Statistically Ranking Teams in the English Premier League

Will Schmutz

Winter 2022

1 Abstract

Soccer is a game that has been around since the 19th century and is beloved by hundreds of millions of fans worldwide. Researchers have been predicting soccer rankings for many years, yet, it's one of the most challenging sports to predict. There are several methods for predicting soccer results, and different sets of information and techniques are employed to create valuable statistics. In this thesis, we propose three new methods based on simple statistical tools like linear regression and logistic regression to predict the final season rankings of the soccer teams after a certain amount of games in the English Premier League. We compare our methods with existing methods like Elo's ranking model and establish the superiority of the performance of our model.

2 Introduction

Soccer is the most-watched sport globally, with approximately 3.5 billion fans and over 200 different leagues. The most interesting part of soccer is the ninety minutes of constant back and forth battle with little to no stoppages. Every player must be in sync for the whole match for teams to succeed. The sheer dynamics on the field and the small number of goals scored during each game add randomness to the result, making it extremely difficult to predict. However, after teams play many games, some patterns in the teams' performance emerge, and the best teams eventually show their strengths and separate as the clear favorites.

One of the most popular soccer leagues globally is the English Premier League which consists of clubs such as Manchester United, Liverpool, Arsenal, etc. The English Premier League was founded in 1888, and the league consists of 20 teams. Currently, each team plays every other team two times for a total of 380 games played throughout the Premier League season. Each team receives 3 points for a victory, 1 point for a tie, and 0 points for a loss. The team with the highest accumulated points throughout the league wins the English Premier League. At the end of the season, the bottom three teams get demoted to the league below the English Premier League, which is called relegation. At the end of the season, the top four teams qualify for the Champions League, which consists of the most prestigious soccer clubs in Europe, while the fifth-place team qualifies for the Europa League.

Predicting the winner of any soccer league is an intriguing and challenging question that has allured many researchers in past years. Common models used throughout the papers are the Elo Rating, Logistic regression, Poisson distribution, and a Bivariate distribution. [Mit20] Elo is a classic model used to rank teams where after each game, teams lose or gain a certain amount of points based on different factors during the match, which includes but is not limited to the result of the match and a statistical judgment for how well the two teams performed during the match. It was initially formulated to rank chess players. Logistic regression uses a logistic function to model the probability of a discrete event taking place, such as the probability a team wins a soccer match. A Poisson distribution is a probability distribution that measures the probability of multiple events occurring during a time interval. Lastly, a Bivariate distribution is the same as a joint distribution which contains two variables and gives the probability of the same outcomes for the two variables occurring. Here are some of the models researchers have developed for predicting soccer rankings. [AH21] advocated using Elo Ratings of teams and ratings of individual players as response variables to a Logit regression model that outputs the probability of the outcome of each game. [Jay18] proposed a method that uses data to calculate the expected goals in the match along with a Poisson distribution of actual goals scored to estimate the likelihood outcome of each match to rank teams. The Poisson distribution was used to avoid outlier scores in the ranking where results occurred but did not represent the overall quality of the two teams. [BK19] used machine learning techniques such as feature engineering and exploratory data analysis to predict the outcome of matches in the EPL. [Con19] predicts soccer outcomes in a league in one country by looking at matches from other countries, by using both dynamic rating and Bayesian networks that use historical match data that teams playing weren't necessarily involved in. [CF13] uses a rating system by looking at the differences in scores between two teams playing and the results of their relative matches earlier in the season since score is considered a good indicator for predicting the outcome of matches. The model proposed in [CDLR02] measures how well teams are at both offense and defense using an independent Poisson distribution to try and predict the probabilities of a home win, away win, and a draw. They aim to outperform betting odds, which bears a certain resemblance to our model performances. Last but not least, [BKM17] uses both inter-arrival-times-based count processes along with a cumulative distribution function to produce a Bivariate distribution to predict the number of goals scored in each game. Although dozens of models have been proposed to predict soccer rankings, it is difficult to achieve good prediction results.

This thesis presents three new methodologies, namely the *Logistic Method*, *Betting Model*, and the *Rank Model*, and to predict the final rankings of the teams in the EPL based on the results after 50, 100, 200, and 300 games (recall that the EPL has a total of 380 games). For our analysis, we have collected data from footystats.org [Dat] that keeps track of dozens of stats for each football match. We looked at every stat in the database and filtered out only the stats that seemed to be the most significant in predicting the outcome of the games. In Section 3, we present the details of the methods. Briefly speaking, the Logistic Method relies on logistic regression by calculating the probability each team will win, lose, or draw a match. The Betting Model is a linear model

focusing on betting odds in each match where betting odds are predictions by sportsbooks about the probable outcome of each match. Lastly, the Rank Model is also based on a linear model but uses other information like on-field statistics and previous season ranking along with betting statistics to predict the final ranking of the teams. In Section 4, we compare performances from our proposed methodologies with other extant techniques. The two existing methods that we compared performances were a baseline and an Elo method (mentioned above). The baseline proposal is what the live Premier League table looks like throughout different points in the season. Based on the empirical evidence in the results section, both the Betting Model and the Rank Model perform well. Still, we can conclude that the Rank Model outperforms other methods in predicting the final ranking of these teams.

