Improving Elo Rankings For Sports Experimenting on the English Premier League

Connor Sullivan CSULL@VT.EDU
Virginia Polytechnic Institute and State University

Christopher Cronin CHRISCRO@VT.EDU
Virginia Polytechnic Institute and State University

Abstract

In this paper we examine the Elo rating system and how it can be applied to the English Premier League for predicting the outcome of matches. Specifically we examined four different methods of modifying the basic Elo formula to improve prediction accuracy. The four methods with which we experimented were incorporating home field advantage, adjusting the K-factor at different points in a season, rewarding and penalizing winning and losing streaks, and rewarding a win proportionally to the margin of the win. By incorporating these additional parameters into the Elo formula, we were able to achieve a notable increase in prediction accuracy over the basic Elo formula.

1. Introduction

The Elo rating system was created by physicist and chess Grand Master Arpad Elo, as an improvement upon existing chess rating systems. The difference in ratings of two players serves as a predictor of the outcome of a match. For example, if the difference in scores of two players is 100, then the stronger player is predicted to have a 64% of winning a match against the weaker player. The formula for predicting a match's outcome is:

$$E[S_A] = \frac{10^{R_A/400}}{10^{R_A/400} + 10^{R_B/400}}$$
$$E[S_B] = \frac{10^{R_B/400}}{10^{R_B/400} + 10^{R_A/400}}$$

After a match each player's rating is updated based on the outcome of the match. If the favored player wins the match, he will gain relatively few points. However if there is an upset and the weaker player wins, his rating gains relatively more points, than the favored player would if he had won. The formula for updating ratings is as follows:

$$R_{\text{post}} = R_{\text{pre}} + K(S - E[S])$$

$$S = \begin{cases} 1, & \text{for a win} \\ .5, & \text{for a draw} \\ 0, & \text{for a loss} \end{cases}$$

In the above equation K serves a weighting term that determines how much the previous match should affect the player's rating. The higher the K-factor, the quicker a player's rating will rise or fall. The value of K depends on the particular application of the Elo system. In our implementation, we used a base value of 20. This is the value used by the World Football Elo system for friendly matches.

When new players enter the system, they are assigned a default rating. This default rating is again application specific. We chose a default rating of 1200 in our implementation to match the value used by the World Football Elo system.

In our experiments, we used a data set from the English Premier League, that contains historical data on matches spanning more than 100 years. We focused on the most recent data by training on recent seasons and predicting outcomes for the latest season.

Lastly, the outcomes of a match are either a win, a draw, or a loss as described previously. Our model predicts a win if the score is at least 0.60 (60% probability that this team will win) and predicts a loss if the score is less than or equal to 0.40 (60% probability that the other team will win. Between 0.40 and 0.60,

Virginia Tech CSx824/ECEx424 final project report.

a draw is the predicted outcome.

2. Experiments

The following subsections detail the four experiments run on individual factors being adjusted. Given the short timeline for this research, we analyzed each factor individually first, and then brought them together into a final model.

Each experiment was modeled off of the following steps:

1. Set the minimum, maximum, and step values for the parameter.
2. Set the number of years for training.
3. For each iteration of training, loop over possible parameter values and record results.
4. Visually examine the recorded best results and choose the best parameter value.

Also due to the short timeline, the robustness of our experiments and training methods may not be optimal. For instance, since the original motivation behind this research is to improve the Elo rating system in order to better predict next season's games, we trained to predict the 2013 season. This means that we trained on the n previous seasons and then evaluated our prediction accuracy on the 2013-14 season.

Throughout our experiments we mention correctness and almost correctness. We define correctness as the percentage of predictions that were correct (e.g. we predicted a win for team A and its true outcome was a win for team A). However, if we predict a draw and the outcome ends up being a close game that resulted in a win (e.g. the final score is 1-0), then our prediction was close. Thus, we define almost correctness as the percentage of predictions that were almost correct, meaning that we are only wrong if we predict a win and the outcome is a loss or vice versa. This metric was used earlier on in our experimentation to get a better idea of how well our model was doing. In our final conclusions, however, we do not reference almost correctness since correctness is the ultimate metric on which we focus (in real-world application, almost correct doesn't count).

