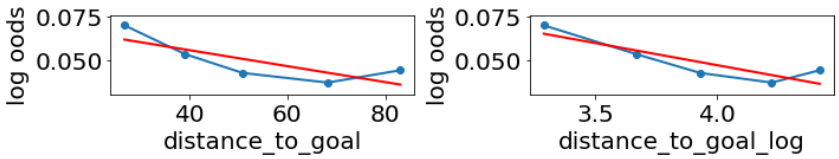


Figure 4.1.6: Dependence between the log odds of the goal chance creation rate and *distance_to_goal* and a logarithmic transformation.

Next, the effect of the distance between the end location of the throw-in and the center of the opposition's goal is analysed, as shown in Figure 4.1.6. The dependence appears to be decaying, meaning that increasing the *distance_to_goal* from an already high value has lower effect on the goal chance creation than increasing *distance_to_goal* from low values. Thus a transformation by taking the natural logarithm of *distance_to_goal* is considered. The dependence after the transformation, presented in the right subfigure, results in a more linear relationship even though there is an increase in log odds for the highest values.

More features and transformations related to goal chance creation were investigated in a similar way as described above, and the rest of the results can be found in Appendix B.

4.2 Model analysis

Based on the feature and transformation analysis, presented in the section above, a number of models with different properties were created. The results of these

models are first shown when predicting possession retention and then for goal chance creation.

4.2.1 Possession retention model

Model comparison

In order to predict possession retention, five models with different properties were constructed. The models and the included features are presented in Table 4.2.1 below.

Table 4.2.1: Possession retention models

Model name	**Included features**
Basic Model	start_x_adj, angle, length, time_since_last
Basic Model Transformed	distance_to_middle, angle, length, time_since_last
Full Model	start_x_adj, angle, length, time_since_last, distance_to_goal, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x
Full Model Transformed	distance_to_to_middle, angle_cubed, length, time_since_last, distance_to_goal_log, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x
Full Model Transformed & Non-Correlated	distance_to_middle, time_since_last, distance_to_goal, distance_to_goal_diff, angle_length

The *Basic Model* refers to the model that includes the fundamental features that describe the basic characteristics of a throw-in and serves as a benchmark for evaluating the performance of more complex models. The *Basic Model Transformed* is the model obtained when replacing features in the *Basic Model* with transformations, however only if the replacement results in better model fitting in terms of AIC. The *Full Model* includes those features which are considered to best describe the retention rate after a throw-in. The feature selection process for this model is conducted by starting with all relevant features and then iteratively removing the feature which results in the largest decrease in AIC, until no improvement in AIC is achieved. The *Full Model Transformed* replaces those features in *Full Model* with the transformed versions that result in higher accuracy. Finally, the *Full Model Transformed & Non-Correlated* includes a set of features from the *Full Model Transformed* that exhibit no strong correlation with each other. This is determined by considering the correlation coefficients and VIFs for the included features, as will be shown further below. All of

the above mentioned models were evaluated in regards to three metrics: the AIC, AUC and the log loss and the results are presented in Table 4.2.2.

Table 4.2.2: Accuracy of various logistic regression models predicting possession retention. Note that the AIC is based on the entire data set, while AUC and log loss were calculated on a test set comprising 25% of the data.

Model	**AIC**	**AUC**	**Log loss**
Basic Model	10875	0.737	0.554
Basic Model Transformed	10866	0.742	0.551
Full Model	10429	0.770	0.529
Full Model Transformed	10423	0.770	0.528
Full Model Transformed & Non-Correlated	10528	0.764	0.534

Starting with the comparison of *Basic Model* and *Basic Model Transformed*, one can see that the transformed model results in better accuracy based on all three metrics. Note, however, that the improvement originated only from the transformation of *start_x_adj* to *distance_to_middle*, as one can deduce from Table 4.2.1. Transforming *angle* to *angle_squared*, *angle_cubed* or *angle_exp* did not result in model improvement in this case, as one could have anticipated based on Figure 4.1.1 above. The AIC for the various transformations can be found in Table C.0.1 in the Appendix.

Including all relevant features, comprising the *Full Model*, results in a significant improvement in all three metrics. Next, the *Full Model Transformed* results in a slight decrease in AIC, which mostly originates from the replacement of *distance_to_goal* with *distance_to_goal_log*. Replacing *angle* with *angle_cubed* results in a minor AIC improvement, while introducing the transformation *distance_to_middle* instead of *start_x_adj* leads to a negligible change in AIC. Nevertheless, *distance_to_middle* is kept in the model as this is considered to be a more reasonable feature for measuring the position along the touchline based on logit plot presented in Figure 4.1.2. All other transformations resulted in an increase in AIC and are thus not included in the *Full Model Transformed*. For example, the exponential transformations *distance_to_goal_diff_exp* and *distance_to_goal_diff_exp_1_1* turned out to be particularly poor as these transformation resulted in a substantial increase in AIC. Once again, for details regarding the change in AIC after introducing the various feature transformation, the reader is referred to Table C.0.1 in the Appendix.

Finally, introducing the model without highly correlated features (*Full Model Transformed & Non-Correlated*) results in a decrease in model accuracy, even though it is still considerably higher than for the *Basic Model*. Before presenting the results of the underlying multicollinearity analysis, the ROC curve for three different

models is presented in Figure 4.2.1. From this plot it becomes clearer that the *Full Model Transformed & Non-Correlated* is only slightly shifted down compared to *Full Model Transformed*, indicating that these two models have similar predictive power. However, compared to the *Basic Model*, the ROC curve for *Full Model Transformed & Non-Correlated* is clearly higher suggesting that this model is more accurate in the predictions than the baseline model.

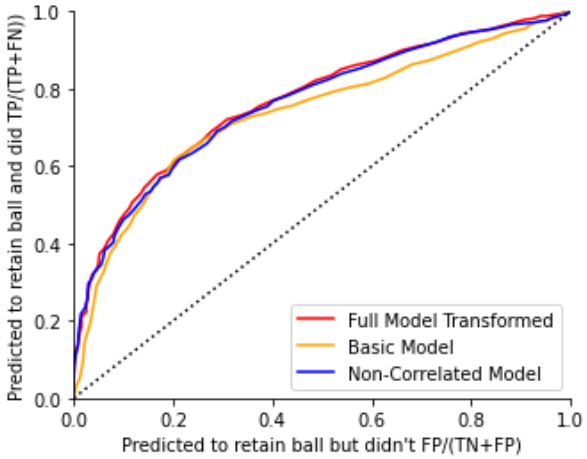


Figure 4.2.1: Comparison of ROC curves for three logistic regression models predicting possession retention. Note that the model named *Non-Correlated Model* in the legend refers to *Full Model Transformed & Non-Correlated*.

Multicollinearity

The formation of the *Full Model Transformed & Non-Correlated* is motivated hereafter by presenting the results of the multicollinearity analysis. The features of the *Full Model Transformed* were removed in an iterative manner based on the VIFs and the correlation coefficients between pairs of features. The removal of features was conducted by identifying sets of correlated features, and then removing the feature that resulted in the least increase in AIC value after its removal.

In order to understand the effect that the feature removal has on the multicollinearity, the correlation matrices and VIFs of the features in the *Full Model Transformed* and *Full Model Transformed & Non-Correlated* are compared. First, the correlation matrices for the two models are shown in Figure 4.2.2.

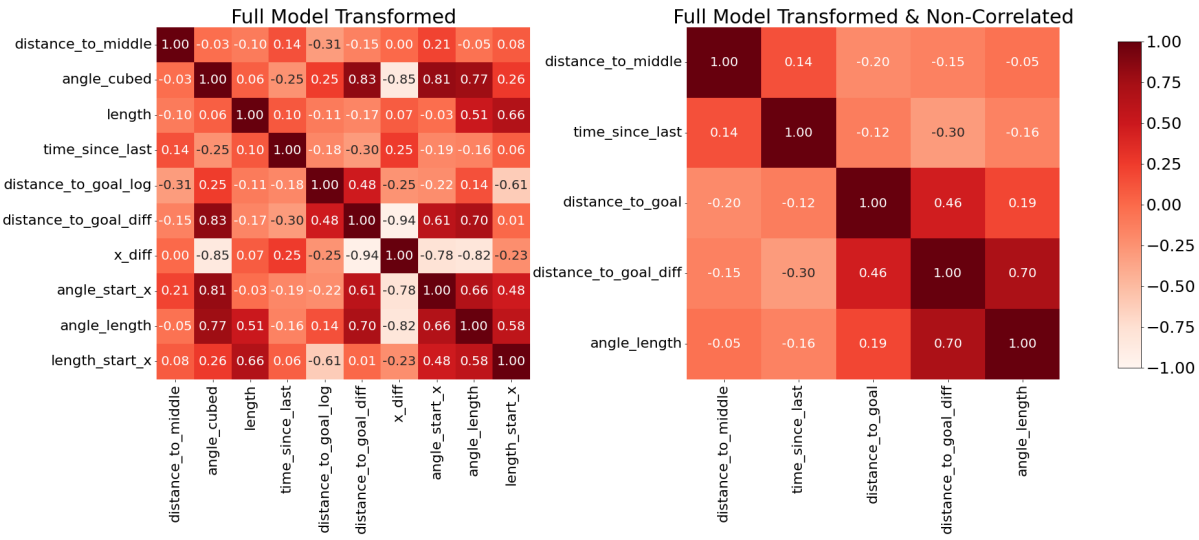


Figure 4.2.2: Correlation matrices of the possession retention models *Full Model Transformed* and *Full Model Transformed & Non-Correlated*.