3 Our methods

3.1 Logistic Method

The Logistic Method is based on a multiple logistic regression model, relying on adding one of the three different equations to predict whether a home team won, an away team won, or a draw occurred. The primary motivation behind using logistic regression is that it is easy to train, interpret, and frequently achieves adequate accuracy on real data. In this model, the Y variable is a categorical variable with three classes Home team win, Away team win, and Tie. The variable for θ consisted of a vector of parameters of length 22. A parameter was represented for each team, a tie parameter, and a home parameter to add an explicit factor for being at home. We used three equations depending on the results of the game.

$$P(Y = \text{Home team win}|\theta) = \frac{e^{h-a+p}}{e^{h-a+p} + e^{a-h+p} + e^{t}} \tag{3.1}$$

$$P(Y = \text{Away team win}|\theta) = \frac{e^{a-h+p}}{e^{h-a+p} + e^{a-h+p} + e^{t}} \tag{3.2}$$

$$P(Y = \text{Tie}|\theta) = \frac{e^{t}}{e^{h-a+p} + e^{a-h+p} + e^{t}} \tag{3.3}$$

Where h is the home team parameter, a is the away team parameter, p is the home team advantage parameter, and t is the tie parameter.

For each extra game, depending on the result, we added one of three equations indicating the probability that the specific game result occurred. Then, to solve for each of the parameters in the logistic expression, we took the log-likelihood of the expression to simplify the calculation and optimized the model, calculating a coefficient value for each team. The one constraint we had on our model was that the 20 coefficients representing each team had to sum to zero. We compare these rankings against the results of the end of the season via Spearman Rank Correlation Coefficient to

determine how well these ranks reflect the end results of each season. The Spearman's correlation coefficient between two sets of ranking is defined as:

$$p = 1 - \frac{\sum_{i=1}^{n} 6(d_i)^2}{n(n^2-1)} \tag{3.4}$$

where d is the difference between the model rank and the end of the season rank for each team i and n is the number of teams in the Premier League which is always 20.

Spearman's Rank Correlation is a common way to measure the strength and direction of ranked variables. The key reason for using Spearman Rank Correlation Coefficient is that it is better tuned to understand and compare two rankings. The Spearman correlation can find linear relationships and determine non-linear relationships like an exponential relationship, which allows it to rank variables well. For example, if a team had a rank of 6, but its actual rank was 3, it would penalize that rank more than other correlation methods. The two sets of ranks we use for the Spearman Correlation are our model's expected rankings of the teams and the end of the season results of the actual Premier League in each season.

The Logistic Method performed the worst out of our three models, and the performance can be seen in Section 4. The Logistic Method performed poorly because we did not look at any real statistics other than goals scored. Also, using an Elo Model would be better than a logistic regression model for looking at the actual results. It acts more like the Premier League table, where the table will change accordingly after each result. Another issue with the Logistic Method is that overfitting occurs because the model gets very complex after many games are played. Specific outlier games can cause a dramatic, unnecessary change in the predicted rankings. Overall, the Logistic Method was the first model we attempted in our research and proved a good stepping stone for producing models with better prediction accuracy.

3.2 Betting Model

The Betting Model was the second model we attempted during our research. When creating the Betting Model, we used only a sample of betting odds for each game. We obtained the betting odds from our data footystats.org. The primary reason for using betting odds is that it offers a unique perspective into the soccer game because betting odds somewhat indicate which team is more likely to win. Betting odds have valuable information that can inherently measure a team's quality and other unseen factors like player injuries or tactical advantages. The betting odds data provides us with the home team odds (i.e., the probability that the home team will win), the away team odds (i.e., the probability that the away team will win), and tie odds (i.e., the probability that the two teams will draw) for each game. In our model, we tried two different response variables.

