A Descriptive Analysis of Football Matches using Logistic Regression

KTH Bachelor Thesis Report

OSCAR GRANKVIST, IVAN-EDVARD BERGMAN
Abstract

The aim of this study was to explore how match-related statistics contribute to winning association football matches. This is relevant for stakeholders in the football industry to facilitate the understanding of what factors contribute to winning matches and can thus be of use when formulating match tactics. A model was constructed through the use of binary logistic regression, where winning/not winning was used as the response variable, and standardized match-related statistics were used as predictor variables. Using the acquired coefficients, it was concluded that, among other variables, the home advantage and the ability of a team to finish on target has a strong correlation with winning games. Further, the study explores the impact of a team’s ability to win football games on the financial landscape of the modern football world. The results show that some of the examined statistics are well correlated to winning a match, but that the tactical useability of these insights is low.

Keywords: Football, Premier League, Logistic Regression, Match-related statistics
Sammanfattning


Nyckelord: Fotboll, Premier League, Logistisk regression, Matchrelaterad Statistik
## Contents

1 Introduction 6  
   1.1 Background 6  
   1.2 The Aim of the Study 7  
   1.3 Purpose and Scope 7  
   1.4 Current Literature 8  
   1.5 Source of Data 8  

2 Methods 9  
   2.1 Binary Logistic Regression 9  
   2.2 Data Handling and Variable Selection 11  
   2.3 Model Validation 14  

3 Results 16  

4 Discussion 18  
   4.1 Reliability of the results 18  
   4.2 Validity of the Results 18  
   4.3 What Conclusions can be Drawn? 19  
   4.4 Economic Gains 20  

References 22
1 Introduction

1.1 Background

Football, or soccer, is the biggest sport in the world both in terms of viewing numbers and total turnover. The 2022 World Cup final, which saw France taking on Argentina, is estimated to have been seen by 1.5 billion people [1], placing it among the biggest sports events to date. The English Premier League is among the biggest national football leagues in the world, being broadcast in 188 countries and accumulating 3.2 billion viewers throughout the 2018/2019 season [2]. The league, in turn, contributes to English society, providing 94,000 jobs and £3.6 billion in direct tax revenue [3].

For a football club, the primary way to gain a following of fans is to win football games. The rise of social media during the last decade has led to more marketing opportunities for modern-day football clubs. Still, ultimately it’s through winning football games a club can retain its relevancy. Winning football games is thus an integral part of the clubs’ value-creation process. The primary sources of income for a football club are broadcast rights, matchday income, tournament prize money, player sales, and merchandise sales. The revenue from the broadcast rights and the prize money is distributed based on league position, and player sales and merchandise sales are dependent on overall team performance, thus the sources of income depend on the teams’ ability to win their games.

The use of data analysis and statistics has become more widespread among all tiers in modern football, being used both to analyze games and to aid in player scouting. This usage has increased over an almost 20-year period, with the old preconception that football is too fluid for the appliance of data-driven aid being disproven [4]. An abundance of statistics is collected electronically for every present-day Premier League game, ranging from the partaking teams’ amount of possession (the percentage of the game the team is in possession of the football) to expected goals (the number of goals the team would be expected to score during a game given the accumulated goal-scoring chances). Understanding and making use of this data gives teams an extra edge in a sport where small margins often are the deciding factor.

With the recent prominence of more data usage, there exists a reluctance to rely too heavily on these sorts of analytics. Expressions like "the ball is round" refers to the fact that a game of football includes a certain amount of unpredictability and that circumstances can change momentarily. In the 1999 Champions League final, which saw Manchester United FC face FC Bayern Munchen, Bayern attained a 1-0 lead in the first half and the game was all but over when Manchester United scored two goals during the three extra-time minutes of the game, winning, what is considered by many to be, the most prestigious club competition in the world. This illustrates the unpredictability of football and suggests that data analysis, however comprehensive, will never solely be able to explain a result.
1.2 The Aim of the Study

The aim of this study is to map match-related statistics to the result of the match, or in other words, to examine which elements of the game can be attributed to winning a match. Examples of match-related statistics are:

- The number of corner-kicks that are taken by either team.
- The number of fouls committed by either team.
- Whether a team is playing at home or away.

