Forecasting Football Corner Odds

Statistical Modelling, Betting Strategies and Assessing Market Efficiency

Marcus Laurens, Gustav Pålsson

Master's thesis
2023:E33

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

Abstract

Statistical modelling could be included in a betting strategy where the value of a bet is assessed by comparing model predictions and market odds. This thesis presents several models based on statistical learning methods for predicting the total number of corners in a football match. Generalised linear regression and decision tree models were developed and their profitability was examined by using historical odds data. The models were trained and tested on recent seasons of the English Premier League. To further test the predictive strength, the models were tested on the German Bundesliga. Since the number of corners in a football match is count data but exhibits overdispersion, negative binomial regression was used to numerically model the number of corners. This approach was accompanied by logistic regression as well as numerical-based and classification-based random forest models. The number of corners could be seen as a classification variable with the classes defined as above or below a certain number of corners, often referred to as the betting line on the over-under odds market.

The explanatory variables used to develop the models were match-by-match statistics from the Premier League, processed by creating averages of different lengths and supplemented by variables representing team capabilities and self-created variables representing current form and motivation. Backward stepwise selection and elastic net were used to select variables to include in the generalised linear regression models. The combinations of model approaches and methods resulted in fifteen possible models, which were assessed using statistical evaluation measures. Level stakes and the Kelly criterion were applied as betting strategies on the best-performing models for each method. Furthermore, the over-under betting market for corners was examined in order to identify potential asymmetries in the offered odds implying an inefficient market.

The results indicated that the best-performing models from each method were all profitable when tested on new data from the Premier League, despite having a low degree of explanatory power. On the contrary, the explanatory power and profitability decreased significantly when the Premier League-based models were tested on the Bundesliga without retraining leading to the majority of the models turning unprofitable. The analysis of the over-under market suggested that the under odds offered in the Premier League matches were generally undervalued, while this undervaluation was not statistically significant for the Bundesliga.

Acknowledgements

We would like to express our deepest gratitude and appreciation to our supervisor Johan Lindström for his invaluable guidance, support and expertise throughout our master's thesis journey. His commitment, knowledge and ability to provide insightful feedback and constructive criticism have been instrumental in refining our research methodology and strengthening our theoretical framework.

Contents

1 Introduction

1.1 Association Football

Association football (hereafter football) is the world's most popular sport both in terms of practitioners and supporters (Rollin et al., 2023). According to the international governing body of football, Fédération Internationale De Football Association (FIFA), 130 000 professional football players are registered in the world (FIFA, 2019). Furthermore, it is estimated that over five billion people support a club or a national team (FIFA, 2021). The worldwide appeal of football has led to a dominant position in relation to other sports.

Football is a fast-paced sport and the outcome of a match can change quickly. Set pieces occur after a stoppage in the game due to a foul, misconduct or if the ball is out of play and is a way to return the ball into play (Rollin et al., 2023). A set piece could be a free kick, a throw-in, a goal kick or a corner. These moments could be of strategic importance, providing an opportunity for the players of a team to coordinate and execute practised tactics. Set pieces have the potential to be a significant goal-scoring opportunity and create a moment of anticipation for both players and fans.

1.2 Sports Betting

By definition, gambling or betting is the act of risking money or any other staked object in the hope of correctly predicting the outcome of an event and making a profit (Glimne, 2023). Sports betting, where a bettor places a wager on an outcome connected to a sports event, has increased vastly in popularity during the last decades (Market Decipher, 2022). The global sports betting market is steadily growing and accounted for USD 85 billion in 2022 and is expected to grow further in the foreseeable future. Betting companies today offer a wide range of betting opportunities for their customers. For example, a money line bet involves the prediction of the outcome of a match, whereas over-under bets involve the prediction of a certain statistic in a match to be over or under a given number (Mollenkamp, 2022; Webber, 2022). This given number will hereafter be referred to as ”the line”. The line is often stated as an integer added by 0.5 to ensure that the outcome strictly will end up over or under. A consequence of the large amount of accessible information and statistics related to sports events, and in particular football, is that data analysis and prediction of matches have increased in use for betting purposes.

1.3 Purpose

The purpose of this thesis is to investigate whether it is possible to predict the aggregated number of corners that occur in a football match based on publicly available match statistics. It will be examined if the corner betting market shows signs of inefficiency and if asymmetry in the pricing could be found. Furthermore, the possibility of creating a profitable strategy based on the models and betting theories will be investigated. The disposition of the thesis will be as follows. Background and related work will be covered in chapter 2, data in chapter 3, theory in chapter 4, method in chapter 5, the result and analysis are presented in chapter 6 and conclusions and future work are found in chapter 7.

1.4 Problem Statement

This thesis aims to create data-driven prediction models based on match-by-match statistics. The models will predict the total number of corners or if the number of corners will be over or under a predetermined line with a certain probability. The models will be trained, validated and tested on matches from the English Football Association Premier League Limited (hereafter Premier League) over eight seasons. To further test the models' predictive ability, they will be used for the same purpose in the German top tier 1. Bundesliga (hereafter Bundesliga). In order to evaluate the profitability of the prediction models, the predicted probabilities will be compared against historical market odds for a certain line on the over-under market for corners.

1.5 Delimitations

To provide a more narrow and focused study, several delimitations have been considered. Firstly, the thesis will only consider corners and no other set pieces situations or possible match outcomes. This is motivated by the fact that the over-under market for corners is considered to be smaller compared to the money line market, but still widely offered, and may therefore be more likely to show signs of inefficiency. Secondly, the study is geographically delimited to develop the prediction models on football matches in England and tested on matches in England and Germany. More precisely, the study focuses solely on the top men's leagues in each country. Finally, the prediction models are exclusively based on publicly available match statistics from previous matches and do not include pre-match market-affecting changes close to kick-off. No subjectivity in the probability assessment of a certain outcome is included, all predictions are based on the mathematical models applied.

2 Background and Related Work

2.1 Concepts of Football

2.1.1 Corner Kicks in Football

A corner kick is a type of set piece situation in football which is awarded to the attacking team when the ball has crossed the defending team's goal line, outside the goal posts or over the crossbar, and is last touched by a player in the defending team (The International Football Association Board, 2022). The game is restarted by placing the ball at the corner of the football pitch closest to where the ball went out of play. The corner kick is performed by one of the players in the attacking team who undisturbed kicks the ball from the designated corner area, generally, towards the defending team's penalty area where players of both teams are positioned.

A corner kick is seen as a potential goal-scoring situation for the attacking team. Since the ball is out of play and restarted from the corner, both the attacking and defending team are able to organise themselves tactically before the corner kick is taken. Thus, some teams may be particularly successful in scoring goals from corner kicks. In an analysis of 18 seasons in the Premier League, 9-14% of the goals per season were scored through a corner kick (Clarke, 2022). Because a corner kick may result in a goal for the attacking team, it can have a significant impact on the dynamic and outcome of the match. The number of corners in a football match could vary significantly from match to match. For example, in the match between Burnley and West Ham United in the season 2021/2022 the aggregated number of corners was 20 (FootyStats, 2023). On the other hand, in the match between Newcastle United and Chelsea in the same season, the total number of corners amounted to two.

2.1.2 Top Leagues in Europe

The Premier League is the top tier of the English football league system for men (Premier League, 2023b). Founded in 1992, the Premier League has become the most viewed football league in the world (Premier League, 2023a). The league is contested by 20 clubs each season in accordance with a promotion and relegation system. The table is determined based on a point system where a team is awarded three points for a win, one point for a draw and no points for a loss (UEFA, 2022). According to the country coefficient ranking system of the governing body of European football, the Union of European Football Associations (UEFA), the Premier League is 2023 regarded as the highest-ranked and most prominent league in Europe followed by the Spanish Liga Nacional de Fútbol Profesional (commonly referred to as LaLiga), the German Bundesliga and the Italian Serie A (UEFA, 2023). Just as the Premier League, LaLiga and Serie A are contested by 20 clubs while the Bundesliga is contested by 18 clubs. In relation to each other, the Premier League is exceptional regarding broadcasting and sponsorship deals making the league the most well-known league globally (Ajadi et al., 2022). Furthermore, the UEFA ranking system also affects the opportunity for the clubs to qualify for European cup competitions, such as the UEFA Champions League (UEFA, 2022). For top-ranked European leagues, there may be as many as six clubs that could qualify for the group stages of the UEFA tournaments, of which the top four automatically qualify for the UEFA Champions League the following season.

A common belief in football is that there exist distinctive differences between the top leagues in Europe in terms of playing style and physicality. A study by Oberstone (2011) aimed to quantitatively map the key differences in the major European football leagues. The author was able to empirically claim that there exist disparities concerning passing, goal scoring and tactical organisation between leagues. For example, the Premier League is particularly known for its fast pacing and physical style of playing, whereas LaLiga is more technically and attacking-oriented. Serie A is according to the study the most possession-oriented league. In a similar and more recent study made by Yi et al. (2019), the results confirmed that there exist differences between the European top leagues with respect to technical aspects, but the results differed somewhat from Oberstone's findings. The differences between the Premier League and LaLiga were not as apparent and the authors concluded that players in Serie A performed worse in passing and tactical organisation in comparison to the other leagues. To further exemplify the differences, the goals per match in Bundesliga in season 2021/2022 were 3.12, while Serie A amounted to 2.87, Premier League to 2.82 and LaLiga to 2.50 (Bundesliga, 2022).

Table 2.1 presents some key metrics of corner kicks in the four leagues during the seasons 2015/2016 to 2021/2022. The variance-to-mean ratio (VMR) represents the variance divided by the mean, where a value above 1 indicates overdispersion. Overall, these numbers confirm that there may exist some differences between the leagues worth considering.

2.2 Bookmakers and Odds

Betting companies, or so-called bookmakers, offer a large variety of bets on all kinds of events (Glimne, 2023). Traditionally, competitive sports events such as horse racing and football matches have attracted betting but bookmakers may also offer bets on political elections and other non-sport-related contests. Unlike other gambling forms, such as lottery or roulette which completely rely

Table 2.1: Statistics of corners in the European top leagues during the seasons 2015/2016 to 2021/2022.

League	**Corner mean**	**Corner variance**	**VMR**
Premier League	10.4	12.0	1.15
Bundesliga	10.3	12.7	1.23
Serie A	9.6	11.3	1.18
LaLiga	9.5	10.7	1.13

on chance, sports betting has a significant element of subjective evaluation. Common for all types of betting is that the probability of a certain outcome is inversely expressed as odds. A high probability relates to low odds. In case of a favourable outcome for the bettor the profit is determined by the odds, otherwise, the betting company retain the wager and the payoff for the bettor is zero. The odds reflect the bookmaker's probability estimation for a certain outcome and can be expressed in various formats (Sohail, 2023). The decimal format is most established in European countries and is sometimes also referred to as "European odds". A decimal odds reflects the total payout of a bet, and the total payoff is calculated as the product of the wager and the odds. Other types of odds formats such as fractional odds and money line odds also occur.

In order to ensure a sustainable profit margin towards the bettors, the offered odds' implied probability differs from the actual probabilities determined by the bookmakers. Hence the sum of the probabilities for related outcomes exceeds 1. This is known as the betting margin. As a consequence, the bettor's profit is less than what the actual probability would generate.

Odds are usually offered to the market several days before the event. These odds constitute the pre-match market. When the match starts, the live-betting market is launched which allows the bettor to place bets with respect to the real-time dynamics of a game. Since the pre-match odds are offered well in advance of the match, probability-affecting circumstances may arise which in turn affect the odds (Smarkets, 2023). An example could be unforeseen changes in a team's starting lineup. The pre-match odds may also fluctuate from the initial odds offered depending on the flow of stakes and market competitors' odds.

2.2.1 Betting Exchanges and Arbitrage

An alternative to placing bets on odds offered by betting companies is to use a betting exchange. In traditional betting, the bookmaker acts as a counterpart and bets that the outcome the bettor places its wager on will not occur. Conversely, a betting exchange enables bettors to both buy (known as back) and sell (known as lay) an outcome (Betfair, 2023b). In that way, the odds offered on the exchange are backed by another bettor who believes in the opposite outcome and hence holds the liability. The range of offered betting alternatives on the betting exchange is hence not limited to the bookmaker's selection. Another difference compared to bookmakers is that the exchange's odds do not include the margin. Instead, the trading venue charges a commission on the net winning (Betfair, 2023a). A prerequisite for the betting exchange to be an attractive alternative to betting companies is that there exists sufficiently high liquidity. The leading and most liquid betting exchange is Betfair which has operated on the betting market since 2000 (Betfair, 2017).

An important aspect when considering several types of betting platforms is arbitrage betting, which involves taking advantage of differences in offered odds on a certain outcome generating a risk-free profit. The differences could occur because of differing expectations of the probability of an outcome, either among bookmakers or among actors on the betting exchange. However, bookmakers actively take precautions in order to avoid arbitrageurs taking advantage of such opportunities, for example by stake limitations or closing the bettor's account (Bet Types, 2023).

2.3 Betting Strategies

In order for a bettor to consider that a bet may be profitable, the offered odds should exceed the odds implied by the predicted probability, either based on subjectivity or quantified analysis. In that case, the bettor could argue that there exists a value in the bet. If the discrepancy between the offered odds and the odds implied by the predicted probability is large and positive, the expected value of the bet becomes large.

The majority of bettors gamble mainly for entertainment purposes (Hultåker, 2022). Regardless of betting for pleasure or with the aim of long-term profit, different betting strategies can be used. A simple betting strategy is to evaluate if there exists a value in the bet and then place a predetermined wager on the outcome (Wheatcroft, 2020). The wager is constant regardless of the betting value. Another, more sophisticated yet simple, betting strategy is to adjust the wager depending on the value. Introduced by Kelly (1956), the strategy has had a significant impact both on betting and within investing strategies on the financial market. The strategy relies on a formula which determines the proportion of the bettor's capital that should be placed on the bet. A large value in the bet relates to a larger wager.

2.4 Related Work

A large amount of data and statistics related to sports events is available today, particularly in football. This has led to a growing interest in prediction and quantitative sports analysis in recent years (Fathima et al., 2018). It is not only applicable for professional analysts hired by clubs to improve performance and gain tactical advantages against competitors, or to use for betting decision-making, but has also gained interest as a research area for statisticians.

Moroney (1956) examined which distribution fits the number of goals scored in a football match best and favoured the negative binomial distribution, whereas Maher (1982) demonstrated that a bivariate Poisson distribution, including the teams' defence and attacking capabilities, accurately describe football goals. Dixon and Coles (1997) continued Maher's findings, but also considered home team advantage and let recent performances have larger weight in their model. Their model was used for betting purposes, and bets were placed on all outcomes where the model predicted a higher probability than the market. The betting strategy turned out to yield a positive return.

Yip et al. (2022) changed the perspective and tried to model corners with respect to overdispersion and the phenomenon of corner clustering. Once again, the most well-fitting distribution was investigated. The authors claimed that the aggregated number of corners was best explained by a negative binomial distribution or a geometrical Poisson distribution. The modelling consisted of regression analysis and led to a profitable betting model. A moving average of the number of corners as well as shots on goal by the home team and away team in the last three matches were included as explanatory variables. In addition, a slightly adjusted mean corner count, the expected total number of goals in the match and the goal supremacy of the home team, derived from a double-independent Poisson model were also considered in the regression model.

Furthermore, machine learning methods have been used to forecast the likelihood of match events. Baboota and Kaur (2018) developed a predictive model for match results in the Premier League. Both random forests and gradient boosting were examined and performed roughly equivalently. The prediction accuracy of the machine learning models was benchmarked against market-leading betting companies, with the conclusion that the methods slightly underperformed the market. Similarly, Alfredo and Isa (2019) used machine learning algorithms to predict the results of Premier League matches. In the analysis, the random forest model outperformed extreme gradient boosting and C5.0, but the authors concluded that their approach of using decision tree-based methods was unable to accurately predict match outcomes.

There exists several studies that examine signs of inefficiency in the over-under market. In addition to creating a prediction model, Yip et al. examined asymmetry in the over-under market in favour of betting on the under selection. Their findings indicated that the market may overvalue the over selection, in other words overestimating the probability of an outcome being over the line and hence systematically increasing the odds for the under selection. Another well-documented phenomenon linked to market inefficiency is the favourite-longshot bias, where the market undervalues high probability events (Constantinou and Fenton, 2013). There are several theories for why this bias occurs, such as some bettors having more risk-loving behaviour and thus affecting the odds. Both the asymmetry in the over-under market in favour of the under selection and the favourite-longshot bias imply that the market occasionally is mispricing betting options which leads to inefficiency in the betting market.

