# RANKING NCAA DIVISION I COLLEGE SOCCER TEAMS

*Vicki Nomwesigwa, Cameron Ratliff, Matt Dockman*

## Introduction

Soccer is undoubtedly the world's most popular sport. It attracts millions of spectators and players of all ages and demographics. Despite its popularity, the application of data-driven methods to soccer has been slow. Though there are models that predict match outcomes to serve the ravenous betting market, relatively little consideration has been given to ranking soccer teams. In fact, no ranking system for college soccer currently exists. This stands in stark contrast to college football and basketball, in which systems such as Massey, KenPom, and Colley matrix have been created to rank teams. For over a decade, such ranking systems were used to help determine the participants in the college football national championship.

Team rankings are important to generate compelling narratives for team performance. Rankings provide an ordinal structure that readily compares teams; when a low-ranked team defeats a high-ranked team, we can confidently declare that an upset occurred. When two high-ranked teams play, we can state that two of the best teams squared off. In short, rankings are essential to creating such narratives. Winning percentage is an option but notably does not consider the level of competition a team has played. If a good team with a difficult schedule defeats a bad team with a much easier schedule, winning percentage would imply an inaccurate narrative that the first team upset the second. Betting odds offer an alternative, but sportsbooks neglect to offer regular betting odds for college soccer games. As a further benefit, rankings can provide means by which to predict game and score outcomes and inform playoff seeding.

These reasons motivate the creation of a ranking system designed specifically for college soccer. In this paper, we create a ranking system for college soccer teams that has several key features that differentiate it from existing methods: (1) the model is derived probabilistically so that we can quantify uncertainty in our ratings, (2) each team's rating is interpretable as the expected number of goals by which it would defeat an average college team, (3) draws are modeled directly, and (4) we use an adjusted margin of victory metric. We demonstrate that our model creates reasonable rankings that conform with our intuition and expert knowledge and outperforms our baseline model, the FIFA ranking system, in match outcome and score prediction. The paper is laid out as follows: in the next section, we introduce the existing methods, move to formulating our model, and then discuss our results before concluding and providing avenues for future work.

## Literature Review

For our purpose, we will discuss football and basketball ranking methodologies in addition to soccer. Football and basketball are entirely different sports from soccer, but their modeling principles can still effectively be applied to soccer. The Colley matrix rankings, developed for football, use only game outcomes to simultaneously adjust each team's winning percentage by strength of schedule. Also developed for football, the Wolfe rankings employ a Bradley-Terry model estimated via maximum likelihood estimation to approximate team strengths. Ley et al. apply similar methods to soccer in using Thurstone-Mosteller and Bradley-Terry type models. Despite their interpretability, these models do not fully account for the strength of a victory. Surely, a 4-0 margin is more significant than a 1-0 margin and should be treated as such. To account for win margins, the Massey football rankings use the score of each game to create a prior distribution for each team's power rating. The actual game outcomes are then used to provide a Bayesian correction to the team' power rating to form the overall rating. Ley et al. include match score information by applying independent Poisson and bivariate Poisson models, improving their models mentioned earlier. Despite offering acceptable performance, these models are too simplistic as they fail to integrate in-game features, which we suspect are strong predictors of team performance. Moreover, they cannot easily be tailored to account for draws, a much more probable outcome in soccer than in soccer or football.

Considering more complex models, the KenPom rankings for college basketball apply the Kalman filter to update offensive and defensive efficiency ratings based on strength of schedule. Though KenPom is too basketball-specific to apply to soccer, its interpretation of each team's rating as its expected margin of victory against an average college team is our inspiration for the interpretation of our ratings. Li et. al. (2020), in their analysis of Chinese Football Super League data, employ a linear support vector machine classifier to match outcomes. Team ratings for each game are then assigned according to the output of this classifier. For our purposes, we will consider a statistical, rather than a machine learning, approach so that we can quantify uncertainty. The most prominent method for ranking soccer teams is the Federation Internationale de Football Association (FIFA) rankings of national teams. The current ranking methodology, updated in 2018, has shown a stark improvement in providing rankings that reflect actual match outcomes. Employing an adjustive-rating system, each team's rating changes according to the difference between its game result and predicted probability of victory with a multiplicative factor for match importance. As the only existing ranking system employed in high-level soccer, it will serve as our baseline upon which we will attempt to improve.

## Methods

### Model Specification

Our goal is to create a ranking system for men's and women's NCAA Division I soccer teams that provides reasonable team rankings, predicts match outcomes and scores, and makes tournament projections. We will consider our ranking system successful if it can accurately predict at least 55% of three-way match outcomes (win, loss, and draw) and at least 85% of the teams that composed the field for the 2021 NCAA tournament. We chose this accuracy benchmark because predictive models for three-way outcomes for professional soccer generally have around 50% accuracy. 85% for tournament projections is a heuristic benchmark that considers the inherent randomness in the NCAA tournament selection process.

Creating sports rankings is difficult because we never observe the true rankings. Instead, we must use a team's performance throughout a season to discover the true rankings. Consequently, we do not have the true labels with which to compare our results to evaluate performance. As a proxy, we will use the United Soccer Coaches poll to determine if our rankings are reasonable.  In order to evaluate predictive performance, the current FIFA model will serve as our baseline model. Aside from being the only existing quantitative ranking system employed in high-level soccer, the difficulty FIFA faces in ranking national teams is very similar to ranking college soccer teams. In the four years between FIFA World Cups, national teams are mostly confined to playing teams within their own regions. Thus, FIFA must find a way to compare teams that never play and may have no common opponents. Similarly, in DI college soccer, 200 men's and 350 women's college soccer teams are grouped into conferences of 10-15 teams with conference schedules accounting for roughly $\frac{2}{3}$ of each team's games, providing few opportunities for crossover.

For clarity, the FIFA model ranks teams by assigning each a rating which is then updated after each game according to:

$$r_A^{t+1} = r_A^t + I\left(Result - p_w\right)$$

$$p_w = \frac{1}{1 + 10^{\frac{-\left(r_A^t - r_B^t\right)}{600}}}$$

where Result is 1 for a win, 0.5 for a draw, and 0 for a loss. $r_A^t$ and $r_B^t$ are the ratings of teams A and B, respectively, at time $t$, $p_w$ is the a priori probability of victory, and I is the match importance factor. FIFA created the initial ratings for their model through a conversion of the ratings from the previous ranking model. To apply their model to college soccer, we will treat the value at which we initialize the ratings as a hyperparameter. As far as the match importance factor, college soccer differs from national team soccer in that there are never meaningless games. Whereas there are international friendlies in national team soccer that differ little from exhibition games, we have no analogue in college soccer. Every game affects either a team's standing in their conference or their standing in the NCAA tournament, so we will set the match importance to be uniform across all games.

Ranking college soccer teams requires adjusting a team's performance by the strength of its opponent. FIFA's model accomplishes this by using a team's and its opponent's ratings to create a probability of victory for each team prior to a game. A team's rating is then updated according to a comparison between how a team was expected to perform, represented by their win probability, and how it actually performed, the result of the game. In this way, a team is rewarded less for winning games they were expected to win and more for games they were not. By updating ratings like this, we compare each team only to itself. This framework allows us to compare teams which may never play and may not even have any common opponents.

FIFA's approach of an adjustive-rating system is desirable for a few reasons. With an adjustive-based system, we have a consistent average rating across time, readily allowing comparison to an average team. This stands in contrast to an accumulation-based system wherein an average team's rating is continuously increasing as the season progresses (with wins resulting in larger increases), making comparison less intuitive. Moreover, adjustive-rating systems allow for a differing number of games for each team, which will be the case for college soccer due to weather and postseason play.