In practice, this will be accomplished through the use of binary logistic regression, where the response variable is either to win or not to win, and the predictor variables are the aforementioned match-related statistics. The aim is thus to identify the underlying factors of a win, in the form of a descriptive study.

1.3 Purpose and Scope

The results attained from the study are meant to supply a basic understanding of what attributes are important to win football games. General trends in games may be identified and this may be used as a basis upon which to form general team tactics.

It is to be considered that a football game includes a stochastic element, which makes results hard to predict. Because of this fact, it is unreasonable to expect the model to identify causations with perfect certainty. The results from the study could be used by stakeholders in the football industry to evaluate team tactics or identify trends in the Premier League. In extension, the results could to some extent be used to predict games, by observing how opposing teams average in the statistics brought forth by the model. Using the model alone within this domain would however be unreliable, as each match plays out differently and teams might take measures to prohibit their opponents from performing their strategy.

The model was constrained to matches within the Premier League, the results could however be used to assess matches within other leagues, but it should be noted that tactics vary among different leagues and that this fact may have an effect on the useability of the results. Questions to be handled are:

(i) Are there any identifiable causalities between match statistics and the outcome of the match itself?

(ii) Are there any noteworthy and unexpected trends to be observed in the Premier League?

(iii) Could these insights be utilized in a viable manner when forming or adjusting team strategy?

(iv) How can a club benefit economically from insights regarding how to win football games?
1.4 Current Literature

With the prominence of data analysis within the football industry over the last decades there exists an abundance of studies that have explored how data can be used to describe and predict football games.

Data analysis is also commonly used to assess individual football players, as an aid in player scouting, for example. Wyscout is an often-used tool among professionals in the industry, both as an aid in player scouting and match preparation, covering 550,000 players and teams at different professional levels [5].

Logistic Regression has previously been used in order to assess football games, both with descriptive and predictive aims, however, with the recent breakthroughs within the field of neural networks, there exists a more extensive methodology for achieving higher accuracy, because of the neural networks’ capability to identify more innate relationships between circumstances related to the football matches [6].

1.5 Source of Data

The dataset [7] used contains information for every Premier League game played between 2010 and 2020, this amounts to 4070 individual observations. The assumption is made that these games and their outcomes are independent. This assumption is necessary for the implementation of the model, but it could be argued that factors like the current form of playing teams could play a part in the result of their match.

Each observation includes 114 different variables, including season, date, teams involved, end-result and also clearances, corners, fouls and cards, offsides, passes, possession, touches, shots on- and off-target, and tackles for each respective team. There are also average team statistics over the entire season available.
2 Methods

2.1 Binary Logistic Regression

Logistic Regression is a type of classification algorithm that aims to explain a relationship between a binary outcome (referred to as a response variable) and some random variables (which are referred to as regressors).

In practice, one denotes the response variable by

\[ Y \in \{0, 1\}, \]

where \( Y = 1 \) denotes a team winning and \( Y = 0 \) denotes losing or drawing in the case of this study. Assuming \( Y \) follows the Bernoulli distribution, that is \( Y \equiv Be(p) \), one gets the probability mass function

\[ P(Y = 1) = p, \quad P(Y = 0) = 1 - p, \quad p \in (0, 1). \] (1)

The definition of expected value for discrete random variables is

\[ E[Y] = \sum_{i=1}^{Y_i} P(Y_i) \]

(1) is utilized to rewrite this as

\[ E[Y|x] = 0 \cdot P(Y = 0|x) + 1 \cdot P(Y = 1|x) = P(Y = 1|x). \] (2)

An expression for \( P(Y = 1) \) is to be found so that \( Y \) is said to follow a logistic regression.