2.1. K-Factor

The first factor examined was altering the K-factor. Altering the K-factor consists of two parts: setting the standard K-factor to be used throughout the season, and setting the initial (higher) K-factor to be used at the beginning of the season. The intuition behind this change in K-factor is that the beginning of the season demonstrates the quality of this season's team, so setting a higher K-factor allows for bigger adjustments to the Elo rankings of the teams. A few examples that affect a team's quality between seasons include acquiring or losing players, changing management, and changing the teams in the league.

As mentioned previously, the initial, standard K-factor was set at 20, reflecting the World Football Elo Ranking system's value for friendly matches. To implement a higher initial K-factor, the point at which the K-factor dropped back to its standard level had to be decided. A single Premier League season consists of 38 matches for each club. We decided to set this change point, call it C, to be 7. This was chosen, because seven is just under two months into the season and the season is almost one-fifth completed[1]. Therefore, the update formula remains the same, but the choice of K must be made.

$$R_{\text{post}} = R_{\text{pre}} + K(S - E[S])$$

$$K = \begin{cases} K_{\text{initial}} & \text{if matches} < \text{C} \\ K_{\text{normal}} & \text{if matches} \geq \text{C} \end{cases}$$

First, we trained our model to find the best K_{normal}. The number of seasons to train on was set to be 1, 2, 3, and 5. The values for K_{normal} ranged from 5 to 60 with a step size of 5. Looking at the outcomes of all four of these, the value of $K_{\text{normal}} = 25$ was chosen. The original choice of 20 was a very good estimate for this value which is expected since the World Elo Rating system is well known.

Second, we trained our model to find the best K_{initial} given a changing point of $C = 7$ and the above K_{normal}. Training on the same seasons and using the same range yielded a $K_{\text{initial}} = 40$ to be the best choice. Figure 1 illustrates improvement we see when training on the single previous season. For this experiment, the training done on just the previous season was the only training that saw improvement by setting different K_{initial} values. Training on more than one season resulted in similar optimal values for K_{initial}, but often included multiple values that shared the same correct percentage. All of these multi-year trainings had a range of less than 0.5%.

It is important to note that both of these experiments

[1] Note: We do acknowledge that this value was determined arbitrarily. In the future, this value could be optimized.

