Utilizing logistic regression to apply the ELO system in forecasting Premier League odds

CLAUDIO THEGELSTRÖM
Acknowledgements

I want to express gratitude to my supervisor, Viktor Nilsson at KTH, for support and guidance during the creation of the project. I also want to thank Erik Rosén and Simon Rulander for meaningful discussions during the creation of the model.
Abstract

This thesis provides insights into the creation of a model for predicting odds in the Premier League. It illustrates how the ELO system and historical odds, in combination with Monte Carlo simulations, can be implemented through logistic regression to predict odds in an unbiased way. The findings are that the model performs generally well, but significantly worse at the beginning and end of the Premier League seasons. For further improvements, it is most likely necessary to factor in variables not available in the current model. Such factors could for example be incentives, injuries, or changes in the squad, all not being accounted for by the model in this case.

Keywords: Premier League, ELO system, Historical odds, Logistic regression, Unbiased prediction
Användning av logistisk regression för att tillämpa ELO-systemet vid prognos
tisering av Premier League-odds

**Sammanfattning**

Detta examensarbete ger insikter om skapandet av en modell för att
förutsäga oddsen i Premier League. Den visar hur ELO-systemet och
historiska odds, i kombination med Monte Carlo-simuleringar, kan im-
plementeras genom logistisk regression för att förutsäga oddsen på ett
opartiskt sätt. Resultaten visar att modellen generellt sett fungerar bra,
men betydligt sämre i början och slutet av Premier League-säsongerna.
För ytterligare förbättringar är det troligtvis nödvändigt att ta hänsyn till
variabler som inte är tillgängliga i den nuvarande modellen. Sådana fak-
torer kan till exempel vara incitament, skador eller förändringar i truppen,
som alla inte tas hänsyn till i modellen i detta fall.

**Nyckelord:** Premier League, ELO systemet, Historiska odds, Logistisk re-
gression, Opartiska prediktioner
Contents

1 Introduction 7
   1.1 Background .................................................. 7
      1.1.1 Football and the Premier League ...................... 7
      1.1.2 Odds and the odds market ............................. 7
   1.2 Project goals ................................................. 7
   1.3 Prior research ............................................... 8

2 Method 8
   2.1 The ELO system implementation ............................ 9
   2.2 Data set and data management .............................. 10
   2.3 Variables .................................................... 10
   2.4 Variable Selection .......................................... 11
   2.5 Cross Validation ............................................. 12
   2.6 Model evaluation ............................................ 12

3 Results 13
   3.1 Initial model ................................................ 13
      3.1.1 Model performance ...................................... 13
      3.1.2 ELO Ratings .............................................. 16
   3.2 Final model .................................................. 17
      3.2.1 Data set ................................................. 17
      3.2.2 Choice of regressors ................................... 17
      3.2.3 ELO Ratings ............................................. 17
      3.2.4 Model Performance ....................................... 20

4 Discussion 22
   4.1 Analysis of Regression Model .............................. 22
   4.2 Efficient market hypothesis ................................ 24
   4.3 Applications of the model ................................ 25

5 Conclusion 26

Bibliography 27
1 Introduction

1.1 Background

1.1.1 Football and the Premier League

Football is the world’s most popular sport, with approximately 250 million players in over 200 countries [1]. The Premier League is the top level of football in England. It is also the most watched sports league in the world, with a potential audience of 4.7 billion people. Over one season, the 20 different teams play against each other twice. As such, every team plays 38 games, and therefore the season consists of 380 matches in total.

1.1.2 Odds and the odds market

There are different types of odds, but in this report "odds" only refer to the "European odds" version. European odds, or decimal odds, are formally defined as the inverse of a probability to occur. As such, an odds of 5 indicates a 20% probability for that specific event.

The global sports betting market is expected to continuously grow at a rate of 10.3% between 2023 and 2030. In 2022 the global sports betting market amounted to USD 83.65 billion [2].