A function \( g \) is said to be a link function if

\[ g(E[Y|x]) = x^T \beta, \] (3)

where \( x \) is a vector containing values of the regressors and \( \beta \) is a vector of corresponding coefficients. In practice, \( x \) will contain an extra variable (referred to as the indicator variable), that takes the value of 1, for all samples. The reason for the addition of this variable is to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Thus, if the model takes \( k \) regressors into account, the \( x \) and \( \beta \) vectors will contain \( k + 1 \) elements. This is

\[ x^T \beta = \sum_{i=0}^{k} x_i \beta_i. \]

With regards to the case of the study, (1) can be rewritten as

\[ P(Y = y) = p^y(1 - p)^{1-y} = e^{\ln \left( \frac{p}{1-p} \right) y + \ln (1-p)}. \]

The fraction \( \frac{p}{1-p} \) from the above expression is known as the odds of success. By taking the logarithm of the odds of success one receives the logit function
\[ \theta := \text{logit}(p) = \ln\left(\frac{p}{1-p}\right), \quad (4) \]

and by inverting (4) one gets

\[ p = \text{logit}^{-1}(\theta) = \frac{e^\theta}{1 + e^\theta}, \quad (5) \]

(5) is rewritten and denoted by \(\sigma(\theta)\)

\[ \sigma(\theta) := \frac{1}{1 + e^{-\theta}}, \quad (6) \]

which is called the logistic function. With regards to (3), one chooses the link function \(g(x) = \text{logit}(x)\), getting

\[ \text{logit}(E[Y|x]) = x^T \beta \]

\[ \iff E[Y|x] = \sigma(x^T \beta) = \frac{1}{1 + e^{-x^T \beta}}, \]

and by (2) that is

\[ P(Y = 1|x) = \frac{1}{1 + e^{-x^T \beta}}. \quad (7) \]

Since the response variable \(Y\) is binary one also gets

\[ P(Y = 0|x) = 1 - P(Y = 1|x) = 1 - \frac{1}{1 + e^{-x^T \beta}} = \frac{e^{-x^T \beta}}{1 + e^{-x^T \beta}}. \]

Fitting the logistic regression model is done by adjusting the values contained within the coefficient vector \(\beta\) with regards to maximizing the likelihood function

\[ L(\beta) := \prod_{i=1}^{m} P(y_i|x_i)^{y_i} (1 - P(y_i|x_i))^{1-y_i}, \quad (8) \]

on a dataset of \(m\) observations. For simplicity of computation, one analyzes the negative log-likelihood function, given by

\[ -l(\beta) = -\ln(L(\beta)) \]

\[ = - \sum_{i=1}^{m} (y_i \ln(P(y_i|x_i)) + (1 - y_i) \ln(1 - P(y_i|x_i))) \]

\[ = - \sum_{i=1}^{m} (y_i \ln(\sigma(x^T \beta)) + (1 - y_i) \ln(1 - \sigma(x^T \beta))) \quad (9) \]

and the maximum-likelihood estimates for \(\beta\) are obtained through minimization of (9), or in particular, setting its partial derivatives with regards to \(\beta\) to zero:

\[ \frac{\delta}{\delta \beta_0} (-l(\beta)) = -y_i (1 - P(y_i|x; \beta)) = 0 \]

\[ \frac{\delta}{\delta \beta_k} (-l(\beta)) = -y_i x_i^T (1 - P(y_i|x; \beta)) = 0 \]
With these calculations, one receives optimal values of \textit{coefficient estimates} \( \beta \) (referred to as \( \hat{\beta}_{ML} \)) with regards to the \textit{maximum likelihood} of the model. The \textit{maximum likelihood estimate} is then assessed with regards to the \textit{goodness of fit} of a model in relation to other models, (with \textit{goodness of fit} referring to how well a model fits a set of data). In practice this is done through the use of the \textit{Akaike information criterion} (\textit{AIC-statistic}), defined as

\[
AIC := 2k - 2\ln(\hat{L}),
\]  

with a lower \textit{AIC} value indicating a better model. Here \( k \) refers to the number of \textit{regressors} used within the model and \( \hat{L} \) refers to the aforementioned maximized \textit{likelihood function}. The \textit{AIC-statistic} prioritizes minimizing information loss while disfavoring models with a large number of \textit{regressors}.

### 2.2 Data Handling and Variable Selection

Before fitting any model to a dataset, it is beneficial to review the data for several different reasons:

- \textit{Face validity}: with the goal of the analysis in mind, one should hypothesize that there is some causality between the data and what one aims to describe. In other words: the data is relevant within the domain of the study.