**Commented [GH1]:** This should be qualified with the team winning games, right?  Otherwise a team just grows in rating no matter what.  If a team accumulates no matter what, you should make it clear that teams that win accumulate more than teams that draw/lose.

To improve FIFA's model, we will make some significant changes. First, their model is not interpretable. Ratings generally range between 750 and 1800; it is unclear with this scale how much better one team is than another. If Belgium has an 1800 rating and England has a 1700 rating, it is not intuitive how often and by how much we would expect Belgium to defeat England. To improve on this, our model will interpret each team's rating as the expected number of goals by which it would beat an average team. Second, their model does not consider the margin of victory. The margin of victory should be informative in that teams that win by more tend to be better than those that scratch out wins. Third, their model does not allow for a home-advantage effect. Home teams win a disproportionate number of games due to the crowd, familiarity with the playing surface, etc. Fourth, the FIFA model does not account for the effect of past performance on the current game. There may be a momentum effect in that a team may be more likely to win its next game given that it won its past few due to an increase in confidence; conversely, there may be an opposing regression to the mean effect so that a team performs closer to its true strength in its next game after overperforming in the past few. Fifth, the FIFA model does not consider draws. The model is formulated by assuming away draws as ½ of a win and ½ of a loss and then modeling solely the probabilities of win and loss. Draws are a far more probable outcome in soccer than other sports, accounting for as much as 20% of outcomes, so we model the probability of a draw directly. Lastly, we introduce probability to quantify uncertainty. This allows us to measure how confident we are in our rating of each team.

To formulate our model, let $X_A$ be a random variable which represents the expected number of goals by which team A would win or lose if it were to play an average college team. We assume $X_A$ has distribution $N(\mu_A, \sigma_A^2)$, as shown in Figure 1, and take $\mu_A$, the mean of this distribution, to be the underlying rating for team A which we will hope to discover. $\mu_A$ represents the actual, long-run average expected number of goals by which a team would beat an average one. The standard deviation from this expectation $\sigma_A$ measures our uncertainty in our rating of team A; it is composed both of uncertainty resulting from the randomness in team performance as well as the randomness in our ability to accurately determine the true rating.

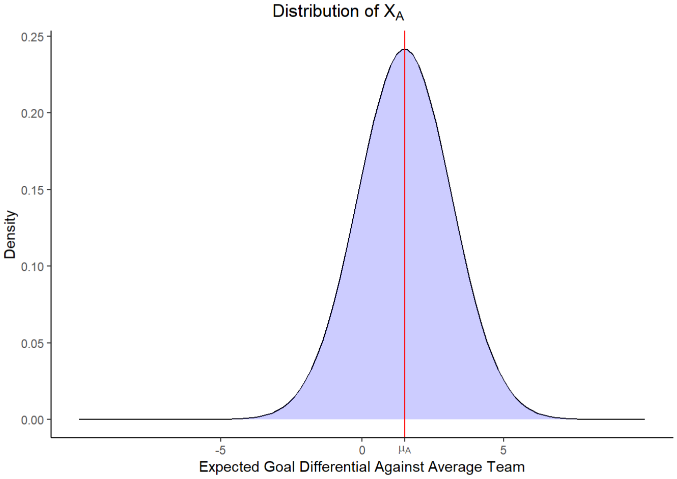


FIGURE 1. We let team A's strength be described by a random variable $X_A$. The mean $\mu_A$ represents the rating of team A while $\sigma_A$ represents the amount of uncertainty we have in team A's rating.

By adding the assumption of transitivity, this framework of comparison to an average team allows for easy recognition of team strength and comparison between teams. If team A were to play team B, we assume that team A would defeat team B, on average, by $X_A - X_B$ goals (if the result is negative, then we expect team B to defeat team A). To account for home-field advantage, we add a hyperparameter $h > 0$ (referred to as the home advantage parameter) to our model so that we expect team A to beat team B by $X_A - X_B + h$ goals at home and $X_A - X_B - h$ goals on the road, i.e., $h$ is added to the home team's strength. Our normal distribution assumption implies that $X_A - X_B \pm h$ is distributed as $N(\mu_A - \mu_B \pm h, \sigma_A^2 + \sigma_B^2)$, as shown in Figure 2, if we also assume that the correlation between these outcomes is 0.

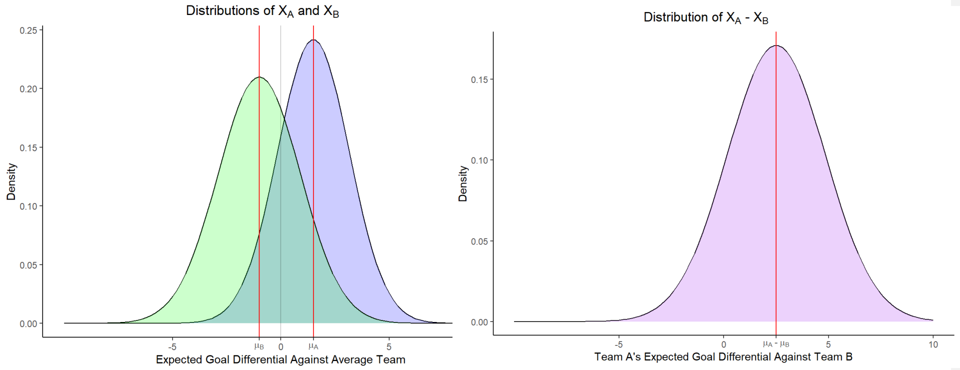


FIGURE 2. Left: Expected goal differentials of teams A and B in a game against an average team. Right: Team A's expected goal differential in a game against team B.

We note that we never actually observe $X_A - X_B$ the expected number of goals by which team A beats team B. Instead, we observe only the realization of one game. We could use the goal differential as a realization of $X_A - X_B$, but doing so seems a poor approximation. Rather, we use the normal distribution implied by $X_A - X_B$ to compute the probability of a win $p_w$, draw $p_d$, and loss $p_l$ for each game, as shown in Figure 3, and use the game result to track match outcome probabilities. In particular,

$$p_w = \mathbb{P}[r_A - r_B \pm h > 1] = \Phi\left(\frac{\mu_A - \mu_B \pm h - 1}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)$$

$$p_d = \mathbb{P}[-1 \leq r_A - r_B \pm h \leq 1] = \Phi\left(\frac{1 - (\mu_A - \mu_B \pm h)}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right) - \Phi\left(\frac{-1 - (\mu_A - \mu_B \pm h)}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)$$

$$p_l = \mathbb{P}[r_A - r_B \pm h < -1] = \Phi\left(\frac{-(\mu_A - \mu_B \pm h) - 1}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)$$

where $\Phi$ is the standard normal CDF.

**Commented [GH2]:** Add a caption to this figure that explains it. Also, label the x-axis the point of transition (-1 and 1) to make it clear where you are considering a draw (-0.5 and 0.5 might also be a natural choice).