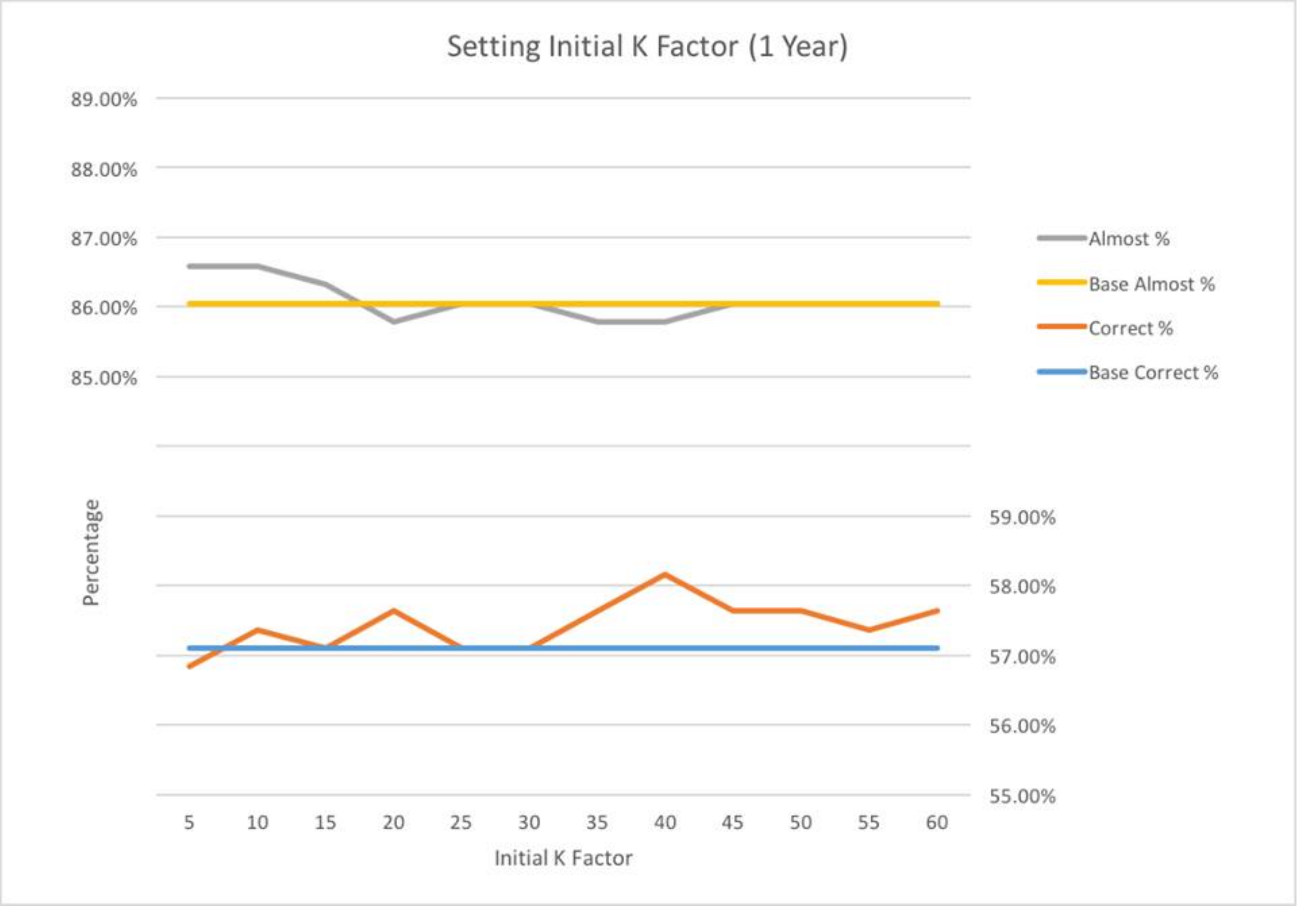

Figure 1. Initial K Factor (1 Season)

were conducted using a home field advantage boost equal to 100. This is the home field advantage mentioned by the World Football Elo rating system. The home field advantage boost will be discussed in the next section. Thus, in the end, setting $K_{\text{initial}} = 40$ and $K_{\text{normal}} = 25$.

Table 1 illustrates the change in prediction accuracy over 1, 2, 3 and 5 seasons. As you can see, incorporating only a different initial K-factor doesn't improve the model very much. However, we still select the above K-factor values, because when other parameters are added, the effects of the K-factor are much greater than 1%.

2.2. Home Field Advantage

Home field advantage was the next feature incorporated. Intuitively, the home team was an increased chance of winning, because they are at home. The reasoning behind this can be attributed to familiarity with the stadium or turf, supporters, sleeping in one's own bed the previous night, etc.; the list goes on. This boost was added to the difference rating.

Yrs	% Increase
1	0.480769
2	-0.934579
3	0.473934
5	2.884615

Table 1. K-Factor Accuracy Increases

Thus, the new prediction equation is

$$E[S_{\text{home}}] = \frac{1}{1 + 10^{-(R_{\text{home}} - R_{\text{away}} + h)/400}}$$

$$E[S_{\text{away}}] = 1 - E[S_{\text{home}}]$$

where h is the home field advantage boost.

Note that the above predictions are from the home team's perspective. This is a reflection of design decisions made during the implementation of our Elo Rating system: always calculate the home team's prediction and then deduce the away team's prediction.