We tried using both odds ratio (home team odds/away team odds) as our response variable and the log of the odds ratio log(home team odds/away team odds), henceforth denoted by $\mathbf{y} \in \mathbb{R}^n$. We can see the results of the two different models in Table 1 at the end of Section 3.1. Here n is

the number of game we train our model on, i.e., as mentioned in Section 2, we run our models for $n = 50, 100, 200, 300$. We create a covariate matrix $\mathbf{X} \in \mathbb{R}^{n \times 21}$. The first column of $\mathbf{X}$ represents the intercept term, and the last 20 columns represent an indicator for each team in the EPL (recall that we have 20 teams in EPL). More precisely, if the m^{th} game consisted of team i and team j, where team i is the home team, and team j is the away team, then we set:

$$\mathbf{X}_{m,k} = \begin{cases} 1, & \text{if } k = 0 \text{ or } k = i, \\ -1, & \text{if } k = j, \\ 0 & \text{otherwise}\,. \end{cases}$$

We fit the linear regression model:

$$\mathbf{y} = \mathbf{X}\beta_0 + \epsilon\,,$$

using the covariate and the responses described above and obtain the OLS estimate:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}\,.$$

Here the coefficient vector β_0 can be interpreted as a quantification of the performance of each team. We rank the teams in terms of the coefficients of $\hat{\beta}$, i.e., the team with the highest coefficient was given rank 1, whereas the team with the lowest coefficient was given rank 20. We compared these rankings against the results of the end of the season ranking using the Spearman Rank Correlation Coefficient.

Table 1: Comparison of Response Variables

Games	Odd Ratio	Log(Odd Ratio)
50	0.710	0.812
100	0.7850	0.834
200	0.8427	0.869
300	0.8874	0.8868

The above table, Table 1, shows the performance of the two different response variables in the Betting Model. We averaged out the Spearman Correlation of the two betting odds model with different response variables by looking at Premier League seasons from 2012-21. We can see that using the Log(Odds Ratio) as the response variable for the Betting Model produces a better Spearman Correlation after each point in the season except for after 300 games. However, after 300 games, the two models have nearly identical results, concluding that the Log(Odds Ratio) performs better than simply using the odds ratio. When discussing the Betting Model, the response variable will always be the Log(Odds Ratio) from this point forward. The intuition behind why we believe Log(Odds Ratio) produces better results is likely because the Log(Odds) produces a more natural spread of numbers for the response variable. When a team is heavily favored to win, the regular odds ratio does not differ significantly from when two teams have similar odds of winning, making it harder for the Betting Model to decipher which team should win. Log(Odds) also allows negative

numbers in the response variable, and it is symmetric around 0, which allows the range of values to be more consistent and spread out compared to the regular odds ratio. The performance of the Betting Model versus the other models in the thesis is present in Section 4.

3.3 Rank Model

Our final and most consistent method is the Rank Model. For this model, we use multiple statistics to create our response variable. The main difference between this model and the Betting Model (Subsection 3.2) is the response variable, Y. As explained earlier, in the Betting Model, the response variable Y is the log of the betting odds ratio. The key reason for considering a different response variable is that the betting odds are based on the information before the match, whereas data during the match (based on the performance of the teams) is also quite valuable for prediction. Therefore, it is intuitive that utilizing both betting odds and other statistics obtained during the match would likely give us a better predictor for the final ranking. We next describe the method for creating the new response variable Y for this Rank Model. We looked at data from seasons 2012-2021 to obtain Y and our procedure to achieve Y consisted of these two steps:

1. Define $Y^{(1)}_{\text{temp}}$ as the ternary variable which takes value 1 if the home team wins, -1 if the away team wins and 0 if the match tied. Define the covariate X as a collection of following statistics from each game:

   - both teams' corner count
   - both teams' shots on target
   - fouls
   - home team possession
   - betting odds ratio
   - both teams' total red cards
   - previous season rankings

   Upon regression Y on X, we select the *significant* predictors with p-value < 0.05 (for testing whether j^{th} coefficient is 0). The selected covariates were consistent across the seasons, which are as follows:

   - shots on target
   - corner kicks
   - betting odds ratio
   - goal differential
   - previous season ranking

   Let $\hat{\beta}^{(1)}$ be the estimated coefficient and $\hat{\beta}^{(1)}_{\text{sel}}$ be its subset with the significant coefficients. Define $\hat{\mathbf{Y}}^{(1)} = \mathbf{X}_{\text{set}}\hat{\beta}^{(1)}_{\text{sel}}$ (see equation 3.5 below for details).