- To find outliers and faulty data points, such as duplicates or missing data.

- To handle possible multicollinearity among variables in the dataset.

All of the aforementioned facts contribute to the level of \textit{reliability} of the study and neglect of these aspects may be detrimental to the overall quality of the study.

The following \textit{regressor variables} were examined in the initial model:
<table>
<thead>
<tr>
<th>Statistic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total goals</td>
<td>- The total number of goals scored in the match.</td>
</tr>
<tr>
<td>Clearances</td>
<td>- The difference between the number of clearances of team 1 and team 2 respectively. A clearance is when a defending team kicks the ball away from their side of the pitch.</td>
</tr>
<tr>
<td>Corners</td>
<td>- The difference between the number of corners taken by team 1 and team 2 respectively.</td>
</tr>
<tr>
<td>Conceded fouls</td>
<td>- The difference between the number of committed fouls by team 1 and team 2 respectively.</td>
</tr>
<tr>
<td>Offsides</td>
<td>- The difference between the number of times a teammate has played another into an offside position by team 1 and team 2 respectively.</td>
</tr>
<tr>
<td>Passes</td>
<td>- The difference between the number of passes completed by team 1 and team 2 respectively.</td>
</tr>
<tr>
<td>Possession</td>
<td>- The difference in possession between the playing teams.</td>
</tr>
<tr>
<td>Touches</td>
<td>- The difference between the number of touches taken by team 1 and team 2 respectively. Includes all touches.</td>
</tr>
<tr>
<td>Yellow Cards</td>
<td>- The difference between the number of yellow cards received by team 1 and team 2 respectively.</td>
</tr>
<tr>
<td>Red cards</td>
<td>- The difference between the number of red cards received by team 1 and team 2 respectively.</td>
</tr>
<tr>
<td>Shots taken</td>
<td>- The difference between the number of shots taken by team 1 and team 2 respectively.</td>
</tr>
<tr>
<td>Clinical</td>
<td>- A measurement of shot accuracy, a fraction is calculated for each team by ( \frac{\text{shots on goal}}{\text{shots taken overall}} ). ( \text{clinical-diff} ) is the difference between this metric of team 1 and team 2.</td>
</tr>
</tbody>
</table>

Table 1: Description of Match Statistics

Multicollinearity is a phenomenon in which a regressor can be linearly predicted by other regressors to a certain degree of precision. The phenomenon is problematic as it increases the variance of the affected coefficients’ estimates, thus complicating the goal of estimating the influence of individual regressors on the response.

The following graph (1) depicts covariation for every pair of regressors included in the initial model.
The aim is for all regressors included in a model $M$ to have a covariation that is less than 0.7, that is

$$|\text{cov}(x_i, x_j)| < 0.7, \forall i, j \in M \mid i \neq j$$

As can be observed in the graph, passes, possession and touches are strongly correlated to each other, therefore passes and touches were removed from the model. The remaining regressors are kept in the model.

**Occam’s razor** is a problem-solving principle that states that better models are often-times the more simple ones. For the purposes of this study, reducing the number of regressors facilitates the understanding of the causalities that contribute to a win. The **Stepwise AIC method** was utilized in the variable selection, wherein the AIC-statistic (10) is calculated for different sub-models, and the sub-model with the lowest AIC is returned and chosen as the optimal model.

**Total goals** was also removed from the model, this was done to reduce bias, as it became apparent that the model remunerated more goals because of the fact that if no goals were scored in a match, no team could possibly win. Further, every team undeniably aims to score goals, so including "to score more goals" among the match-winning factors appears superfluous and of inconsequential significance.

**Standardization** of the regressors was utilized to aid in the comparison of the significance of the regressors. This is because of the different units of the regressors included in the model - through standardization one can compare the relative size of the corresponding regressor coefficients. The formula for standardizing a variable is

$$z_i = \frac{x_i - \mu_x}{\sigma_x},$$

where $z_i$ is the standardized data point, $x_i$ the data point and $\mu_x$ and $\sigma_x$ are the mean of $x$ and the standard deviation of $x$ respectively.