3 Data

3.1 Match Statistics

Match-by-match statistics were collected from the football statistics provider FootyStats. The downloaded statistics included all relevant information about the played matches including date, home team name, away team name, goals in the first and second half, corners, free kicks, possession and shots on and off target for each team. The data set also included pre-match odds for the home-draw-away market for each match. The statistics were collected for all the Premier League seasons from 2013/2014 to 2022/2023. The season 2020/2021 was assumed to be heavily affected by restrictions during the COVID-19 pandemic and not to be completely representative of seasons played under normal circumstances, hence it was excluded from modelling and testing. Matches played after the 13th of March during the 2019/2020 season were excluded for the same reason. For the season 2022/2023, all matches until the turn of the year were included. The data from the seasons 2014/2015 to 2019/2020 were used for modelling and the data from the seasons 2021/2022 to 2022/2023 were used for testing. The season 2013/2014 and 2020/2021 were used to obtain initial values for the averages for seasons 2014/2015 and 2021/2022, respectively. Match data of the corresponding test seasons were collected from the Bundesliga. Table 3.1 presents the minimum, median and maximum values of the variables in the model data set.

3.2 FIFA Ratings

As a measure of each team's capabilities, the FIFA index was used. The FIFA index is a player database used in the video games series FIFA, developed by EA Sports, based on actual players (Britannica, 2023). Ratings for each team and season were collected. Defence, midfield and attacking statistics based on the team roster's skills were weighted together to create an overall rating of a team's capabilities. A new edition of FIFA is released every year, usually in September. FIFA also updates the ratings continuously during the year, approximately twice a month.

The major European leagues have two transfer windows, a summer transfer window and a winter transfer window. A transfer window is a determined

Table 3.1: Summary of match data for model data seasons. The home team's corners (goals) were interpreted as the away team's conceded corners (conceded goals) and vice versa. These statistics will be identical when considering all matches, but not identical when calculating variable averages for a specific team.

Variable Name	**Min**	**Median**	**Max**
Home team corners	0	5	19
Away team corners	0	4	16
Home team goals	0	1	9
Away team goals	0	1	9
Home team goals first half	0	0	5
Away team goals first half	0	0	5
Home team shots on target	0	5	19
Away team shots on target	0	4	17
Home team shots off target	0	6	25
Away team shots off target	0	5	21
Home team fouls	1	10	23
Away team fouls	1	11	26
Home team possession	18	51	82
Away team possession	18	49	82
Home team points per game	0	1	3
Away team points per game	0	1	3
Game week	1	19	38
Odds full time home team win	1.05	2.26	23.00
Odds full time draw	2.73	3.70	20.50
Odds full time away team win	1.10	3.41	42.75

period of time which allows the clubs to trade and register new players on their team roster. In England, the summer window is open from the 10th of June to the 1st of September, while the winter transfer window is open from the 1st of January to the 31st of January (Transfermarkt, 2023). In Germany, the winter transfer window is open during the same dates, while the summer transfer window is open from the 1st of July to the 1st of September. To account for changes in the teams' capabilities, FIFA ratings were collected for each team just before each season starts and updated after the winter transfer window closed. Table 3.2 presents the minimum, median and maximum FIFA ratings of the Premier League clubs contesting the seasons used for modelling.

3.3 Historical Odds

It could be assumed that the betting market is arbitrage-free and that the odds from a betting exchange are representative of the betting market as a whole. Hence, historical odds from Betfair were downloaded in order to perform a profitability analysis of the prediction models. Odds for over and under 10.5

Table 3.2: Summary of FIFA ratings for the teams in the Premier League seasons used for modelling.

Rating Name	**Min**	**Median**	**Max**
Attack	66	78	89
Midfield	66	77	88
Defence	65	76	85
Overall	66	77	86

corners were collected for each match in the Premier League and the Bundesliga, ranging from the season 2018/2019 to 2022/2023. Since this thesis aims to further examine the tendency of asymmetries in favour of the under selection on the over-under market, more odds data than covered by the matches in the test set was downloaded.

The unprocessed odds data showed how the odds changed until kick-off. Because bets can be placed long before the start of the match and since they can be affected by new information reaching the market, the last over and under odds placed before kick-off were chosen. The odds data set contained matches where no bets had been placed on the betting exchange, hence no historical odds were available. If there existed betting volume on only one of the options or no betting volume at all, the match was treated as a missing value. The odds data was connected to the match data by team names and match dates. Some odds were not able to be connected to a match due to conflicts in the match date which could be a consequence of match rescheduling. These matches were filtered out. For the Premier League, 14.8% of the matches were excluded and for the Bundesliga 35.8% of the matches were excluded.

4 Theory

Previous studies show that predicting the outcomes of a match is difficult. It could be assumed that there is both a statistical pattern and a randomness in how many corners occur in a football match. Hence, statistical learning methods could be applied. Generalised linear regression and tree-based methods are used to investigate which football statistical variables affect the aggregated number of corners. To approach the problem, statistical learning models are developed and combined with betting strategies to, hopefully, create profitability.

4.1 Generalised Linear Regression

There are several different statistical learning methods which could be classified as generalised linear regression methods (James et al., 2013). Depending on which distribution within the exponential family the dependent variable is assumed to follow, a particular generalised linear regression method will be preferred. Generalised linear regression is characterised by a link function g which creates a mapping between the explanatory variables and the conditional expectation of Y given X,

$$g(E(Y|X)) = X\beta, \tag{4.1}$$

or reformulated as

$$E(Y|X) = \mu = g^{-1}(X\beta). \tag{4.2}$$

4.1.1 Poisson Regression

Since the number of corners in a match always is a positive integer or zero, the dependent variable Y should be treated as count data. Therefore, Poisson regression could be an adequate method to fit a model,

$$Y \sim Po(\lambda). \tag{4.3}$$

The number of corners then is assumed to follow a Poisson distribution as

$$Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \tag{4.4}$$

where $\lambda = E(Y) = Var(Y) > 0$ (James et al., 2013). The mean is then written as a function of the explanatory variables as

$$E(Y|X) = \mu = \lambda(X) = e^{X\beta}, \tag{4.5}$$

hence Poisson regression uses the following link function

$$g(\mu) = \ln \mu. \tag{4.6}$$

The β-coefficients are estimated by maximising the likelihood function

$$l(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}, \tag{4.7}$$

resulting in parameter estimate β^0 and approximate covariance matrix Σ for the estimated parameters.

4.1.2 Negative Binomial Regression

However, by studying table 2.1 the variability in the number of corners is greater than the mean, thus

$$\frac{Var(Y)}{E(Y)} > 1. \tag{4.8}$$

The observed data hence shows signs of overdispersion, a finding that is consistent across the leagues. Poisson regression strictly assumes that $E(Y) = Var(Y)$. For that reason, negative binomial regression could instead be used since it is especially suitable when handling overdispersed count data,

$$Y \sim NB(\lambda, \theta), \tag{4.9}$$

where θ represents the dispersion parameter allowing for a higher variability than what is assumed in the Poisson distribution (James et al., 2013). Negative binomial regression uses the same logarithmic link used in Poisson regression.

The number of corners is then assumed to follow a negative binomial distribution

$$Pr(Y = k) = \frac{\Gamma(k+\theta)}{\Gamma(k+1)\Gamma(\theta)} \left(\frac{\theta}{\theta+\lambda}\right)^{\theta} \left(\frac{\lambda}{\theta+\lambda}\right)^{k}, \tag{4.10}$$

according to a common parameterisation with mean E(Y) = λ and variance Var(Y) = $\lambda + \lambda^2\theta^{-1}$ (Ismail and Jemain, 2007). $\Gamma(\cdot)$ is the gamma function.

In correspondence to eq. (4.7), the regression coefficients are obtained with maximum likelihood estimation as

$$l(\beta) = \prod_{i=1}^{n} \frac{\Gamma(y_i+\theta)}{\Gamma(\theta)y_i!} \left(\frac{\theta}{\theta+\lambda(x_i)}\right)^{\theta} \left(\frac{\lambda(x_i)}{\theta+\lambda(x_i)}\right)^{y_i}, \tag{4.11}$$

resulting in parameter estimate β^0 and approximate covariance matrix Σ for the estimated parameters.

Since the output of the Poisson or negative binomial regression are predictions of the number of corners, it has to be processed further to be interpreted as a probability that the total number of corners will be above the line and to use the results in a betting strategy.

4.1.3 Logistic Regression

Logistic regression is used when modelling the probability that a certain outcome is either true or false (James et al., 2013). The dependent variable is then assumed to follow a Bernoulli distribution and hence may only take a value of either 0 or 1,

$$Y \sim Be(p). \tag{4.12}$$

On the other hand, the explanatory variables are allowed to be continuous or binary. Instead of fitting a linear line to the data, an S-shaped link function which goes from 0 to 1 is used for the fitting.

The link used in logistic regression, called logit-link, has the form

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) \tag{4.13}$$

(Nelder and Wedderburn, 1972). Accordingly, the conditional expectation of Y given X is

$$E(Y|X) = Pr(Y = 1|X) = \mu = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{4.14}$$

and the coefficients are estimated by maximising the likelihood function

$$l(\beta) = \prod_{i:y_i=1} p_i(x_i) \prod_{i:y_i=0} (1 - p_i(x_i)) \tag{4.15}$$

resulting in parameter estimate β^0 and approximate covariance matrix Σ for the estimated parameters (James et al., 2013).

Applied to predicting corners, the classification variable considers if the number of corners is greater or smaller than the line. This can be intuitively linked to the offers that are available on the over-under betting market for a certain match statistic. Unlike negative binomial regression, a probability prediction if the number of corners will be above or below the line will be directly obtained and the output does not need further processing.

A potential drawback of using logistic regression is the loss of information within the classes when classifying the outcomes. The model then does not count for the variability of the outcomes as it does when using numerical variables. Although, this could also be seen as a potential advantage as outliers do not affect the model to the same extent.

4.2 Variable Selection

The number of explanatory variables is large and many of the variables are correlated. To evaluate which variables should be included in the model, some form of variable reduction has to be applied in order to obtain a reduced data set of manageable size.

4.2.1 Lasso Regression, Ridge Regression and Elastic Net

One option to reduce the number of variables is to use a so-called shrinkage method (James et al., 2013). In this concept, there are mainly two techniques, lasso regression and ridge regression, which both penalise variables that are weakly informative by forcing their coefficients towards zero. An important difference between lasso regression and ridge regression is that the former will set weak predictors to exactly zero while ridge regression will shrink the coefficient close to zero but never exactly to zero.

In ridge regression, a penalising term with a tuning parameter $\lambda \geq 0$ is added to the maximum likelihood expression. The penalty is only applied to the β-coefficients that relate to an explanatory variable and not the intercept, according to

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} L(y_i, X\beta) + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{4.16}$$

where L is the negative log-likelihood contribution of observation i, n is the number of observations and p is the number of coefficients.

Depending on the value of λ, a set of coefficients will be generated. A higher value of λ will lead to a decreased prediction variance, but a more biased estimation. Ridge regression is particularly favourable as a variable selection method when the explanatory variables exhibit a high level of multicollinearity (NCSS, 2023).

Lasso regression is similar to ridge regression, but is stricter in the penalisation of variables. Because lasso regression can set non-informative variables to zero, it increases the interpretability of the model. The expression to be minimised is largely consistent with ridge regression but instead uses the sum of the coefficient's absolute values instead of the squared values (James et al., 2013)

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} L(y_i, X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{4.17}$$

There is no definitive answer as to which of the shrinkage methods should be preferred. Both ridge regression and lasso regression lead to a reduction in the variance of the coefficients at the expense of increased bias, even though the reduction tends to be slightly larger for ridge regression.

Instead of choosing between the two options it is possible to combine them. This method is called elastic net, which weighs the trade-off between high model interpretability and increased prediction performance. The minimisation problem solved in the elastic net can be expressed as the combination of eq. (4.16) and eq. (4.17) with some reformulations. A parameter α is introduced which controls the penalisation of the respective shrinkage method, while λ sets the general penalty strength (Hastie et al., 2023). The optimisation problem can be formulated as

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} L(y_i, X\beta) + \lambda \left[\frac{(1-\alpha)}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right]. \tag{4.18}$$

Since a new set of coefficients arises generated for each value of λ, cross-validation can be used to choose an appropriate model. Two adequate ways to do this are either to find the λ-value that gives the lowest mean cross-validated error (λ_{min}) or the λ-value which yields the most shrunken model and where the cross-validated error is still within one standard error from the minimum (λ_{1se}).

4.2.2 Stepwise Selection

An alternative to shrinkage methods is to use stepwise selection to reduce the number of variables (James et al., 2013). Forward stepwise selection starts with a model with no explanatory variables and successively adds new variables to the model. Backward stepwise selection starts in the other end and iteratively removes the least informative variable. All variables are included in the model from the beginning and the method excludes one variable at a time. A regression is performed and the variable with the highest p-value is excluded. That is, the least statistically significant variable to explain the dependent variable is removed.

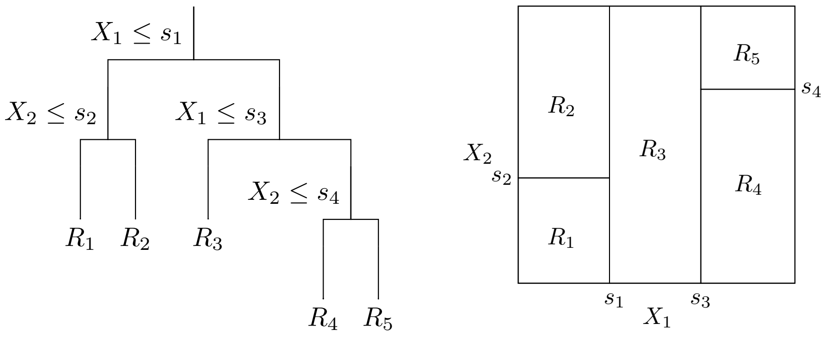
4.3 Tree-Based Methods

4.3.1 Decision Tree

Tree-based methods are another way to examine the relationship between the explanatory variables and a dependent variable. These models are built on decision trees which can be applied to both regression problems and classification problems (James et al., 2013). A decision tree consists of a root, internal nodes and leaves, also called terminal nodes. The tree is often visualised upside down with the root at the top and the leaves at the bottom, see fig. 4.1a.

The root represents the entire data set and could be seen as the predictor space. For each internal node the predictor space is divided into different regions, see fig. 4.1b. The internal nodes represent a decision which is based on one of the variables. The path that connects nodes is called a branch. The leaves represent the final result of the algorithm and are either a numerical value or a class.

The construction of a decision tree consists of two steps. Firstly, the predictor space is divided into distinct and non-overlapping regions. The predictor space is the set of all possible values for the variables, $X_1, X_2, ..., X_p$ and is divided into $R_1, R_2, ..., R_J$ regions. For regression trees, the idea when dividing



(a) Decision tree

(b) Predictor space

Figure 4.1: Illustrations of a generic decision tree and predictor space. The predictor space is divided into the regions R_1, R_2, R_3, R_4, R_5 determined by decisions based on the variables X_1, X_2 and the cut points s_1, s_2, s_3, s_4.

the predictor space is to minimise the Residual Sum of Squares (RSS) where $\hat{y}_{R_j}$ is the mean response from the training data in the region j,

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2. \tag{4.19}$$

Due to computational expense, it is often preferable to only optimise the split one step at a time. The variable with the most impact is chosen and the split is made to minimise RSS. This method is called recursive binary splitting. In every step, a predictor X_j and a cut point s have to be determined. The predictor space is divided into two new regions

$$R_1(j,s) = \{X|X_j < s\} \quad \text{and} \quad R_2(j,s) = \{X|X_j \geq s\}. \tag{4.20}$$

X_j and s are chosen to minimise

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2. \tag{4.21}$$

At the root, the predictor space is divided into two regions. In the next step, one of the regions is divided into two new regions. The region, predictor and cut point are chosen to minimise RSS. The process is repeated until a stop criterion is reached, for example, a maximum number of predictors in each region.