2. We next run another regression with a new response variable $Y^{(2)}_{\text{temp}}$, the goal differential of the match. For example, if the home team defeated the away team 4-1, $Y^{(2)}_{\text{temp}}$ equaled 3, but if the away team defeated the home team 4-1, $Y^{(2)}_{\text{temp}}$ equaled -3. The covariate X is same as the previous regression. We found that the same set of covariates are again selected

as significant with different values of coefficients as seen in equation (3.6). Let $\hat{\beta}^{(2)}$ be the regression coefficient and $\hat{\beta}^{(2)}_{\text{sel}}$ be the subset of $\hat{\beta}^{(2)}$ consists of the significant coefficients. As before, define $\hat{Y}^{(2)} = \mathbf{X}_{\text{sel}}\hat{\beta}^{(2)}_{\text{sel}}$ (see equation (3.6)) and set our response $\mathbf{Y} = (\hat{Y}(1) + \hat{Y}^{(2)})/2$.

A key difference between the Rank and Betting Models was that to avoid overreacting to a single game result, we set a max $\mathbf{Y}$ value of 5 or -5 for each game. For example, if a team won 7-1, they would achieve a coefficient value of more than 5, but since it was only one game and various outside factors could have caused that result, like an early red card or significant player injuries, we standardized the $\mathbf{Y}$ value by giving it a max value.

$$\begin{aligned}\hat{Y}(1) = (0.073s_h) + (-0.086s_a) + (0.06c_h) + (-0.047c_a) + \\ (0.0085b) + g + (0.095p_h) + (-0.0095p_a)\end{aligned} \tag{3.5}$$

$$\begin{aligned}\hat{Y}(2) = (0.189s_h) + (-0.213s_a) + (0.068c_h) + (-0.085c_a) + \\ (0.034b) + g + (0.018p_h) + (-0.0198p_a)\end{aligned} \tag{3.6}$$

where:

- s_h = home team shots
- s_a = away team shots
- c_h = home team corner count
- c_a = away team corner count
- b = betting odd ratio
- g = goal differential
- p_h = previous season home team ranking
- p_a = previous season away team ranking

Finally, we run a regression with $\mathbf{Y}$ (obtained via two steps described above) and the same covariate matrix $\mathbf{X}$ as used in the betting model. We obtained coefficient values for each team when running both regressions and then ranked the team based on the increasing order of coefficient values. The results are presented in Section 4.

Remark 3.1. *In the Betting Model, we find that the predictive performance of the trained model reaches a plateau at around 100 games. Even if we train the model with more games, the prediction accuracy remains similar. In contrast, the prediction performance of the Rank Model with multiple*

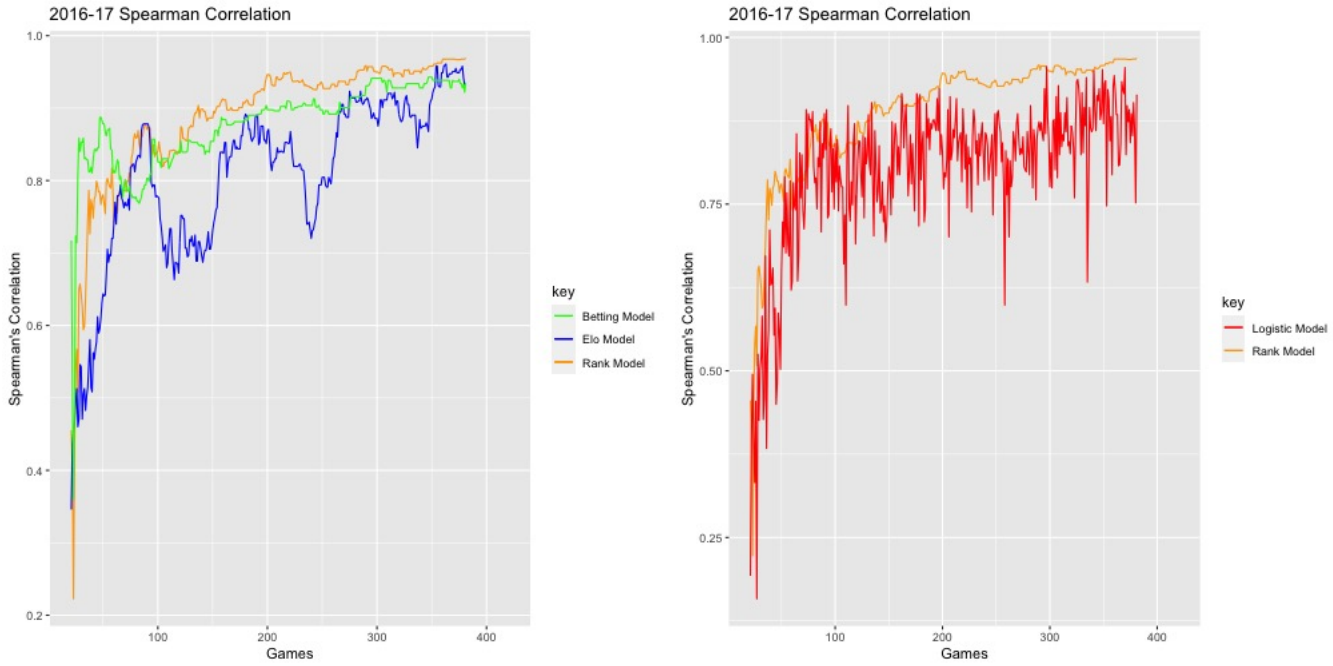
statistics keeps improving as we increase the number of games in our training model. One possible explanation for this is that most betting odd systems do not use data from the current season, like our Rank Model, but instead mostly use data from previous years to determine the outcome of predictions. Overall, in Section 4 we can see that the different graphs and tables indicate that the Rank Model was superior for predicting final season results after 100 games.