Yrs	% Increase with base K-factor	% Increase with optimal K-factor
1	9.497207	17.877095
2	14.835165	17.582418
3	18.888889	20.000000
5	17.679558	18.232044

Table 2. Home Field Prediction Accuracy Increases

The range for home field advantage was from 0 to 200 with a step size of 10. Training for the home field advantage boost was performed twice: first with the base K factors and then with the optimal K-factors discussed previously. The models were trained on 1, 2, 3, and 5 seasons.

When the home field boost was optimised based on the original, base K-factor of 20, the increase in prediction accuracy was significant with an average increase over 15%. These values are tabulated in Table 2. The optimal home field boost for this case was determined to be 130.

When the optimal altering K-factors were used from the previous experiment, the optimal home field boost value returned was 120. The increases in prediction accuracy due to home field advantage are very significant: increasing by over 17% for each training set. These values are also shown in Table 2

2.3. Modified Point Assignment

The next factor we considered when modifying the Elo formula is the number of points assigned for a win, loss, and draw. The intuition behind this, is that a victory or a loss by a large margin should have a larger impact on the team's rating, than a win or loss by a small margin. To account for this in the Elo rating update, the score assigned for a win or loss becomes proportional to the margin of the victory. The score assignment used in the Elo update equation then becomes:

$$S = \begin{cases} 1 + (|s_m| - 1)F_m, & \text{for a win} \\ .5, & \text{for a draw} \\ 0 - (|s_m| - 1)F_m, & \text{for a loss} \end{cases}$$

where F_m is the marginal score factor and s_m is the marginal score, the difference of the home team and away team's match scores. With this formula, if a team wins a match by the minimum possible margin of 1, they are rewarded with the standard Elo score of 1. However, for each victory with a margin of 2 or more, the team is rewarded with 1 plus some bonus. To examine the impact this modification had on the model's predictive performance, the value of F_m was varied from 0 to 2 in steps of 0.01. Figure 2 shows the predictive performance for various values of F_m trained on 1 season of data.

Modifier	Correct	Almost
0.00	52.63158%	81.05263%
0.10	58.42105%	81.05263%
0.20	57.63158%	81.05263%
0.30	57.63158%	81.84211%
0.40	58.68421%	82.63158%
0.50	57.63158%	82.36842%
0.60	57.36842%	81.57895%
0.70	56.84211%	81.05263%
0.80	56.84211%	81.31579%
0.90	57.10526%	80.78947%
1.00	56.31579%	80.00000%
1.10	56.57895%	79.47368%
1.20	57.10526%	79.21053%
1.30	57.36842%	79.21053%
1.40	57.10526%	80.00000%
1.50	56.57895%	80.00000%
1.60	56.31579%	80.00000%
1.70	55.78947%	79.73684%
1.80	55.78947%	79.73684%
1.90	56.05263%	79.73684%
2.00	56.31579%	80.00000%

Table 3. Modified Score Assignment (3 Seasons)

Examining this figure, we see that for nearly all values of F_m there is an improvement in the predictive performance of the model. Examining the numerical data, we found that a value of 0.57 gave the best accuracy of 60.00%, which is a roughly a 10% increase in accuracy over the basic Elo model performance.

We ran similar experiments where the model was trained on 3 years and 5 years of data. These experiments produced similar results to the first. This data is summarized in Table 3 and Table 4.

2.4. Streaks

The final factor that we incorporated into the Elo model, is win and loss streaks. The intuition behind adding this is: if a team is on a winning streak, it is likely that they will continue the streak and win the next game. Likewise, if a team is on a losing streak, it is likely that they will lose the next match. An existing paper studying the addition of momentum to the Elo system proved useful when implementing our method (Bester & von Maltitz, 2013). This idea was incorporated into the model in a similar manner to