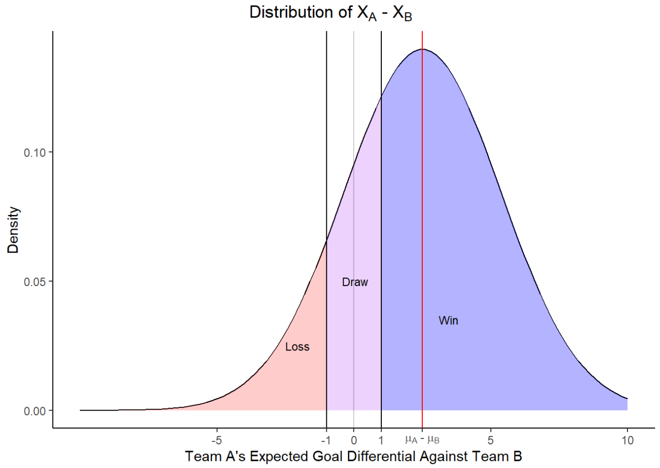


FIGURE 3. We formulate the probabilities of win, draw, and loss by taking any difference in expected goal differential less than -1 to be a loss, any difference greater than 1 to be a win, and anything in between as a draw.

Using these probabilities, the outcome of each game can be modeled as a multinomial distribution with 1 trial and probability vector given by $p = (p_w, p_d, p_l)$ . By further assuming that game outcomes are independent, we can find the likelihood of observing the actual game outcomes as

$$\mathcal{L}(\mu, \sigma) = \prod_{i=1}^{n} p_w^{1(y_i=1)}\, p_d^{1(y_i=0.5)}\, p_l^{1(y_i=0)}$$

where $\mu$ is a vector composed of $\mu_i, i = 1, \ldots, T$, the ratings for all $T$ teams, $\sigma$ is the vector composed of all $\sigma_i, i = 1, \ldots, T$, $1(\cdot)$ is the indicator function, $y_i$ is an indicator for the result for game $i$ with 1 for win, 0.5 for draw, and 0 for loss, and $n$ is the number of total games played.

To estimate $\mu_i$ and $\sigma_i$, we perform maximum likelihood estimation by conducting online gradient descent on the negative log-likelihood. This is appealing because the result is an adjustive-rating system, the merits of which we have discussed previously. In performing online gradient descent, we take each successive game a team plays, compute the gradients for the rating and its volatility using the prior values and the information from that game, and update our prior estimates. For a game between team A and team B, the gradients for team A are:

**Commented [GH3]:** It is great to include all of this; it may be better suited for the appendix, and to reference the appendix for more detail here.

For $y_i = 1$,

$$\nabla_{\mu_A}\ell = \frac{\phi\left(\frac{\mu_A - \mu_B \pm h - 1}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)}{\Phi\left(\frac{\mu_A - \mu_B \pm h - 1}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)\sqrt{\sigma_A^2 + \sigma_B^2}}$$

$$\nabla_{\sigma_A}\ell = \frac{-\phi\left(\frac{\mu_A - \mu_B \pm h - 1}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)(\mu_A - \mu_B \pm h - 1)\sigma_A}{\Phi\left(\frac{\mu_A - \mu_B \pm h - 1}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)(\sigma_A^2 + \sigma_B^2)^{3/2}}$$

For $y_i = 0$,

$$\nabla_{\mu_A}\ell = \frac{-\phi\left(\frac{-(\mu_A - \mu_B \pm h) - 1}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)}{\Phi\left(\frac{-(\mu_A - \mu_B \pm h) - 1}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)\sqrt{\sigma_A^2 + \sigma_B^2}}$$

$$\nabla_{\sigma_A}\ell = \frac{-\phi\left(\frac{-(\mu_A-\mu_B\pm h)-1}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)(-(\mu_A-\mu_B\pm h)-1)\sigma_A}{\Phi\left(\frac{-(\mu_A-\mu_B\pm h)-1}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)(\sigma_A^2+\sigma_B^2)^{3/2}}$$

For $y_i = 0.5$,

$$\nabla_{\mu_A}\ell = \frac{\frac{1}{\sqrt{\sigma_A^2+\sigma_B^2}}\left[-\phi\left(\frac{1-(\mu_A-\mu_B\pm h)}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)+\phi\left(\frac{-1-(\mu_A-\mu_B\pm h)}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)\right]}{\Phi\left(\frac{1-(\mu_A-\mu_B\pm h)}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)-\Phi\left(\frac{-1-(\mu_A-\mu_B\pm h)}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)}$$

$$\nabla_{\sigma_A}\ell$$

$$= \frac{-\phi\left(\frac{1-(\mu_A-\mu_B\pm h)}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)\frac{(1-(\mu_A-\mu_B\pm h))\sigma_A}{(\sigma_A^2+\sigma_B^2)^{3/2}}+\phi\left(\frac{-1-(\mu_A-\mu_B\pm h)}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)\frac{(-1-(\mu_A-\mu_B\pm h))\sigma_A}{(\sigma_A^2+\sigma_B^2)^{3/2}}}{\Phi\left(\frac{1-(\mu_A-\mu_B\pm h)}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)-\Phi\left(\frac{-1-(\mu_A-\mu_B\pm h)}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)}$$

where $\phi$ is the pdf of the standard normal and $\ell$ is the log-likelihood. Looking at these expressions seems daunting, but a few observations become clear when examining them analytically. Teams that win when they are expected to win, i.e., are rated higher than their opponent beforehand, receive a smaller update than those that are not as shown in Figure 4. When a team performs how we expect them to perform, i.e., win when they are expected to win, our uncertainty in their rating decreases. When we are more certain about a team's or its opponent's rating, its rating update is larger in magnitude regardless of win or loss. Intuitively, if we are more certain that a team's opponent is really good, then that team should receive a larger update for winning. Conversely, if we are more unsure about a team's rating, we would like to be more conservative with our updates. There is a greater chance that team may actually be much better than their rating suggests, so we do not want to penalize it as much for a loss. Lastly, when a team performs as we expect, the decrease in volatility is larger in magnitude as our prior certainty in that rating increases. This is to say that we become increasingly confident in our ratings when teams perform as we expect. On the other hand, when teams do not perform as we expect, the increase in volatility is larger as our prior uncertainty increases.

An underlying assumption we made in our model formulation is that game results are independent. This may not be true in practice. Recent past performance seems to have a significant effect on a team's current performance. A team that has played extremely well in their past 2-3 games may be much more likely to win their next game than their overall performance over the entire season would suggest. On the contrary, a team may win a few games in a row due to lucky bounces in which case that team might be less likely to win its next game than its current rating would suggest. To account for momentum and regression to the mean, we add a term for prior performance to our update. This term adds or subtracts, depending on the sign, a percentage $\beta$ (referred to as the prior performance parameter), of the prior update to or from the current update so that recent performance is better reflected in a team's rating. If $\beta > 0$ and a team has had a few victories in a row, the term serves as a momentum effect that increases that team's rating so that its win probability is higher for its next game. If $\beta < 0$ and a team has had a few victories in a row, the term serves as a regression to the mean effect that damps its rating increase so that its win probability is lower for its next game. We tune $\beta$ using positive and negative values to determine which of the momentum or regression to the mean effects is more relevant for college soccer. Specifically, we update a team's rating and its volatility after each game according to

**Commented [GH4]:** Again, what evidence do you have for this? Is this born out with your data?

$$r_A^{t+1} \leftarrow r_A^t - \alpha \nabla_{r_A} \ell + \beta\left(r_A^t - r_A^{t-1}\right)$$

$$\sigma_A^{t+1} \leftarrow \sigma_A^t - \kappa \nabla_{\sigma_A} \ell$$

where $\alpha$ and $\kappa$ serve as the step size parameters. $\kappa$ is a hyperparameter that we tune (referred to as the volatility update parameter); we also treat the initialization of the volatility, which is initialized uniformly across all teams, as a hyperparameter (referred to as the initial volatility parameter).