Starting with the *Full Model Transformed*, the left subfigure of Figure 4.2.2 shows that this model includes a number of highly correlated features. For example, the features *distance_to_goal_diff* and *x_diff* have a correlation of -0.94. After removing highly correlated features, the features *distance_to_middle*, *time_since_last*, *distance_to_goal*, *distance_to_goal_diff* and *angle_length* are left. This model exhibits a lower degree of multicollinearity and the feature pair with the highest correlation is *angle_length* and *distance_to_goal_diff* with a value of 0.70.

Next, the VIFs of the two models are presented in Tables 4.2.3 and 4.2.4.

Table 4.2.3: VIF for *Full Model Transformed,* predicting possession retention.

Feature	VIF
distance_to_middle	5.75
angle_cubed	19.82
length	209.78
time_since_last	2.72
distance_to_goal_log	12.63
distance_to_goal_diff	37.14
x_diff	142.41
angle_start_x	31.02
angle_length	281.33
length_start_x	19.57

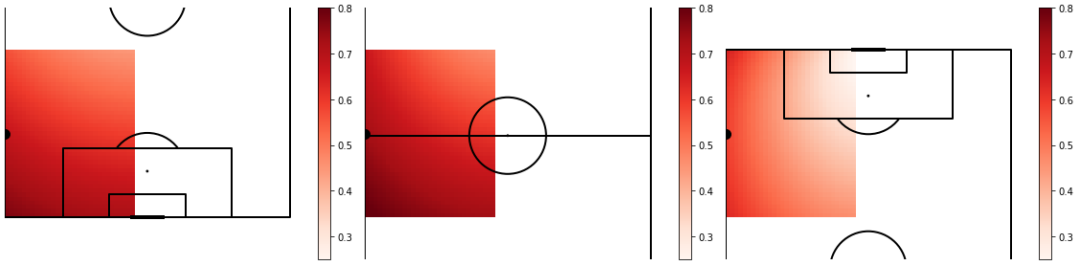
Table 4.2.4: VIF for *Full Model Transformed & Non-Correlated*, predicting possession retention.

Feature	VIF
distance_to_middle	3.75
time_since_last	2.64
distance_to_goal	3.64
distance_to_goal_diff	2.20
angle_length	3.30

Tables 4.2.3 shows once again that *Full Model Transformed* exhibits a high degree of multicollinearity, as seen by the high VIFs and Table 4.2.4 show the decreased VIFs for *Full Model Transformed & Non-Correlated*. As the *Full Model Transformed & Non-Correlated* exhibits no strong multicollinearity, this model will be further interpreted and evaluated below.

Result visualization

First, the results of the non-correlated model *Full Model Transformed & Non-Correlated* are evaluated visually, as presented in Figure 4.2.3.



(a) Defensive (b) Middle (c) Attacking

Figure 4.2.3: Shows the probability of retaining the ball possession for three different scenarios, i.e. throw-ins taken from the defensive, middle and attacking region of the pitch using logistic regression. The color of each heat map reflects the probability of retention given the end location of the throw-in. The start location of the throw-in is marked with a dark half-circle along the touchline and the time since the ball went out of the touchline is set to 10 seconds.

The results displayed in Figure 4.2.3 are obtained when training the non-correlated model on the entire data set. The subfigures display three situations when a throw-in is taken, and the location of the throw-in is marked with a dark half circle. The figures show the probability of possession retention for different end locations of the throw-in, as lighter areas of the heat maps correspond to regions where it is harder to retain the possession. Comparing the heat maps, one can see that according to the model it is harder to retain the ball for a throw-in taken in the attacking zone as seen by the lighter color in (c), especially if the ball is thrown towards to opposition's goal. It is also seen from (c) that the highest retention rate for attacking throw-ins is achieved when throwing backwards or towards the corner. Moreover, even though the maps of the defensive (a) and mid-pitch (b) throw-ins are similar, one can deduce a slightly darker color for the middle scenario. For both the defensive and middle situations, the model shows that the highest retention rate is obtained when throwing backwards, which aligns with the results in Figure 4.1.1.

Feature importance

Next, the importance of the included features in the non-correlated model is presented. This is done by standardizing the feature values before fitting the model, and then plotting the magnitude and confidence interval of every feature coefficient in the model, as seen in Figure 4.2.4.

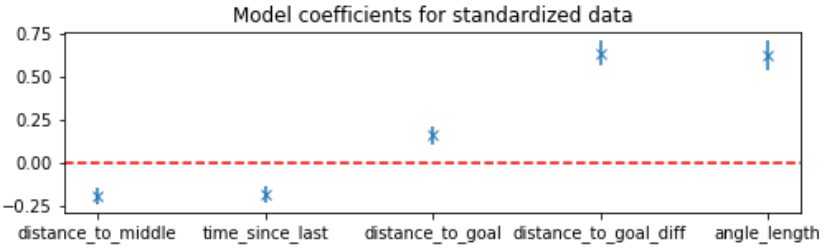


Figure 4.2.4: Standardized model coefficients, for non-correlated model predicting possession retention.

The larger the magnitude of a coefficient, the greater is the impact of the corresponding feature on the model's predictions. As observed in the plot, the features with the highest impact are *distance_to_goal_diff* and *angle_length*, both of which have a positive effect on the retention rate when being increased. The other three features, namely *distance_to_middle*, *time_since_last* and *distance_to_goal* seem to have approximately equal magnitudes of impact, although *distance_to_goal* has a positive effect while the other two have negative effects for increased feature values.

The next part of the results shows the effect that each feature has on the retention rate, when increasing the feature value with one unit.

Table 4.2.5: Odds ratio for unscaled coefficients, for non-correlated model predicting possession retention.

Regressor	Odds ratio	95 % confidence interval
distance_to_middle	0.828	[0.790, 0.868]
time_since_last	0.840	[0.800, 0.881]
distance_to_goal	1.17	[1.12, 1.24]
distance_to_goal_diff	1.89	[1.76, 2.03]
angle_length	1.86	[1.72, 2.02]

Table 4.2.5 shows the odds ratio, defined in Equation 2.5, of each feature in the non-correlated model, together with a 95 % confidence interval which is based on the confidence interval of the model coefficients. As one could also see in Figure 4.2.4, increased feature values of *distance_to_middle* and *time_since_last* result in lower retention rates, while increasing *distance_to_goal*, *distance_to_goal_diff*

and *angle_length* leads to higher possession retention rates. For example, if *time_since_last* is increased with one second, the odds of retaining the ball will decrease with 16 % as the multiplicative factor is 0.84.

4.2.2 Goal chance creation model

In this section, the results for the model predicting the goal chance creation is presented, similarly to the presentation of the results of the retention possession model in the previous section. Note that the heat maps and correlation matrices are displayed in shades of blue in order for the reader to more easily navigate between the different sections of the report.

Model comparison

Once again, the different models are defined, as shown in Table 4.2.6. The motivation of the models is the same as for the models of possession retention, but note that other sets of features are used compared to before. The accuracy of the models in regards to AIC, AUC and log loss is presented in Table 4.2.7.

Table 4.2.6: Chance creation models

Model name	**Included features**
Basic Model	start_x_adj, angle, length, time_since_last
Basic Model Transformed	start_x_adj_squared, angle_exp_0_2, length, time_since_last
Full Model	start_x_adj, angle, length, distance_to_goal, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x
Full Model Transformed	start_x_adj_squared, angle_exp_0_2, length, distance_to_goal_log, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x
Full Model Transformed & Non-Correlated	angle_exp_0_2, length, distance_to_goal_log

Table 4.2.7: Accuracy of various logistic regression models predicting goal chance creation. Note that the AIC is based on the entire data set, while AUC and log loss were calculated on a test set comprising 25% of the data.

Model	AIC	AUC	Log loss
Basic Model	3640	0.627	0.193
Basic Model Transformed	3621	0.621	0.193
Full Model	3601	0.609	0.196
Full Model Transformed	3597	0.611	0.195
Full Model Transformed & Non-Correlated	3592	0.627	0.193

When comparing the results of the *Basic Model* with the *Basic Model Transformed*, it can be observed that the AIC decreases after introducing feature transformations. In this case, *start_x_adj* and *angle* are transformed to *start_x_adj_squared* and *angle_exp_0_2* respectively, as shown in Table 4.2.6. However, the AUC decreases slightly after the transformations, suggesting lower predictive power, while the log loss remains unchanged. When introducing the *Full Model*, the AIC decreases again, however the accuracy decreases slightly based on the AUC and log loss. When fitting the transformed model *Full Model Transformed*, i.e. transforming *start_x_adj* to *start_x_adj_squared*, *angle* to *angle_exp_0_2* and *distance_to_goal* to *distance_to_goal_log* negligible accuracy improvement is obtained with respect to all three metrics. After removing correlated features to create the *Full Model Transformed & Non-Correlated*, the AIC decreases further slightly, resulting in the lowest AIC among the tested models for predicting goal chance creation. Note that the effect that each transformation has on the AIC is presented in Tables C.0.3 and C.0.4 in the Appendix.

Before presenting the results of the multicollinarity comparison between the models *Full Model Transformed* and *Full Model Transformed & Non-Correlated*, the ROC curve is presented for three of the models, as shown in Figure 4.2.5. In general, the ROC curves show that predicting the goal chance creation using logistic regression model is rather challenging as the ROC curves are relatively close to the random chance line, shown by the dotted diagonal.