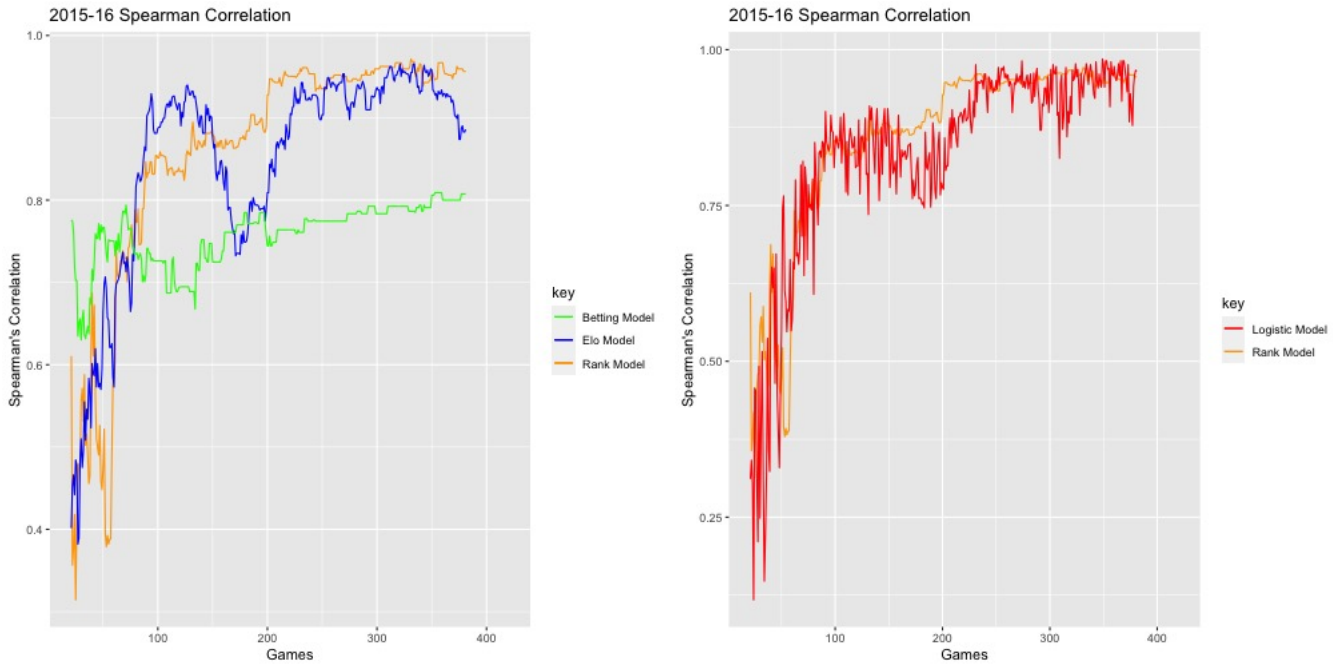
4 Results

In this section, we are presenting three types of prediction results. In section 4.1, we compare all the models we have looked at in our research and evaluate the pros and cons of each model. In section 4.2, we compare our Rank Model against the betting odds system to see how well the Rank Model compares against the best models that have been developed. In section 4.3, we compare the Rank Model versus the Baseline Model in predicting how well the Rank Model can predict which teams are champions of the Premier League and qualify for the Champions League or are relegated from the Premier League. A more in-depth description can be found at the beginning of the section.