By construction, the gradient depends only on the game outcome and the prior estimates. As a result, a winning team receives a positive update regardless of how that outcome was achieved and vice versa for a losing team. To use margin of victory in our rating updates, we incorporate an adjusted measure of goal differential into $\alpha$, denoted by AGD. We use an adjusted measure because goal differential does not always provide an accurate depiction of a team's performance. Since soccer is a low-scoring sport, outcomes can have high variance. A team may have possession for much of the game and dominate its opponent but only win 1-0. Using goal differential alone would not adequately reward this team for its performance. To adapt our model to such outcomes, we create an adjusted goal differential metric by using in-game statistics, such as shots, corners, and saves, to predict what the goal differential should have been. Since in-game statistics should predict goal differential well, the predicted goal differentials from such a model offer a reasonable adjustment to the observed goal differentials. We will discuss the specifics of this model in the variable selection and modeling section. Moreover, to further integrate in-game features more directly into our model, we also use shot differential, denoted by SD, as part of the step size. Thus, for the rating update step size, we have

$$\alpha = \gamma_{AGD}\sqrt{AGD} + \gamma_{SD}\sqrt{SD}$$

where $\gamma_{AGD}, \gamma_{SD}$ are hyperparameters (referred to as the goal differential and shot differential update parameters), and we use the square root to moderate the effect of outsized goal or shot margins.

One technical detail that must be addressed is the identifiability of our model. Using maximum likelihood estimation when the model specification is unidentifiable tarnishes any ability that we have to interpret our parameter estimates. Since the game outcome probabilities are specified in terms of the difference in ratings, it is easy to see that adding any fixed constant will yield an equivalent parametrization. To enforce identifiability, we constrain the mean of the ratings to be 0 so that an average team corresponds to a rating of 0. A convenient property of our model specification without the prior performance term is that rating updates are zero-sum, i.e., every team's gain in rating corresponds to another team's loss of rating (incorporating the prior performance term yields a negligible difference from 0). This can be easily shown using the fact that any normal distribution is symmetric about its mean. Thus, in practice, we can enforce identifiability by initializing each team's rating at 0. This initialization makes sense anyway because we have no reason at the outset to designate some teams as better or worse than others. By initializing everyone at 0 though, this implies that every team has roughly the same (not exactly the same due to the home-advantage effect) chance of winning its first game. Consequently, the model may need a substantial number of games before the ratings are well-calibrated and accurately reflect each team's strength. We need to be cognizant of this concern and analyze our results over time to understand the effects of this assumption. To mitigate this effect and since the same teams tend to be good or bad from year to year, for each season after the first, we initialize each team's rating at its final rating from the previous season.

**Commented [GH5]:** This isn't a clear connection.  Are you referencing the updates to the means two equations ago here?  In all cases, clarify this connection.

To apply our model to match outcome and score prediction is a relatively simple task. We model the probabilities of win, loss, and draw directly, so we predict the outcome of any game as the outcome which corresponds to the highest predicted probability. Moreover, since our ratings represent the expected number of goals by which a team beats an average one, we can predict the goal differential as the difference in the prior ratings.

We note that our hyperparameters cannot be tuned through a rigorous method such as cross-validation. We create the win probabilities and rating updates jointly; the win probabilities cannot be created until the ratings are updated, and the ratings cannot be updated until the win probabilities are determined. As a result, we trained our model on the entire 2020 season, using a grid of values for our hyperparameters and iterating through each possible combination, ultimately selecting the combination that corresponded to the minimum mean squared error between our predicted goal differentials and the observed goal differentials because we would like the ratings to represent the predicted score margin as accurately as possible. For the hyperparameters of FIFA's model, since their model is not applicable to score prediction, we used the Brier score, which measures the mean squared deviation of the probabilities from the outcome.

**Commented [GH6]:** Why not use the Brier score for both?

## Data

To implement our model, we have obtained team-level and player-level data for DI men's and women's college soccer teams for the 2020-2022 seasons. The player-level data pertains to individual player statistics collected throughout each game of the season, including goals, shots, and fouls, for example. The team-level data consists of the player-level data aggregated across the entire team for each game of the season. Though the data has been cleaned and relatively few values are missing, this is because many missing values were recorded as 0 instead. This is particularly true of the player-level data. For this reason, we discarded player-level statistics. Otherwise, we are fortunate to have team-level stats and results for every game for every team for the seasons listed.

## Preprocessing

We found that for 1,333 games, about 9% of the games, the number of shots was not equal to the number of shots on target plus the number of shots off target. Fortunately, this error was due to missing values for shots off target for all but 70 games and was imputed as the difference in the number of shots and the number of shots on target. For those 70 remaining games (0.5% of games), the number of shots were recorded, but neither the shots on target nor shots off target were recorded. We imputed the number of shots on target as the floor of the overall percentage of shots on target for all games multiplied by the number of shots. We chose not to employ a more in-depth imputation model given the small number of missing observations. The shots off target were inferred as before. We then used the shot data and the number of goals to infer saves and save percentages for each team and its opponent. These numbers were largely the same except for a few instances. We turn now to our model to create the margin of victory metric.

## Variable Selection and Modeling

After conducting preprocessing, we determined which features were salient for predicting goal differential. This was done through Bayesian linear regression. Using Zellner's g-prior with α as the number of games in our data, we performed a search over all possible linear models and found the marginal likelihood of each model given the data. From this, we determined the marginal inclusion probability of each variable and selected the following variables using a threshold of 40%: shots, shots on target percentage, offsides, goal kicks, saves, and save percentage, where each variable is for both the team and its opponent. We used a low threshold to include any variables that could potentially be informative in predicting goal differential.

For our model of goal differential, we opted to use a random forest since it offered strong predictive performance while also allowing us to have some measure of variable importance. We tuned its hyperparameters via ten-fold cross-validation on the data for the prior seasons. For the 2021 season, the 2020 data was used to select the hyperparameters; for the 2022 season, the 2020 and 2021 data were used. For the 2020 season, we used the true goal differentials in our ratings updates to allow for sufficient data to be gathered to adequately predict goal differential. For the 2021 and 2022 seasons, we used the data from all prior seasons to train the goal differential model for the current season. The predicted goal differentials from these models then served as our margin of victory metric. While we considered re-training the model after each week of games, we decided against this to avoid an excessive computational burden. Though we worried this might lead to underfitting, we found our training and test $R^2$ values to all be around 0.95. From this, we conclude that our margin of victory metric is very close to the observed goal differential the majority of the time but helps us account for those games where in-game statistics paint a different picture of a team's performance.

A drawback of using a machine learning algorithm within our rating updates is that we lose the ability to directly see how our updates are made. Though we may gain the ability to better reflect a team's performance, we cannot see how in-game statistics are aggregated within the random forest to create the adjusted goal differential metric. Fortunately, with a random forest, we can at least use gini impurity to create a measure of variable importance, which then allows us to see which variables are dictating the magntiude of the update. From Figure 4, it is clear that save percentage and the number of saves and shot on target percentage and the number of shots (as we would expect) are the most influential factors in the magnitude of the margin of victory metric.

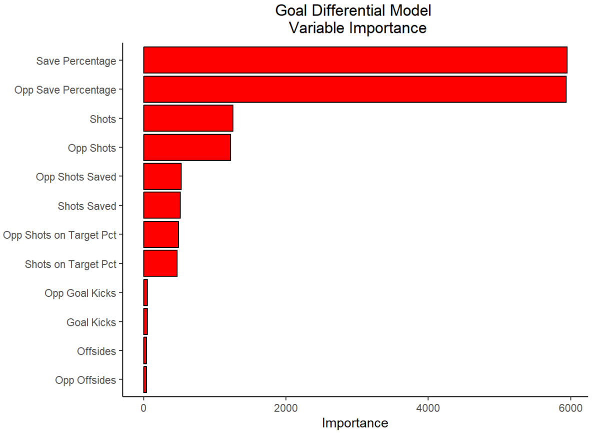


FIGURE 4. Saves, shots, and their respective percentages dictate the margin of victory.