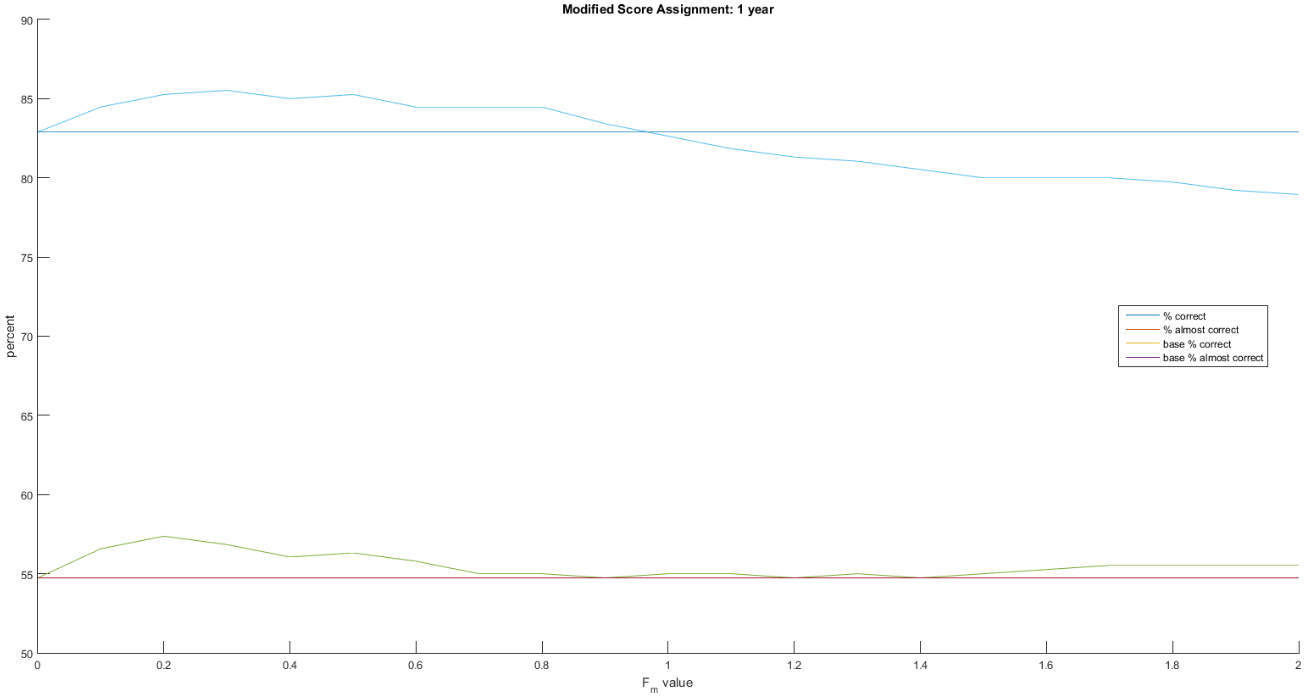

Figure 2. Modified Score Assignment (1 Season)

Modifier	Correct	Almost
0.00	54.47368%	81.05263%
0.10	58.15789%	81.31579%
0.20	56.84211%	81.57895%
0.30	57.89474%	81.31579%
0.40	58.42105%	81.57895%
0.50	56.84211%	81.31579%
0.60	57.36842%	80.78947%
0.70	57.36842%	80.26316%
0.80	56.57895%	79.73684%
0.90	56.57895%	80.00000%
1.00	56.31579%	79.47368%
1.10	56.57895%	79.21053%
1.20	56.31579%	79.21053%
1.30	56.57895%	79.21053%
1.40	56.57895%	79.47368%
1.50	56.84211%	78.42105%
1.60	56.57895%	78.15789%
1.70	56.57895%	78.42105%
1.80	55.78947%	78.42105%
1.90	56.05263%	78.15789%
2.00	56.31579%	78.15789%