In this study, three models were derived based on different datasets:
• **Model A** is fitted on *standardized* data from the perspective of the home team. Meaning that the home/away team advantage is incorporated in the coefficient of the *indicator variable*. The *stepwise AIC method* was used to select the *regressors* used within the model.

• **Model B** consists of two *sub-models*, one fitted on *standardized* data from the perspective of the home team and one fitted on *standardized* data from the perspective of the away team. The *stepwise AIC method* was used to select the *regressors* used within the model. This model allows for observing eventual differences between the influence of match-related statistics from both perspectives.

• **Model C** is fitted on a combined *standardized* dataset, consisting of data from the away team perspective added onto the data from the home team perspective. In practice, this means the dataset used for *model C*, includes twice the number of data points of *model A*, and in addition, pairwise duplicates of every datapoint: one datapoint of *team 1* facing *team 2*, and another of *team 2* facing *team 1*. This model includes an *home/away team regressor*, allowing for conclusive measurement of the influence of this advantage. The *stepwise AIC method* was used to select the *regressors* used within the model.

### 2.3 Model Validation

*Statistical model validation* is the process of evaluating whether a statistical model is appropriate for the purposes of the study. For *regression models* a convenient way to validate the model is through *cross-validation*. *Cross-validation* consists of three steps:

(i) Splitting the dataset into two samples: a *training-sample* and a *test-sample*.

(ii) Fitting the model utilizing the *training-sample*.

(iii) Validating the model on the *test-sample*.

To take these measures means one alleviates the risk of model *overfitting* meaning that the model is effective on general data, on which it has not been fitted.

In the case of this study, the dataset was split randomly into *training-sample* and *test-sample* with proportions 0.7 : 0.3. The *model validation* was, in part, performed with the use of *confusion matrices*. 

---

14
The aim is to classify the data points in the test sample using the derived model and, based on the prediction of the class and their actual class, categorize them within the confusion matrix. A handful of statistics can then be derived from the table: (True negative = TN, False positive = FP, False negative = FN, True positive = TP)

- **Accuracy** = \( \frac{TN + TP}{TN + FP + FN + FP} \) (overall accuracy of the model)
- **Precision** = \( \frac{TP}{TP + FP} \) (when predicted class = 1, how often does the model classify correctly?)
- **Sensitivity** = \( \frac{TP}{TP + FN} \) (when actual class = 1, how often does the model classify correctly?)
- **Specificity** = \( \frac{TN}{TN + FP} \) (when actual class = 0, how often does the model classify correctly?)

Model validation was also conducted using the receiver operating characteristic curve (hence referred to as ROC curve). The ROC curve is an illustrative plot describing the diagnostic ability of a binary classifier as its discriminant threshold is varied. The curve is generated by plotting the sensitivity against the probability of false alarm (which is 1 − sensitivity). A well-fit model is indicated by the integral of the ROC curve being close to 1.

**McFaddens’s pseudo-R squared** is defined as

\[
R^2_{McFadden} = 1 - \frac{\log(L_c)}{\log(L_0)},
\]

where \( L_c \) denotes the maximized likelihood value for the fitted model and \( L_0 \) denotes the corresponding value for the null model - a model with an intercept and no regressors. A higher McFaddens’s pseudo-R squared indicates a more well-fitted model.
3 Results

The three suggested models presented in section 2.2 had their regressors reduced with regards to their AIC through the step-AIC method. Through the use of cross validation, it was concluded that results of Model A, trained on data for the home team solely, and Model B, using the two sub-models, were superior to those of Model C. With the intention to better quantify and analyze the effect of the home-advantage, Model B was chosen in order to observe potential differences in trends of teams playing at home and teams playing away. The two sub-models are hereby denoted by the home-model and away-model respectively and all models presented have had their variables reduced using the step AIC method, meaning that variables have been included/excluded with regards to the AIC statistic of the models.