Before we move to our results, a few nuances of the goal differential model must be considered. First, a team's and its opponent's adjusted goal differential must be equal in absolute value. The random forest cannot impose this condition, so we enforce it by using the average of the absolute value of the two adjusted goal differentials. The average of the two is only marginally different from the previous values and likely does not affect our results, but we enforce this condition for the sake of thoroughness. Second, we impose the condition that the result reflected by the adjusted goal differential is the same as that reflected by the observed goal differential. To be rigorous, we impose the condition that the signs of the adjusted and observed goal differentials are equal. This ensures that teams are always rewarded for winning and penalized for losing. If a team has differing signs for observed and adjusted goal differential, we

**Commented [GH7]:** What are the units of "Importance" in the figure? If there are none, is it worth rescaling so that the max is 1?

set the adjusted goal differential to be the sign of the observed goal differential multiplied by 0.5. For example, suppose a team wins a game by one goal but has an adjusted goal differential of -0.5 goals. Given the differing signs, we can say that this team won a game that it actually should have lost. However, we believe that we should actually reward this team for its victory. This is because strong teams find ways to win games that they should lose whereas weak teams blow winnable games. At the same time, we need to penalize this team for playing poorly, which is achieved by the 0.5 factor. This factor decreases the margin of victory so that the team receives a smaller, though positive, update. In our example, the team would have a margin of victory of 0.5, which we believe finds a balance between reward for a win and penalty for poor performance.

A consequence of imposing this sign condition is that the margin of victory for a draw is always 0, which corresponds to an update of 0. This implies that a draw contains no information about a team's rating, which is not always the case. When a team is much better than their opponent, then a draw is effectively a loss. In fact, the gradient for a draw asymptotically approaches the gradient for a loss as the ratings difference between a team and its opponent becomes large as we can see in Figure 5.

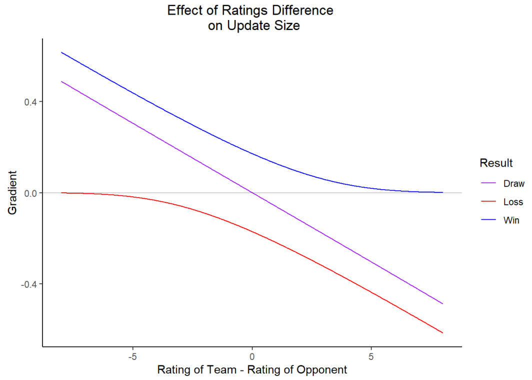


FIGURE 5. A draw asymptotically approaches a win as its ratings difference with its opponent decreases and asymptotically approaches a loss as its rating difference increases.

To account for this, we set the margin of victory for a draw to be 0.5. The update will still be small when the ratings difference is small, as before (from Figure 5 the gradient will be very close to 0), but will move away from 0 when the ratings difference becomes large.

**Commented [GH8]:** It would be helpful to have the corresponding figure from your presentation here to help explain

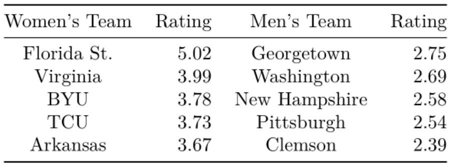
## Results

After tuning our model on the 2020 season for both the men's and women's models, we found our hyperparameters to be 0.75 for the goal differential update parameter, 0.5 for the volatility update parameter, 0.25 for the shot differential update parameter, and -0.05 for the prior performance parameter for both models. That the prior performance term is negative indicates that the regression to the mean effect outweighs the momentum effect for college soccer generally. Interestingly, the initial volatility parameter was 2 for the men's model and 1.75 for the women's model. This difference likely arises from the smaller talent disparity in men's soccer which creates more similarly rated teams and higher uncertainty in each team's rating. We also found the home advantage parameter to be 0.4 for the women's model but 0.2 for the men's model. Women's college soccer is more popular and tends to have higher attendance which creates a larger home-advantage effect, on average.

Our first and foremost goal is to have a model that is interpretable; we believe that this advantage is what most sets our model apart from FIFA's model. Though our model's interpretability has a strong foundation in probability, we need to ensure our results conform with our expectations, i.e., we need to make sure each team's rating provides a reasonable value for the average number of goals by which it would defeat an average college team. Below in Tables 1 and 2, we have the five highest-rated and five lowest-rated teams from the models for the women's 2021 season and the men's 2021 season.

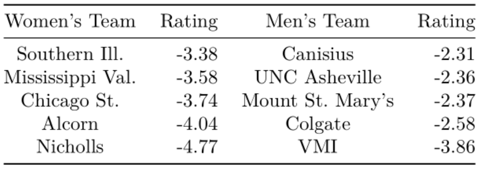
**Commented [GH9]:** Can you caption the figure and reference it?  It is unclear which is men's and which is woman's and the figure has become interwoven into the text (which is fine if there would be a caption)

Table 1: Top 5 Teams of 2021



| Women's Team | Rating | Men's Team | Rating |
|---|---|---|---|
| Florida St. | 5.02 | Georgetown | 2.75 |
| Virginia | 3.99 | Washington | 2.69 |
| BYU | 3.78 | New Hampshire | 2.58 |
| TCU | 3.73 | Pittsburgh | 2.54 |
| Arkansas | 3.67 | Clemson | 2.39 |

Table 2: Bottom 5 Teams of 2021



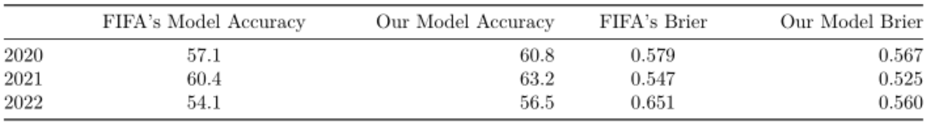
| Women's Team | Rating | Men's Team | Rating |
|---|---|---|---|
| Southern Ill. | -3.38 | Canisius | -2.31 |
| Mississippi Val. | -3.58 | UNC Asheville | -2.36 |
| Chicago St. | -3.74 | Mount St. Mary's | -2.37 |
| Alcorn | -4.04 | Colgate | -2.58 |
| Nicholls | -4.77 | VMI | -3.86 |

We can see that the ratings generally lie in the interval [-4,4] for the women's and [-3,3] for the men's with a few exceptions. With 350 women's DI college soccer teams, this implies that the best women's teams would beat the 175th-best team by about four goals and the worst teams by about eight goals. Similarly, with 200 men's teams, the best men's teams would beat the 100th-best team by about three goals and the worst teams by about six goals. This matches our intuition and provides a very easy comparison of teams, both to average and each other. A rating of 5.32 might seem very high for Florida St., but we believe it is very reasonable given that Florida St. regularly beat very strong teams by multiple goals on its way to the 2021 national championship.

Having demonstrated that our ratings are interpretable, we can move to examining our model's predictive performance. The most critical component of our model is ensuring that the outcome probabilities are well-calibrated. For our model to make accurate updates, the model probabilities must accurately represent the true probabilities of the three outcomes. One way to measure whether these probabilities are well-calibrated is by analyzing what they imply as far as predicted outcomes. In Tables 3 and 4, we have the accuracy for our model and the baseline FIFA model for 3-way match outcomes.