Considering the popularity of football and the sheer size of the global betting market, one realizes that there are several reasons to understand both the sport, but also the market as a whole.

To understand the betting market better, its similarities with the financial markets can be considered. They are both associated with risk-taking to get returns since outcomes are unknown. Furthermore, information is crucial in the sense that actors in the respective markets can gain advantages and as such positive returns.

Another comparison worth making is the decision-making process and the price setting. In both the financial and betting markets, the best-informed investor or bettors will have an edge. Considering the price setting, trades on financial markets require a buyer and a seller. This is applicable in the betting industry as well, where there are exchanges where bettors with different opinions can stake each other [3].

1.2 Project goals

This project aims to create a predictive model that uses quantitative data, specifically historical odds, to estimate future odds in the Premier League. The focus is on assessing the model’s accuracy in predicting future odds. By using objective quantitative data instead of subjective human judgments, the goal is to develop a precise and dependable method for making predictions. The
project results will provide insights into the challenges involved in developing a sophisticated odds prediction model and determine whether human evaluations are necessary to improve its accuracy.

These insights could benefit any party looking to implement a predictive model for the Premier League, for example, private bettors, but also for bookmakers.

1.3 Prior research

Given the popularity and size of both football as a sport and its corresponding betting market, there has been a significant amount of previous research conducted on related topics. Many of the previous reports or articles on the subject take different approaches for the same goal - the prediction of outcomes.

Hvattum & Arntzen \(^4\) wrote an article in 2009 with similarities to this project. The authors used the ELO system mostly associated with chess, to predict the outcome of football matches. Hvattum & Arntzen later investigated the return for different betting strategies using their obtained model. Finally, they concluded that there "remains an open question as to which covariates are needed in an ordered logit regression model to create predictions which are on par with the market odds, or even whether this is possible at all."

2 Method

For the creation of the model, logistic regression will be used. The logit function is given by:

\[
\text{logit}(p) = \log \left( \frac{p}{1-p} \right),
\]

(1)

where \( p \) is the probability of a positive outcome - the probability of victory. This will be used to train the model by fitting the model to the log odds of the positive outcome:

\[
\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n,
\]

(2)

where \( \beta_0 \) is the intercept or bias variable, and the rest of the \( \beta_i \)'s are the coefficients of the different predictor variables, the \( X_i \)'s. The \( X_i \)'s, in this case, will be ELO differences between the teams in matches. Thereafter \( L(\beta) \) is minimized with regards to the negative log-likelihood of the data to estimate the values of the coefficients as the following:

\[
L(\beta) = - \sum_{i=1}^{N} \left[ y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right],
\]

(3)
where $N$ is the number of observations, $y_i$ is the observed binary outcome for the $i$-th observation, and $p_i$ is the predicted probability of the $i$-th observation belonging to the positive class calculated using the logit function in equation (1) and the logistic regression model in equation (2).

Lastly, given a set of predictor variables, the probability of a positive outcome (victory) can be predicted for new data using the following equation:

$$\hat{p}(X) = \frac{1}{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}}$$

(4)

A different number of predictor variables will be assessed concerning significance level and final model performance. The initial model will contain several predictor variables of ELO differences between the teams. With these variables, the goal is to model the relative strengths of the teams in different time frames. That is, a more sensitive ELO rating, meaning more volatile, is expected to focus more on recent matches being played. In contrast, a less volatile ELO rating is expected to be less sensitive to recent matches. Finally, the goal is for these variables to predict the probability of each outcome: Home win, draw, or away win.

### 2.1 The ELO system implementation

In chess, the ELO system is used to rate players. Every individual player is assigned an ELO rating based on previous performance, which in turn represents their relative strengths. After every game, the players’ ELO ratings are adjusted according to two factors: the difference in their ratings and the outcome of the game. The differences between the players imply an expected score which could be interpreted as an expected win probability. Thus, the system is used to make probabilistic guesses of the probabilities of outcomes in a game of chess [5].