Figure 2: The confusion matrices for the home- and away-models respectively.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Home-model</th>
<th>Away-model</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>2871.6</td>
<td>2501.5</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.7549</td>
<td>0.7932</td>
</tr>
<tr>
<td>Precision</td>
<td>0.7400</td>
<td>0.7216</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.7949</td>
<td>0.9050</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.7065</td>
<td>0.5455</td>
</tr>
<tr>
<td>$R^2_{McFadden}$</td>
<td>0.2715</td>
<td>0.2812</td>
</tr>
</tbody>
</table>

Table 2: Statistics relevant to model validation.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Home-model</th>
<th>Away-model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.30911</td>
<td>-1.27852</td>
</tr>
<tr>
<td>Clearances</td>
<td>0.97166</td>
<td>0.89325</td>
</tr>
<tr>
<td>Possession</td>
<td>0.16467</td>
<td>0.22193</td>
</tr>
<tr>
<td>Yellow cards</td>
<td>-0.08364</td>
<td>-0.10513</td>
</tr>
<tr>
<td>Red cards</td>
<td>-0.25428</td>
<td>-0.32423</td>
</tr>
<tr>
<td>Shots</td>
<td>1.01053</td>
<td>1.00585</td>
</tr>
<tr>
<td>Clinical</td>
<td>0.94611</td>
<td>1.07588</td>
</tr>
</tbody>
</table>

Table 3: Variable selection and the standardized variable coefficient magnitudes of Model B.

The home- and away-model, brought forth using the step-AIC method, initially included different regressors, but in order to facilitate the comparison
between the two, *regressors* were added to the home-model because it had an insignificant impact on the *AIC-statistic*.

Figure 3: *ROC-plot: Home-model.*

\[ \int dp = 0.8378 \]

Figure 4: *ROC-plot: Away-model.*

\[ \int dp = 0.8479 \]

For the purposes of this study, the relative magnitudes of the coefficients and the properties of the match-related statistics are to be analyzed in order to be able to quantify the effect of these statistics on the result of a match. *Violin plots* are used to better understand the influence of these statistics.

**Home team:**

![Home team Violin plots](image)

**Away team:**

![Away team Violin plots](image)

The above plots show the distribution of each selected *standardized regressor variable* when the corresponding result was a win (*blue*) and when it was not (*red*). The included box plots display the median and quartiles for ease of
interpretation. The displacement of the bulk of the variables for either game result suggests some influence and a wider figure represents larger variance.

4 Discussion

4.1 Reliability of the results

Standard reproducible procedures within the field of regression analysis were utilized in every step of the derivation of the final model, such as:

- Reviewing, cleaning and standardization of data
- Variable selection
- Model validation

Every measure taken with regard to the derivation of the model was motivated by the analysis of relevant statistics in order to derive the best-fitted model possible using the data. Some decisions regarding the inclusion of some variables were taken to increase the validity of the model, at the expense of model accuracy. For example, the total goals regressor was removed as this created a bias in the model.

4.2 Validity of the Results

(i) Are there any identifiable causalities between match statistics and the outcome of the match itself?

One can conclude that the model brought forth performs sufficiently in predicting the result of a match based on the regressors included in the model. The home- and away models correctly classify the game results in 75.5% and 79.3% of the cases respectively, with both models having sufficient McFadden’s pseudo-R squared ratings > 0.25. The integral of the ROC curve of the respective models, at 0.838 and 0.848, are indications of the models’ accuracy in being able to predict wins.

In analyzing the metrics related to the confusion matrices, the only outlier among the statistics is the specificity of the away model at 0.546%, in practice this denotes the away model’s ability to classify lost or drawn matches as lost or drawn. The reason for this discrepancy could be that it is not uncommon for an away team to be pushed down and outplayed, only to score through an unexpected counterattack, like in the aforementioned Manchester United - Bayern Munich match. Football is a psychological sport, and a lot of factors not included in our model can play a part in the result of a match.

One can conclude that there exists a causal link between the regressors included in the final model and the results of matches, meaning there is high internal validity. Because the risk of overtraining has been mitigated through the practice of cross-validation, one could generalize these results to be used on other seasons than those used to fit the model in this study (external validity).
As noted earlier, it is unclear how applicable the results found are to other football leagues than the Premier League, because of different strategies and trends in these leagues.

4.3 What Conclusions can be Drawn?

Because of the standardization, one can compare the proportions of the coefficients in the model and thus see the relative influence of each statistic. The magnitudes of the coefficients can be observed either in table 3, or in the violin plots.