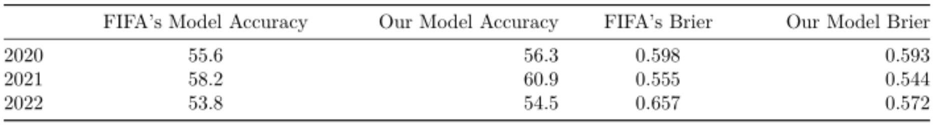
**Commented [GH10]:** Given that FIFA only makes a 2-way prediction, is it possible that the added benefit of your model is solely coming from the addition of draws? Or would you still outperform fifa if you did not consider draws? You should be clear about were your model is deriving its benefit (i.e. the benefit comes from the general approach or solely from the inclusion of draws)

Table 3: Women's Model Results



| | FIFA's Model Accuracy | Our Model Accuracy | FIFA's Brier | Our Model Brier |
|---|---|---|---|---|
| 2020 | 57.1 | 60.8 | 0.579 | 0.567 |
| 2021 | 60.4 | 63.2 | 0.547 | 0.525 |
| 2022 | 54.1 | 56.5 | 0.651 | 0.560 |

Table 4: Men's Model Results



| | FIFA's Model Accuracy | Our Model Accuracy | FIFA's Brier | Our Model Brier |
|---|---|---|---|---|
| 2020 | 55.6 | 56.3 | 0.598 | 0.593 |
| 2021 | 58.2 | 60.9 | 0.555 | 0.544 |
| 2022 | 53.8 | 54.5 | 0.657 | 0.572 |

Our model meets our performance benchmarks and outperforms the FIFA model. Moreover, the model's performance improves from the 2020 season to the 2021 season, providing evidence that the model can better learn each team's strength as it accumulates more games. There appears to be a dip in performance from 2021 to 2022, but this is due to an NCAA rule change that removed the overtime period from regular season games. This increased the proportion of draws from 11% to 22%, and draws are extremely difficult to predict. Nonetheless, providing such high accuracies with a relatively simple, interpretable model is almost astonishing, especially given that the model's main purpose is not outcome prediction. For comparison, Ulmer et al. achieve an accuracy of around 50% for 3-way outcomes using various machine learning models on English Premier League data (though the proportion of draws in the Premier League is 29%). Another method to check whether our probabilities are well-calibrated is using the Brier score. The Brier score is given by

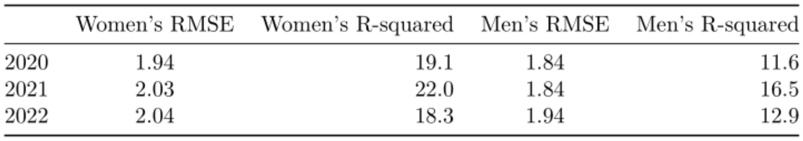
$$BS = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}(f_{ti} - o_{it})^2$$

where N is the number of instances, R is the number of classes for each instance, $f_{ti}$ is the predicted probability for class i for instance t, and $o_{it}$ is an indicator for if instance t belongs to class i. This metric measures the mean squared difference between the outcome of each instance and its predicted probabilities. Consequently, a smaller value indicates better calibrated probabilities. We can see that our model has better calibrated probabilities when compared to the FIFA model. With these results, we can be confident that our model can be applied to match outcome prediction, and our win probabilities are well-calibrated, so our model is accurately updating ratings.

**Commented [GH11]:** I know your model never has the max likelihood on a draw which will help defend it’s robustness in the accuracy figures, but from the point above, is the increased brier score solely due to the inclusion of draws in your model?  Or would it be better than FIFA even without this?

Now that we have examined our match outcome prediction results, we can move to our score prediction results. As stated previously, we use the difference in a team’s and its opponent’s prior ratings as the predicted goal differential for that game. We have no similar model with which to compare results because FIFA’s model is not applicable toward match score prediction. Moreover, it would not make sense to compare performance with a model specifically trained for this task because our model is not. Without a comparable model, we will have to evaluate our results in absolute terms. In Table 5, we can see our performance as measured by the root mean squared error and the $R^2$ value.

Table 5: Score Prediction Results



| | Women’s RMSE | Women’s R-squared | Men’s RMSE | Men’s R-squared |
|---|---|---|---|---|
| 2020 | 1.94 | 19.1 | 1.84 | 11.6 |
| 2021 | 2.03 | 22.0 | 1.84 | 16.5 |
| 2022 | 2.04 | 18.3 | 1.94 | 12.9 |

Given the low scoring nature of soccer games, it is clear that our results may leave something to be desired, but this was not unexpected. Match score prediction is a difficult task, especially given that we use solely a team’s and its opponent’s overall strengths. Though we do not capture the exact goal differential particularly well, we do capture the general trend, i.e., as the predicted goal differential increases, the observed goal differential increases, on average. Given that we capture this relationship, we believe that our model provides utility in predicting match scores.

**Commented [GH12]:** This is a good defense for tracking win probabilities rather than goal differential in your original model as you are justifying it above.

A natural question to ask is what aspect of our model is driving the increase in performance. To answer this, we tested numerous variations of our model in which we subtracted the various improvements we made. We found that the majority of the performance improvement is a result of the home-advantage effect while the adjusted margin of victory metric and the prior performance term both result in slight improvements. If we model solely two outcomes (win and loss) as the FIFA model does, we find that our accuracy and RMSE actually improve. This is a result of the fact that modeling draws is difficult and introduces more uncertainty in prediction. However, we still prefer to use the model that handles draws directly because its Brier score is significantly lower than the model that does not. This is particularly true of the 2022 season in which the proportion of draws increased greatly and will be the case moving forward.

We would be remiss to not mention the tournament projections implied by our model. With the impact of the COVID-19 pandemic, half of the 2020 season was held in the fall of 2020, and the other half, including postseason play, was held in the spring of 2021. Given the shortened schedules and odd nature of the season, we did not make projections for the 2020 NCAA tournament. Likewise, we only have about 60% of the games for the 2022 season, so we did not make projections for this season either. For the 2021 women's NCAA tournament, we accurately predicted 22 out of 31, or 71%, conference winners. Given the randomness of conference tournaments, since they are only a few games in length, this is fairly strong performance. Of the nine teams that we did not accurately predict as winning their conference tournaments, all were ranked second in their respective conferences except for Santa Clara who was ranked third. With respect to the 2021 men's NCAA tournament, we predicted 14 out of 23, or 61% of conference winners. Of the nine incorrect predictions, seven of the nine true winners were ranked second in their conference. The other two, Mercer and Notre Dame, were ranked third and sixth, respectively. The fact that Notre Dame was ranked sixth and won their conference is not a cause for concern given the Atlantic Coast Conference's depth and strong reputation for soccer. Overall, these results indicate that we generally capture the correct within-conference rankings.