For the different players, say player A and player B, the expected score $E_A$ and $E_B$ is calculated by:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

(5)

$$E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}}$$

(6)

where $R_A$ and $R_B$ are the ELO ratings of the different players. These scores are adjusted according to each possible outcome, where a win is scored as a 1, a draw as 0.5, and a loss as 0. After each match, the ratings are updated concerning the outcomes, by the following formula [6]:

$$R'_A = R_A + K \cdot (S_A - E_A),$$

(7)
where $K$ is a constant, determining how many points can be lost or gained after each game, $S_A$ is the outcome score and $E_A$ is the expected score for team A. A higher $K$ value means more points can be gained after fewer games and consequentially more points can be lost after fewer games. In chess, different $K$ values are used for different tiers of players; the wide majority outside the utmost elite has $K = 32$.

In this project, a similar approach to the above will be implemented, but with a few adjustments. The outcomes will be modeled according to their implied "true" probabilities, which are the inverse odds provided by oddsmakers. For example, a team that has odds of 2 will have a 50% probability of winning. With this, a Monte Carlo simulation will be implemented such that the outcomes are random variables with a probability distribution based on the odds of each outcome. Each premier league match will be simulated 10000 times, and the mean of the new ratings based on equation (7) over these simulations will be added to the ratings. By running this simulation, the estimated outcome is thought to converge to a more accurate value than if a single outcome were to be used. This way, the ELO ratings of the different teams will be determined by the odds of the previous matches played.

### 2.2 Data set and data management

The data set is collected from football-data.co.uk, which is a website that provides historical data, such as odds, results, and other statistics for past seasons in the Premier League.

Data for the past ten Premier League seasons will be used in the project. Furthermore, necessary edits to the data will be evaluated to identify systematic errors in the model. Since the model will be purely based on historical quantitative data, there is reason to believe that it will perform worse at both the beginning and end of a season. In the beginning, there is limited historical information for the odds-makers, as new teams have joined the league and teams have changed players. Furthermore, towards the end of the season, different scenarios can arise where some teams have nothing to play for. Such an example could be a team that has won the league with certainty with a few matches left. Then it is possible that incentives and lineups would differ from what would be expected, which in turn could skew the odds to more unpredictable values.

### 2.3 Variables

The predictor variables will be ELO differences between teams in a specific match. For example, a team with an ELO of 1500 facing a team with 1200 ELO would result in a +300 ELO difference for the home team.

Different $K$ values will be used and evaluated to find the best combination of variables. Consequentially, the predictor variables will be ELO differences
between the teams extracted from ELO ratings with different $K$ values. The goal is for these ELO differences to correlate with the probability of victory for the different teams, and combined, show the relative strengths of the teams for different time aspects.

The response variable will be the probability of a home win and the probability of an away win for each match. Two different regressions will be used to fit the data since there is a known home-field advantage. As a result, an ELO difference for the home team will not imply the same probability for outcomes if the teams swapped which were playing home and which were playing away. The probability for a draw will be calculated as the complementary event,

$$1 - (P_A + P_B),$$

where $P_A$ & $P_B$ are the probabilities of the teams winning.

The response variables will be calculated as the inverse of the payout-adjusted odds. This means that the bookies’ odds will be adjusted for their inverses to sum up to 1. The need for this comes from the fact that the gambling companies have a payout of less than 100% which is the company’s margin and can be seen as a direct cost of placing the bet [9].

The formula for calculating the payout for each match, and making these odds payout-adjusted so that the probabilities sum up to 1 is as follows:

$$I_{TP} = \frac{1}{O_H} + \frac{1}{O_D} + \frac{1}{O_A},$$

$$M_p = \left(1 - \frac{(I_{TP} - 1)}{I_{TP}}\right),$$

where $I_{TP}$ is the implied total probability of the odds. $O_H$, $O_D$, and $O_A$ are the odds for home, draw and away win outcomes. Lastly, $M_p$ is the margin payout of the bookmaker. Now, dividing the original odds by $M_p$ will lead to their implied probabilities summing up to 1.