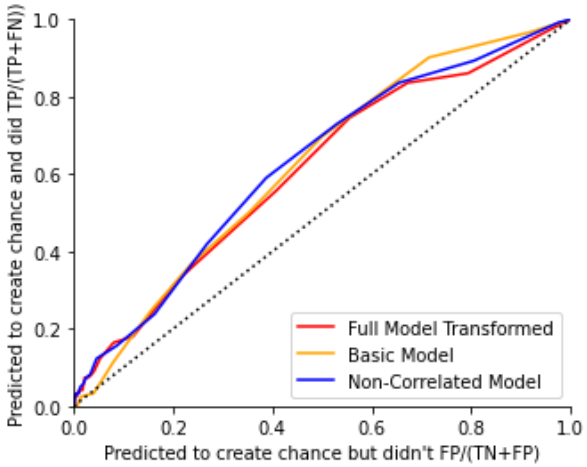


Figure 4.2.5: Comparison of ROC curves for three logistic regression models predicting goal chance creation. Note that the model named *Non-Correlated Model* in the legend refers to *Full Model Transformed & Non-Correlated*.

Multicollinearity

As for the possession retention model, the results of the multicollinarity results are presented next.

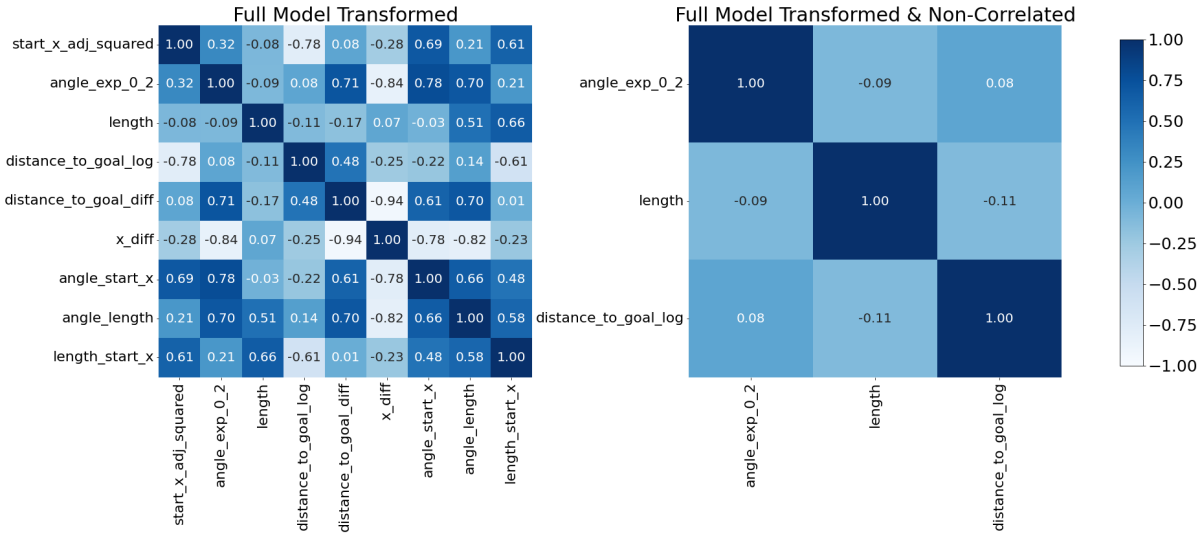


Figure 4.2.6: Correlation matrices of the goal chance creation models *Full Model Transformed* and *Full Model Transformed & Non-Correlated*.

Figure 4.2.6 shows that the *Full Model Transformed* includes several feature pairs with high correlation, the largest pair once again being *distance_to_goal_diff* and *x_diff* with a value of -0.94. The right subfigure shows the presence of correlation after removing correlated features. Among these three features, no pair exhibits any strong correlation.

Next, the VIFs of the two models are presented. Table 4.2.8 shows that the features

Table 4.2.8: VIF for *Full Model Transformed*, predicting chance creation

Feature	VIF
start_x_adj_squared	21.92
angle_exp_0_2	140.30
length	262.26
distance_to_goal_log	100.63
distance_to_goal_diff	24.32
x_diff	123.45
angle_start_x	19.29
angle_length	323.32
length_start_x	36.50

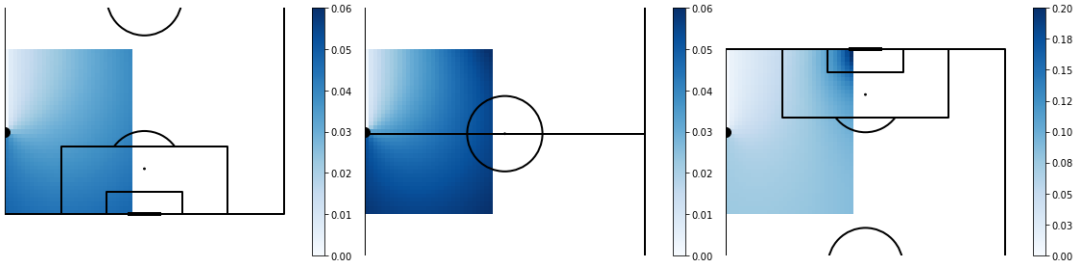
Table 4.2.9: VIF for *Full Model Transformed & Non-Correlated*, predicting chance creation

Feature	VIF
angle_exp_0_2	23.02
length	5.05
distance_to_goal_log	24.50

in the *Full Model Transformed* in general have large VIFs, indicating that this model has strong presence of multicollinearity. Table 4.2.9 indicates that there are still two features in *Full Model Transformed & Non-Correlated* with rather high VIFs. However, as the confidence intervals of the coefficients in the *Full Model Transformed & Non-Correlated* are tight, which is presented in Figure 4.2.8 below, the features *angle_exp_0_2*, *length* and *distance_to_goal_log* are kept in the non-correlated model.

Result visualization

As for the possession retention model, the non-correlated model is further investigated and the results are first presented visually in three different situations in the field, as shown in Figure 4.2.7.



(a) Defensive (b) Middle (c) Attacking

Figure 4.2.7: Shows the probability of creating a goal chance opportunity for three different scenarios, i.e. throw-ins taken from the defensive, middle and attacking region of the pitch using logistic regression. The color of each heat map reflects the probability of chance creation given the end location of the throw-in. The start location of the throw-in is marked with a dark half-circle along the touchline. Note that the color scale is different in *Attacking*, compared to *Defensive* and *Middle*.

Firstly, note that the color scale in Figure 4.2.7 is different in (c) compared (a) and

(b) in order to clearer interpret the results. From these figures, one can see that the goal-scoring opportunity increases as the throw-in is taken further up in the pitch. For the defensive and middle situations, the goal chance creation rate is higher for throws that are directed backwards than forwards. Apart from this, it can be noticed that a relatively high chance creation rate is obtained when throwing a long ball straight ahead. For the attacking situation, the highest chance of creating a goal-scoring opportunity is obtained when throwing the ball close to the opposition's goal. Also, one can note in (c) that throwing the ball backwards results in higher goal-scoring opportunity than throwing the ball towards the corner flag.

Feature importance

Next, the importance among the features in the non-correlated model is presented.

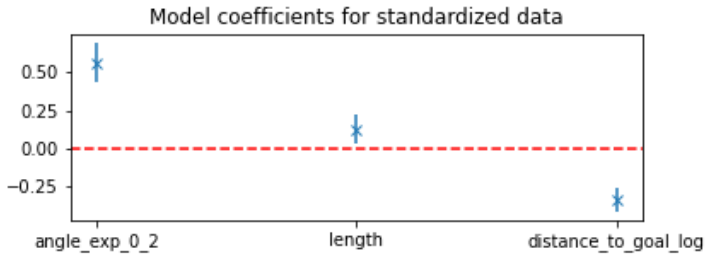


Figure 4.2.8: Standardized model coefficients, for non-correlated model predicting chance creation.

Figure 4.2.8 shows that the feature having the most impact on the goal chance creation is *angle_exp_0_2*. The second most influential feature is *distance_to_goal*, while the least important feature among the three is the length of the throw-in. Increasing the feature *distance_to_goal* has a negative effect on goal chance opportunity creation, while *angle_exp_0_2* and *length* have a positive effect. Table 4.2.10 shows how the goal chance creation rate changes when increasing each of the features with one unit.

Table 4.2.10: Odds ratio for unscaled coefficients, for non-correlated model predicting chance creation.

Regressor	Odds ratio	95 % confidence interval
angle_exp_0_2	21,3	[10,7, 42,6]
length	1.02	[1.00, 1.03]
distance_to_goal_log	0.52	[0.44, 0.60]

The information presented in Table 4.2.10 can be interpreted in the following way. When the length of the throw is increased by one meter, the odds of creating a goal

chance increase by a factor of 1.02 and similar reasoning can be applied to the other two features. Note that odds ratio for *angle_exp_0_2* differs substantially from the others. This can be explained by the angle being measured in radians, and thus a one unit increase is a substantial change in angle.

4.3 Results for gradient boosting model

This section presents a comparison to a more advanced model, namely gradient boosting. The same feature combinations are examined as for the possession retention and goal chance creation models. However, the models are only evaluated in regards to AUC and log loss, as AIC is more suitable for statistical methods with a well defined likelihood function.

Table 4.3.1: Accuracy of various gradient boosting models predicting the possession retention.