For the at-large bids for the women's tournament field, we mistakenly gave at-large bids to Gonzaga, Oklahoma St., Utah Valley, and West Virginia. Of the 33 teams to receive at-large bids, they were ranked 21st, 27th, 29th, and 22nd, respectively. We did not give at-large bids to Alabama, LSU, NC State, Ohio St., SMU, and Wisconsin who were ranked 122nd, 59th, 80th, 78th, 65th, and 70th among all teams, respectively. For reference, the last at-large bid was given to St. John's who was ranked 57th overall. We note that the difference in the number of teams comes from having conference winners in our projected field who did not actually make the tournament and then giving at-large bids to the actual conference winners. For the men's at-large bids, we incorrectly gave at-large bids to Belmont, Cornell, James Madison, and Stanford. Of the 25 teams to receive at-large bids, they were ranked 12th, 23rd, 16th, and 15th, respectively. We did not give at-large bids to Charlotte, Creighton, Louisville, Portland, UCLA, and Villanova who were ranked 48th, 63rd, 71st, 74th, 81st, and 99th, respectively, among all teams. For reference, the last at-large bid was given to St. John's who was ranked 46th overall. With these results, we see that the teams that we incorrectly did not include were reasonably close to making the projected field, and the teams we incorrectly included were generally on the cusp of not making the projected field. Overall, we projected 56 out of 64 women's teams correctly, or 87.5%, and 40 out of 48 men's teams correctly, or 83.3%. Our projected fields have a strong correspondence to the true ones, which assures us that our model can make tournament projections, and our rankings resemble the true ones.

# Discussion

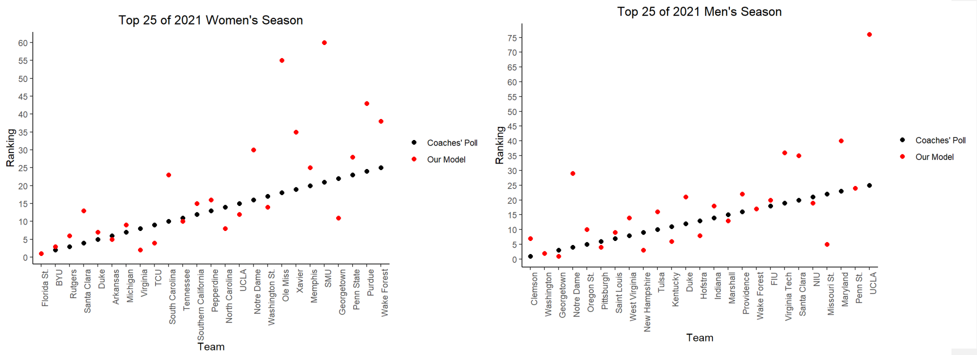


FIGURE 6. Left: The top 25 rankings from the United Soccer Coaches poll and our model from the final week of the 2021 women's season. Right: The top 25 rankings from the United Soccer Coaches poll and our model from the final week of the 2021 men's season.

After showing our model has the necessary applications, we can now examine the rankings themselves. To determine whether our rankings are reasonable, we compare the top 25 ranked teams in the United Soccer Coaches poll at the end of the 2021 season with the rankings from our men's and women's model in Figure 6. Due to the subjective nature of the coaches' poll, we do not want to match its rankings exactly, but we do want our rankings to lie reasonably close since these rankings incorporate expert knowledge. We can see that our rankings are positively correlated with the coaches' poll; most teams lie close to the coaches' poll with a few teams ranked significantly higher or lower. In particular, the correlation between our rankings and the coaches' poll is 0.63 and 0.73 for the men's and women's model, respectively. Following the approach of Colley, we can measure the difference in rankings using the mean percentage difference, which is given by $\eta = e^{\frac{1}{25}\sum_{i=1}^{25}|log(i_M)-log(i_C)|}$. Colley demonstrates that this measure of the difference in rankings is much better behaved than the mean absolute difference. For the men's and women's model, η is 1.85 and 1.67, respectively, so our rankings differ by 85% on average for the men's model and 67% on average for the women's model. Thus, around rank 10, we differ by about 8 spots for the men's model and 6 spots for the women's model. In quantifying the difference, we can see that despite the positive correlation, our rankings seem to differ significantly between our model and the coaches' poll. However, the seemingly large difference is a result of only a few large deviations. η is 1.48 and 1.44 for the men's and women's models if we remove the four largest percent deviations, which seems to be a fairly moderate difference. These large deviations are mostly a function of the NCAA tournament results. For example, Notre Dame's men's team had a tremendous run to the final four of the tournament, which led to it being ranked 4th in the coaches' poll. However, our model uses a holistic view of the entire season, not just the past few games, and therefore, Notre Dame was ranked 29th given its poor performance earlier in the season. Even if our rankings completely disagreed, it is unclear whether disagreement with the coaches' poll is necessarily bad. The coaches' poll is biased as coaches tend to overrate the teams within their own conference to make their own team look better. In general, coaches tend to overrate teams in typically stronger conferences and underrate teams in typically weaker conferences. Given the moderate percent deviation and fairly strong positive correlation though, we believe these results provide evidence that our model provides reasonable rankings that conform with common sense.

Two properties our model must possess are sensitivity to unexpected outcomes and stability. If a team underperforms for the first portion of the season but then rights the ship and starts to perform much better, we want our model to recognize that change in performance and make large, positive updates to that team's rating to reflect that it is much better now than their previous ratings suggested. At the same time however, we do not want updates to always be large. If this were the case, then our rankings would change drastically from week to week as the ratings move erratically. This is undesirable because we will lose the information from games much earlier in the season, and the rankings will be dependent upon the nuances of the schedule.

To determine our model's ability to balance these two properties, we have found the size of each team's change in ranking and rating over the 2021 season. These results are plotted below in Figure 7.

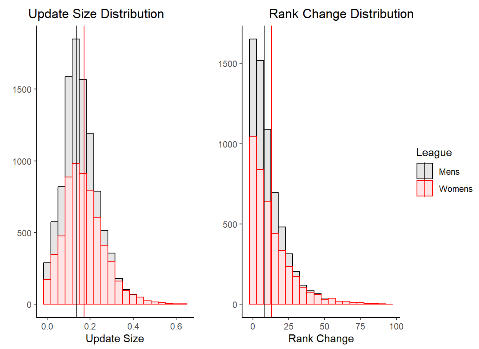


FIGURE 7. Left: Distribution of the absolute value of the change in ratings after each game for all teams. Right: Distribution of the absolute value of the weekly change in rankings for all teams.

For the change in ratings, we see that the vast majority are concentrated between 0 and 0.2. Using our ratings interpretation, this implies that each team's rating generally changes by between 0 and a fifth of a goal, which seems fairly small. Similarly, the majority of rank changes are between 0 and 10 for the men's model and 0 and 15 for the women's model. Given that there are 200 men's teams and 350 women's teams, this range appears modest. In sum, our model captures the stability property. At the same time, we can also see long tails in the distributions. These tails encapsulate those unexpected outcomes that led to significant rating changes of upwards of half of a goal and ranking changes of 50-75 spots. The tail appears relatively light though, providing evidence for the stability of our model. Overall, the average change in ranking from week to week is 9.1 for the men's model and 13.4 for the women's model (very similar to the FIFA model results), which we believe represents a good balance between sensitivity and stability.

**Commented [GH13]:** This could also be a good place to reference a study where you reset the rankings and see how long they take to agree again if you didn't reset them (supposing this study makes sense… it may not)

One limitation of our model is the starting point of the ratings for each season. We cannot track player turnover between seasons, so we do not know how to adjust the final rating from the prior season to reflect offseason changes. Initializing at last year's final rating is not ideal because we do not want a team's rating to be beholden to their performance from the prior season. If updates are not aggressive enough, then a team's rating will be dependent upon its performance from the prior year. To check how quickly our ratings adjust to the new season, we have plotted our predictive performance over time for the 2021 and 2022 seasons below.