Table 4. Modified Score Assignment (5 Seasons)

home field advantage, in that each team has a boost added to their predicted outcome if they are on a winning streak, and a penalty subtracted if they are on a losing streak. To add streaks to the model we first rewrite the prediction equation as:

$$E[S_A] = \frac{1}{1 + 10^{-(R_A - R_B)/400}}$$
$$E[S_B] = 1 - E[S_A]$$

We then apply bonuses N_A and N_B to the difference of the team's ratings, to get:

$$E[S_A] = \frac{1}{1 + 10^{-(R_A - R_B + N_A - N_B)/400}}$$
$$E[S_B] = 1 - E[S_A]$$

where

$$N_{A,B} = \begin{cases} C, & \text{for a win streak} \\ 0, & \text{no streak} \\ -C, & \text{for a loss streak} \end{cases}$$

C is some constant boost and was varied from -100 to 100 in steps of 10 to find the optimal value. An additional parameter that had to be considered for this experiment was the threshold for a sequence of wins or losses to be considered a streak. We varied this

Modifier	Correct	Almost
-100.00	56.05263%	84.73684%
-90.00	55.26316%	85.52632%
-80.00	55.78947%	86.05263%
-70.00	55.78947%	85.78947%
-60.00	56.05263%	86.05263%
-50.00	56.57895%	86.05263%
-40.00	58.15789%	86.05263%
-30.00	58.94737%	85.26316%
-20.00	58.94737%	85.52632%
-10.00	58.94737%	85.52632%
0.00	58.15789%	84.73684%
10.00	57.89474%	84.47368%
20.00	56.57895%	84.73684%
30.00	56.31579%	84.21053%
40.00	56.05263%	83.94737%
50.00	56.31579%	83.94737%
60.00	56.57895%	83.68421%
70.00	56.05263%	83.42105%
80.00	55.26316%	82.89474%
90.00	54.73684%	82.63158%
100.00	54.21053%	82.10526%

Table 5. Streaks (1 Season)

Modifier	Correct	Almost
-100.00	55.00000%	83.94737%
-90.00	55.26316%	85.26316%
-80.00	55.52632%	85.52632%
-70.00	54.47368%	85.78947%
-60.00	54.47368%	85.78947%
-50.00	55.00000%	86.57895%
-40.00	55.78947%	85.52632%
-30.00	56.31579%	85.52632%
-20.00	57.36842%	85.52632%
-10.00	57.63158%	84.73684%
0.00	56.84211%	84.47368%
10.00	57.10526%	84.21053%
20.00	56.31579%	84.47368%
30.00	55.78947%	84.21053%
40.00	54.47368%	83.94737%
50.00	53.94737%	84.21053%
60.00	53.94737%	83.68421%
70.00	53.42105%	83.15789%
80.00	53.94737%	82.89474%
90.00	54.21053%	82.63158%
100.00	54.21053%	81.57895%

Table 6. Streaks (3 Seasons)

Modifier	Correct	Almost
-100.00	54.47368%	83.94737%
-90.00	54.47368%	84.47368%
-80.00	55.26316%	85.26316%
-70.00	55.00000%	85.52632%
-60.00	54.73684%	85.52632%
-50.00	55.00000%	85.52632%
-40.00	56.31579%	85.00000%
-30.00	56.57895%	84.47368%
-20.00	57.89474%	84.21053%
-10.00	58.42105%	84.21053%
0.00	57.63158%	83.94737%
10.00	56.84211%	84.21053%
20.00	56.57895%	83.94737%
30.00	56.31579%	83.68421%
40.00	55.52632%	83.15789%
50.00	55.26316%	83.15789%
60.00	54.73684%	82.63158%
70.00	55.26316%	82.63158%
80.00	55.26316%	82.10526%
90.00	55.00000%	81.84211%
100.00	56.05263%	81.31579%

Table 7. Streaks (5 Seasons)

threshold from 2 to 4, and while all thresholds gave some improvement when trained on a single season, a value of 2 produced the best overall results. We then trained a model with these parameters on 3 seasons and then 5 seasons. In each case the model performed better than the basic Elo model. Analyzing the results we found a value of -10 to be optimal. This would indicated that our original intuition about streaks was incorrect. The model seems to indicate that a team is more likely to break the streak in their next game than to continue it. The results of these experiments are summarized in the tables 5, 6 and 7 below.