Model	**AUC**	**Log loss**
Basic Model	0.766	0.532
Basic Model Transformed	0.758	0.537
Full Model	0.770	0.528
Full Model Transformed	0.768	0.529
Full Model Transformed & Non-Correlated	0.768	0.529

First, the different models for predicting possession retention are investigated. From Table 4.3.1 one can see that the transformations do not seem to have any clear positive effect on the model fitting accuracy. Comparing the two basic models, the transformed version has a lower AUC and the log loss increases slightly indicating poorer predictive power. When adding more features to the model and thus creating the *Full Model*, the model fitting is slightly improved. Introducing the transformations in the full model does not change the AUC or log loss noticeably compared to the non-transformed version of the full model and same the applies to the non-correlated model.

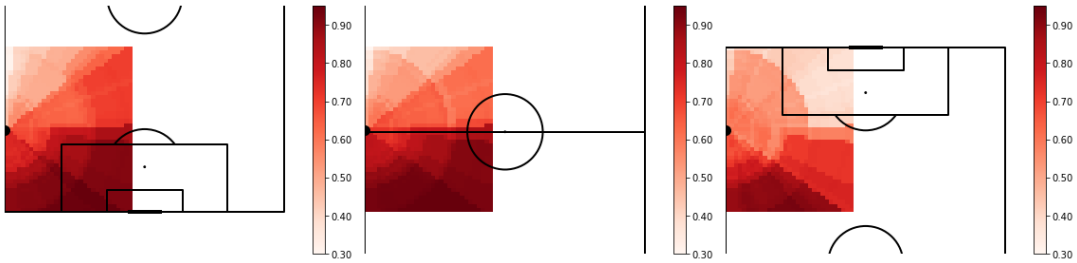
Moving on to the model that predicts goal chance creation, Table 4.3.2 presents the results of the model fitting for different sets of features.

Table 4.3.2: Accuracy of various models predicting goal chance creation using gradient boosting.

Model	AUC	Log loss
Basic Model	0.611	0.194
Basic Model Transformed	0.611	0.194
Full Model	0.600	0.194
Full Model Transformed	0.600	0.194
Full Model Transformed & Non-Correlated	0.592	0.195

Table 4.3.2 shows once again that feature transformations do not have any positive impact in regards to AUC and log loss when using gradient boosting to predict the goal chance creation after a throw-in. Moreover, it is interesting to note that the basic models are slightly better in regards to AUC than the rest of the models while the log loss barely changes across the models.

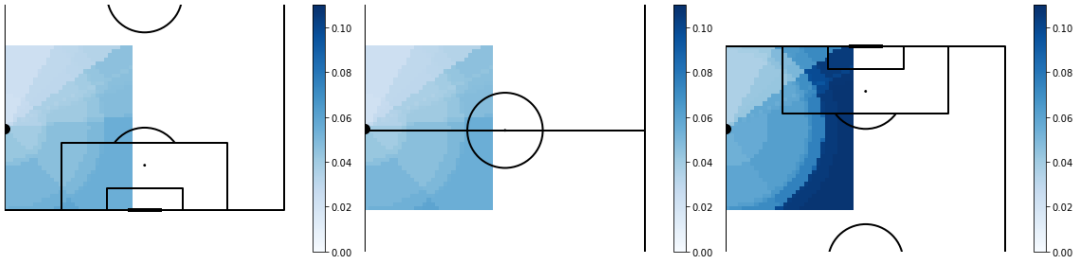
Since the the transformations and addition of features did not contribute to any clear improvement in terms of AUC and log loss, the visual representation of the results are presented based on the *Basic Model*. As for the logistic regression case in the previous section, the results are presented in three different scenarios for each model.



(a) Defensive (b) Middle (c) Attacking

Figure 4.3.1: Shows the probability of retaining the ball possession for three different scenarios, i.e. throw-ins taken from the defensive, middle and attacking region of the pitch using gradient boosting. The color of each heat map reflects the probability of retention given the end location of the throw-in. The start location of the throw-in is marked with a dark half-circle along the touchline and the time since the ball went out of the touchline is set to 10 seconds.

Figure 4.3.2 shows the probability of retaining the possession depending on the end location of the throw, according to gradient boosting. In all three situations the retention rate is the highest when throwing the ball backwards. In the attacking zone, the retention rate as clearly lower when the end location of the throw-in is close to the opposition's goal. The corresponding results for the model predicting the goal chance creation is presented below.



(a) Defensive (b) Middle (c) Attacking

Figure 4.3.2: Shows the probability of creating a goal chance opportunity for three different scenarios, i.e. throw-ins taken from the defensive, middle and attacking region of the pitch using gradient boosting. The color of each heat map reflects the probability of retention given the end location of the throw-in. The start location of the throw-in is marked with a dark half-circle along the touchline and the time since the ball went out of the touchline is set to 10 seconds.

Figure 4.3.2 shows that when taking the throw-in in the attacking zone, throwing the ball further away results in higher goal chance creation rate as seen in (c). Also, one can note from the defensive (a) and middle (b) situation that the ball should be thrown backwards in order to have greater chance of creating a goal-scoring opportunity, even though the probability of creating a goal-scoring opportunity is low in these situations.

4.4 Team analysis

This section will present the results of the comparison of the Allsvenskan 2022 teams in regards to their throw-in strategies. Using this it will be possible to investigate if there are teams that stand out when it comes to successful throw-ins and in that case what distinguishes these teams from the rest.

The results will be presented using club emblems and thus a specification of the club emblems and the corresponding team names is given in Table 4.4.1.

Table 4.4.1: The Allsvenskan 2022 teams and corresponding emblems.

	AIK		IFK Göteborg
	BK Häcken		IFK Norrköping FK
	Degerfors IF		IFK Värnamo
	Djurgården		IK Sirius FK
	GIF Sundsvall		Kalmar FF
	Hammarby		Malmö FF
	Helsingborgs IF		Mjällby AIF
	IF Elfsborg		Varbergs BoIS FC

Figure 4.4.1 shows how a team's average throwing angle affects the possession retention and goal chance creation rate for throw-ins taken in the attacking third of the pitch. As a reminder, an angle of 0 degrees represents a throw-in directed straight up towards the opposition's half, while an angle of 180 degrees corresponds to a throw-in straight down the pitch towards the own half. The left subfigure of Figure 4.4.1 shows that, in general, the larger the average throwing angle of a team is, the higher is the rate of retaining the ball after a throw-in.

The right subfigure of 4.4.1 demonstrates how the a team's average throwing angle affects the goal chance creation rate during attacking throw-ins. Notably, Varbegs BoIS is a team that stands out as this team throws the ball with the smallest angle, i.e. most forward, and creates most goal-scoring chances. By observing both subfigures it can also be noticed that teams which were successful at retaining the ball, such as Kalmar FF and BK Häcken, are relatively less successful at creating goal-scoring opportunities.

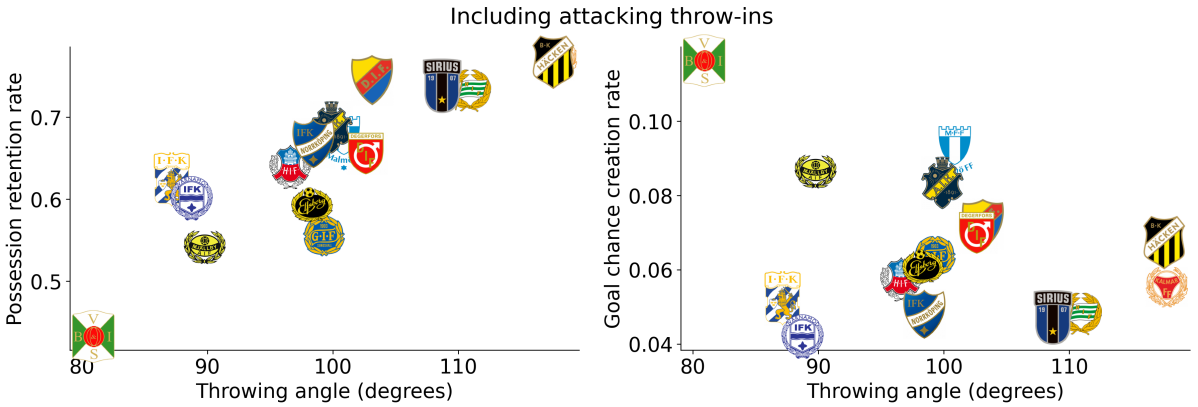


Figure 4.4.1: Shows the relationship between a team's average throwing angle and the average success rate of the throw-ins, with respect to possession retention to the left and goal chance creation to the right. Note that only throw-ins taken in the attacking third of the pitch are included here. The higher the throwing angle, the more backwards is the throw-in directed.

Figure 4.4.2 shows how the average throwing angle affects the success rate for throw-ins taken in the defensive and middle thirds of the pitch. Once again, it it clear that teams that throw the ball backwards are also better at keeping possession of the ball. When it comes to the angle's effect on the goal chance creation, there appears to be a positive correlation between the angle and the chance creation rate. Teams with high chance creation rate, such as Hammarby, BK Häcken and Kalmar FF, have a relatively large angle on their throw-ins. In contrast, teams with small throwing angles, such as Varbergs BoIS and IFK Norrköping tend to create less chances.