The idea is the same for a classification tree, but RSS can not be used as a criterion for the splits. Instead, the classification error rate could be minimised. The classification error rate is defined as the proportion of the observations from the training data that do not belong to the most common class in a region

$$E = 1 - \max_k(\hat{p}_{mk}), \tag{4.22}$$

where $\hat{p}_{mk}$ is the proportion of the training observations in region m that are from class k. A more sensitive criterion is the Gini index. The Gini index measures the variance across K classes and should be minimised,

$$G = \sum_{i=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}). \tag{4.23}$$

The method described, for both regression trees and classification trees, is likely to overfit the tree to the training data. A way to reduce this risk is to have a smaller tree with fewer splits. Hopefully, this will lead to lower variance but with the downside of potential bias. This is called tree pruning.

Secondly, a prediction is determined in each region based on the training data. The new observations that end up in the region are assigned that prediction. How to determine the prediction differs between a regression tree and a classification tree. In a regression tree, the prediction is determined as the mean of the observations in the region for the training data, while in a classification tree the prediction is determined as the majority class in the region.

4.3.2 Bagging and Random Forest

The downside of decision trees is the high variance and the risk of overfitting (James et al., 2013). A possible improvement is the bootstrap aggregation (bagging) method. The idea is to build a forest with many different decision trees. The trees are constructed from bootstrapped training samples. Bootstrap means that several training sets are sampled from the original training set. Some observations may then occur more than once since the sampling is performed with replacement. The observations are selected randomly with the same probability of being drawn. By constructing a decision tree for each bootstrapped training sample, one gets a forest of trees. Each tree is constructed without pruning, hence it has low bias and high variance. Since there are a lot of trees, the variance can be reduced by averaging across the trees. For the regression problem, the prediction is the mean of all predictions. For the classification problem, the prediction could be made by taking the most frequently predicted class from the trees. A risk with bagging is that the trees become correlated. If one variable has high importance, almost all of the trees will use that predictor for the first split. This will give correlated trees and the variance reduction will be limited. A method which considers this problem is the random forest. To decorrelate the trees, only m out of all p predictors are considered for each split when creating the decision trees. The m predictors are chosen as a random sample of p. The split is then performed based on one of the predictors in m. m new predictors are sampled for every split. This reduces the correlation since the most important predictors can not be considered in every split. It is common to use $m = \sqrt{p}$. A small value of m often improves the result when there are many correlated variables. Bagging could be seen as a special case of random forest where $m = p$. Hence, both methods are hereafter referred to as random forest.

No cross-validation is needed to test the performance of a random forest model. In the bootstrap process, approximately two-thirds of the data is used for each tree, since the probability that one observation is in the bootstrap sample is

$$P = 1 - \left(1 - \frac{1}{n}\right)^n. \tag{4.24}$$

An observation that is not included in a tree is called an out-of-bag (OOB) observation. Hence, it is suitable to use the OOB observations for the evaluation of the prediction quality. Since approximately one-third of the observations are OOB in a tree, the same proportion of predictions are obtained.

For a large number of trees, it is no longer clear which parameters are the most important. Node purity is a measure of how well a variable is able to split the data. For regression trees, the total difference in RSS resulting from splits for a specific variable, for all trees, is calculated. The sum of these differences indicates the importance. A large value means an important variable. For classification trees, the same could be done but instead using the decrease in Gini index for the splits for the variable.

4.4 Model Evaluation Measures

For variable selection, a decision rule is needed in order to know how many variables to exclude from the model without losing too much information and model quality. The Akaike Information Criterion (AIC) is a model selection criterion based on maximum likelihood estimation (James et al., 2013). The criterion considers the risk of overfitting and underfitting, as well as penalising models with too many explanatory variables. The AIC value for a certain model can be calculated as

$$AIC = -2\ln(l) + 2k, \tag{4.25}$$

where l is the likelihood function and k is the number of parameters in the model.

The Bayesian Information Criterion (BIC) is similar to AIC but, in general, penalises models with many variables more than AIC. The BIC value for a certain model can be calculated as

$$BIC = -2\ln(l) + k\ln(n), \tag{4.26}$$

where n is the number of observations.

By comparing equation 4.25 and 4.26, the natural logarithm in the second term in BIC penalises a large number of variables more than the corresponding term in AIC. For both AIC and BIC, a lower value means that the model fits the data better.

4.4.1 Numerical Model Evaluation

In order to evaluate and compare the predictive strength of the developed numerical models, several statistical measures can be used.

R^2 measures the fit of the model to the data as a proportion of the explained variance in relation to the total variance calculated as

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y_i} - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{4.27}$$

where $\hat{y_i}$ denotes the prediction of observation i, $\overline{y}$ is the average of the observations and y_i is the actual outcome of observation i.

Root mean squared error (RMSE) is calculated as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}. \tag{4.28}$$

A small RMSE indicates that the predicted value is close to the true outcome.

As a complement to RMSE, mean absolute error (MAE) could also be used and is computed as

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y_i}|}{n}. \tag{4.29}$$

MAE is the average distance between the observed values and the model predictions (Willmott and Matsuura, 2005). In comparison to RMSE, MAE does not increase to the same extent when a few predictions deviate significantly from the observations.

4.4.2 Classification Model Evaluation

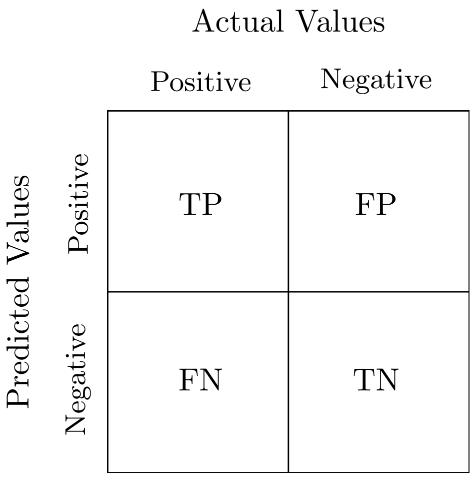
The actual outcomes and predictions of the number of corners could be defined by a classification variable. The number of corners is then binary classified in relation to the line. The confusion matrix, illustrated in fig. 4.2, presents the possible relationships between the prediction and the actual outcome for an arbitrary classification problem (James et al., 2013). The matrix consists of four components. True positive (TP) is when the model correctly predicts the outcome being positive. False positive (FP) is when the model falsely predicts the outcome to be positive. True negative (TN) is when the model correctly predicts the outcome being negative. False negative (FN) is when the model falsely predicts the outcome to be negative.

F1-score and Matthew's Correlation Coefficient (MCC) could be used as evaluation metrics of the model performance in a classification problem (Matthews, 1975).

$$F1\ score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{4.30}$$

and can take values between 0 and 1 where a score of 1 indicates perfect model precision in relation to the outcome.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{4.31}$$



		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

Figure 4.2: Illustration of a generic confusion matrix.

and the metric takes values in the range of -1 to 1 where 1 indicates perfect model precision in relation to the outcome. F1-score gives reliable results for balanced data but is misleading for imbalanced data. MCC manages to give reliable results also for imbalanced data (Chicco and Jurman, 2020).

4.5 Betting Strategies

The predicted probabilities from the models should be combined with a betting strategy to optimise the profitability. Two betting strategies were taken into consideration, namely level stakes and the Kelly criterion.

The level stakes strategy suggests placing one unit on every bet where the offered odds is greater than the odds implied by the predicted probability, in other words where there is value in a bet. The amount placed on the bet is the same regardless of how large the value is (Wheatcroft, 2020). The wager normalised to 1 is defined as

$$W = \begin{cases} 1 & if \quad o > o_p \\ 0 & if \quad o \leq o_p \end{cases}, \tag{4.32}$$

where o_p is the odds implied by the predicted probability for an outcome and o is the offered decimal odds for the same outcome. For the predicted probability p, the implied odds o_p is defined as

$$o_p = \frac{1}{p}. \tag{4.33}$$

Unlike the level stakes strategy, the Kelly criterion suggests that the number of units placed on a bet should be dependent on the size of the value (Thorp, 2008). A large difference between the predicted probability and the odds implied probability suggests a larger wager. The fraction of wealth placed on a particular outcome is calculated as

$$W = \max\left(p - \frac{1-p}{o-1}, 0\right). \tag{4.34}$$

In order to make the betting strategies comparable, the stakes proposed by the Kelly criterion for match i can be normalised as

$$s_i = c \cdot W_i, \tag{4.35}$$

with normalising constant c, for m matches, such that

$$\frac{1}{m}\sum_{i=1}^{m} c \cdot W_i = 1. \tag{4.36}$$

In their study, Yip et al. (2022) found that the market may be underestimating the probability of the under selection. This means that the odds are overvalued and indicating a potential profitability opportunity in consistently betting on the under selection. If this theory holds, a strategy could be to bet on the under alternative without any modelling. The simplest strategy is to use the level stakes strategy and always place a bet on the under alternative. Another option could be to combine this theory with Kelly criterion using the models. Then a bet could be placed on the under alternative only when the model predicts that betting value exists and the wager size is determined according to the Kelly criterion.

5 Method

The dependent variable could be described as the number of corners in the match or as a classification variable. In the latter case, the dependent variable takes the value 1 if the number of corners is less than 10.5 and 0 otherwise. A logistic regression can then be used to estimate the probability that the number of corners will be above or below the line. Related to the confusion matrix, the value 1 corresponds to a positive outcome and prediction, while 0 corresponds to a negative outcome and prediction. An alternative to the regression approach when modelling this type of event is to treat the data as a time series, one time series for each team and season. However, the number of time series would then have become large and too short which would complicate the handling of the data and the ability to obtain reliable results. Hence, time series modelling was not applied.

One way to model the number of corners in the match would be to divide the problem into modelling the home team and away team corners separately, with individual regressions and the team's individual corner count as a dependent variable. Then the predictions could be added to get the prediction for the total corner count. Alternatively, the total number of corners may be modelled directly. For the classification problem, the line has to be determined to be able to classify the dependent variable, hence only the classification variable of the total corners could be used as the dependent variable.

The modelling was divided into two main methods, generalised linear regression methods and tree-based methods. For each main method, one numerical and one classification modelling approach was applied. The variable processing was made in MATLAB (R2022b) and the modelling as well as testing in R Statistical Software (version 4.2.2).

5.1 Data Processing

5.1.1 Data Splitting and Cross-Validation

The model data consisted of Premier League seasons from 2014/2015 to 2019/2020, while the seasons 2021/2022 and 2022/2023 formed the test set. For the test data set covering the Bundesliga, the seasons 2021/2022 and 2022/2023 were used. The model data from the Premier League was used to find appropriate models which were then tested on the test data from the Premier League and the Bundesliga. The splitting of model data and test data season-wise was used to test the model's ability to predict on completely new data.

By using cross-validation, the risk of overfitting the model to the training data decreased and instead strengthened the prediction model's capability to handle new data. This may also increase the interpretability of the model and the stability of the predictions, since it focuses on the underlying structure rather than particular patterns of noise (James et al., 2013). k-fold cross-validation is commonly used which means that the data set is divided into k subsets, where $k-1$ subsets are used as training data and the remaining subset is used for validation. This procedure is repeated k times, where all subsets are used as validation data once. A 10-fold cross-validation was used on the model data when training the model. The number of folds was determined to get an appropriate size of the subsets. When it was not possible to divide the data set into exactly equal subsets, the remaining observations were included in the last subset. When performing the 10-fold cross-validation, the data were randomly rearranged and divided into subsets and later sorted back to the original order. For the generalised linear regression models, the same randomisation was made for all models. For the random forest problem, the function *rfcv* in the *randomForest* package was used for the cross-validation (Liaw et al., 2004). Hence, the cross-validations were not performed with identical subsets for the generalised linear models and the random forest models.

5.1.2 Variable Processing

Based on the match data set, moving averages with lengths of three, four and five matches were computed for each variable and team. A season average of all matches played so far during the current season was also computed, in tables and figures denoted as *avg* while the moving averages are denoted as *avg3*, *avg4* and *avg5*. Corresponding moving averages and seasonal averages were also made separately for the team's home and away matches, creating home and away averages for each variable and each team. This resulted in that every variable, for the home team and away team, respectively, had eight subvariables which together form a variable family, see fig. 5.1. For the averages, values for the initial matches for each season were missing. One option to deal with this problem was to exclude enough matches in the first rounds of play each season to obtain a data set capable of providing values up to the fifth moving average in each variable. Since moving averages based only on home and away matches were also taken into account, this meant that a considerable number of matches would have been needed to be excluded from the data set. At the same time, it was important to have enough matches in each season to maintain the robustness of the data set. For that reason, the problem was addressed by using the previous season's mean value for each variable during the first rounds of play. Thus, for the 2014/2015 season, which was the first season in the model data, team averages from the 2013/2014 season were used as initial values. As the matches in the first game weeks were played, the moving averages were built up and the averages of the previous seasons faded out. For the clubs that were promoted, average values of the variables of the clubs placed in the bottom quarter the previous season were used as initial values.

$$\text{Home team} = \begin{cases} \text{Home team season average} \\ \text{Home team average 3} \\ \text{Home team average 4} \\ \text{Home team average 5} \\ \text{Home team home season average} \\ \text{Home team home average 3} \\ \text{Home team home average 4} \\ \text{Home team home average 5} \end{cases}$$

$$\text{Away team} = \begin{cases} \text{Away team season average} \\ \text{Away team average 3} \\ \text{Away team average 4} \\ \text{Away team average 5} \\ \text{Away team away season average} \\ \text{Away team away average 3} \\ \text{Away team away average 4} \\ \text{Away team away average 5} \end{cases}$$

Figure 5.1: Variable family for a home team and an away team. The averages are of different lengths and divided to represent all matches played by a team, as well as home and away matches separately.

5.1.3 Additional Variables

In addition to using actual historical match data and the teams' FIFA ratings, other potentially helpful variables were created. This was done in order to supplement the available match statistics with the purpose of further explaining the randomness in the number of corners.

Two variables that could possibly affect a team's performance in a match are the team's form based on recent matches and its motivation to perform. Both form and motivation largely depend on psychological factors of players and staff rather than physical, technical and tactical abilities which otherwise are easier to measure. For example, if a team is in good form as a consequence of being on a winning streak or has incentives which lead to boosted motivation, it may enter a match with enhanced confidence which could spill over into the overall performance and willingness to take risks.

Form Variable

The approach of the form variable was to create a point system, similar to the standard point system of the match outcome. The form point system should reflect whether the match result in terms of win, draw or loss was expected or an upset. The difference in the teams' overall FIFA rating indicated how big a potential upset would be if the lower-rated team beat the higher-rated team, which gave the so-called FIFA rating compensation. The form point system also compensated the team playing away. The form points a team received from a match were determined based on the points the team received from the match result multiplied by the compensation factors. A team's form was then determined based on moving averages of the points of three, four or five matches. The form points a team was given in a match were defined as

$$\text{form points} = \text{match points} \cdot \text{away compensation} \cdot \text{FIFA compensation}. \tag{5.1}$$

The away compensation for a team was defined as

$$\text{away compensation} = \begin{cases} 1 & \text{if team is home} \\ 1.3 & \text{if team is away} \end{cases}, \tag{5.2}$$

and was derived as the mean of points obtained by the home teams divided by the points obtained by the away teams of the model data.

The FIFA compensation was defined as

$$\text{FIFA compensation} = \begin{cases} 1 & \text{if } \Delta_{FIFA} < 4 \\ 1.2 & \text{if } 4 \leq \Delta_{FIFA} < 8 \\ 1.4 & \text{if } \Delta_{FIFA} > 8 \end{cases}, \tag{5.3}$$

where Δ_{FIFA} was the difference in the overall FIFA rating between the opponent team and the team considered. Hence, the weaker team was rewarded but the stronger team was not penalised.

Motivation Variable

The motivation variable was assumed to be particularly important at the end of the season and was created to take into account the teams' table position and the importance of the match. The variable was only valid for the last five game weeks. For the other matches, the teams were considered to have a normal level of motivation. The motivation variable was binary and thus only took the values 0 or 1. In the cases where a team had the value 1 in the final game weeks, it was considered to have increased motivation due to the risk of relegation, its chances of qualifying for a UEFA tournament or its chances of winning the league title. The motivation variable for a team in a match was defined as

$$M = \begin{cases} 1 & \text{if } ML \leq 5 \text{ and } T_T \leq 7 \text{ or } T_B \leq 5 \\ 0 & \text{otherwise} \end{cases}, \tag{5.4}$$

where ML was matches left the current season, T_T was table placement from top and T_B was table placement from bottom.