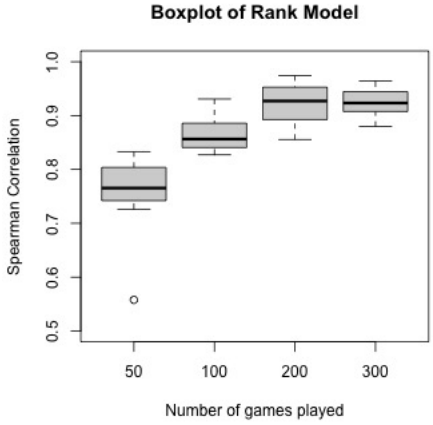
4.1 Comparison of Models

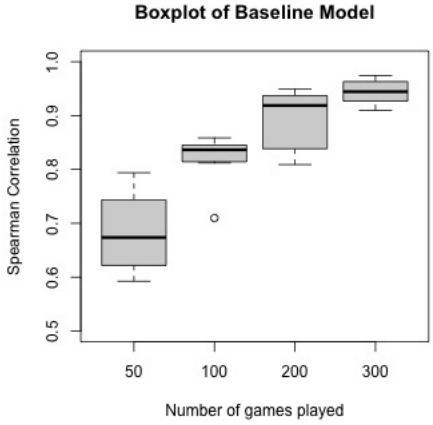
In this section, we compared the five different models in our research against one another. Based on our analysis, we found that the Rank Model (Subsection 3.3) outperforms other models in terms of the predictive performance of the final ranking of the teams.

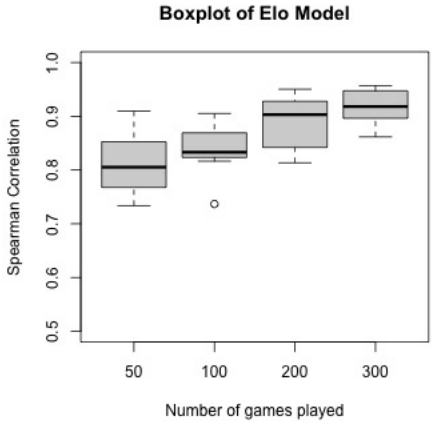




These four figures above outline the change in Spearman Correlation throughout a single season. The X-axis corresponds to the number of games, and the Y-axis corresponds to the Spearman's correlation of the predicted ranking of the teams at the end of the corresponding number of games and the actual final ranking of the teams. The Rank Model is the most consistent by improving continuously throughout the season and never reaches a plateau. The Betting Model in both seasons reaches a plateau around 100 games and stops improving dramatically in the Spearman Correlation after this 100-game mark. The Elo Model also performs well on the data but is inconsistent and can underperform the Rank Model significantly during different parts of the season. One possible explanation is that a single match can influence the Elo Model more than the Rank Model, making it non-robust. Lastly, the Logistic Method also performs well but has overfitting issues as this model relies too heavily on the last games the model looked at. For example, if the Logistic Method wanted to obtain a Spearman Correlation after 100 games, the games between 90 and 100 would be more significant in the model's final results than games between 1 and 10, even though each game should have the same predictive performance.







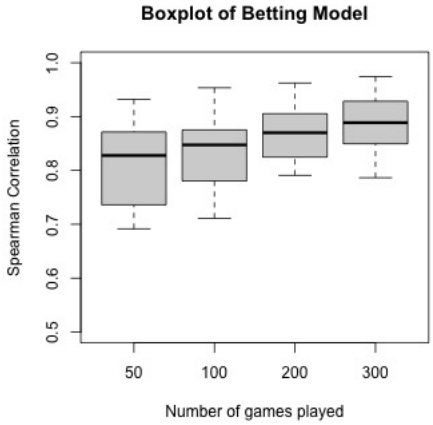


Table 2: Spearman Correlation Averages of 2021-22 seasons					
Games	Betting	Logistic	Elo	Rank	Baseline
50	0.812	0.5432	0.8118	0.7517	0.6831
100	0.834	0.7167	0.8367	0.8654	0.82
200	0.869	0.8143	0.8887	0.9218	0.8932
300	0.8868	0.8786	0.9177	0.9242	0.9442

The boxplots and Table 2 above illustrate the Spearman Correlation from seasons 2012-21 after 50, 100, 200, and 300 games. The key takeaway from these boxplots is the consistency and stability of the Rank Model, producing the best results throughout various times in the Premier League season. We can see that after 50 games, the Elo Model and Betting Model have a slightly higher Spearman Correlation, likely because the Elo Model can achieve its peak performance after few results since it's only looking at the results of the games. However, it's hard for Elo to continuously improve throughout the season because, as stated earlier because single matches have too much influence on the Elo Model. Also, it makes sense that the Betting Model would perform better than the Rank Model after 50 games. As discussed in Section 3.2, betting odds have more inherent information and don't need as much data to rank teams successfully because betting odds use lots of data from previous seasons that we do not have access to. However, betting odds don't seem to adjust too much throughout the season. They don't react too much to in-season results compared to the Rank Model, which could be an issue with odds-makers focusing more on the previous season results when calculating betting odds. After 100 and 200 games, we can see that the Rank Model performs the best. It seems that the Rank Model needs data from each team playing about ten games to be able to start successfully ranking teams.