### 2.4 Variable Selection

There are several ways to tackle the difficulty of choosing appropriate variables for the model. Forward or backward selection methods or all possible regressors can be used to find the best combination of regressors. The former methods are preferably used when there are many regressors, or when the computer power is limited. With this data set it is reasonable to use all possible regressors. The number of regressors is not determined beforehand, but the expectation is that the number will be in the range of 2-10 variables, where these regressors are different ELO ratings based on different $K$-values.
2.5 Cross Validation

To avoid overfitting and since the goal is prediction, cross-validation will be used. By having separate training and validation sets, the idea is that the model becomes more accurate in its predictions using new data, which is the goal.

2.6 Model evaluation

To evaluate the model, mainly MAE (Mean Absolute Error) will be used. The definition in this project will be the absolute value of the ”correct" probability of an outcome (based on the odds), in comparison to the predicted probability. As such, more formally it can be described mathematically:

\[ \text{Absolute error} = |y_i - \hat{y}_i| \]  \hspace{1cm} (11)

\[ \text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \]  \hspace{1cm} (12)

Where \( y_i \) is the inverse of the payout adjusted odds, and \( \hat{y}_i \) is the predicted probability.

An additional metric to evaluate the model is by using the unadjusted odds, from the bookmakers, and checking how often the model predicts a probability that is within the interval where the expected value of placing a bet is negative. To illustrate the method, consider a fair coin. The two outcomes are equally possible, and for a bettor to have an expected value of zero or less, the odds should be 2.0 for each outcome. Considering that bookmakers take a fee for placing a bet, more reasonable odds are 1.9 for each outcome. In this case, the maximal true probability of a specific outcome is \( \frac{1}{1.9} \approx 0.53 \), since a higher probability would mean that the bet could be placed with a positive expected return. With similar reasoning, the lowest probability a specific outcome can have is one minus the implied probability of the complement. In the example with the coins, and with odds of 1.9, this would imply a probability of \( 1 - \frac{1}{1.9} \approx 0.47 \). Now there is an interval, based on the bookmaker’s odds, in which the true probability should lie under the assumption that no bets can be placed with a positive expected return. Therefore, a telling metric for the strength of the model is how often the predicted odds are within such intervals.

Furthermore, the models’ accuracy will be evaluated concerning the percentage of correctly predicted favorites in matches.
3 Results

3.1 Initial model

The first model was created with the initial idea of the classical ELO with a $K$ value of 32, complemented with another ELO variable with a $K$ value of 200. Here, the underlying reasoning was to create ELOs with different sensitivities to recent events, with the purpose of better predicting the winning probabilities of the teams. The ELO values were calculated based on the start of the 2012/2013 season, where each match was simulated 10000 times to make the ELO representative of the odds. Lastly, the first two seasons were removed from the training set to let the ELO values converge to more representative values. For example, in the beginning, every team was assigned an ELO of 1500, and this unrepresentative data was now avoided.

3.1.1 Model performance

With the first model, the aim was to test the viability of the concept, in combination with assessing its weaknesses and make adjustments to make the final model more reliable and precise.
In figure 1 and figure 2, the 500 largest errors for the home teams are plotted over the seasons. The errors are defined in equation (11). What can be noted is that at the beginning and end of each season (August and May), the error is more often high than in the middle of the season. This indicates that model performance can be improved by removing these outlying data points. Therefore, the final model will not be trained or evaluated on data from the first month of the season, and the last month of the season.
Figure 3: Violin plot of the home (1), draw (2) and away (3) errors, showing also the mean value & distribution in a box plot.