5.2 Modelling Framework

An overview of the modelling framework is illustrated in fig. 5.2, explaining the process starting from the selection of averages in the variable families and ending with the predicted probabilities of the models. For the generalised linear regression models, the number of variables was further reduced with either backward stepwise selection or elastic net. The modelling framework resulted in 15 different model predictions.

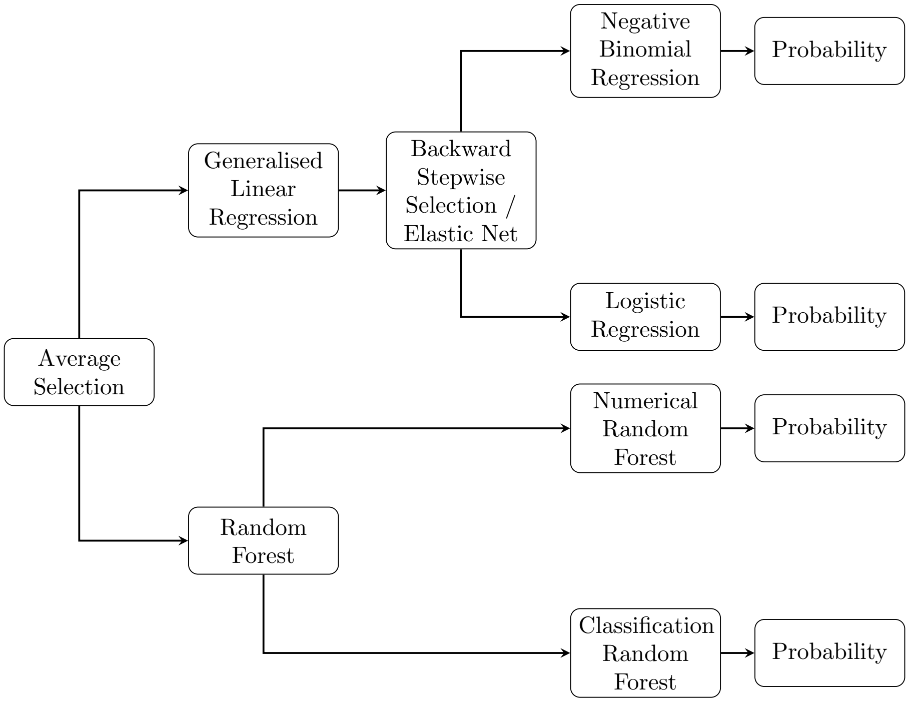

Average Selection
Generalised Linear Regression
Backward Stepwise Selection / Elastic Net
Negative Binomial Regression
Probability
Logistic Regression
Probability
Random Forest
Numerical Random Forest
Probability
Classification Random Forest
Probability

Figure 5.2: Modelling framework with used selection and statistical learning methods, leading to 15 model predictions.

5.3 Average Selection

After the variable processing has been conducted, the full data set included 198 explanatory variables. The large amount mainly depended on that eight moving averages of different lengths were included from each variable family. Therefore, a methodical approach of selecting the average from each variable family that best explains the dependent variable was used.

The purpose of creating averages of different lengths was to identify the average of a certain length that best could relate to the dependent variable. The averages of different lengths were naturally strongly correlated, hence it was unnecessary to include more than one moving average per variable family. Variables that did not belong to a specific variable family, namely game week, FIFA ratings, match odds and the motivation variable were left out of this primary variable selection. In the case of predicting corners with negative binomial regression, a negative binomial regression was first performed on each variable family independently and the average with the lowest p-value was chosen to represent that variable family in the model. For the logistic regression case, each family underwent a logistic regression in order to choose the most informative average. In the logistic case, the built-in function *glm* was used and in the negative binomial case, the function *glm.nb* in the package *MASS* was used (Venables and Ripley, 2002). The procedure was repeated for all dependent variables. This led to the number of variables being reduced to 38 for each dependent variable.

5.4 Generalised Linear Regression

When the first variable filtering was completed, the backward stepwise selection and the elastic net were performed in parallel. In the backward stepwise selection, one variable was removed at a time until the smallest AIC and BIC values were found, thus two different models were obtained. To perform the backward stepwise selection, the built in-function *step* was used. The backward stepwise selection was conducted with both negative binomial regression and logistic regression. The elastic net was used as a variable selection method and the α-value was set to 0.5, weighing lasso regression and ridge regression equally. The *glmnet* package in R was used to fit a regression model with the elastic net (Friedman et al., 2010). With the partially reduced data set of 38 variables, 100 λ-values were generated and thereby 100 different models. In the case of negative binomial regression, a Poisson distribution was used due to the absence of the option of negative binomial distribution in *glmnet.* Note that the Poisson distribution was only used in the elastic net in order to select which variables should be included in the final negative binomial regression. Models obtained from a Poisson-based elastic net are still referred to as negative binomial models. In the classification case, binomial distribution was used. Based on the cross-validation, the models corresponding to λ_{min} and λ_{1se} were identified. New negative binomial and logistic regressions were conducted based on the variables included in the chosen models.

Cross-validation was performed for all models. The best-performing negative binomial regression model and logistic regression model were selected for the numerical and classification problem with respect to the evaluation metrics. For the selected models, predictions were subsequently made on the test data. From the predicted values, evaluation metrics were determined to evaluate the models' performance on test data.

5.4.1 Probability Sampling

Based on the regression for the number of corners, a probability that the number of corners in a particular match ended up over or under the line had to be identified. This step had to be done for the negative binomial regression models, but not for the logistic regression models since the sought probabilities then were obtained directly.

The following steps show how the predicted corner distribution for the dependent variable t and match i was derived. The first step was to generate a sample β_t from a normal distribution from the regression coefficient estimates β_t^0 and the variance Σ_t. When estimating the regression coefficients with maximum likelihood estimation, the assumption of asymptotic normality gave that the estimate was distributed as

$$\beta_t \sim \mathcal{N}(\beta_t^0, \Sigma_t). \tag{5.5}$$

The mean μ_{ti} of the sampling was then, in line with eq. (4.5), computed as

$$\mu_{ti} = e^{X_{ti}\beta_t} \tag{5.6}$$

where X_{ti} is the explanatory variables for dependent variable t and match i. The number of corners y_{ti} for dependent variable t in match i was drawn from a negative binomial distribution as

$$y_{ti} \sim NB(\mu_{ti}, \theta), \tag{5.7}$$

where θ is the estimated dispersion parameter. The procedure was repeated 1000 times for each match.

Finally, when modelling home team corners and away team corners, respectively, the aggregated number of corners was given by the pairwise summation of the samples. The probability of the number of total corners being less than the line of 10.5 was given by calculating the proportion below the line. When modelling the total corner count directly, the proportion could be calculated without pairwise summation.

5.5 Bagging and Random Forest

Random forest was conducted in the same way for both the regression problem and the classification problem. The package *randomForest* in R was used for the regression (Liaw and Wiener, 2002). The number of trees and the number of variables considered in each split, m, were inputs to the random forest algorithm. The number of trees was set to 500, which was the default value in the *randomForest* package. To perform the cross-validation, the function *rfcv* was used. m ranged from $m = p$ to $m = 1$. The value of m which resulted in the smallest cross-validation error was chosen. The evaluation metrics were determined from the cross-validated predictions and the best-performing numerical model was selected to make predictions on test data. Only the model predicting total corner count was optional for the classifier.

The function *randomForest* was used to train the final model on the entire model data and then predictions were made on the test data. The number of variables considered in each split, m, from the cross-validation which resulted in the smallest cross-validation error was used. The evaluation metrics were determined based on the predicted values to evaluate the models' performance on test data.

To determine the importance of the variables, the function *importance* was used. Importance is defined as the total decrease in node impurities from splitting on the variable, averaged over all trees. For classification, importance is based on the Gini index and for regression, it is based on RSS.

The predicted probability for having over or under 10.5 corners for the regression problem was obtained by the proportions over-under 10.5 corners of all trees from each prediction. In the same way, the predicted probability was obtained in the classification problem based on the proportions of predictions over and under the line.

5.6 Under Odds Theory

As described in chapter 2, there are findings indicating an asymmetry in the over-under market in favour of betting on the under selection. To examine this theory, a hypothesis test was performed where the null hypothesis and alternative hypothesis were formulated as

$$\begin{aligned} H_0 &= \text{under odds market are correct or overvalued,} \\ H_1 &= \text{under odds market are undervalued.} \end{aligned} \tag{5.8}$$

The outcomes were seen as a classification variable where under 10.5 corners corresponded to 1 and over 10.5 corners corresponded to 0. The odds implied probability for the under market for match i, p_{1i}, was calculated as

$$p_{1i} = \frac{o_{1i}}{o_{1i} + o_{0i}}, \tag{5.9}$$

where o_{1i} was the odds for the under market and o_{0i} was the odds for the over market for match i. For sampling t, an outcome for match i was sampled from the implied probability as

$$Y^t_{p_{1i}} \sim Be(p_{1i}). \tag{5.10}$$

The expected value of the sampled outcome t minus the implied probability for match i was

$$E[Z^t_i] = E[Y^t_{p_{1i}} - p_{1i}] = 0. \tag{5.11}$$

The mean of the sampled outcomes t minus the implied probabilities for N matches had an expected value

$$E[Z^t] = E\left[\frac{\sum_{i=1}^{N} Z^t_i}{N}\right] = 0. \tag{5.12}$$

The sampling was repeated 1000 times for each implied probability. The means of the sampled outcomes minus the implied probabilities for N matches belong asymptotically to a normal distribution with zero mean due to the Central Limit Theorem. Hence,

$$Z^t \sim \mathcal{N}(0, \sigma^2) \tag{5.13}$$

where σ^2 is an arbitrary variance. The mean of the actual outcomes minus the implied probabilities was calculated as

$$Z^{obs} = \frac{\sum_{i=1}^{N} Y_i^{obs} - p_{1i}}{N} \tag{5.14}$$

where Y_i^{obs} was the actual outcome for match i. If the under market odds were undervalued, Z^{obs} should be sufficiently larger than zero. A p-value was obtained from a one-sided Z-test based on Z^{obs} and a fitted normal distribution to determine if the null hypothesis could be rejected as

$$p\text{-}value = P(Z^{obs} > Z) = 1 - \Phi\left(\frac{Z^{obs} - 0}{\sigma}\right), \tag{5.15}$$

where $\sigma = \sqrt{Var[Z]}$ is estimated based on the samples Z^t. The Z-test is illustrated in fig. 5.3.

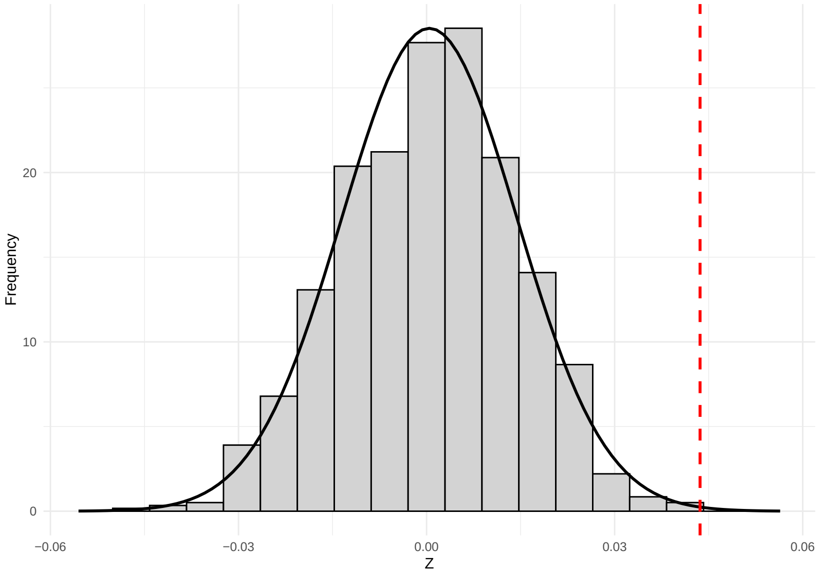


Figure 5.3: A histogram of Z^t values and a fitted normal distribution. The red dashed line represents Z^{obs} which should be examined whether it belongs to the distribution.

5.7 Profitability Analysis

In order to evaluate the profitability of the models, the predicted probabilities were combined with the betting strategies presented. Since the theory of consistently betting on the under selection was analysed, the models and betting strategies were also combined with only playing on the under selection. The profitability analysis was made on the test data both for the Premier League and the Bundesliga. The profitability was evaluated both on profit in units and return on investment (ROI). ROI measures the return on placed bets only, while total profit considers the overall profitability taking all available matches into account. These two ways of measuring profitability differ, and the difference can become particularly apparent when applying conditional betting strategies. In 2023, the commission for Swedish customers at Betfair was 2% on net profits, which was considered in the calculated profitability (Betfair, 2023a).

6 Result and Analysis

The modelling was accomplished as described by theory and methods. The result is demonstrated by first presenting which variables that were included in each model. Then model evaluation metrics are presented which are the foundation for the model selection. Finally, the profitability analysis is presented where the best-performing models are combined with the betting strategies.

6.1 Variable Selection

6.1.1 Average Selection

The first step in the variable selection process is to select one average from each variable family. This process is performed by using negative binomial regression and logistic regression, respectively, for each of the dependent variables. How the averages are selected is illustrated in fig. 6.1. As can be seen, the season average is dominant, especially when modelling the home team and away team corner count, respectively. The regular average is preferred over the home-away-dependent average.

6.1.2 Generalised Linear Regression

Further variable selection is made either by elastic net or backward stepwise selection. In the elastic net, the λ-value is chosen either as λ_{min} or λ_{1se}. Figure A.1, fig. A.2, fig. A.3 and fig. A.4 in appendix A provide illustrations of the tuning parameter selection. In the figures, the mean cross-validated error is plotted against λ-values. The red vertical lines mark the values for λ_{min} and λ_{1se}. For all regression models considered, the smallest mean cross-validated error is found for a relatively low λ-value. Since the curves for the negative binomial regression models have a markedly positive slope after finding the minimum, λ_{1se} is related to a comparatively higher mean cross-validated error. The curve for the logistic regression models instead flattens out after finding the minimum, indicating a lower sensitivity for selecting the tuning parameter. Both for negative binomial and logistic regression, modelling the total number of corners, λ_{1se} results in models where the elastic net shrinks the number of explanatory variables to zero. It is again mentioned that the Poisson distribution was used

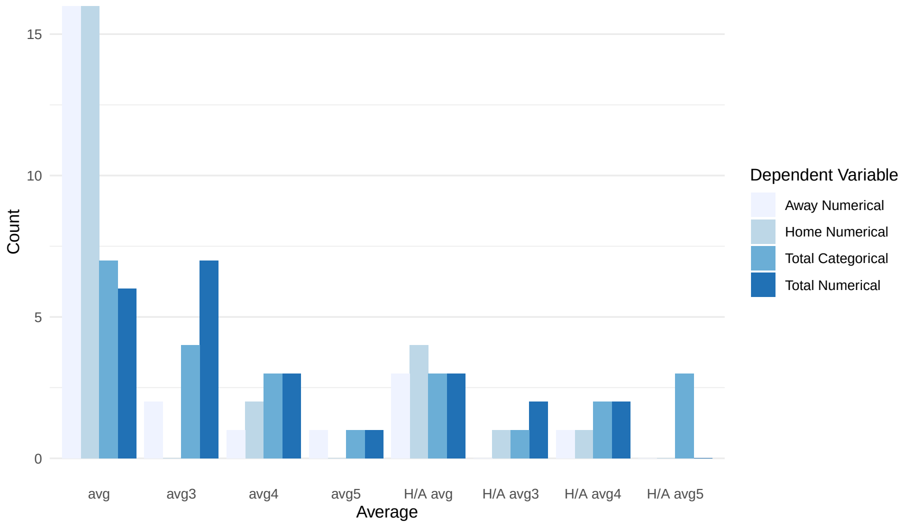


Figure 6.1: Selection of variables from the variable families. Each variable family consists of eight averages, where H/A denotes if the average is only taking home or away matches into account.

as input in *glmnet*, instead of the non-optional negative binomial distribution, which may have led to a sub-optimal choice of the tuning parameter.