However, after 300 games, the Baseline Model performs slightly better than the Rank Model. This is likely because since there are only 380 games in the season, the table doesn't adjust too much throughout the last 80 games of the season. In the Rank Model, after 300 games, the model begins to perform equally as well as other models, likely because we still rely too heavily on previous season results. We mean that after 300 games, the previous season's results should have little to no impact on the current season's results because we have enough data from the current season to predict the rank of the teams. Regarding the first method, the Logistic Method struggles to compete with other models, has very high variance, and performs poorly until the 300-game mark.

4.2 Betting Prediction

In this section, we take an in-depth look at the Rank Model and see how it compares against betting odds systems.

Table 3: Betting Predictions

Seasons	Betting-Odds-200-Games	Rank-Model-200-Games	Betting-Odds-100-Games	Rank-Model-200-Games
20-21	4-4	3-5	8-1	9-0
19-20	4-4	4-4	6-2	6-2
18-19	4-3	6-1	8-1	8-1
16-17	5-3	5-3	3-4	3-4
15-16	2-4	3-3	6-2	6-2
14-15	5-3	5-3	4-3	4-3
13-14	8-2	10-0	5-2	4-3
12-13	6-2	6-2	4-2	5-1
Total	39-25	42-22	44-18	45-17

We wanted to create this table to compare the Rank Model's predicting accuracy against actual betting odds and its prediction accuracy. Sportsbooks are generally considered very good at predicting the outcome of sports matches. Therefore we aim to compare the prediction of the Rank Model to the betting prediction. One remark here is that the actual betting odds (Subsection 3.2) do not relate to the Betting Model, and we are looking at the betting odds given to us in the data. It's different because, in the Betting Model, we fit the betting odds using linear regression and then ranked the teams, producing different results than the actual betting odds. This table was created by looking at games 101-110 and 201-210 in each season and predicting which team would win each game. For the Betting-Odds-200 and Betting-Odds-100 columns, we predicted each of the ten games after the 100 game mark and 200 game mark by looking at the game's betting odds and predicting that the team with the lower betting odds would win the game. If the game ended in a draw, then we didn't count it in our predictions which is why each season often has less than ten games. We looked at just ten games at each point because each team plays once in a ten-game span.

For example, in Season 20-21, the Betting-Odds-200-Games was 4-4. This means that in games 201-210 in the season, the team with the lower betting odds won four of the games, the team with the higher betting odds won four of the games, and two of the games ended in a draw. This means that the betting odds predicted the winning team correctly four times and incorrectly four times. It's significant to look at the betting odds because the team with the lower betting odds is the team that the sportsbook is projecting to win the match, so that sportsbooks would rank the team with the lower betting odds as the better team. The total for Betting-Odds-200 was 39-25, which means the team with the better odds won 39 games, the team with the worse odds won 25 games, and the total for the Betting-Odds-100 was 44-18. For the Rank-Model-200, we used the Rank

Model to get the coefficient values of each team after 200 games played. Then, we updated these coefficient values based on statistics we knew pre-match, like whether the team was at home and the previous season's ranking. Next, we predicted that the team with the higher coefficient value would win the match and compared that against the actual results of the game. As we can see from the table, after both 100 and 200 games, the Rank Model performed slightly better than the betting odds, which shows that the Rank Model can predict the outcomes of each match similarly and possibly slightly better than real sportsbooks. The Betting Predictions table illustrates that the Rank Model is just as good at predicting the results of a match as betting odds by a sportsbook.

4.3 Predicting tiers

A key aspect of prediction in soccer is predicting what tier a team will fall in. There are three main tiers that researchers look at that are valuable for prediction in the English Premier League. The first tier predicts the winner of the Premier League. The second tier predicts if a team will make the Champions League, a top-four position in the Premier League. The final tier predicts if a team will get relegated, which is if they finish in the bottom three spots in the Premier League. These three tiers are important because teams gain substantial money by making the Champions League or winning the Premier League and lose a substantial amount of money if they get relegated or demoted to the lower league.