2.5. Grid Search

The final experiment that we conducted was to run a grid search over the entire parameter space with the exception of the K-factor. We held the K-factor fixed at the initial value of 40 and normal value of 25. This experiment is obviously inefficient due to the size of the parameter space, but can provide insight on how the different factors we investigated work together for the single best model. The grid search was run with training on 1, 2, 3, and 5 seasons previous to the 2013-2014 season.

We chose the bounds for our parameters based on the experiments we ran on the individual parameters. The

Experiment	Optimal
K-factor	$K_{\text{initial}} = 40$
	$K_{\text{normal}} = 25$
Home Field	$h = 200$
Modified Points	$F_m = 0.50$
Streaks	$C = -40$

Table 8. Summary of Experiment Results

Parameter	Original	Optimal
K_{initial}	20	40
K_{normal}	20	25
Home Field, h	N/A	120
Modified Points, F_m	N/A	0.20
Streaks, C	N/A	-10

Table 9. Original and Optimal Model Parameters

home field boost ranged from 120 to 300 with a step size of 10, the win streak factor was between -50 and 75 with an increment of 0.1, and the win loss margin factor for score assignment ranged from 0 to 60 by 1.

The experiment exported every parameter permutation with correct and almost metrics to a csv file. This file was sorted with respect to the correct percentage, and then by the almost correct percentage. We then examined the file to see the trends across the different training seasons and to select optimal parameter values.

The highest correct accuracies were found with home field advantage boosts ranging from 180 to 260, but the center of these appeared to be 200. Thus the optimal home field advantage boost was chosen to be 200. The win loss factor optimal value fell in the range of 0.40 to 0.60. Thus, choosing the center of these yields the optimal win loss factor to be 0.50. Lastly, the win streak factor showed a surprising range of -50 to 0, with the distribution skewed to the right. The value -40 was chosen for the optimal win streak.

3. Results and Conclusions

3.1. Results

Table 8 summarizes the results found in each of the experiments. Table 9 shows the original values for the base Elo model and the optimal values found in the grid search.

Premier League (380 matches)

Yrs	Original	Optimal	% Increase	% Correct
1	179	229	27.932961	60.26316
2	182	227	24.725275	59.73684
3	180	228	26.666667	60.00000
5	181	222	22.651934	58.42105

Table 10. Prediction Accuracy for Premier League

Combining all of the adjustments made to the Elo rating system for each parameter together yields the following equations for prediction and update.

$$R_{\text{post}} = R_{\text{pre}} + K(S - E[S])$$

$$S = \begin{cases} 1, & \text{for a win} \\ .5, & \text{for a draw} \\ 0, & \text{for a loss} \end{cases}$$

$$K = \begin{cases} K_{\text{initial}} & \text{if matches} < \text{C} \\ K_{\text{normal}} & \text{if matches} \geq \text{C} \end{cases}$$

$$E[S_{\text{home}}] = \frac{1}{1 + 10^{-(R_{\text{home}} - R_{\text{away}} + N_{\text{home}} - N_{\text{home}} + h)/400}}$$

$$E[S_{\text{away}}] = 1 - E[S_{\text{home}}]$$

$$N_{A,B} = \begin{cases} C, & \text{for a win streak} \\ 0, & \text{no streak} \\ -C, & \text{for a loss streak} \end{cases}$$

3.2. Conclusions

These values support most of the hypotheses we made at the beginning of our research. altering the K-factor, boosting the home team's rating, and factoring in the margin of victory or defeat all improved the prediction accuracy. Our hypothesis about the win streak factor was not supported though. However, when we allowed for this factor to be negative, it did improve the prediction accuracy. Intuitively this means that a team on a win streak is less likely to keep the streak going.

The optimized model was ran against the original model to determine the increase in prediction accuracy. The optimized model improved the original model's prediction accuracy by over 20%. Table 10 displays the comparisons for training on 1, 2, 3 and 5 seasons.

This research shows a very significant improvement for predicting the outcome of soccer matches. The factor that contributed the most to the improvement was the home field advantage.