6.1.3 Bagging and Random Forest

The number of variables considered in each split, m, in the random forest models is chosen to generate the smallest cross-validation error. Random forest is considered for the four types of dependent variables. In all cases $m = p$ is chosen in order to minimise the cross-validation error, this means bagging is performed.

6.1.4 Analysis of Included Variables

The number of included variables in the generalised linear regression models clearly varies both regarding to the variable selection method and the dependent variable used in the model. For many of the models, the variables selected are logical to describe the number of corners in relation to the dependent variable, while some combinations of selected variables may be perceived as counterintuitive. This can be explained by the fact that many explanatory variables in the data are correlated, which can cause some variables to be selected over other variables that may be more intuitive. The analysis will therefore focus on identifying logical patterns in the variable selection and comparing similarities and differences between the models. A complete presentation of all variables in each model can be found in table B.2, table B.3, table B.4, table B.5, table B.6, table B.7, table B.8 and table B.9 in appendix B.

Studying the variables selected in the negative binomial regression models when modelling the total number of corners, the away team's overall FIFA rating and the home team home average goal count in the first half occurs most frequently. When modelling the teams' corners separately with negative binomial regression, the money line winning odds occurs multiple times as well as the corner count and the opponent team's conceded corner count. In logistic regression, the home team corner season average and the away team possession season average are included in several models.

In numerical random forest, most variables have a relatively high level of importance. When modelling the total number of corners with random forest, the away team's season average for shots off target in away games as well as both teams' goal count in the first half get the highest importance. For the home team, the latter variable concerned the average of matches played at home, which is in line with the findings in negative binomial regression. When modelling the number of corners for the home team in numerical random forest, the money line odds for the home team win has a remarkably high importance and the odds for the away team win has the second highest importance. The importance is evenly spread across most of the remaining variables. Worth noting is that the motivation variable has by far the lowest level of importance, both for the home team and away team. Similar observations can be made when modelling the away team's corner in numerical random forest. The motivation and form variables are however included in several generalised linear regression models. On the whole, the results indicate a pattern that numerical random forest and negative binomial regression lead to similar variable selection. For classification random forest, variables such as the team's season average of corners, fouls and possession have considerable importance and the motivation variables have again negligible importance. This is also in line with the classification equivalent in generalised linear regression modelling.

Moreover, variables connected to the home team and away team respectively are evenly distributed among the models. The number of included variables in negative binomial and logistic regression is larger for the backward stepwise variable selection with AIC and the elastic net using λ_{min} compared to the backward stepwise variable selection with BIC and the elastic net using λ_{1se}, which is in line with the expectations from theory.

6.2 Model Performance Evaluation

For negative binomial regression and numerical random forest, the modelling results are presented both for the case when the total number of corners is predicted and when the teams' corners are predicted separately and then combined. For logistic regression and classification random forest, only the total number of corners is modelled. The predictions are obtained by 10-fold cross-validated data. Furthermore, the results include the variable selection methods leading to eight different model alternatives for negative binomial regression and four model alternatives for logistic regression. Two model alternatives are obtained when using numerical random forest and one for classification random forest. Altogether, fifteen models are developed and analysed. The best-performing models for each regression method are highlighted in bold in table 6.1. The negative binomial and logistic regression models with elastic net using λ_{1se}, modelling the total number of corners, are reduced to solely a constant due to heavy penalisation. Hence, these models will hereafter not be considered in the analysis. For comparison, a naive numerical model and a naive classification model are added. The naive numerical model uses the arithmetic mean of the model data as a prediction and the naive classification model uses the share of matches with under 10.5 corners as the predicted probability. Naturally, the naive models are the same as the disqualified constant models. The naive models differ considerably since the mean of total corners is 10.55 while the share of matches with less than 10.5 corners is 52.4%, hence their predictions are each other's opposites. When cross-validation is performed, the predictions do not need to be constants for the whole data set since all subsets do not necessarily have the same distribution.

6.2.1 Model Evaluation Measures

When evaluating the models, F1-score and MCC are used to make the results comparable between the numerical and the classification methods. The numerical methods have been supplemented with RMSE, R^2 and MAE for internal comparison. Since the data can be considered to be balanced, F1-score is favoured over MCC due to its simplicity and higher interpretability.

6.2.2 Model Comparison

The results show that the F1-score is consistently larger for all classification models compared to the numerical models. It can also be seen that the F1-score is generally larger for the generalised linear regression models than for the random forest models. On the other hand, MCC is larger for the numerical models compared to the classification models and is also larger for the generalised linear regression models compared to the random forest models. The MCC value for the classification random forest is notably low but still has a competitive F1-score. The naive classification model gets an undefined MCC value since no true negatives or false negatives are obtained. On the other hand, it has the largest F1-score. The naive numerical model has both a low F1-score and MCC value. This finding confirms that F1 and MCC differ and may not lead to identical results.

RMSE, R^2 and MAE for the numerical models indicate that the explanatory power for the models is low. Still, there are differences in the model performances worth considering. For negative binomial regression, separate modelling using λ_{1se} is competitive since it exhibits the second highest R^2 and the second lowest RMSE and MAE, only surpassed by modelling the total number

Table 6.1: Performance of all models for the Premier League model seasons, supplemented by a naive numerical model and a naive classification model.

Model	**F1**	**MCC**	**RMSE**	$\mathbf{R^2}$	**MAE**
Numerical Naive	0.173	0.011	3.482	0.000	2.764
Total Numerical NegBin Stepwise AIC	0.536	0.082	3.452	0.017	2.730
Total Numerical NegBin Stepwise BIC	0.520	0.055	3.461	0.012	2.739
Total Numerical NegBin λ_{min}	0.530	0.072	3.460	0.013	2.740
Total Numerical NegBin λ_{1se}	0.173	0.011	3.482	0.000	2.764
Home/Away Numerical NegBin Stepwise AIC	0.533	0.074	3.462	0.012	2.736
Home/Away Numerical NegBin Stepwise BIC	0.519	0.084	3.462	0.012	2.736
Home/Away Numerical NegBin λ_{min}	0.513	0.054	3.474	0.005	2.749
Home/Away Numerical NegBin λ_{1se}	0.533	0.082	3.456	0.015	2.733
Classification Naive	0.688	-			
Total Classification Logistic Stepwise AIC	0.606	0.059			
Total Classification Logistic Stepwise BIC	0.621	0.070			
Total Classification Logistic λ_{min}	0.605	0.056			
Total Classification Logistic λ_{1se}	0.688	-			
Total Numerical Random Forest	0.514	0.054	3.476	0.003	2.755
Home/Away Numerical Random Forest	0.490	0.054	3.500	-0.010	2.773
Total Classification Random Forest	0.578	0.020			

of corners with AIC. Regarding the numerical random forest models, RMSE and MAE are higher and R^2 is lower for modelling the total number of corners compared to separate modelling. The numerical random forest models perform worse across all metrics compared to the negative binomial regression models. For separate modelling using random forest, the R^2 is negative implying that it performs worse than a constant model. Worth mentioning is that the R^2 values when modelling home team corners and away team corners, separately, are significantly higher than when the two models are merged. The predictive strength is reduced when the separate predictions are summed since the variance of the merged model is the sum of the variance and covariance of the separate models. RMSE, R^2 and MAE for the models of the home team corners and the away team corner could be found in table C.1 in appendix C.

6.2.3 Method Comparison

Concerning the elastic net versus the backward stepwise variable selection, the results suggest that the latter may be preferred for handling the large number of explanatory variables. The generalised linear regression models perform better across the metrics compared to the random forest models. It can also be seen that modelling with a classification approach is preferred over a numerical procedure. The results indicate that modelling the total number of corners should be preferred to modelling the teams' corners separately. An advantage of modelling the total corner count is also the opportunity to transfer the dependent variable from numerical to classification.

6.2.4 Qualified Models

Modelling the total number of corners with stepwise variable selection based on AIC performs best among the negative binomial regression models, while modelling with BIC is preferred in the logistic regression. Modelling the total number of corners is also preferred in the numerical random forest. These three models together with the classification random forest are qualified for further investigation. Altogether, the results based on model data show that the logistic regression model using stepwise variable selection with BIC is superior to all other models. The naive numerical model performs worse than the negative binomial models and the numerical random forest when modelling the total number of corners. The naive classification has a better F1-score than the other classification models but has an undefined MCC value.

The qualified negative binomial regression model and logistic regression model with their included variables and coefficient estimates are presented in table 6.2 and table 6.3.

Table 6.2: Best performing negative binomial regression model and its estimated parameters.

Model	**Included variables**	$\boldsymbol{\beta}$
Total Numerical NegBin Stepwise AIC	intercept	2.8028
	away team OVR	-0.0078
	away team avg 4 goal count	0.0198
	home team avg conceded goal count	-0.0275
	away team avg conceded goal count	0.0451
	home team avg home goal count first half	0.0863
	home team avg4 corner count	0.0129
	home team avg conceded corner count	0.0177
	away team avg3 away conceded corner count	-0.0062
	away team avg possession	-0.0016

Table 6.3: Best performing logistic regression model and its estimated parameters.

Model	**Included variables**	$\boldsymbol{\beta}$
Total Classification Logistic Stepwise BIC	intercept	-0.3731
	home team avg corner count	-0.1303
	away team avg possession	0.0231

6.3 Model Testing

The performance evaluation metrics are presented in table 6.4 for all qualified models when the Premier League test data is applied. Some differences in the evaluation measures arise when studying how the models handle new data. All qualified models increase in F1-score and the improvement is especially apparent for the random forest models. This is not an expected result. It can either depend on a coincidence or that the test data is easier to predict on. The generalised linear regression models get a worse MCC value. The MCC value for classification random forest increases considerably compared to when evaluated on model data, and an improvement is also apparent for the numerical random forest. In fact, the highest F1 and MCC values hitherto are found for the classification random forest model. Regarding RMSE and MAE for the numerical models, the former is improved while the latter is slightly worsened compared to the results in table 6.1. At the same time, the R^2 values for both numerical models are negative. When comparing the models in table 6.4, the random forest models outperform the generalised linear regression models with respect to F1 and MCC values, both for numerical and classification modelling respectively, which is opposite to the findings in table 6.1. The naive numerical model performs better than the other numerical models according to RMSE and R^2 while is in between according to MAE. The F1-score for the numerical naive model is 0 since the predictions result in no true positives. The naive classification model has the highest F1-score among the classification models.

Table 6.4: Performance of qualified models for the Premier League test seasons.

Model	**F1**	**MCC**	**RMSE**	**R^2**	**MAE**
Numerical Naive	0	-	3.413	-0.004	2.752
Total Numerical NegBin Stepwise AIC	0.566	0.058	3.429	-0.014	2.743
Classification Naive	0.717	-			
Total Classification Logistic Stepwise BIC	0.642	0.062			
Total Numerical Random Forest	0.558	0.100	3.435	-0.017	2.757
Total Classification Random Forest	0.671	0.120			

6.4 Profitability Analysis

6.4.1 Under Odds Theory

To assessing the under odds theory, the statistical test described in section 5.6 is performed. The aim is to investigate whether the under odds are systematically undervalued and hence there would exist inefficiency in the considered over-under market for corners. Odds from the seasons 2019/2020 to 2022/2023, with only matches in 2023 included in the latest season, are used in the test. Due to missing data, the Premier League data set includes 1211 matches. The p-value is determined to be 0.0010. This means that the null hypothesis can be rejected at a low significance level. The test result is illustrated in fig. 5.3.

6.4.2 Betting Analysis

The best-performing models were combined with the betting strategies. As part of analysing the asymmetry in the over-under market in favour of betting on the under alternative, the models were also combined with the betting strategies under the condition of only betting on the under alternative if value exists. This led to each model having four different betting approaches. The result for consistently betting on the under alternative, without first assessing the potential betting value, is presented along with the models' profitability. This strategy is the same as the naive classification model combined with level stakes, thus the naive classification model is included in the analysis. Since the betting format is over-under, the naive classification model is assumed to be more relevant than the naive numerical model, hence the latter is not included in the analysis.

In table 6.5, it can be seen that all models generate positive returns when the models are tested on new Premier League data, consisting of odds from 544 matches. The Kelly criterion is outperforming level stakes for all models both in terms of total profit and ROI. Considering when the models are allowed to place bets either on the over or under alternative, the total profit is in most cases nearly twice as high using the Kelly criterion compared to level stakes. Classification random forest then yields the highest profit, followed by the logistic regression model. On the other hand, the negative binomial regression model under level stakes performs worst but is still profitable. Consistently betting on the under alternative generates a positive profit. It performed worse than the models in combination with the Kelly criterion but in most cases better than the models in combination with level stakes.

Generally, the ROI and the share of successful bets increase significantly when betting on the under alternative only when the model predicts that value exists on the under odds. The percentage of placed bets then decreases from 94-97% to 45-53% and the share of wins increases from 51-53% to 57-59%. Classification random forest using the conditional Kelly criterion exhibits the highest ROI. This could be explained by the fact that there, in general, exists more value in the under odds offer. Still, the model has a lower total profit compared to when it is combined with the unconditioned Kelly criterion. This observation is consistent with the other models as well.

Figure 6.2, fig. 6.3, fig. 6.4 and fig. 6.5 illustrate how the profit of the models develops over time. It is noteworthy that for many models and betting strategies, an initial short negative trend is visible during the first 50 matches, followed by a recovery period and a horizontal trend during the following 200 matches. Then a clearly positive trend is apparent during the remaining matches of the test data set. The initial negative trend could be due to the usage of average statistics from the previous season, which is faded out as the matches are played and may not be representative to yield enough predictive strength. On the other hand, the test data consists of two seasons, with the second season starting at match 327, and no such negative trend can be discerned in either of the charts. The profit graph from consistently betting on the under alternative follows the trends of the other graphs quite well apart from a scaling factor.

Figure 6.5 demonstrates clearly how well classification random forest performs in combination with the Kelly criterion, both when using it conditionally and unconditionally to the under odds. Contrary to the other model and betting strategy combinations, the positive trend starts already after 150 matches played. Another interesting aspect can be seen in fig. 6.3 and fig. 6.4 where the models combined with the Kelly criterion actually generates negative return during the first 150 matches, but then recover and eventually yield the highest returns among the betting strategies applied on the respective model. This may indicate that the Kelly criterion is largely dependent on being combined with a prediction model that is competitive in assessing the value of the bet in order to be successful. Classification random forest may be better at assessing the value of the bet compared to numerical random forest and logistic regression. This should result in a wager that better reflects the risk in the bet and eventually higher profitability.

Furthermore, the logistic regression model in combination with level stakes exhibits a persistent negative return during the majority of the matches, but later improves and generates profit. When comparing the numerical models to the classification models, the spread of the profitability curves is larger for the classification models. This could be explained by how large the difference is between the model prediction and the line, which in turn affects the wager size when betting under the Kelly criterion. Classification modelling, regardless of whether it is based on generalised linear regression or random forest, is then more successful in assessing the potential value in the bet leading to more profitable models compared to numerical modelling.

Table 6.5: Profitability analysis on the Premier League test seasons.