Table 4: Predicting Winner/Champions League/Relegation

.	Rank	Rank	Rank	Baseline	Baseline	Baseline
Games	Champion	Champions League	Relegation	Champion	Champions League	Relegation
50	5/8	22/32	12/24	2/8	18/32	11/24
100	4/8	24/32	17/24	4/8	23/32	15/24
200	4/8	27/32	17/24	4/8	28/32	14/24
300	3/8	26/32	19/24	8/8	27/32	19/24

In this section, we aggregate our results across all eight seasons, which is different from the other results we have produced in previous subsections where we focus on each season individually. Table 4 compares the prediction results of the Rank Model and the Baseline Models compared to the end of the season results. We looked at eight seasons when predicting these results, so there were eight champions, 32 teams who qualified for the Champions League, and 24 teams who were relegated to the lower league.

We now explain the performance of our models: the Rank model and the Baseline model. Recall that the Baseline model is what the live Premier League table looks like throughout different points in the season. After 50 games, the Rank Model predicts the champion five out of eight times correctly, i.e., the top team in the Rank Model was the true champion in that specific season five times out of the eight seasons we looked at. The Baseline, after 50 games, predicted the champion two out of eight times. This means that if we looked at the live Premier League table after 50

games, the team winning the Premier League at that point only won the league two out of the eight seasons.

The Rank Model is also significantly better at predicting the three categories (Champion, Champions League, Relegation) after 50, 100, and 200 games. However, as we've seen with earlier tables in the results section, after 300 games, the Rank Model begins to perform worse than the Baseline Model. Some interesting results from this table are that after 50 games, the Rank Model predicts five of the eight champions compared to two out of the eight champions in the Baseline Model, which is the best at any point in the season. The success of the Rank Model indicates that the previous season's result can be an essential factor in predicting the champion at an early stage. Another interesting result is that after 300 games, our Rank Model predicted three out of eight games, and the Baseline Model predicted eight out of eight games. This shows that our model can sometimes over-fit at the end of the season compared to the Baseline Results. Towards the end of the season, the ranking becomes more stable, and game statistics become less informative for prediction, i.e., the ranking at the end of 300 games is the same as at the end of the season (i.e., 380 games).

5 Conclusion

In this project, we compare several models' performance for predicting the teams' final rank based on a few initial games. Of all the methods, the Rank Model has the most stable and overall best performance in terms of Spearman correlation. This performance can probably be attributed to the Rank Model using a combination of statistics throughout the match. As we can see from Table 2 in Section 4, compared to other methods, the output of the Rank Model trained using 100 games has the maximum correlation with the true final ranks. However, after 300 games, the baseline performed slightly better because the season is only 380 games, so looking at the actual results after 300 games would be the best way to predict the end of the season results. Some ways to improve the Rank Model would be to adjust the coefficients for each variable along the season (i.e., fitting some time-varying regression model). For example, at the start of the season, the previous season's results should play more of a factor than in-match statistics. However, as the season goes along, statistics and the results of each match should tell us more about the strength of teams than the previous season rankings because teams adapt and improve as the season goes along, and previous season rankings do not help with that information.

In soccer, the superior team often loses or draws the game because it's hard to predict when goals are scored, and they often occur at random times. This makes it easier for inferior teams to score and win the game than in other sports where more points are scored, like basketball, or teams have clear offensive and defensive positions like American football. In other sports, the best teams often have enough time to overtake the worse team because they have more opportunities. In soccer, it is often not possible to judge the quality of a team just by looking at the number of goals scored, especially in the first few matches. Other detailed information about the match is more valuable for understanding the true caliber of a team. Therefore, an efficient aggregation of this information is expected to be more revealing for predicting the final ranks of teams. So, to conclude our investigation of methods for predicting soccer results, the Rank Model is the most efficient at producing regularly accurate predictions, and after 300 games, one should use a Baseline model as they are very stable at the end.

References

[AH21] Halvard Arntzen and Lars Magnus Hvattum. Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, 21(5):449–470, 2021.

[BK19] Rahul Baboota and Harleen Kaur. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 35(2):741–755, 2019.

[BKM17] Georgi Boshnakov, Tarak Kharrat, and Ian G McHale. A bivariate weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466, 2017.

[CDLR02] Martin Crowder, Mark Dixon, Anthony Ledford, and Mike Robinson. Dynamic modelling and prediction of english football league matches for betting. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51(2):157–168, 2002.

[CF13] Anthony Costa Constantinou and Norman Elliott Fenton. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, 9(1):37–50, 2013.

[Con19] Anthony C Constantinou. Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, 108(1):49–75, 2019.

[Dat] Premier league table amp; stats — footystats.

[Jay18] Jayboice. How our club soccer predictions work. *FiveThirtyEight*, Aug 2018.

[Mit20] Raghav Mittal. What is an elo rating? *Medium*, Nov 2020.