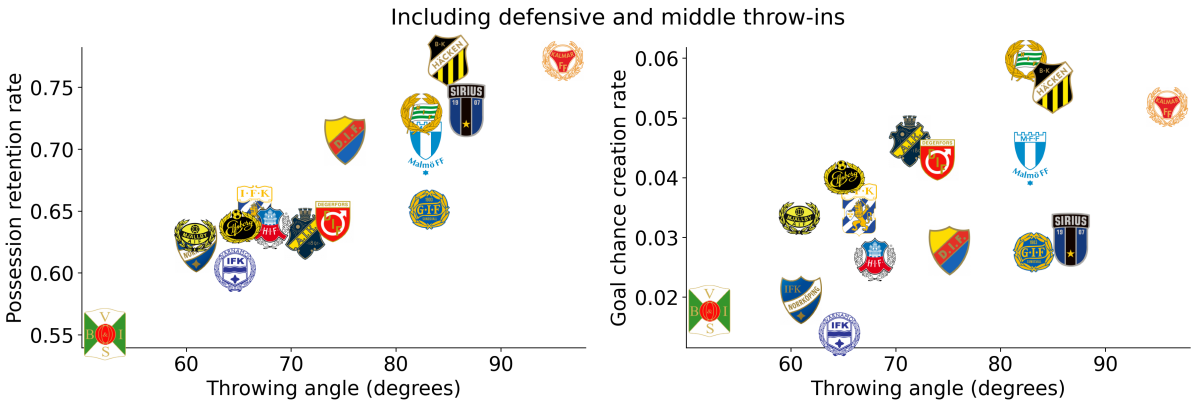


Figure 4.4.2: Shows the relationship between a team's average throwing angle and the average success rate of the throw-ins, with respect to possession retention to the left and goal chance creation to the right. Note that only throw-ins taken in the defensive and middle thirds of the pitch are included here. The higher the throwing angle, the more backwards is the throw-in directed.

Next, the effect of the average throwing length on the success rate is presented. In Figure 4.4.3 it is seen how the throwing length affects the success rate for throw-ins

taken in the attacking third of the pitch. In the left subfigure it can be noticed that teams which take long throw-ins are also worse at retaining the ball. In particular it can be seen that Varbergs BoIS is clearly the team that throws the ball the furthest, and are also least successful when it comes to keeping possession of the ball. In contrast, Kalmar FF throws the shortest throw-ins on average, and they are one of the most successful teams at keeping possession after a throw-in.

By considering the right subfigure of Figure 4.4.3, it can be seen that Varbergs BoIS is clearly the most successful team when it comes to creating goal chance opportunities after a throw-in in the attacking third of the pitch. Moreover, it is seen that Mjällby AIF has the second longest throw-ins among the teams and that they are relatively successful at creating goal-scoring opportunities.

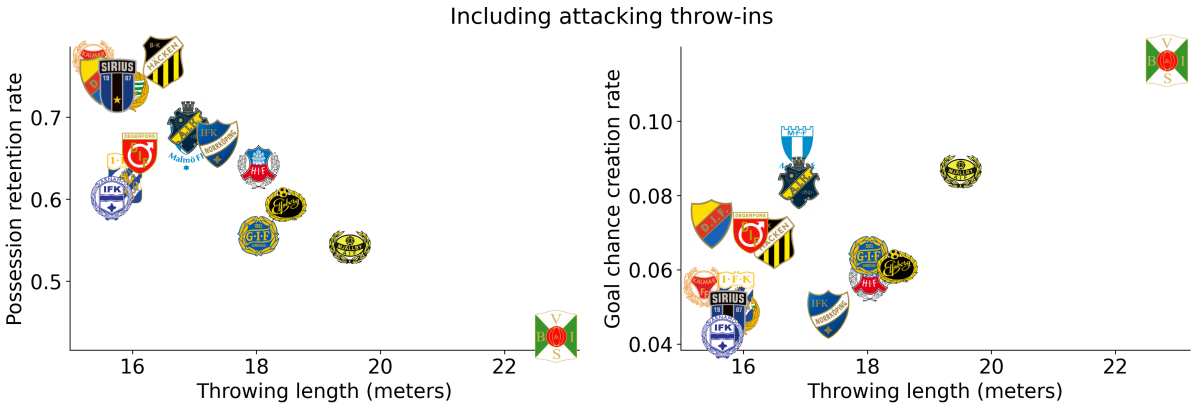


Figure 4.4.3: Shows the relationship between a team's average throwing length and the average success rate of the throw-ins, with respect to possession retention to the left and goal chance creation to the right. Note that only throw-ins taken in the attacking third of the pitch are included here.

Finally, Figure 4.4.4 considers throw-ins taken in the defensive and middle thirds of the pitch, and shows how the average throwing length affects the success rate. By considering both subfigures, it can be noticed that team's which have a higher rate of successful throw-ins tend to take relatively short throw-ins. However, note that there are also teams with short average throw-ins which are less successful.

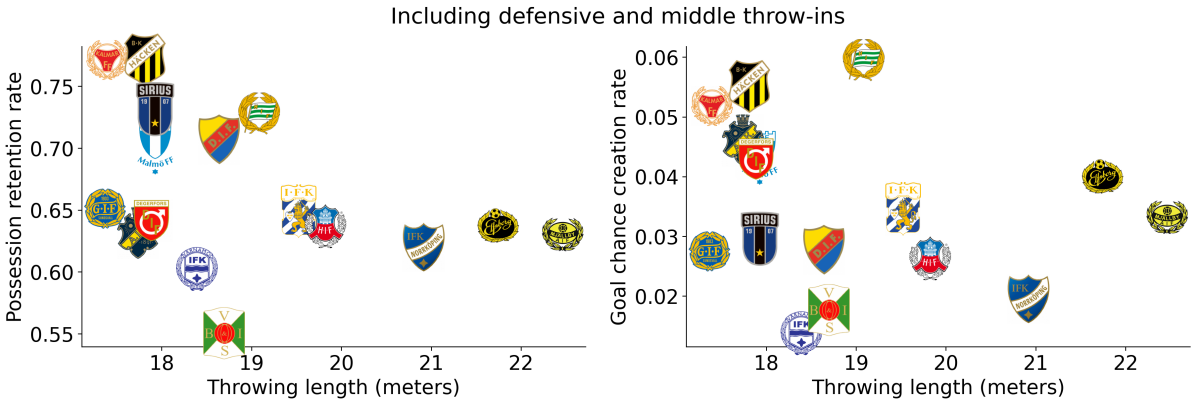


Figure 4.4.4: Shows the relationship between a team's average throwing length and the average success rate of the throw-ins, with respect to possession retention to the left and goal chance creation to the right. Note that only throw-ins taken in the defensive and middle thirds of the pitch are included here.

Chapter 5

Discussion

This section contains a discussion and analysis of the obtained results. First, the logit plots are evaluated as a tool for finding relevant features and transformations. Secondly, the results of the logistic regression models are discussed and interpreted. Thirdly, the predictive power of the logistic regression is compared to the gradient boosting method and finally, the team analysis is further discussed and analysed.

5.1 Logit plots and transformations

To start with, the approach of determining the feature target dependence prior to fitting a logistic model, as presented in Section 4.1, is discussed in more detail. These plots serve for two objectives: the first being to understand whether a feature actually has an impact on the target, and secondly how the relationship looks like between the feature and target, i.e. if it is linear or non-linear. With the feature importance plots for the two targets in mind, i.e. Figure 4.2.4 and 4.2.8, the usefulness of the logit plots when it comes to finding impactful features can be evaluated. According to Figure 4.2.4, *distance_to_goal_diff* and *angle_length* are the most important features in the non-correlated model for predicting possession retention. This is consistent with the logit plots for these two features, as seen in Figure 4.1.3 and B.0.1 in the Appendix, which show relatively high variation in the log odds when varying these features. Same can be seen for *angle_exp_0_2*, see Figure 4.1.4, which was the most impactful feature in the non-correlated model predicting the goal chance creation. Based on this, the logit plots can be highly useful when trying to find important features for a logistic regression model.

Regarding the use of these kind of plots for determining potential transformations the following can be concluded. The transformations used in the models were based on the logit plots, and in several cases they resulted in better model fitting, as could be

seen in Tables 4.2.2 and 4.2.7, suggesting that the logit plots served an effective way of linearising the feature target dependencies, improving the model fitting. However, not all feature transformations resulted in better model fitting, even though one could have thought that they would based on the logit plots. For example, the transformation of the feature *distance_to_goal_diff* to an exponential with base 1.1 or 1.2 appeared to be an effective transformation in order to obtain a linear relationship, according to the logit plots in Figure 4.1.3. However, when adding these transformations to the possession retention model, the model fitting results became worse, and thus these features were not included in the transformed versions of the models. One potential reason for this is that the feature transformations were based on the logit plots, which only approximate how the log odds depend on the feature values before fitting a model, as described by Figure 3.3.1. A possible way of improving the approximation would be acquire more data of taken throw-ins, such that the data could be divided into a larger number of bins making the approximation more precise.

The results of the feature transformations imply that caution should be taken when using the logit plots as a way of determining a potential transformation, even though this might in many times be a advantageous procedure. Also, when conducting the transformations there is also a trade-off between improving the model accuracy and decreasing the interpretability of the features in the model. Furthermore, the feature selection and feature transformation should be conducted together with one's own understanding and intuition of football. To sum up, the plots presented in Section 4.1 could serve as a guidance when trying to find relevant features and transformations for a logistic regression model.