Champions League (552 matches)

Yrs	Original	Optimised	% Increase	% Correct
1	180	243	35.000000	44.02174
2	198	246	24.242424	44.56522
3	204	242	18.627451	43.84058
5	202	249	23.267327	45.10870

Table 11. Prediction Accuracy for Champions League

League 1 (552 matches)

Yrs	Original	Optimised	% Increase	% Correct
1	198	253	27.777778	45.83333
2	195	245	25.641026	44.38406
3	204	241	18.137255	43.65942
5	209	242	15.789474	43.84058

Table 12. Prediction Accuracy for League 1

An interesting note about the home field advantage boost is that home field advantage appears to have a bigger impact previously than it does currently. This observation results from models that trained farther back in time returned a higher home field advantage boost.

3.3. Applying Our Model to Other Leagues

The data set on which we trained our data included multiple English football leagues. Up to this point, we have focused exclusively on the English Premier League. Now, however, we look at the other leagues included: the Champions League, League 1, and League 2 as they are currently known.

We ran the same comparison that we ran for the Premier League on these leagues: testing the optimized model against the original model. All the leagues saw significant improvement as well, although the improvement was not as much as the Premier League. This observation lends itself to the fact that the model may be over fit to the Premier League exclusively, rather than fit for all football leagues. This is understandable, however, since this paper details the research on fitting a model to the Premier League. More discussion on this thought can be found in the Further Development section.

Tables 11 to 13 display the comparisons for training on 1, 2, 3 and 5 seasons for the Champions league, League 1, and League 2, respectively. Even though these predictions weren't improved as much as the Premier League predictions, the smallest increase was still over 15%, which is significant.

League 2 (552 matches)

Yrs	Original	Optimised	% Increase	% Correct
1	172	214	24.418605	38.76812
2	169	205	21.301775	37.13768
3	174	205	17.816092	36.95652
5	172	204	18.604651	36.95652

Table 13. Prediction Accuracy for League 2

4. Further Development

The biggest roadblock throughout our research into improving Elo rankings for sports was the time. However, this was understood at the onset of the research. We completed all of our original plans for our research into this project. We had further goals in place if time permitted, but we were only able to begin looking into these goals. Primarily, the next step for our research would have been to develop an algorithm to train an optimal model on all of these parameters efficiently. To do this, a grid search can not be employed, and thus the parameter space would need to be examined further. Particularly, if any parameters were found to be convex, that would greatly aid in the creation of such an algorithm.

During our experimentation, it was noticed that home field appeared to be convex (ignoring slight noise since it is real world data). This held throughout the experiments, even when combined with the other parameters. Therefore, the home field boost could be found using standard optimization methods to arrive at a good local optimum.

The modified points assignment for win/loss margin appeared to vary drastically. A common shape was not easily discernible directly from the experimental results. Further investigation into this parameter would be needed.

The streaks parameter also posed difficulty. The fact that the parameter was negative can be explained, but isn't necessarily intuitive. We held the streak threshold constant throughout all of our experiments at a value of 2. The streak threshold is the number of games a club must consecutively win (or lose) to be considered on a streak. This streak threshold could be researched as well in the future.

Similarly, the changing point for transitioning between the initial and normal K-factor was set and held constant at 7, and, thus, could be researched in the future.

In conclusion, the goals that we set out to complete were completed, but there is plenty left to do with these results. For instance how well will this transition to other football leagues, or even other sports?

Acknowledgments

We would like to thank Dr. Bert Huang for providing us with guidance and feedback during this research project and for teaching us about machine learning throughout this semester.

References

Bester, D.W. and von Maltitz, M.J. Introducing momentum to the elo rating system. *University of the Free State: Department of Mathematical Statistics and Actuarial Science*, 2013. URL http://natagri.ufs.ac.za/dl/userfiles/Documents/00002/2069_eng.pdf.

Glickman, Prof. Mark E. A comprehensive guide to chess ratings. URL http://www.glicko.net/research/acjpaper.pdf.