(ii) Are there any noteworthy and unexpected trends to be observed in the Premier League?

For the home model, the largest contributors to winning a match are the teams’ shots taken, clearances, and clinicalness. It is hardly surprising that a team taking on more shots than their opponents would be more likely to win, nor that clinicalness, i.e. the teams’ ability to score from said shots, would make an impact. Clearances however are often associated with suppressed teams, whereas a stronger team normally tries to play out from the back, even under pressure. A team having more clearances would also signify them defending more than their opponents, which, according to the model, is beneficial to the game result from their perspective.

The same trend can be observed in the away model, only that the clinicalness overtakes the shots taken as the biggest contributing factor. This is to be observed in the violin plots, where the three aforementioned statistics have their bulk displaced to the right of $x = 0$ for the winning teams and to the right for the non-winning teams.

Increased control of the possession of the football is beneficial from both perspectives with coefficient values of 0.16 and 0.22 respectively. This is the only factor that has a clear connection to the playing philosophy of the teams, where some teams may strive for higher possession to control a game and others to be more passive and focus on counterattacks on the former.

Yellow and red cards, unsurprisingly, have a negative impact on the team, with red cards seemingly being the most detrimental to all factors brought forth by the model. A slight shift in the violin plots is to be observed for yellow cards from the home and away perspective, meaning home teams are slightly more likely to receive them than the away team. As for red cards, receiving more than one in a game is highly unlikely for either the home or away team, as is shown in the violin plots.

The intercept differs by 0.97, with $-0.309$ and $-1.279$ for the home and away teams respectively. It was hypothesized that a large part of this constant encapsulates the home advantage, i.e. the fact that the home team plays in front of their fans and at their own stadium. How big this advantage is is unclear, but if the assumption is made that the other hidden factors encapsulated by the indicator variable are the same for the home and away team, it is 0.97, making it the third biggest winning factor for the home team and the most detrimental factor for the away team, though this is quite speculative.
(iii) Could these insights be utilized in a viable manner when forming or adjusting team strategy?

While the model is successful in identifying the factors that contribute to winning football matches, it is questionable whether the insights can be used in a viable way. The aim of a football team is to score more goals than their opponents, thus winning the game. Clearly, if a player has a chance to score by taking a shot, he will, no matter the tactics. Thus calling "taking more shots" a viable tactic is pointless, as is "get fewer red cards" as players would if they had the opportunity.

The remark to be made in this context is that there seems to be a causality between higher possession and winning games. This is something a team can incorporate tactically, however, it is very broad, and building a tactic around possession is definitely feasible, but it is how to do this tactically that is challenging.

To be able to make tactical remarks on strategies of teams in the Premier League a more comprehensive model would have to be derived, one that includes more specific game-related statistics that can be linked to game tactics.

4.4 Economic Gains

(iv) How can a club benefit economically from insights regarding how to win football games?

Winning football matches is the primary objective of a football team and its manager, and in turn, the teams' ability to do so has a high impact on the income of the club as a whole. A winning team is of interest to multiple stakeholders in a football club, among others the fans and sponsors. This is because fans are more likely to follow a team in good form, resulting in higher match attendance, i.e. higher matchday income from ticket sales and other services, and also higher broadcast viewing numbers. This elevated attention makes the club more attractive to sponsorships and advertisers, thus generating more income.

Other income streams are broadcasting rights and tournament prize money. In the case of the Premier League, parts of the broadcasting rights revenue are distributed based on league position [8], and placing among the top seven league positions allows for participating in the European tournaments: the UEFA Champions League, UEFA Europa league and UEFA Europa Conference League. Participation in these tournaments is thus decided by accomplishments in the previous season, and not on the historical merits of the clubs. The European tournaments offer prize money and are lucrative for the participants.

Winning matches is thus interconnected to all revenue streams of the club and winning is therefore of great importance. A model that aids in understanding how to influence a team’s ability to win is thus of great benefit economically.
Acknowledgments

We would like to thank our project supervisor Viktor Nilsson for his continuous input and guidance during the writing of the thesis. We would also like to thank fellow student Oden Allen for his input during the development of models used within this thesis.
References