5.2 Logistic regression models

Next, the possession retention model is discussed in more detail. Starting of with the heat maps presented in Figure 4.2.3, the obtained results are rather expected. The heat map for the defensive and mid-field situations illustrate that the ball should primarily be thrown backwards in order to keep possession of the ball. This is expected as the opposition team is naturally positioned up the pitch. For the attacking situation it is considerably more difficult to retain possession when throwing towards the opposition's goal, while the highest retention rate is obtained when throwing close to the touchline. Another feature that has an impact on the retention rate, but that can not be observed from the heat maps, is how fast the throw-in is taken, denoted by the feature *time_since_last*. The longer time the thrower waits, the harder it is to keep possession after a throw-in as shown by the odds ratio presented in Table 4.2.5. An explanation for this could be that the opposition players then have more time to organize.

Another observation that could be made in the possession retention model is that a transformed feature could have a positive impact on the model when being part of one set of features, while in another set of features it could have the opposite effect. For example, the transformed feature *angle_cubed* resulted in lower AIC when applied to the Full Model, while when applying it to the Basic Model it resulted in the model having a higher AIC value. One possible explanation for this could be the presence of multicollinearity in the data sets, particularly in the one used for the *Full Model*. Remember that if a set of features have a strong degree of multicollinearity, a small change in the data set could result in a substantial change in the model coefficients and, consequently, the model's output. For this reason, replacing *angle* with *angle_cubed* in the *Full model* could have a substantial impact on the model's result and in this case, the effect appears to be positive based on the improvement in AIC.

Comparing the possession retention model with the goal chance creation model, one can observe that the latter results in poorer accuracy in terms of AUC and log loss as seen from Tables 4.2.2 and 4.2.7. Note that the AIC values are not comparable for the possession retention and goal chance creation models. From the AUC and log loss values one can conclude that, using logistic regression, it is harder to predict whether a throw-in results in a goal chance compared to predicting possession retention. One reason for this could be that it is relatively rare to create a goal chance opportunity directly from a throw-in, which makes the data set unbalanced in regards to the target variable *chance_created*. This makes it harder to learn how a throw-in should be taken in order to create a goal-scoring opportunity.

Continuing the comparison, one can observe that the feature transformations included in the model for predicting goal chance creation have a positive impact on the model fitting, although the improvement for the *Full Model* was negligible as seen by Table 4.2.7. Also, not all transforms resulted in an improved model fitting, as previously stated. Another interesting observation is that the model which was created after removing correlated features, resulted in the lowest AIC as seen in Table 4.2.7.

Furthermore, the heat maps presented for the goal chance creation model provide some interesting insights. The heat maps, but also the odds ratio in Table 4.2.10 suggest that it is more effective to throw the ball backwards in defensive and middle parts of the pitch in order to create a goal-scoring opportunity. This could appear counterintuitive as one could think that it is better to throw the ball towards the opposition's goal in order to score. This is however true for the attacking situation, as seen in Figure 4.2.7. An explanation of this result can be that the model has two features, *angle_exp_0_2* and *distance_to_goal_log* with opposite effects, and when the end location location of the throw-in is close to the opposition's goal, *distance_to_goal_log* becomes dominant. Furthermore, interestingly, the heat map for the attacking situation shows that it is more effective to throw the ball backwards

than throwing towards the corner flag as the region around the corner flag has a slightly lighter color.

To sum up, the results of the non-correlated logistic regression models show that when the aim is to keep the possession of the ball, the rate of success is relatively high when throwing the ball backwards, and this applies for all three investigated positions of the pitch. For attacking situations, the rate of possession retention is also high when throwing the ball forward but close to the touchline. In contrast, if a team's goal is to create a goal-scoring opportunity, the model suggests that the ball should be thrown forward towards to opposition's goal if the throw-in is taken in the attacking part of the pitch. However, when taking the throw-in from the middle or defensive parts of the pitch, it is surprisingly more effective to throw the ball backwards than forwards. A conclusion that can be made consequently is that keeping possession is an important factor in order to create an goal-scoring opportunity.

5.3 Comparison to gradient boosting method

In order to further evaluate the possession retention and goal chance creation models, a comparison to a more complex model is made, namely gradient boosting. Compared to logistic regression in the context of predicting possession retention, the gradient boosting results in noticeably higher accuracy for the *Basic Model*. The logistic regression achieved an AUC of 0.737 compared to 0.766 of gradient boosting. However, when adding more features to the models, both methods performed rather equally. For example, the *Full Model* of both methods obtained an AUC of 0.770. When predicting goal chance creation, logistic regression surprisingly produces more accurate results in terms of AUC for all models. This is unexpected since gradient boosting is considered to be a more complex method with higher predictive power. For example, the *Basic Model* of the logistic regression achieved an AUC of 0.627 while the corresponding result for the gradient boosting method was 0.611. This could suggest that the gradient boosting algorithm is more sensitive to imbalanced data compared to logistic regression.

One approach to handle the class imbalance would be to resample the data set by oversampling the goal chance creation class so that both classes were approximately equally frequent [2], which could potentially improve the results of gradient boosting. The logistic regression models can also be adjusted to be more suitable for imbalanced data sets. One way is to introduce weights to the likelihood function so that misclassifications of the minority class are penalized more [27]. Including these techniques for handling imbalanced data sets could give more accurate results for both models and thus a deeper understanding of how to throw the ball in order to maximize the chance of scoring.

Continuing with the comparison, one can observe that feature transformations play a less important role when using gradient boosting, as presented in Tables 4.3.1 and 4.3.2. This is rather expected as gradient boosting is a non-linear method, and thus this method can learn non-linear relationships without the need of transformations. Also, note that the hyperparameter optimization was conducted based on the features in the *Basic Model* and thus it is possible that better accuracy results would have been obtained for the non-basic models if separate optimization was conducted for every gradient boosting model.

Having elaborated on the differences in accuracy of logistic regression and gradient boosting, the two methods are further compared by considering the heat maps of the success rate depending on where the ball is thrown. Note that the heat maps of the logistic regression model were based on the features in *Full Model Transformed & Non-Correlated* while the features in the *Basic Model* were used for generating the heat maps for gradient boosting, since the addition and transformation of features did not result in any clear accuracy improvement for gradient boosting.

If first comparing the results for possession retention, by considering Figures 4.2.3 and 4.3.2, both methods suggest that throwing the ball backwards results in relatively high retention rate. However, the results of gradient boosting show a clearer angle dependence, as the variation in retention rate is larger for different angles. Both methods show that for attacking throw-ins, it is most difficult the retain the ball when it is thrown towards the opposition's goal. However, the logistic regression method also suggests high retention rate when throwing the ball forward towards the corner, which gradient boosting does not do.

Moving on to the comparison of the heat maps of the goal chance creation rate, shown in Figure 4.2.7 and 4.3.2, the following observations can be made. Starting with the attacking throw-in, both methods suggest that throwing a long ball towards the opposition's goal results in considerably higher goal chance creation rate. Also, the results of both methods show that short throws directed forward lead to slightly lower goal chance creation rate, compared to short throws directed backwards. However, a major difference between the two methods is that gradient boosting also suggests that high retention rate is achieved for long throws directed backwards. Comparing the results for the middle and defensive throw-ins, one can firstly notice that the variation in goal chance creation rate is relatively low for both methods. This indicates that no matter of how the throw-in is taken, it is difficult to create a goal-scoring opportunity from a throw-in in these parts of the pitch. However, if the chances of scoring should be maximized, the results of both methods suggest that the ball should not be thrown with a small angle, i.e. forward.

5.4 Team analysis

Finally, the conducted team analysis for the teams participating in the 2022 season of Allsvenskan is discussed. Comparing these results to the fitted models also makes it possible to evaluate whether the conclusions drawn from the models are reasonable and have support.

Combining the effect of the average throwing angle and length on the success rate among the Allsvenskan teams, presented in Section 4.4, can give some insightful conclusions. If the aim is to create a goal-scoring opportunity from a throw-in in the attacking third of the pitch, the results of the team analysis demonstrate that the most effective strategy is to take a long throw-in with a relatively small angle, i.e. a throw-in directed to the oppositions' goal as suggested by Figures 4.4.1 and 4.4.3. As shown by the figures, Varbergs BoIS is the team that clearly takes the longest throw-ins and with the smallest angle, in the attacking third. Having a goal chance creation rate of 11.6 %, Varbergs BoIS is the most successful team in creating goal-scoring opportunities from attacking throw-ins. Mjällby AIF uses a similar strategy for their attacking throw-ins and have a goal chance creation rate of 8.4 % which makes them the third most successful team in the league. Throwing long balls towards the opposition's goal in order to increase the chance of scoring is consistent with the inferences drawn from the logistic regression and gradient boosting models which gives support to the models.

These results can be valuable for teams that aim to create more chances from their attacking throw-ins. For example, teams like IFK Göteborg and IFK Värnamo also throw their throw-ins with a relatively small angle in the attacking third, however the throw-ins are in general rather short. Thus, the results of this project suggest that these teams would have to increase the length of their throw-ins in order to be more successful. Note however that taking long throw-ins requires players who actually have the physical ability to throw the ball long. Also, long balls towards the opposition's goal could demand that the players in the penalty area are good at winning aerial duels or that the team develops a certain player movement strategy among the attacking players. For this reason, it is not straightforward for a team to increase their goal chance creation rate in the attacking zone as this could require special training for throwing further or recruiting players with the ability of throwing long and winning aerial duels.

Throwing long and with a small angle does however not seem to be a successful approach if the aim is to retain possession of the ball. Varbergs BoIS and Mjällby AIF are the least successful teams in the league at keeping possession of the ball after taking their attacking throw-ins. This also agrees with the results of the logistic regression and gradient boosting models, which suggested that it was most difficult to keep possession

of the ball when it was thrown towards the opposition's goal. Thus there seems to be a trade-off for attacking throw-ins: throwing long with a small angle is effective for increasing the chance of scoring, but ineffective for keeping control of the ball.