Figure 3 shows how the errors are distributed for the different probabilities. For the home errors, one can see that there is a large spread, but most of the errors are in the range 0.005 to 0.10, with a mean value of 0.0363. For the draw probabilities, there is a noticeable difference where the range is smaller, with values concentrated in the range 0.005 to 0.04, and a mean value of 0.0158. Lastly, regarding the away probabilities the distribution of errors resembles that of the home probabilities. The range is almost the same, but the values are concentrated in the lower range of 0.005 to 0.07, and has a lower mean value of 0.0336.
3.1.2 ELO Ratings

Figure 4: ELO for a few teams, K=32

Figure 5: ELO for a few teams, K=200
In the two plots, 4 & 5 the ELO ratings for a few teams in the Premier League are plotted. What can be concluded from the different plots, is that the ELO with a higher $K$ value is more volatile.

### 3.2 Final model

After the initial model was created. The number of variables was extended to 8 with the following order of the $K$ values in equation (7): $K = 32, K = 100, K = 125, K = 150, K = 175, K = 200, K = 400, K = 500, K = 2000$.

#### 3.2.1 Data set

From the initial model, it was evident that the model had difficulties at the beginning and end of the seasons. Therefore the first and last month of every season was removed from the data set for the final model. Furthermore, to make a fair and unbiased evaluation of the final model, the season of 2022/2023 was removed from the data set, to provide a test set. Therefore the final model will be trained and validated on the data set from season 2014/2015 until season 2021/2022. The ELOs are calculated in the same way, for the whole unstripped data set from season 2012/2013 until season 2022/2023.

#### 3.2.2 Choice of regressors

For the final model, the choice of regressors was determined by the all possible regressors method. All of the possible combinations of regressors were trained and validated over the data set. Every possible set of regressors was evaluated over the MAE over the different folds using cross-validation. The combination of regressors which resulted in the lowest average MAE over the validation sets was chosen. This was done both for the home win probability, and the away win probability.

The best combination of regressors, for both the home win probability and away win probability was the ELOs using the following K-values: $K = 32, K = 200, K = 500, K = 2000$.

![Table 1: MAE comparison between the initial model and the best model, calculated as the average error over the validation sets.](image)

#### 3.2.3 ELO Ratings

Figure 6, 7 & 8 illustrates how the different K-values in equation (7) affect the volatility in the ELO ratings. For the final model, different ELOs with K-values
ranging from the standard $K = 32$, to $K = 2000$ were deemed the best.
Figure 8: ELO for a few teams, K=2000

<table>
<thead>
<tr>
<th>Team</th>
<th>ELO(K=32)</th>
<th>ELO(K=200)</th>
<th>ELO(K=500)</th>
<th>ELO(K=2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man City</td>
<td>1651</td>
<td>1638</td>
<td>1646</td>
<td>2795</td>
</tr>
<tr>
<td>Liverpool</td>
<td>1561</td>
<td>1499</td>
<td>1509</td>
<td>1981</td>
</tr>
<tr>
<td>Arsenal</td>
<td>1536</td>
<td>1567</td>
<td>1547</td>
<td>1679</td>
</tr>
<tr>
<td>Chelsea</td>
<td>1517</td>
<td>1517</td>
<td>1491</td>
<td>1192</td>
</tr>
<tr>
<td>Man United</td>
<td>1488</td>
<td>1524</td>
<td>1540</td>
<td>1763</td>
</tr>
<tr>
<td>Tottenham</td>
<td>1473</td>
<td>1473</td>
<td>1504</td>
<td>2556</td>
</tr>
<tr>
<td>Brighton</td>
<td>1441</td>
<td>1490</td>
<td>1434</td>
<td>2080</td>
</tr>
<tr>
<td>Newcastle</td>
<td>1433</td>
<td>1492</td>
<td>1467</td>
<td>691</td>
</tr>
<tr>
<td>West Ham</td>
<td>1399</td>
<td>1409</td>
<td>1372</td>
<td>933</td>
</tr>
<tr>
<td>Aston Villa</td>
<td>1388</td>
<td>1393</td>
<td>1365</td>
<td>1963</td>
</tr>
<tr>
<td>Brentford</td>
<td>1368</td>
<td>1402</td>
<td>1425</td>
<td>1555</td>
</tr>
<tr>
<td>Wolves</td>
<td>1367</td>
<td>1389</td>
<td>1416</td>
<td>1281</td>
</tr>
<tr>
<td>Leicester</td>
<td>1367</td>
<td>1373</td>
<td>1452</td>
<td>1089</td>
</tr>
<tr>
<td>Leeds</td>
<td>1354</td>
<td>1380</td>
<td>1456</td>
<td>1804</td>
</tr>
<tr>
<td>Everton</td>
<td>1349</td>
<td>1361</td>
<td>1344</td>
<td>1879</td>
</tr>
<tr>
<td>Crystal Palace</td>
<td>1345</td>
<td>1333</td>
<td>1356</td>
<td>1302</td>
</tr>
<tr>
<td>Fulham</td>
<td>1342</td>
<td>1367</td>
<td>1413</td>
<td>1406</td>
</tr>
<tr>
<td>Nott’m Forest</td>
<td>1342</td>
<td>1279</td>
<td>1195</td>
<td>947</td>
</tr>
<tr>
<td>Southampton</td>
<td>1330</td>
<td>1330</td>
<td>1344</td>
<td>1073</td>
</tr>
<tr>
<td>Bournemouth</td>
<td>1294</td>
<td>1299</td>
<td>1306</td>
<td>1627</td>
</tr>
</tbody>
</table>