Model	**Betting Strategy**	**Bets Played (%)**	**Wins (%)**	**Total profit**	**ROI (%)**
	Under	100	56.6	29.5	6.1
Total Numerical NegBin Stepwise AIC	Level stakes	96.7	50.9	14.1	3.0
	Kelly criterion	96.7	50.9	32.0	6.9
	Level stakes U	47.5	57.2	23.5	10.3
	Kelly criterion U	47.5	57.2	34.1	14.9
Total Classification Logistic Stepwise BIC	Level stakes	94.4	50.5	17.2	3.8
	Kelly criterion	94.4	50.5	54.2	11.9
	Level stakes U	44.8	57.4	26.3	12.2
	Kelly criterion U	44.8	57.4	49.9	23.1
Total Numerical Random Forest	Level stakes	97.1	50.6	16.9	3.6
	Kelly criterion	97.1	50.6	36.4	7.8
	Level stakes U	44.8	57.4	25.4	11.8
	Kelly criterion U	44.8	57.4	35.8	16.6
Total Classification Random Forest	Level stakes	96.3	53.0	35.7	7.7
	Kelly criterion	96.3	53.0	74.0	15.9
	Level stakes U	53.1	59.0	36.9	14.4
	Kelly criterion U	53.1	59.0	65.9	25.8

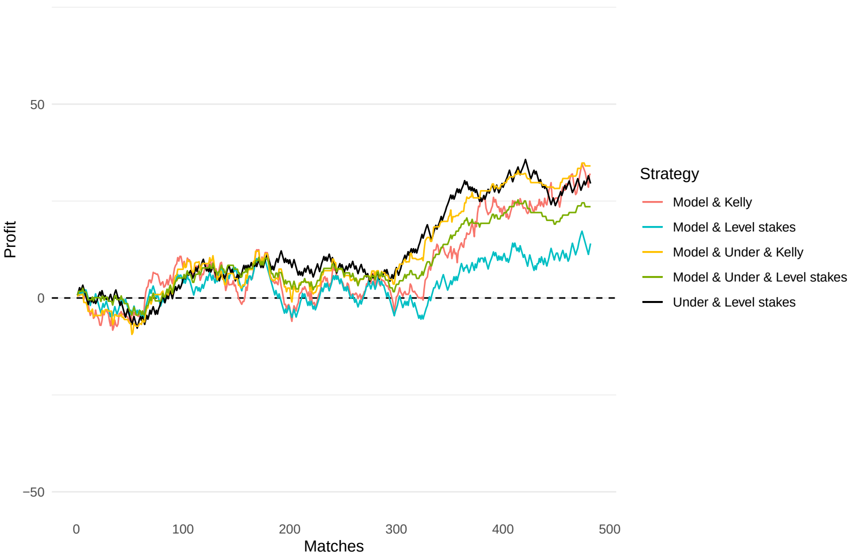


Figure 6.2: Profitability chart for Total Numerical NegBin Stepwise AIC on the Premier League test seasons.

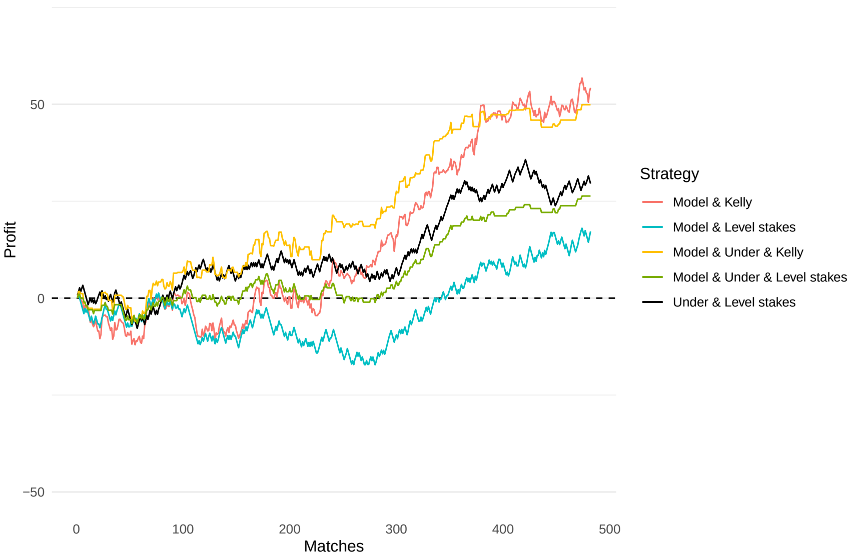


Figure 6.3: Profitability chart for Total Classification Logistic Stepwise BIC on the Premier League test seasons.

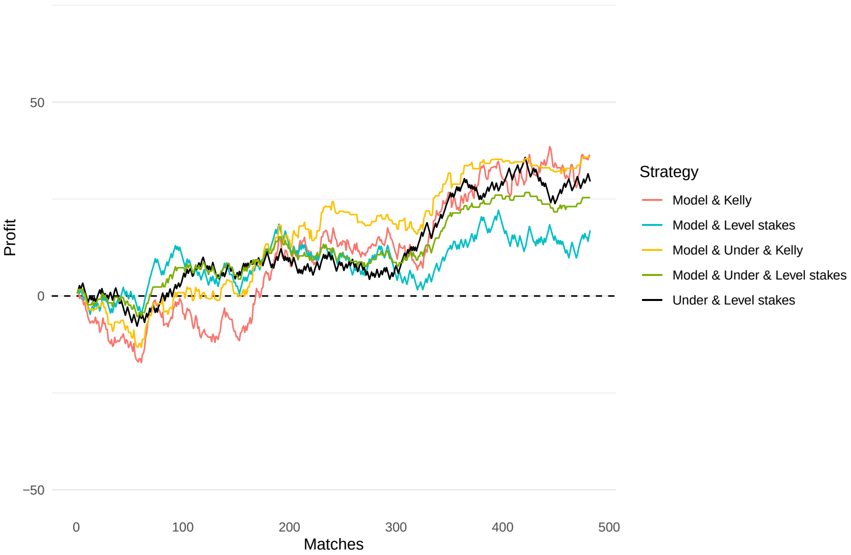


Figure 6.4: Profitability chart for Total Numerical Random Forest on the Premier League test seasons.

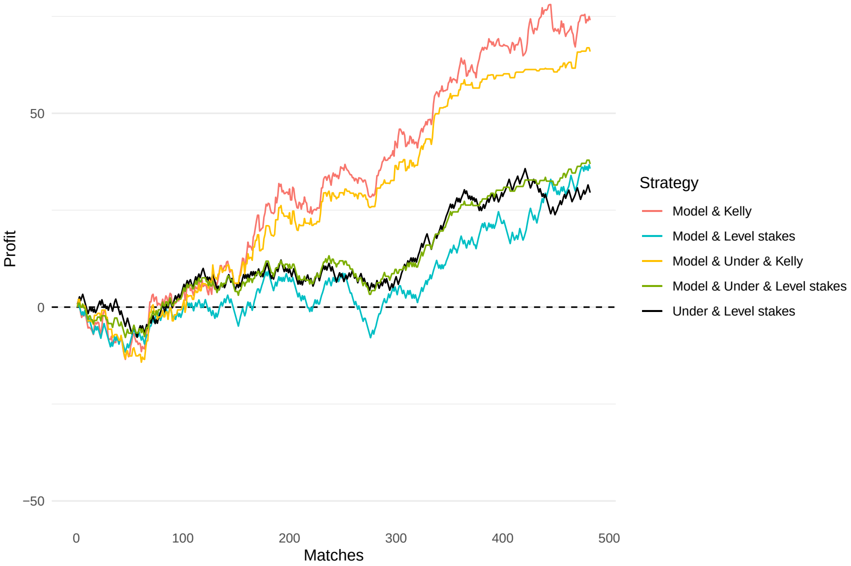


Figure 6.5: Profitability chart for Total Classification Random Forest on the Premier League test seasons.

6.4.3 The Value of F1 and MCC

The model selection is mainly based on the models' F1-score and MCC value, where the F1-score is prioritised. To evaluate whether this is a good decision criterion or not, the total profits for the test data for all betting strategies and models, except the disqualified constant models, are calculated. Total profits are plotted against the F1-score from the test data in order to examine the relationship. The plot and the trend line can be seen in fig. 6.6. The betting strategies are colour coded, which illustrates the profitability spread for the different betting strategies. It can be seen that there is a positive correlation between F1 and profit. The MCC values are also plotted against the profits, see fig. 6.7. There is a positive correlation between MCC and profit as well, although the correlation is weaker. The positive correlations indicate that F1 and MCC are good metrics when choosing a model. The data points in the plots do not follow the line exactly which indicates that there may be additional factors that affect the profit as well.

When investigating the relationship between profit from test data and metrics from model data, the profits and F1-scores are positively correlated while there is a negative correlation for the MCC values. This result contradicts the reasoning above. The negative correlation between profit and MCC values could be a consequence of differences between the test data and model data. The models' performance differs for model data and test data. The differences in performance could depend the low degree of explanatory power in the models. Small differences in the data set could thus affect the model performance. The share of matches in the test data where the total number of corners is less than 10.5 is 55.9% compared to 52.4% in the model data. It is quite small difference, but it affects the F1-score. The profitability is dependent on the odds and which matches have missing odds which could affect the relationship between model performance and profit. Still, F1-score and MCC values are considered as good performance metrics based on the positive correlation with the profit for the same data.

6.5 Bundesliga

In order to evaluate how well the qualified models handle differences between leagues, the models are tested on the Bundesliga. The same procedure of evaluating the models on test data together with a profitability analysis is performed.

6.5.1 Model Testing

The test results from the Bundesliga are generally worse compared to the Premier League, see table 6.6. The MCC values have decreased and turned negative. The majority of the F1-scores are worsened, but the F1-score for the logistic regression model is higher compared to its counterpart in the Premier League result. RMSE, R^2 and MAE are all lower compared to the results in table 6.4.

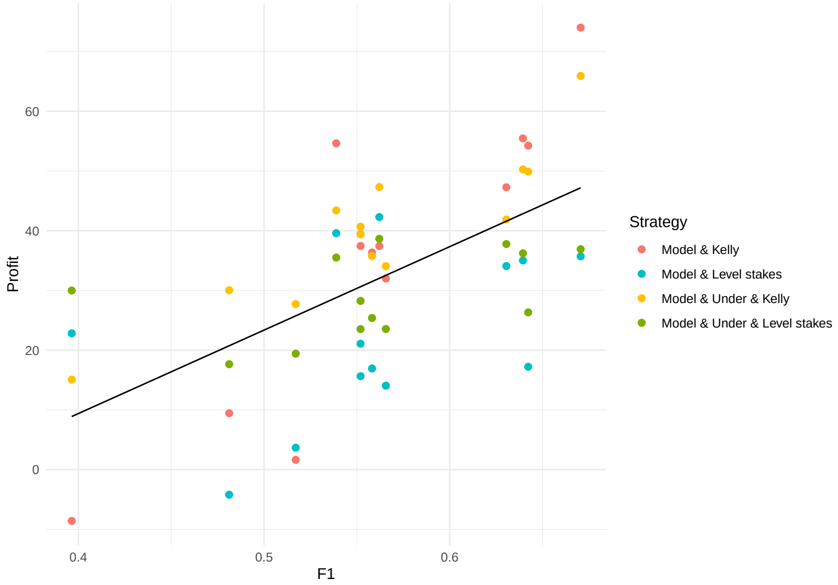


Figure 6.6: Profit plotted against F1-scores for the Premier League test data for all models. The figure also illustrates the profitability spread for model and betting strategy combinations.

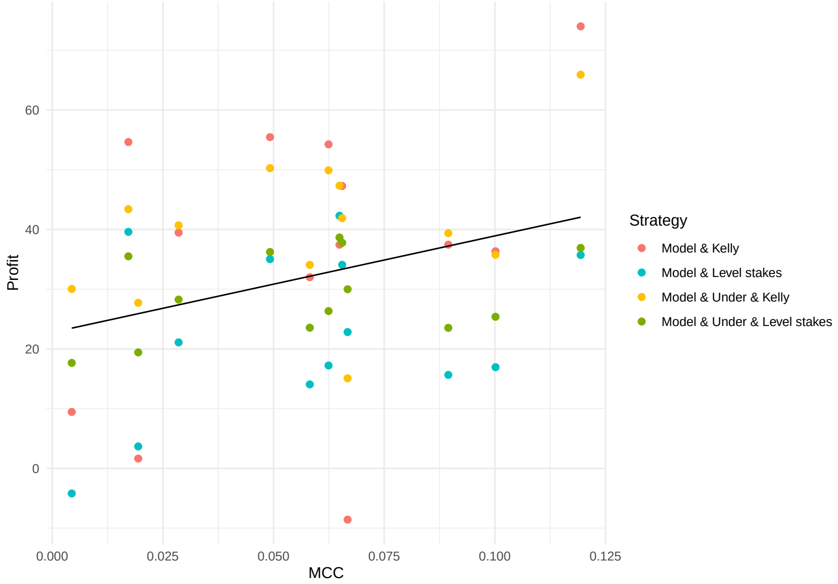


Figure 6.7: Profit plotted against MCC values for the Premier League test data for all models. The figure also illustrates the profitability spread for model and betting strategy combinations.

When the models are compared to each other in table 6.6, the two numerical models perform roughly equally. The classification models have a competitive F1-score but the classification random forest model has a notably low MCC value. Overall, the logistic regression model with backward stepwise selection based on BIC performs best among the statistical methods. The performance metrics of the naive models follow the same pattern as in table 6.4.

Table 6.6: Performance of qualified models for the Bundesliga test seasons.

Model	**F1**	**MCC**	**RMSE**	$\mathbf{R^2}$	**MAE**
Numerical *Naive*	0	-	3.307	-0.072	2.710
Total Numerical NegBin Stepwise AIC	0.439	-0.066	3.445	-0.163	2.788
Classification *Naive*	0.754	-			
Total Classification Logistic Stepwise BIC	0.679	-0.036			
Total Numerical Random Forest	0.423	-0.074	3.460	-0.173	2.829
Total Classification Random Forest	0.657	-0.099			

6.5.2 Profitability Analysis

When the profitability of the models is evaluated on the test data from the Bundesliga, consisting of odds from 441 matches, considerable differences compared to the results obtained from the Premier League data can be found. As can be seen in table 6.7, the majority of the model and betting strategy combinations now generate negative returns. When the models are used with the unconditioned betting strategies they turn out to be markedly unprofitable, with the Kelly criterion applied on numerical random forest generating the largest loss. When the betting strategies conditional on the under odds are applied, the models are more restrictive and only place bets in 9-23% of the matches, compared to 98-99% for unconditional betting strategies, which results in the profit turning positive for some models. The logistic regression model conditional on only betting on the under alternative with level stakes is the most profitable model both with respect to total profit and ROI. However, all models are outperformed in total profitability by consistently betting on the under alternative for all matches. Figure D.1, fig. D.2, fig. D.3 and fig. D.4 in appendix D illustrate how the profit of the models develops over time.

The strategy of betting one unit on the under market generates positive profit, but has lower ROI and a less distinct trend compared to the Premier League. The same statistical test is performed as in section 6.4.1, with data from the same seasons consisting of odds from 589 matches. The p-value is determined to be 0.1909. Hence, the null hypothesis cannot be rejected for the Bundesliga. However, the relatively low p-value for the Bundesliga indicates that the odds may still be somewhat undervalued.

Table 6.7: Profitability analysis on the Bundesliga test seasons.

Model	**Betting Strategy**	**Bets Played (%)**	**Wins (%)**	**Total profit**	**ROI (%)**
	Under	100	63.1	9.0	3.3
Total Numerical NegBin Stepwise AIC	Level stakes	98.5	37.8	-21.2	-7.8
	Kelly criterion	98.5	37.8	-19.3	-7.1
	Level stakes U	8.8	54.2	-0.9	-3.6
	Kelly criterion U	8.8	54.2	1.4	5.7
Total Classification Logistic Stepwise BIC	Level stakes	97.4	39.0	-17.9	-6.7
	Kelly criterion	97.4	39.0	-15.9	-5.9
	Level stakes U	11.7	62.5	2.7	8.4
	Kelly criterion U	11.7	62.5	1.9	5.9
Total Numerical Random Forest	Level stakes	97.8	35.1	-37.9	-14.1
	Kelly criterion	97.8	35.1	-44.9	-16.8
	Level stakes U	10.6	41.4	-7.8	-26.9
	Kelly criterion U	10.6	41.4	-6.5	-22.6
Total Classification Random Forest	Level stakes	98.9	40.6	-17.7	-6.5
	Kelly criterion	98.9	40.6	-21.0	-7.8
	Level stakes U	22.6	58.1	0.1	0.1
	Kelly criterion U	22.6	58.1	1.8	2.9

7 Conclusions and Future Work

7.1 Conclusions

This thesis has aimed to predict the aggregated number of corners of the two teams playing in a football match using numerical and classification statistical learning methods. It has aimed to evaluate the profitability of the created models by using historical odds and applying established betting strategies. In addition, it has been investigated whether the market for over-under corner odds is asymmetric causing market inefficiency.

The results suggest that the models trained on the Premier League data are profitable when tested on new data from the same league. On the contrary, when the models are tested on the Bundesliga they become unprofitable and the degree of explanation decreases. The assessment of the asymmetry for the over-under corner odds market displays a significant undervaluation of the under market in the Premier League, which generates profitability for the test data. The same trend could be seen in the Bundesliga but not as clearly and the undervaluation is not statistically significant, although it generates positive profit for the test data.