If a throw-in is instead taken in the middle or defensive parts of the pitch and the aim is to create a goal-scoring opportunity, the team analysis indicates that teams which throw the ball backwards to a greater extent are in general better at creating goal-scoring opportunities. These results align with the inferences drawn from the fitted models. Remember that the logistic regression and gradient boosting models both suggested that throwing the ball backwards in middle and defensive throw-ins was more successful for creating goal-scoring opportunities than throwing forwards, even though the chance was still comparably low compared to attacking throw-ins. Consequently, these results could be of high value for teams which throw the ball with a small angle in the middle and defensive thirds of the pitch, hoping to increase their chance to score. For example, the results presented in Figure 4.4.2 suggest that if teams like Varberg BoIS and IFK Norrköping would throw the ball with a larger angle, i.e. more backwards, they could become more successful at both keeping possession of the ball and creating goal chances. The effect of the throwing length on the success rate of middle and defensive throw-ins is less obvious and seems to have a smaller impact on the throw-in outcome, based on Figure 4.4.4.

When it comes to retaining the ball, Figures 4.4.1 and 4.4.2 show that teams which tend to throw the ball with a larger angle, are also more successful at keeping the possession of the ball after the throw-in, regardless of the location of the pitch. This finding aligns rather well with the results obtained from the logistic regression and gradient boosting models, which both suggested that throwing the ball backwards increased the rate of possession retention. The logistic regression model also suggested that throwing a short ball forward resulted in an increased retention rate as well, especially for attacking throw-ins. This is less clear from the team analysis results. However one can see a tendency for this by considering the results for IFK Göteborg and IFK Värnamo in Figure 4.4.1. These two teams throw relatively short throw-ins with a small angle in the attacking third, and compared to Varbergs BoIS and Mjällby AIF, which also throw with a small angle but longer distance, IFK Göteborg and IFK Värnamo demonstrate a higher rate of retaining ball possession compared to the other teams mentioned.

To sum up, the results of the team analysis align rather well with the inferences drawn from the fitted models. It is clear that the optimal strategy highly depends on whether the goal is to keep possession of the ball or to create a goal-scoring opportunity.

Chapter 6

Conclusions

To conclude, this project has provided findings which can be useful when designing a logistic regression model and insights regarding throw-in strategies which can be valuable for football teams.

It has been shown how logit plots could serve as a guidance when trying to find relevant features and transformations for a logistic regression model. In particular, logit plots can be highly helpful for finding features with a high impact on the target. The feature transformations have in several cases lead to improved model accuracy. However, it is important to note that not all transformations have resulted in improved accuracy, and when improvements have been observed, they have often been marginal. The substantial accuracy improvements have often originated from finding relevant features and adding them to the models, especially for the model predicting possession retention. This indicates that in this project it has been more useful to identify useful features based on intuition of football than searching for appropriate transformations. Nevertheless, the transformations have still had an important role in fine-tuning and optimizing the models.

Moving on to the football related results, the non-correlated logistic regression models, together with the conducted team analysis, have given a couple of insights that could be useful for football players and coaches. First of all, if a team wishes to keep possession of the ball, the results suggest that it is effective to throw the ball backwards regardless of where the throw-in is taken from. However, if a team prioritises to increase the chance of scoring, the results indicate that the optimal tactics depend on where on the pitch the throw-in is taken from. When taking a throw-in in the attacking third of the pitch, the most effective strategy has shown to be throwing a long ball towards the opposition's goal. If the throw-in is instead being executed from the middle or defensive parts of the pitch, the chance of scoring after the throw-in is relatively low. However, in order to optimize this chance, the results of the logistic regression model

have shown that it is more effective to throw the ball backwards than forwards. This suggests that it is more advantageous to play safe and keep possession of the ball in order to create a goal-scoring opportunity from middle and defensive throw-ins. This finding can offer teams valuable insight that can improve their chances of scoring after a throw-in and winning football games.

Comparing the logistic regression models to gradient boosting, a couple of conclusions can be drawn. When predicting possession retention, which is based on a relatively balanced data set, the gradient boosting clearly outperforms the logistic regression when only including the fundamental features describing the throw-in. However, when including more relevant features, the logistic regression and gradient boosting perform rather similarly. When instead predicting goal chance creation, the accuracy of the models decreases. Notably, the logistic regression outperforms gradient boosting, even though the difference is not substantial. As discussed, a possible explanation to this might be that gradient boosting is more sensitive to class imbalance. For this reason, a potential improvement of this project could be to incorporate techniques for class imbalance in order to improve the accuracy of the models used to predict goal chance creation.

Summarizing the comparison of logistic regression to gradient boosting, this project has shown that even though logistic regression is considered to be a simple method, it obtains a predictive power comparable to gradient boosting if valuable features and transformations can be found. In this case, logistic regression becomes a powerful method at the same time as having a high interpretability and keeping the computational speed high.

Bibliography

[1] Bentéjac, Candice, Csörgő, Anna, and Martinez-Muñoz, Gonzalo. “A comparative analysis of gradient boosting algorithms”. In: *Artificial Intelligence Review* 54 (2021), pp. 1937–1967.

[2] Cahyana, Nurheri, Khomsah, Siti, and Aribowo, Agus Sasmito. “Improving imbalanced dataset classification using oversampling and gradient boosting”. In: *2019 5th International Conference on Science in Information Technology (ICSITech)*. IEEE. 2019, pp. 217–222.

[3] Carter, Jane V, Pan, Jianmin, Rai, Shesh N, and Galandiuk, Susan. “ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves”. In: *Surgery* 159.6 (2016), pp. 1638–1645.

[4] Cavanaugh, Joseph E and Neath, Andrew A. “The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 11.3 (2019), e1460.

[5] Daubechies, Ingrid, DeVore, Ronald, Fornasier, Massimo, and Güntürk, C Sinan. “Iteratively reweighted least squares minimization for sparse recovery”. In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 63.1 (2010), pp. 1–38.

[6] Fifa. *Pitch dimensions and surrounding areas*. [Accessed: May 4, 2023]. n.d. URL: https://publications.fifa.com/en/football-stadiums-guidelines/technical-guideline/stadium-guidelines/pitch-dimensions-and-surrounding-areas/.

[7] Guinness World Records. *Farthest distance football (soccer) throw-in (male)*. [Accessed: May 3, 2023]. n.d. URL: https://www.guinnessworldrecords.com/world-records/longest-throw-in-(football)/.

[8] Harris, Jenine K. “Primer on binary logistic regression”. In: *Family Medicine and Community Health* 9.Suppl 1 (2021).

[9] Lago-Peñas, Carlos, Lago-Ballesteros, Joaquin, Dellal, Alexandre, and Gómez, Maite. “Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league”. In: *Journal of sports science & medicine* 9.2 (2010), p. 288.

[10] Marin, Mike. *Logistic Regression: Checking Linearity*. [Accessed: May 3, 2023]. Youtube. URL: https://www.youtube.com/watch?v=yYffUVTEPl4&t=283s.

[11] McKinley, Eliot. “Game of throw-ins”. In: *American soccer analysis* (2018). [Accessed: May 23, 2023]. URL: https://www.americansocceranalysis.com/home/2018/11/27/game-of-throw-ins.

[12] Midi,
Habshah, Sarkar, Saroje Kumar, and Rana, Sohel. “Collinearity diagnostics of binary logistic regression model”. In: *Journal of interdisciplinary mathematics* 13.3 (2010), pp. 253–267.

[13] Montgomery, Douglas C, Peck, Elizabeth A, and Vining, G Geoffrey. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[14] O’brien, Robert M. “A caution regarding rules of thumb for variance inflation factors”. In: *Quality & quantity* 41 (2007), pp. 673–690.

[15] Perlich, Claudia, Provost, Foster, and Simonoff, Jeffrey. “Tree induction vs. logistic regression: A learning-curve analysis”. In: (2003).

[16] Scikit-learn. *Gradient Boosting for classification*. [Accessed: May 10, 2023]. n.d. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html.

[17] Scikit-learn. *GridSearch*. [Accessed: May 10, 2023]. n.d. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[18] Senaviratna, NAMR, Cooray, TMJA, et al. “Diagnosing multicollinearity of logistic regression model”. In: *Asian Journal of Probability and Statistics* 5.2 (2019), pp. 1–9.

[19] Shaw, Laurie and Gopaladesikan, Sudarshan. “Routine Inspection: A Playbook for Corner Kicks”. In: *Machine Learning and Data Mining for Sports Analytics*. Ed. by Ulf Brefeld, Jesse Davis, Jan Van Haaren, and Albrecht Zimmermann. Cham: Springer International Publishing, 2020, pp. 3–16. ISBN: 978-3-030-64912-8.

[20] Stone, Joseph Antony, Smith, Adam, and Barry, Anthony. “The undervalued set piece: Analysis of soccer throw-ins during the English Premier League 2018–2019 season”. In: *International Journal of Sports Science & Coaching* 16.3 (2021), pp. 830–839.

[21] Sumpter, David et al. *Introducing expected goals*. [Accessed: May 4, 2023]. 2022. URL: https://soccermatics.readthedocs.io/en/latest/lesson2/introducingExpectedGoals.html.

[22] The Football Association. *Law 15 - The Throw-In*. [Accessed: March 6, 2023]. n.d. URL: https://www.thefa.com/football-rules-governance/lawsandrules/laws/football-11-11/law-15---the-throw-in.