Table 2: The Premier League teams sorted by ELO(K=32) per 2023-04-16
Table 2 shows the current ranking of the Premier League teams in the 2022/2023 season. What can be noted is a difference between the values of the regressors. For example, for a $K$ value of 32, Chelsea is ranked higher than Man United. However, Man United is ranked higher for all of the more sensitive ELOs ($K = 200, K = 500$ & $K = 2000$). This indicates that over a longer period, Chelsea has been the better-performing team, but in a more recent time frame, Man United has performed better. Essentially, this means that the odds market is valuing Man United as a stronger team than Chelsea recently.

3.2.4 Model Performance

To test the final model’s performance, the tests were conducted for the 2022/2023 season of the Premier League. As such, the test set is separated from both the training and validation sets.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Error (MAE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>0.029267813</td>
</tr>
<tr>
<td>Draw</td>
<td>0.012708912</td>
</tr>
<tr>
<td>Away</td>
<td>0.027811403</td>
</tr>
</tbody>
</table>

Table 3: MAE for the final model

Table 3 shows the mean absolute errors for the final model on the 2022/2023 season of the Premier League. The average absolute home error and away error are similar, with an average of around 3 percentage points. For the draw probability, the average error is lower around 1.3 percentage points.
Figure 9: Violin plot of the home (1), draw (2) and away (3) errors of the final model, showing also the mean value & distribution in a box plot.

In figure 9, the errors of the final model are plotted in violin plots. The characteristics of each plot are similar to figure 3 from the initial model. For both the home and away probabilities, the spread of the errors is larger than the draw probabilities. Overall, the mean error, and the spread of the errors, are lower for all three probabilities in the final model.
Table 4: Confusion matrix showing how often the favorite, and underdog, are estimated correctly

Table 4 shows how often the model finds the correct favorite and underdog in matches for the 2022/2023 season. There is a noticeable difference, where the model finds a higher percentage of the correct favorites, in comparison to the underdog for a given match.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Within interval</th>
<th>Outside interval</th>
<th>% correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>65</td>
<td>191</td>
<td>25.4%</td>
</tr>
<tr>
<td>Draw</td>
<td>128</td>
<td>128</td>
<td>50%</td>
</tr>
<tr>
<td>Away</td>
<td>76</td>
<td>180</td>
<td>29.7%</td>
</tr>
</tbody>
</table>

Table 5 shows how often the model predicts a probability for each outcome that is within the interval that can be generated from the unadjusted odds of the bookmakers. Again, the model is more accurate in the estimation of the probability of a draw.