Based on the reasoning in model performance evaluation, it is questionable if the processed match-by-match statistics applied on generalised linear regression and tree-based models are sufficiently explanatory to predict the total number of corners in a football match. The match statistics-based models are not significantly better than the naive models. This result suggests that the created models seem not to be sufficient to explain the number of corners. Since the over-under market for corners shows signs of inefficiency, it is nevertheless possible to generate profit and it is clear that the models actually improve the profitability despite having low explanatory power. The winning with the models versus the naive model in the profitability analysis is that the stake can be varied depending on the value of the bet. In general, the profitability for the naive model with level stakes is better than the statistical model with level stakes but worse than the statistical model with varying stakes. Due to the persistent randomness in the models, it is difficult to draw indisputable conclusions regarding which model that should be preferred.

When the models are trained and validated, the conclusion is that generalised linear regression models are favoured, while the model testing shows the opposite. Regarding numerical or classification modelling, the result is vague when evaluating the models on model data but is in favour of classification modelling on the test data. The profitability analysis shows that all qualified models end up being profitable when tested on the Premier League data, no matter if a conditional or unconditional betting strategy is applied. The Kelly criterion is superior as a betting strategy, exhibiting consistently higher total profit and ROI compared to level stakes. Since the share of bets played decreases when using conditional betting strategies the net gain of applying the Kelly criterion conditional to only playing on the under odds is small, but the ROI is significantly higher. The findings also indicate that there is a positive correlation between the models' profit and F1-score, as well as profit and MCC when the metrics are calculated on the same data. Since the most profitable model is not the best-performing model in the model selection, it is not obvious that the model selection should only be based only on the considered evaluation measures. An explanation of why the profitability is consistent across models could be the inefficiency of the over-under market for corner odds in the Premier League. It is possible that if the models were aiming to predict a match statistic connected to a highly efficient odds market, some models could potentially be unprofitable. Except for the market odds implied probabilities, the profitability depends on the commission the betting exchange charges or the margin the bookmaker odds. This means that the possibility of profit could vary between which betting platform the bettor is using.

To summarise, it is difficult to predict the number of corners in a football match. Due to inefficiencies in the over-under corner odds market, profitability can be obtained either by using statistical models, taking advantage of the asymmetry in the over-under corner odds market or combining both parts.

7.2 Future Work

Since the model performance and profit are significantly lower when the models are tested on the Bundesliga, it could be examined in future research whether this modelling framework is successfully applicable by training, testing and evaluating the profitability in another league than the Premier League. It could also be assessed if refitting the models to the Bundesliga with the same variable selection as for the Premier League-trained models could give better model performance and profitability. Another suggestion for future work is to determine the model profitability on model data and use it as a complementing decision criterion together with the existing model evaluation metrics. Furthermore, although modelling the teams' corners separately led to promising results the profitability could not be assessed due to the lack of available historical odds data. This could potentially increase the profitability since the explanatory power of the models is higher.

References

T. Ajadi, A. Clarke, S. Dhillon, G. Gardner, D. Garg, T. Hammond, A. Malcolm, J. Pang, J. Pugh, and D. Jones. *A new dawn.* Annual Review of Football Finance 2022. Deloitte, Aug 2022.

Y. Alfredo and S. Isa. Football Match Prediction with Tree Based Model Classification. *International Journal of Intelligent Systems and Applications*, 11: 20–28, Jul 2019. doi: 10.5815/ijisa.2019.07.03.

R. Baboota and H. Kaur. Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, 35, Mar 2018. doi: 10.1016/j.ijforecast.2018.01.003.

Bet Types. What is arbitrage betting? https://bet-types.com/guide/arbitrage-betting/, Jun 2023. Accessed: 2023-06-02.

Betfair. About Us. https://www.betfair.se/aboutUs/, 2017. Accessed: 2023-05-03.

Betfair. Betfair Charges. https://www.betfair.se/aboutUs/Betfair.Charges/, 2023a. Accessed: 2023-06-13.

Betfair. Placing a Lay Bet. https://betting.betfair.com/how-to-use-betfair-exchange/beginner-guides/placing-a-lay-bet-010819-51.html, 2023b. Accessed: 2023-05-10.

Britannica. FIFA (electronic game series). In *FIFA (electronic game series)*. Encyclopaedia Britannica, Mar 2023. https://www.britannica.com/topic/FIFA-game-series, Accessed: 2023-05-02.

Bundesliga. How did the Bundesliga compare with Europe's other top 5 leagues in 2021/22? https://www.bundesliga.com/en/bundesliga/news/comparison-europe-s-top-5-leagues-premier-league-la-liga-serie-a-ligue-1-18449, Jun 2022. Accessed: 2023-05-03.

D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, Jan 2020. doi: 10.1186/s12864-019-6413-7.

A. Clarke. Season trends: Corners a growing threat. https://www.premierleague.com/news/2638312, Jun 2022. Accessed: 2023-05-03.

A. C. Constantinou and N. E. Fenton. Profiting from arbitrage and odds biases of the European football gambling market. *Journal of Gambling Business and Economics*, 7:41–70, 2013. doi: 10.5750/jgbe.v7i2.630.

M. J. Dixon and S. G. Coles. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 46(2):265–280, 1997. ISSN 00359254, 14679876.

S. A. Fathima, V. P. Sumathi, and S. Sumanth. Data analytics in football sport to identify gaps for the improvement of quality opportunities throughout world-wide teams. *International Journal of Recent Technology and Engineering*, 7(4S2), 2018.

FIFA. *Professional Football Report 2019*. FIFA Professional Football Department and International Centre for Sports Studies, FIFA-Strasse 20 P.O. Box 8044 Zurich Switzerland, first edition, 2019.

FIFA. The Football Landscape. https://publications.fifa.com/en/vision-report-2021/the-football-landscape/, 2021. Accessed 28-04-2023.

FootyStats. Download Soccer / Football Stats Database to CSV and Excel. https://footystats.org/download-stats-csv, Mar 2023. England, Premier League, match stats, seasons 2013/2014 to 2022/2023. Germany, Bundesliga, match stats, seasons 2015/2016 to 2022/2023. Accessed: jan, feb, mar, apr 2023.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1): 1–22, 2010. doi: 10.18637/jss.v033.i01.

D. Glimne. Prevalence of principal forms. In *gambling*. Encyclopedia Britannica, Mar 2023. https://www.britannica.com/topic/gambling, Accessed: 2023-04-28.

T. Hastie, J. Qian, and K. Tay. *An Introduction to glmnet*. Stanford Graduate School of Education, California, USA, 2023.

O. Hultåker. Rapport till Spelinspektionen. Market research, Spelinspektionen, Nov 2022. Reference number S3SEP22.

N. Ismail and A. A. Jemain. Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. *Casualty Actuarial Society*, Forum Winter 2007:103–158, 2007.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer, New York, first edition, 2013. ISSN 1431-875X ISBN 978-1-4614-7137-0 ISBN 978-1-4614-7138-7 (eBook) DOI 10.1007/978-1-4614-7138-7.

J. L. Kelly. A New Interpretation of Information Rate. *The Bell System Technical Journal*, pages 917–926, Jul 1956. Reproduced with permission of AT&T.

A. Liaw and M. Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

A. Liaw, V. Svetknik, C. Tong, and T. Wang. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In F. Rioli, J. Kittler, and T. Windeatt, editors, *Multiple Classifier Systems*, volume 5th International Workshop, MCS 2004, pages 334–343. Springer, Merck & Co., Inc. P.O. Box 2000, Rahway, NJ 07065, USA, 2004.

M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36 (3):109–118, 1982. doi: https://doi.org/10.1111/j.1467-9574.1982.tb00782.x.

Market Decipher. Sports Betting Market Size, Statistics, Growth Trend Analysis and Forecast Report, 2022-2032. *Market Research Report 2022*, Sep 2022. Accessed: 2023-05-03.

MATLAB (R2022b). MATLAB version: 9.13.0, 2022.

B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975. ISSN 0005-2795. doi: https://doi.org/10.1016/0005-2795(75)90109-9.

D. T. Mollenkamp. Money Line Bet. In *Business Leaders, Maths and Statistics.* Investopedia, Dec 2022. https://www.investopedia.com/money-line-bet-5217219, Accessed: 2023-05-02.

M. J. Moroney. *Facts from figures.* Pelican books: A 236. Penguin, 1956.

NCSS. *Chapter 335 Ridge Regression.* NCSS Statistical Software, 329 North 1000 East Kaysville, Utah 84037 USA, 2023.

J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society*, 135(3):370–384, 1972.

J. Oberstone. Comparing Team Performance of the English Premier League, Serie A, and La Liga for the 2008-2009 Season. *Journal of Quantitative Analysis in Sports*, 7(1):1–18, 2011.

Premier League. The Fans. https://www.premierleague.com/this-is-pl/the-fans/686489?articleId=686489, 2023a. Accessed 28-04-2023.

Premier League. Origins. https://www.premierleague.com/history/origins, 2023b. Accessed 28-04-2023.

R Statistical Software (version 4.2.2). R: A Language and Environment for Statistical Computing, 2022.

J. Rollin, P. C. Alegi, B. Joy, R. C. Giulianotti, and E. Weil. football. In *football.* Encyclopedia Britannica, Apr 2023. https://www.britannica.com/sports/football-soccer, Accessed 2023-04-28.

Smarkets. Why do betting odds change? https://help.smarkets.com/hc/en-gb/articles/214559905-Why-do-betting-odds-change-, 2023. Accessed: 2023-05-03.

S. Sohail. The Math Behind Betting Odds and Gambling. In *Trading Skills, Trading Psychology.* Investopedia, Mar 2023. https://www.investopedia.com/articles/dictionary/042215/understand-math-behind-betting-odds-gambling.asp, Accessed: 2023-05-05.

The International Football Association Board. Laws of the Game 22/23. Münstergasse 9, 8001 Zurich, Switzerland, Jul 2022.

E. O. Thorp. Chapter 9 - The Kelly Criterion in Blackjack Sports Betting, and the Stock Market. In S. Zenios and W. Ziemba, editors, *Handbook of Asset and Liability Management*, pages 385–428. North-Holland, San Diego, 2008. ISBN 978-0-444-53248-0. doi: https://doi.org/10.1016/B978-044453248-0.50015-0.

Transfermarkt. Summer Transfer Window, Confederation UEFA. https://www.transfermarkt.com/statistik/transferfenster, May 2023. Transfermarkt GmbH & Co. KG. Wandsbeker Zollstraße 5a. 22041 Hamburg. Accessed: 2023-05-04.

UEFA. Regulations of the UEFA Champions League. https://documents.uefa.com/r/Regulations-of-the-UEFA-Champions-League-2022/23-Online, Aug 2022. Accessed: 2023-05-10.

UEFA. Association Club Coefficients 2022/2023, Country Coefficients. https://www.uefa.com/nationalassociations/uefarankings/, May 2023. Accessed: 2023-05-03.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

M. R. Webber. Over-Under Bet: Definition, Types, and Examples. In *Wealth, Lifestyle Advice.* Investopedia, May 2022. https://www.investopedia.com/over-under-bet-5217714: :text=In, Accessed: 2023-04-28.

E. Wheatcroft. A profitable model for predicting the over/under market in football. *International Journal of Forecasting*, 36(3):916–932, Jan 2020. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2019.11.001.

C. J. Willmott and K. Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30:79–82, Dec 2005.

Q. Yi, R. Groom, C. Dai, H. Liu, and M. Ruano Gómez Ángel. Differences in Technical Performance of Players From The Big Five European Football Leagues in the UEFA Champions League. *Frontiers in Psyhcology*, 10(2738), Dec 2019. doi: 10.3389/fpsyg.2019.02738.

S. Yip, Y. Zou, R. Tsz Hin Hung, and K. Fai Cedric Yiu. Forecasting number of corner kicks taken in association football using overdispersed distribution, 2022.

Appendix

A Tuning Parameter Selection

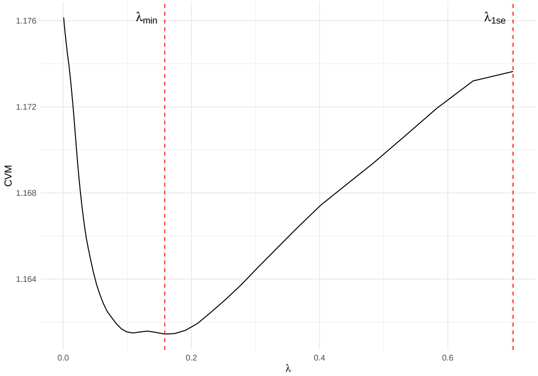


Figure A.1: Mean cross-validation error plotted against tuning parameter λ for modelling total corners with a negative binomial regression.

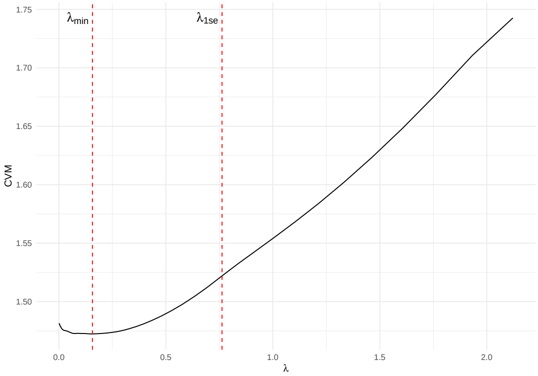


Figure A.2: Mean cross-validation error plotted against tuning parameter λ for modelling home team's corners with a negative binomial regression.

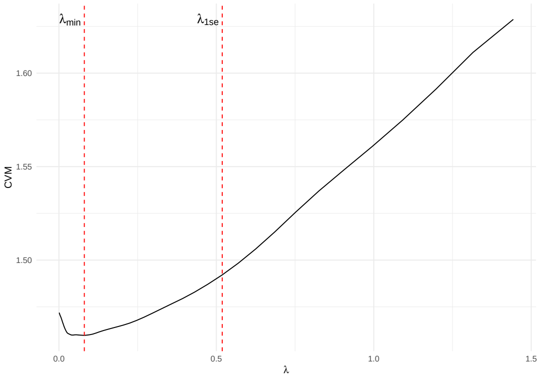


Figure A.3: Mean cross-validation error plotted against tuning parameter λ for modelling away team's corners with a negative binomial regression.

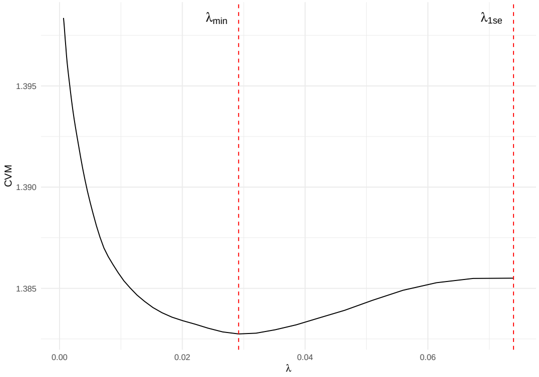


Figure A.4: Mean cross-validation error plotted against tuning parameter λ for modelling total corners with a logistic regression.

B Variable Selection

Table B.1: Abbreviations of average names. The number indicates the length of the moving average. Home and away averages consider home or away matches separately for a team.

Abbreviation	**Average**
A	avg
A3	avg3
A4	avg4
A5	avg5
HA	home avg
HA3	home avg3
HA4	home avg4
HA5	home avg5
AA	away avg
AA3	away avg3
AA4	away avg4
AA5	away avg5

Table B.2: Overview of the variable selection for total negative binomial regression. X indicates an included variable. Abbreviations are stated in table B.1.

Variables	**Total Numerical NegBin**			
	AIC	**BIC**	$\lambda_{\mathbf{min}}$	$\lambda_{\mathbf{1se}}$
game week				
home team ATT				
home team MID				
home team DEF			X	
home team OVR				
away team ATT				
away team MID				
away team DEF				
away team OVR	X	X	X	
odds home team win			X	
odds draw				
odds away team win				
home team motivation			X	
away team motivation				
home team points per game				
away team points per game				
home team goal count				
away team goal count	A4			
home team conceded goal count	A		A3	
away team conceded goal count	A		A	
home team goal count first half	HA	HA	HA	
away team goal count first half				
home team conceded goal count first half				
away team conceded goal count first half				
home team corner count	A4		A4	
away team corner count				
home team conceded corner count	A			
away team conceded corner count	AA			
home team shots on target				
away team shots on target				
home team shots off target				
away team shots off target				
home team fouls				
away team fouls				
home team possession			A3	
away team possession	A			
home team form			A3	
away team form				

Table B.3: Overview of the variable selection for home negative binomial regression. X indicates an included variable. Abbreviations are stated in table B.1.