[23] Wyscout. *Ball possession*. [Accessed: May 4, 2023]. n.d. URL: https://dataglossary.wyscout.com/ball_possession/.

[24] Wyscout. *Pitch coordinates*. [Accessed: May 4, 2023]. n.d. URL: https://dataglossary.wyscout.com/pitch_coordinates/.

[25] Wyscout. *xG*. [Accessed: May 4, 2023]. n.d. URL: https://dataglossary.wyscout.com/xg/.

[26] Yiannakos, A and Armatas, V. “Evaluation of the goal scoring patterns in European Championship in Portugal 2004.” In: *International Journal of Performance Analysis in Sport* 6.1 (2006), pp. 178–188.

[27] Zhang, Lili, Geisler, Trent, Ray, Herman, and Xie, Ying. “Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function”. In: *Journal of Applied Statistics* 49.13 (2022), pp. 3257–3277.

Appendix - Contents

Appendix A

Feature definitions

Appendix A presents a description of the features which have not been explicitly described in the Method section of the report.

Table A.0.1: Definition of investigated features, which were not described in the Method section

Feature	**Description**
angle_cubed & angle_squared	angle α to the power of 3 and 2 respectively, i.e. α^3 and α^2.
angle_exp & angle_exp_0_2 & angle_exp_0_3	exponential transformation of the angle α with the base e, 0.2 and 0.3 respectively, i.e. e^{α}, 0.2^{α} and 0.3^{α}.
angle_length	multiplication of angle α and the length of the throw-in.
angle_length_squared	angle_length to the power of two, i.e. $(angle\_length)^2$.
angle_start_x	multiplication of angle α and the x position of the location from which the throw-in is taken.
angle_throw_goal_cubed & angle_throw_goal_squared	angle β to the power of 3 and 2 respectively, i.e. β^3 and β^2.
distance_to_goal_diff_exp & distance_to_goal_diff_exp_1_1 & distance_to_goal_diff_exp_1_2	exponential transformation of the feature $distance\_to\_goal\_diff$ with the base e, 1.1, and 1.2 respectively, i.e. $e^{distance\_to\_goal\_diff}$, $1.1^{distance\_to\_goal\_diff}$ and $1.2^{distance\_to\_goal\_diff}$.
distance_to_goal_log	natural logarithm of the variable $distance\_to\_goal$, i.e. $\ln(distance\_to\_goal)$.
length_start_x	multiplication of the length of the throw-in and the start x position of the throw-in.
match_minute	The match minute when the throw-in was taken.
start_x_cubed & start_x_squared	the variable $start\_x\_adj$ to the power of 3 and 2 respectively, i.e. $start\_x\_adj^3$ and $start\_x\_adj^2$.
x_diff_exp_0_9 & x_diff_exp_1_1	exponential transformation of the feature x_diff with the base 0.9 and 1.1 respectively, i.e. 0.9^{x_diff} and 1.1^{x_diff}.

Appendix B

Logit plots

Appendix B presents the logit plots of features which were not included in the Results section, first for the possession retention model, and then for the goal chance creation model.

B.0.1 Logit plots for possession retention

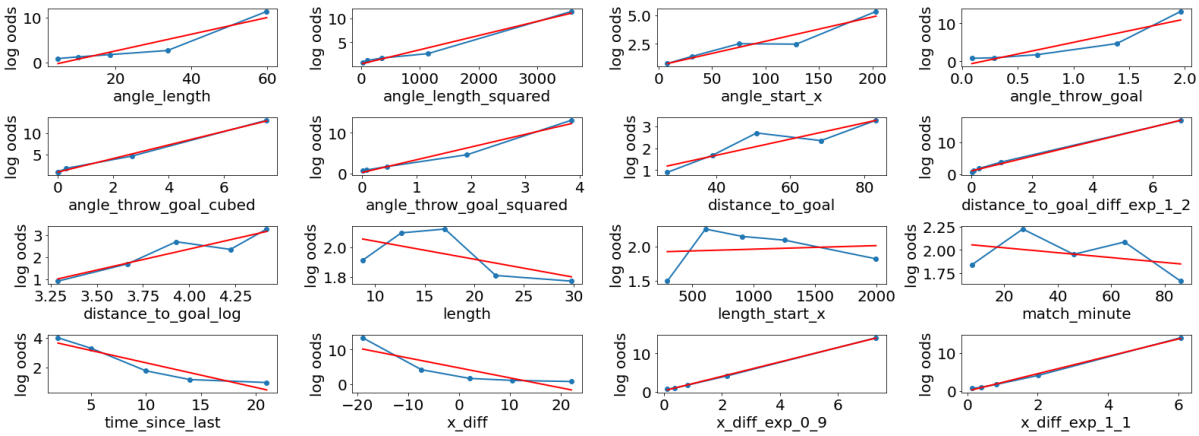


Figure B.0.1: Logit plots for features considered in the possession retention model, which are not included in the Results section of the report. Note that the scale is not the same for the different plots.

B.0.2 Logit plots for goal chance creation

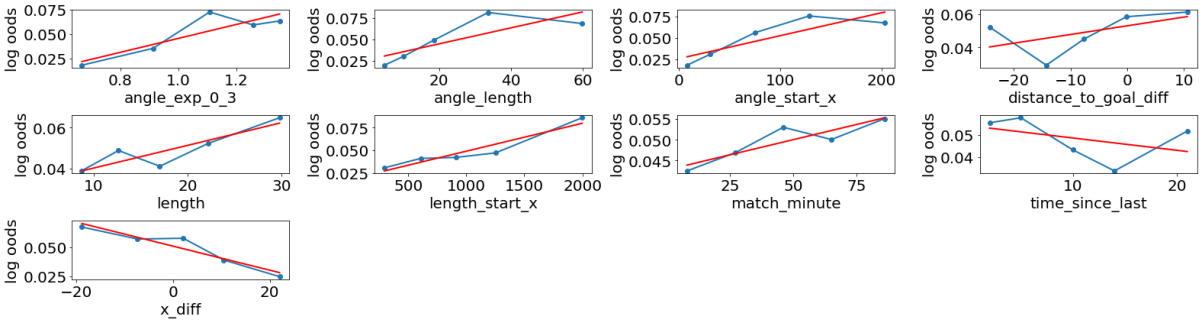


Figure B.0.2: Logit plots for features considered in the goal chance creation model, which are not included in the Results section. Note that the scale is not the same for the different plots.

Appendix C

Feature selection

Appendix C presents the AIC after introducing various feature transformations, which were investigated when determining the transformed models in the project.

C.0.1 Feature transformations for possession retention model

Basic Model Transformed

Table C.0.1: Choosing features for *Basic Model Transformed* for the possession retention model. The model fit is evaluated using AIC.

Features	**AIC**
start_x_adj, angle, length, time_since_last	10875
distance_to_middle, angle, length, time_since_last	10866
distance_to_middle, angle_squared, length, time_since_last	10868
distance_to_middle, angle_cubed, length, time_since_last	10954
distance_to_middle, angle_exp, length, time_since_last	10916

Full Model Transformed

Table C.0.2: Choosing features for *Full Model Transformed* for the possession retention model. The model fit is evaluated using AIC.

Features	**AIC**
start_x_adj, angle, length, time_since_last, distance_to_goal, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x	10429
distance_to_to_middle, angle, length, time_since_last, distance_to_goal, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x	10429
distance_to_to_middle, angle_squared, length, time_since_last, distance_to_goal, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x	10434
distance_to_to_middle, angle_cubed, length, time_since_last, distance_to_goal, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x	10428
distance_to_to_middle, angle_cubed, length, time_since_last, distance_to_goal_log, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x	10423
distance_to_to_middle, angle_cubed, length, time_since_last, distance_to_goal_log, distance_to_goal_diff_exp, x_diff, angle_start_x, angle_length, length_start_x	10529
distance_to_to_middle, angle_cubed, length, time_since_last, distance_to_goal_log, distance_to_goal_diff_exp_1_1, x_diff, angle_start_x, angle_length, length_start_x	10540
distance_to_to_middle, angle_cubed, length, time_since_last, distance_to_goal_log, distance_to_goal_diff, x_diff_exp_0_9, angle_start_x, angle_length, length_start_x	10439
distance_to_to_middle, angle_cubed, length, time_since_last, distance_to_goal_log, distance_to_goal_diff, x_diff_0_9, angle_start_x, angle_length_squared, length_start_x	10455

C.0.2 Feature selection for goal chance creation model

Table C.0.3: Choosing features for *Basic Model Transformed* for the goal chance creation model. The model fit is evaluated using AIC.

Features	**AIC**
start_x_adj, angle, length, time_since_last	3640
start_x_adj_squared, angle, length, time_since_last	3636
start_x_adj_cubed, angle, length, time_since_last	3636
start_x_adj_squared, angle_exp_0_2, length, time_since_last	3621

Table C.0.4: Choosing features for *Full Model Transformed* for the goal chance creation model. The model fit is evaluated using AIC.

Features	**AIC**
start_x_adj, angle, length, distance_to_goal, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x	3601
start_x_adj_squared, angle, length, distance_to_goal, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x	3601
start_x_adj_cubed, angle, length, distance_to_goal, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x	3601
start_x_adj_squared, angle, length, distance_to_goal_log, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x	3599
start_x_adj_squared, angle_exp_0_2, length, distance_to_goal_log, distance_to_goal_diff, x_diff, angle_start_x, angle_length, length_start_x	3597