4 Discussion

4.1 Analysis of Regression Model

Based on the first model with only two regressors, a lot of improvements could be made to the final model. First, there was a significant increase in error at the beginning and end of each season. There could be several different reasons for this increase in error for these months. Most likely, at the beginning of the season, the larger errors are a direct consequence of new teams joining the league and teams changing squads during the transfer windows. The summer transfer window opens at the beginning of June and ends on the first of September.
Therefore, new teams are added to the league and all of the teams potentially change their lineups during the first month. It is then reasonable that the ELO values need a few matches at the beginning of the season to more accurately assess the relative strengths of these "new" teams.

The larger errors for the end of the season are a bit more difficult to explain. However, it is reasonable that when the season reaches its end, some teams do not have anything to play for. Such factors could be considered in the odds for the last month, making them seem "unreasonable" for the model since the ELO values can not predict sudden changes in the odds in such a short time.

Considering that there was a large spread between the errors, as shown in figure 3 and these errors were found concentrated around the beginning and end of the seasons, it made sense to remove these data points.

Moving on to the final model, additional regressors in the form of ELO-values with different sensitivities were added to increase the complexity of the model, and potentially make it more accurate. When evaluating the different combinations of the regressors, it was found that two additional ELO values, with increased sensitivity, provided the best model. One takeaway from this is that the odds in the Premier League are volatile throughout a season, and to create an efficient model this needs to be considered. Additionally, the odds being volatile during the season should imply that the teams' relative strengths shift throughout the season, at least according to the odds market.

Table 2 shows how the different ELO-values differ between the teams of the current 2022/2023 season. An illustrating example of how the teams have performed historically, in comparison to recently, is Aston Villa. Over the long term, and less volatile ELO ($K = 32$), they are ranked 10th in the league. But looking at the more volatile ELO with $K = 2000$ the team is ranked in the top five. What this could mean is that as of recently, the odds market has considered Aston Villa as a stronger team than historically. Coincidentally, as of today (2022-04-21), they have won their last five matches [11].

In table 3 there is a rather large difference between the average error of the different probabilities. One reason for the estimated draw probability being more precise in the final model is that its range is a lot smaller, between an odds of 2.95 to 14.96. In comparison, the home odds are in the range of 1.07 to 21.51. Therefore, the draw probability might be easier to predict since it is within a smaller range of values.

Considering table 4 there is also an explanation for the model having more difficulties in predicting the underdog compared to predicting the favorite. Since the draw probability can not be a favorite (unless it is an extreme case), it is either the home team or away team that is favorite [12]. Consequently, regarding the underdog, there are more possibilities since every outcome could be the underdog. As such, it is reasonable that the favorite is easier to predict.

A performance metric that shows below-par results for the model is the fre-
quency at which it accurately predicts probabilities falling within the odds interval. Although this is not the most important metric for this particular project, it is a measure where the models’ efficiency could not be argued, since the intervals represent the span of the "true" probabilities based on the odds market. To increase the models’ performance for this specific metric, one possible adjustment could be to fit the model to the mean of the intervals, rather than the adjusted probabilities of the odds. This would most likely increase the MAE however since the model would adjust to the new response variable.

4.2 Efficient market hypothesis

By definition, the efficient market hypothesis (EMH) states that the share prices of stocks reflect the underlying asset’s true value. This means that consistent overperformance (alpha) is impossible since the share prices reflect all available information [13].

In the article, *Information, prices, and efficiency in an online betting market*, Guy Elaad, J.James Reade & Carl Singleton (2020) [14] study the efficiency of odds by analyzing data for over 16,000 football matches in England since 2010. The authors found that there is no statistically significant evidence that could dismiss the EMH for matches in England. In addition to that, they discuss how the betting market overall is characterized by readily available, cheap, and easily accessible information. Between the different levels, the highest (Premier League) is more likely to have market actors that are more informed due to the larger volumes of money. Furthermore, there is a fixed time for when a bet is set, which is different from the financial markets. This makes it easier to evaluate the betting market concerning the efficiency of its odds.