Variables	**Home Numerical NegBin**			
	AIC	**BIC**	$\lambda_{\mathbf{min}}$	$\lambda_{\mathbf{1se}}$
game week				
home team ATT				
home team MID			X	
home team DEF				
home team OVR				
away team ATT				
away team MID			X	
away team DEF				
away team OVR				
odds home team win	X	X	X	X
odds draw				
odds away team win	X	X	X	X
home team motivation	X			
away team motivation	X		X	
home team points per game			A	
away team points per game	A		A	
home team goal count				
away team goal count	A			
home team conceded goal count	A		A	
away team conceded goal count			A	
home team goal count first half				
away team goal count first half	A		A	
home team conceded goal count first half				
away team conceded goal count first half				
home team corner count	A	A	A	A
away team corner count			AA	
home team conceded corner count			A	
away team conceded corner count	A		A	A
home team shots on target				
away team shots on target				
home team shots off target	HA		HA	
away team shots off target	AA		AA	
home team fouls				
away team fouls				
home team possession				
away team possession	AA	AA	AA	AA
home team form			A5	
away team form				

Table B.4: Overview of the variable selection for away negative binomial regression. X indicates an included variable. Abbreviations are stated in table B.1.

Variables	**Away Numerical NegBin**			
	AIC	**BIC**	$\lambda_{\mathbf{min}}$	$\lambda_{\mathbf{1se}}$
game week				
home team ATT				
home team MID				
home team DEF				
home team OVR				
away team ATT				
away team MID	X			
away team DEF	X			
away team OVR				
odds home team win			X	X
odds draw				
odds away team win	X	X	X	X
home team motivation			X	
away team motivation			X	
home team points per game				
away team points per game	AA		AA	
home team goal count				
away team goal count			A	A
home team conceded goal count			A	
away team conceded goal count				
home team goal count first half				
away team goal count first half	A5	A5	A5	
home team conceded goal count first half				
away team conceded goal count first half				
home team corner count	A			
away team corner count	AA	AA	AA	AA
home team conceded corner count	A	A	A	A
away team conceded corner count			A	
home team shots on target				
away team shots on target				
home team shots off target	HA		HA	
away team shots off target	A		A	
home team fouls				
away team fouls	AA4		AA4	
home team possession	A		A	
away team possession			A3	
home team form			A3	
away team form				

Table B.5: Overview of the variable selection for logistic regression. X indicates an included variable. Abbreviations are stated in table B.1.

Variables	**Total Classification Logistic**			
	AIC	**BIC**	$\lambda_{\mathbf{min}}$	$\lambda_{\mathbf{1se}}$
game week				
home team ATT				
home team MID				
home team DEF			X	
home team OVR			X	
away team ATT				
away team MID				
away team DEF	X			
away team OVR			X	
odds home team win			X	
odds draw				
odds away team win			X	
home team motivation				
away team motivation				
home team points per game				
away team points per game				
home team goal count				
away team goal count				
home team conceded goal count				
away team conceded goal count				
home team goal count first half				
away team goal count first half				
home team conceded goal count first half				
away team conceded goal count first half				
home team corner count	A	A	A	
away team corner count				
home team conceded corner count	A5			
away team conceded corner count				
home team shots on target				
away team shots on target	A			
home team shots off target				
away team shots off target				
home team fouls				
away team fouls				
home team possession				
away team possession	A	A	A	
home team form			A5	
away team form				

Table B.6: Overview of total numerical random forest model and variable importance.

Total Numerical Random Forest	
Included variables	**Importance**
away team avg away shots off target	1214.55
away team avg goal count first half	1043.93
home team avg home goal count first half	1005.10
home team avg conceded corner count	980.96
away team avg4 away fouls	975.74
away team avg3 possession	967.18
home team avg points per game	967.10
home team avg home shots on target	960.16
away team avg conceded goal count first half	942.52
home team avg5 possession	895.28
odds home team win	892.00
home team avg4 corner count	880.20
away team avg4 away corner count	871.46
home team avg3 fouls	844.10
away team avg points per game	818.60
away team avg conceded goal count	818.57
away team avg3 away shots on target	776.25
away team avg3 away conceded corner count	775.91
away team avg5 form	775.60
odds draw	744.03
game week	729.87
home team avg3 shots off target	706.49
home team avg3 form	668.22
away team avg4 goal count	567.38
odds away team win	544.10
home team avg3 conceded goal count	513.11
away team ATT	477.56
home team avg3 goal count	476.16
away team DEF	410.70
home team ATT	396.89
home team avg3 conceded goal count first half	394.15
home team DEF	364.58
home team MID	328.40
away team MID	313.03
away team OVR	297.53
home team OVR	238.35
home team motivation	37.00
away team motivation	32.88

Table B.7: Overview of home numerical random forest model and variable importance.

Home Numerical Random Forest	
Included variables	**Importance**
odds home team win	3006.08
odds away team win	1024.91
away team avg conceded corner count	731.48
home team avg home shots off target	720.99
away team avg away shots off target	705.76
away team avg away corner count	688.22
away team avg goal count first half	687.66
away team avg away possession	645.83
away team avg shots on target	640.63
home team avg corner count	630.46
home team avg conceded goal count first half	622.81
home team avg3 home fouls	615.54
odds draw	611.83
home team avg conceded goal count	607.42
home team avg conceded corner count	597.60
away team avg conceded goal count first half	585.11
home team avg shots on target	574.93
away team avg4 away fouls	551.04
home team avg goal count first half	546.78
away team avg goal count	540.54
home team avg points per game	526.70
home team avg5 form	514.35
away team avg conceded goal count	512.70
away team avg5 form	504.67
away team avg points per game	501.75
home team avg goal count	498.00
home team avg possession	492.96
game week	478.97
home team ATT	257.22
away team DEF	245.37
away team ATT	238.18
home team MID	186.32
home team DEF	176.71
away team MID	165.19
home team OVR	163.73
away team OVR	143.69
away team motivation	37.94
home team motivation	19.62

Table B.8: Overview of away numerical random forest model and variable importance.

Away Numerical Random Forest	
Included variables	**Importance**
odds home team win	1120.10
odds away team win	826.88
home team avg conceded corner count	690.92
away team avg away corner count	579.38
away team avg shots off target	565.96
away team avg goal count	554.46
away team avg conceded corner count	548.41
home team avg possession	504.64
home team avg conceded goal count	493.02
away team avg3 possession	492.37
home team avg corner count	483.94
away team avg conceded goal count first half	464.11
home team avg3 fouls	463.12
home team avg home shots off target	462.37
home team avg goal count first half	460.97
away team avg shots on target	457.55
away team avg4 away fouls	456.35
odds draw	452.08
home team avg conceded goal count first half	448.16
home team avg shots on target	439.62
away team avg5 form	426.68
away team avg away points per game	420.52
home team avg points per game	413.53
away team avg conceded goal count	402.11
home team avg goal count	386.24
game week	378.87
home team avg3 form	366.46
away team avg5 goal count first half	269.20
away team MID	198.59
away team ATT	186.87
home team ATT	179.12
home team MID	166.59
away team DEF	149.65
home team DEF	131.15
away team OVR	116.78
home team OVR	113.92
away team motivation	15.45
home team motivation	13.12

Table B.9: Overview of classification random forest model and variable importance.

Total Classification Random Forest	
Included variables	**Importance**
home team avg corner count	23.65
home team avg home fouls	23.12
away team avg fouls	21.20
away team avg possession	20.45
away team avg away corner count	20.27
away team avg away shots off target	19.93
home team avg shots off target	19.44
home team avg conceded goal count	18.85
home team avg conceded goal count first half	18.57
home team avg3 possession	18.53
away team avg away conceded goal count first half	17.97
home team avg shots on target	17.89
away team avg away goal count	17.54
home team avg home points per game	17.24
away team avg conceded corner count	17.24
home team avg home goal count	17.18
home team avg5 form	17.04
away team avg conceded goal count	17.04
home team avg5 conceded corner count	17.03
away team avg shots on target	16.77
game week	15.73
away team avg points per game	15.30
away team avg3 form	14.69
odds home team win	14.15
odds draw	13.49
odds away team win	10.42
away team avg3 away goal count first half	8.74
home team avg3 goal count first half	8.67
home team ATT	7.39
away team ATT	6.49
home team DEF	6.08
home team MID	5.79
away team DEF	5.41
away team MID	4.97
home team OVR	4.29
away team OVR	3.91
away team motivation	0.99
home team motivation	0.84

C Modelling Home and Away

Table C.1: Model evaluation statistics for cross-validated data, numerical modelling of home and away teams' corners separately.

Model	**RMSE**	$\mathbf{R^2}$	**MAE**
Home Numerical NegBin Stepwise AIC	2.873	0.163	2.275
Home Numerical NegBin Stepwise BIC	2.878	0.160	2.275
Home Numerical NegBin λ_{min}	2.881	0.159	2.279
Home Numerical NegBin λ_{1se}	2.876	0.161	2.275
Away Numerical NegBin Stepwise AIC	2.519	0.115	2.015
Away Numerical NegBin Stepwise BIC	2.527	0.110	2.021
Away Numerical NegBin λ_{min}	2.529	0.108	2.026
Away Numerical NegBin λ_{1se}	2.526	0.110	2.022
Home Numerical Random Forest	2.916	0.138	2.304
Away Numerical Random Forest	2.528	0.109	2.025

D Profitability on the Bundesliga

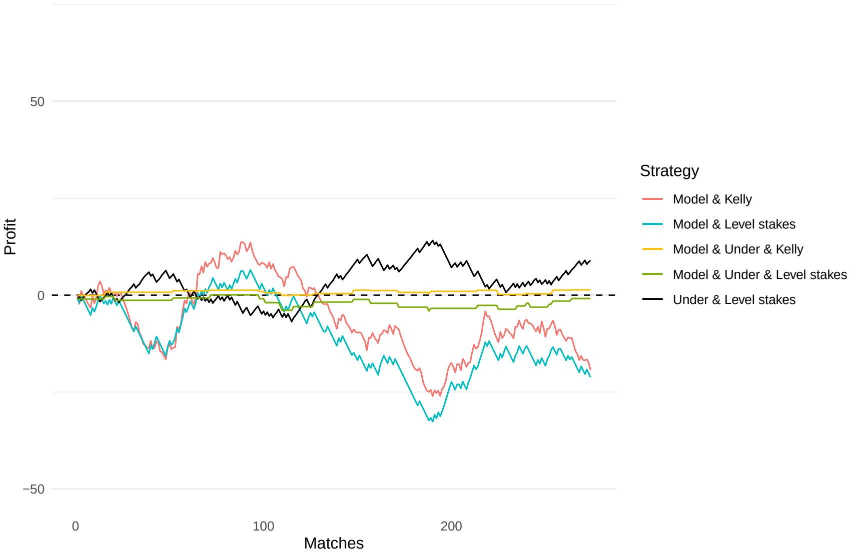


Figure D.1: Profitability chart for Total Numerical NegBin Stepwise AIC on the Bundesliga test seasons.

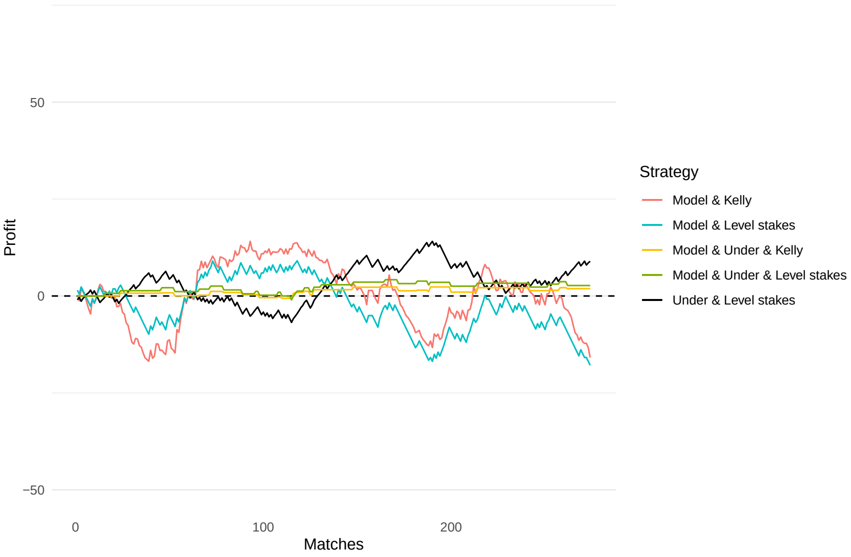


Figure D.2: Profitability chart for Total Classification Logistic Stepwise BIC on the Bundesliga test seasons.

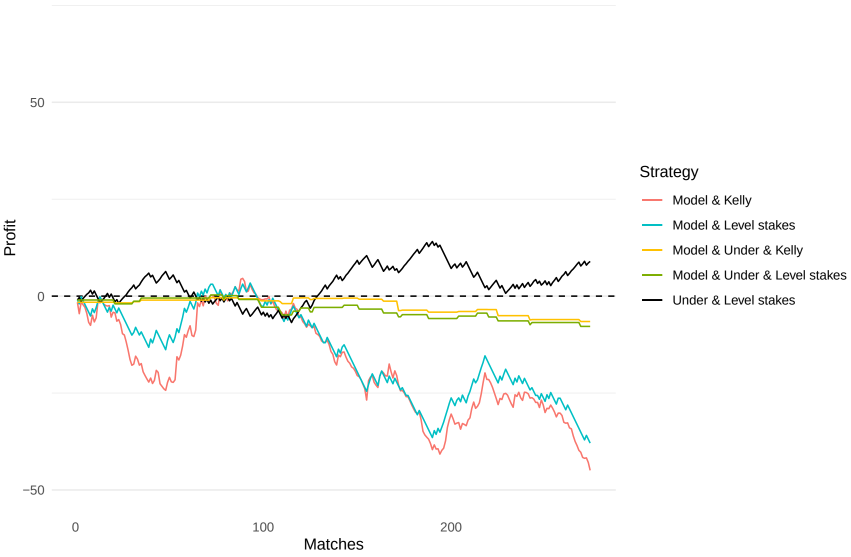


Figure D.3: Profitability chart for Total Numerical Random Forest on the Bundesliga test seasons.

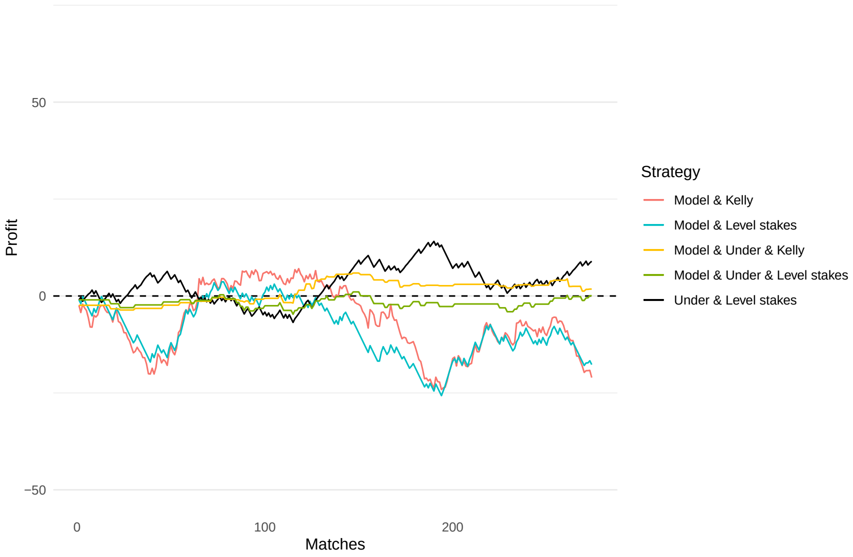


Figure D.4: Profitability chart for Total Classification Random Forest on the Bundesliga test seasons.

Master’s Theses in Mathematical Sciences 2023:E33
ISSN 1404-6342

LUTFMS-3476-2023

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

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