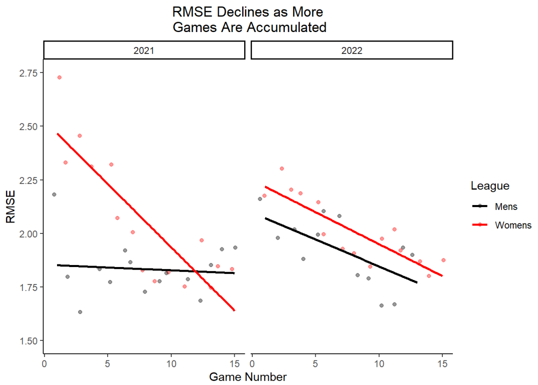


FIGURE 8. Evolution of the goal differential error, measured as the root mean squared error of the difference in prior ratings from the observed goal differential, as the season progresses.

After 4-6 games, there is a clear dip in the goal differential error, which suggests that the model needs several games to learn each team's new strength. For this reason, we should avoid interpreting ratings until five or six games have occurred.

To test this carryover assumption more rigorously, we perturbed the final ratings for the 2020 season using mean zero Gaussian noise with a standard deviation of 0.25. Using ten different random initializations, we found the mean percentage difference, η, between the final rankings from the random initialization and the final rankings from our model for the 2021 season. Averaging over all ten random initializations, we found η = 1.04 for the women's model, and η = 1.08 for the men's model. Thus, the overall difference in rankings is very small, so the rankings generally are not very sensitive to their initial values over the course of an entire season. In addition, the average deviation of the ratings with random initialization from our model's ratings decreases from 0.2 at the beginning of the season to 0.10 and 0.07 by the end of the season for the men's and women's model, respectively. That the difference shrinks indicates that all ratings are converging to the same value, the true rating for each team. Ideally, the difference would shrink to 0, but in practice, this will not occur with only around 20 games per team. Nevertheless, the convergence of the difference seems slightly slow. Thus, if a team changes drastically over the offseason, then their rating for the next season may not be a great depiction of their actual strength.

**Commented [GH14]:** This is great! How many games did it take to see the reduction?

## Conclusion

In conclusion, we have created a ranking system for college soccer that provides reasonable rankings, predicts match outcomes and scores, and makes tournament projections. The model is interpretable, efficient, and exhibits strong performance. We incorporate draws directly and use in-game features to yield a holistic view of team performance. Unlike other ranking systems, the model is probabilistic, enabling us to quantify uncertainty, and our modeling principles are generalizable to different sports, permitting adaptation of the model outside of college soccer. Though we believe strongly in our model, there are several avenues for improvement. Integrating off-season knowledge to better initialize the model each season would improve the model significantly. Furthermore, teams have varying home-field advantages, so estimating a different home-field advantage effect for each team could better reflect team performance. It would also allow us to determine the toughest places to play across the country and provide interesting narratives for which teams have the most raucous fans. Lastly, incorporating player-level data is a worthy pursuit. Adjusting a team’s rating based on a star player, like FiveThirtyEight does with quarterbacks in their NFL model, may provide intriguing results since star players have an outsized effect on team performance in soccer.

## References

2018. Revision of the FIFA/Coca-Cola World Rankings. FIFA. [accessed 2022 Oct 10]. https://digitalhub.fifa.com/m/f99da4f73212220/original/edbm045h0udbwkqew35a-pdf.pdf.

Colley, WN. 2002. Colley's bias free college football ranking method: The Colley matrix explained. Colley Rankings. [accessed 2022 Oct 5]. https://www.colleyrankings.com/matrate.pdf.

Ley C, Van de Wiele T, Van Eetvelde. 2019. Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. Statistical Modelling. 19(1): 55-73.

Li Y, Ma R, Goncalves B, Gong B, Cui Y, Shen Y. 2020. Data-driven team ranking and match performance analysis in Chinese Football Super League. Chaos, Solitons, and Fractals. 141(1).

Massey, K. 2001. Massey ratings description. Massey Ratings. [accessed 2022 Oct 5]. https://masseyratings.com/theory/massey.htm.

Pomeroy, K. 2016. Ratings methodology update. Advanced Analysis of College Basketball. [accessed 2022 Oct 3]. https://kenpom.com/blog/ratings-methodology-update.

Ulmer, B., Fernandez, M., and Peterson, M. Predicting soccer match results in the English Premier League. Doctoral dissertation, Ph. D. dissertation, Stanford (2013).

Wolfe, P. 1992. About these ratings. College Football Ratings. [accessed 2022 Oct 3]. http://prwolfe.bol.ucla.edu/cfootball/descrip.htm.

## Technical Documentation

Table of Contents

### Introduction

The Arria Boost project is designed to rank college soccer teams, focusing on NCAA Division 1 men's and women's teams. This project has been developed in three programming languages: Python, R, and JavaScript. It comprises two primary components: the machine learning prediction, which constructs a model to rank the teams, and the visualization dashboard, which displays narratives of the results generated by the model.

### Repository Structure

**src**: Contains the source code for the core Arria Boost machine learning component. The python files only have the data preparation stage.

**boost-sports-soccer-dashboard**: Contains the source code for the core visualization dashboard. The dashboard has a backend and frontend.

**docs**: Houses the project documentation, including this technical guide.

**.gitignore:** This file specifies which files or directories should be excluded from version control when using Git.

**.gitmodules:** This file specifies the git modules information for the project

**README.md:** This file provides any technical documentation about the project.

**requirements.txt:** specifies python library dependencies for the project.

### System Requirements

1. R Studio
2. Python 3.8 and above
3. Docker

### Installation Guide

1. Clone the Gitlab repository. You can use SSH or HTTPs for this.
    a. *git clone https://gitlab.oit.duke.edu/duke-mids/workingprojectrepositories/arria-boost.git*
    b. *Go into* **boost-sports-soccer-dashboard folder** e.g., *cd boost-sports-soccer-dashboard* and clone the dashboard repository e.g., *git clone https://github.com/missvicki/boost-sports-soccer-dashboard*

2. Machine Learning Prediction component
    a. Open R studio on your machine

b. Open the project in the destination you cloned your repository and execute all R files in the src folder not entitled "Run Everything".

c. Execute the "Run Everything" script to run all models.

Note: You will have to install library dependencies on this project.

3. Visualization dashboard

1. Begin by creating *.env* files within the backend and frontend folders.
   - Add the following lines to the frontend/.env file.
     - REACT_APP_MAIN_VERSION=api
     - REACT_APP_API_URL=http://localhost:8000
   - Add the following lines to the backend/.env file.
     - FLASK_APP=app.py
     - AWS_ACCESS_KEY_ID=hadkjdkajueooajd
     - AWS_SECRET_ACCESS_KEY=ajjdkajd8039288492
     - AWS_S3_BUCKET=boost-sports-duke
     - ENVIRONMENT=dev

2. Run the application.

a. Option with docker compose.

i. Go into the *boost-sports-soccer-dashboard* folder.
ii. Start running docker on your machine.
iii. Run *docker compose up –build -d* to run the application.
iv. Run *docker compose logs -f* to see the application logs.
v. Run *docker compose stop –volumes* to stop running the application.
vi. Run *docker system prune* to clean images and volumes.

b. Option 2 run the backend and frontend separately.

- Backend
  - Go into the *backend* folder.
  - Create a python virtual environment.
  - Activate the python virtual environment.
  - Install python dependencies with *pip install -r requirements.txt.*
  - Run *flask run –reload –host=localhost –port=8000.*
- Frontend
  - Go into the *frontend* folder.
  - Run *npm install.*
  - Run *npm start.*

**Contact and Support**

If you need additional assistance or have questions that are not covered in this documentation, feel free to reach out to the Arria Boost development team.

We hope this technical documentation serves as a valuable resource in helping you understand and utilize the Arria Boost project to its fullest potential. Happy coding!