Ruud H.Koning & Renske Zijm also concluded in their article *Betting market efficiency and prediction in binary choice models* (2022) [15], that the Premier League odds are priced correctly. Meaning that there is no systematic or predictable error in the pricing of the odds.

Considering the characteristics of the betting markets, it is, as Guy Elaad, J.James Reade & Carl Singleton touch on in their paper [14], expected that the closing odds are effective in predicting outcomes of matches. It is a market with low transaction costs and easily accessible information, and there is a lot of liquidity in the top leagues. With these factors in mind, it can be argued in the sense of the EMH, that the closing odds can not be beaten, since then all information is available. If there would exist odds that were too high, bettors would rush to those odds until they were adjusted and the expected return became negative, all in accordance with the efficient market hypothesis.

With the starting point that the betting markets in general can be seen as effective markets, one realizes the need for bookmakers to develop consistently effective models to determine their odds. For the different companies that provide odds to the public, the ones more comfortable in their pricing of matches can in turn take on more volume, and increase the expected revenue per game.
Such an example is Pinnacle, which is a bookmaker known for being confident in its odds pricing, making its business model focused on volume rather than margins [16].

Companies comfortable with their model would be able to have a higher payout, that is, decreasing the fee for their customers. This would mean that the average odds on such a site would be higher, which in turn should attract more bettors to their site since the available odds are higher. Because of this, it is of high interest for bookmakers to have efficient odds pricing models in order to not be at a competitive disadvantage.

Considering the other perspective, bettors who can create a consistently accurate model can have the chance to place bets with a positive expected return, capitalizing on wrongly priced odds.

Also worth mentioning is the arbitrage aspect. Since odds are volatile and changing, correctly predicting the closing odds would provide arbitrage opportunities. For example, there are sites where smaller actors can offer bets [17]. This makes it easy to capitalize on knowing the closing line, you could offer the other "side" of the bet to someone else, while placing the opposite yourself, acting as a bookmaker, and then having a positive expected return.

In general, the odds markets are deemed to be efficient when it comes to pricing in the big leagues, and bookmakers together with private bettors have a strong incentive to make precise predictions about the actual odds of football matches.

4.3 Applications of the model

The final model in this project can be a supportive insight for anyone interested in making an accurate estimation of closing odds in the Premier League. It is based purely on historical odds in combination with a modified version of the ELO system. Even though the model is flawed in the sense that its results have an error in a larger range than preferred, it provides an unbiased estimation based on effective market data. By incorporating some adjustments, the model could become suitable for making future predictions on Premier League odds. These adjustments may have a degree of bias, but they would assist the model by incorporating other metrics, such as injuries, incentives, and other factors that the model currently cannot adjust to.

Implementing an improved version of the model could then be used to take advantage of wrongly priced odds and arbitrage opportunities. Furthermore, it could be used to make qualified guesses in the probabilities of outcomes in matches in general. With this in mind, it could provide value for both the bookmaker and private bettor, looking to develop a new or improve an existing model.
5 Conclusion

The final model is rather effective in predicting the odds of the matches in the Premier League. However, it shows the great difficulty in predictions at the beginning and end of the seasons. In addition to that, the spread between the errors is rather large, making its accuracy unpredictable. It does show, however, that the ELO system can be implemented to predict probabilities in other areas than chess. In addition to that, the results also provide hope for a quantitative unbiased model that is capable of predicting the closing odds in the Premier League.

Furthermore, the odds market can be seen as an effective market, making the closing odds imply a true probability of an outcome. Moreover, actors with an accurate estimation of closing odds have a market advantage over other actors, possibly leading to positive expected returns and arbitrage opportunities.

Further improvements on the model would for example be in measuring injuries, motivation, and other factors that could impact the odds and that would not be captured by the model. Therefore, the model provides a good starting point but should be further improved using more available information.
Bibliography


