**RANKING NCAA COLLEGE SOCCER DIVISION I TEAMS**

*Vicki Nomwesigwa, Rashaad Ratliff-Brown, Matt Dockman*

**Introduction**

Soccer is undoubtedly the world’s most popular sport. It attracts millions of spectators and players across all ages and demographics. Despite its popularity, the application of data-driven methods to soccer has been slow to occur. Though models to predict match outcomes have been created to serve the ravenous betting market, relatively little consideration has been given to ranking soccer teams. This stands in stark contrast to football and basketball in which systems such as Massey, KenPom, and Colley matrix have been created to rank college football and basketball teams. For over a decade, some of these ranking systems were used as a component to determine the participants in the college football national championship. On the other hand, no such system has been created for college soccer, which has suffered from a lack of attention.

Team rankings are important in order to generate compelling narratives for team performance. Rankings provide an ordinal structure that readily allows us to compare teams; when a low-ranked team defeats a high-ranked team, we can confidently declare that an upset occurred. When two-high ranked teams are scheduled to play, we can state that two of the best teams are about to square off. Without such rankings, it is unclear how we could create 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. Moreover, rankings can provide means by which to predict game and score outcomes and inform playoff seeding. These reasons combined with the neglecting of college soccer motivate the creation of a ranking system designed specifically for college soccer. To begin, we must first consider the existing ranking systems.

**Literature Review**

For our purpose, we will discuss football and basketball ranking methodologies in addition to soccer. Needless to say, football and basketball are completely 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 application to 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 simplicity and 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 adapt to this fact, 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, offering an improvement over their aforementioned models.

For 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. 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. Perhaps 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.

**Methods**

**Model Specification**

Our goal is to create a ranking system for men’s and women’s NCAA Division I soccer teams to generate compelling narratives, predict match outcomes, and make 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 64-team field for the 2021 NCAA tournament.

The difficulty in creating sports rankings is that we do not observe the true rankings. Instead, we must use a team’s performance over the course of a season to try to discover the true rankings. Consequently, there is no clear benchmark with which to compare our results to evaluate performance. The United Soccer Coaches offer weekly college soccer rankings for men and women but are subject to the biases of the coaches involved. Rather than use subjective rankings, the current FIFA model will serve as our baseline model. This model has shown promising predictive performance while retaining interpretability. Moreover, the problem FIFA faces in ranking national teams is analogous to that of 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 of each team’s games. For clarity, the FIFA model ranks teams by assigning each team a rating which is then updated after each game according to

where Result is 1 for a win, 0.5 for a draw, and 0 for a loss. and are the ratings of teams A and B, respectively, is the a priori probability of victory, and I is the match importance factor. The initial ratings were created through a conversion of the ratings from the previous ranking model, but we will not discuss these details since they are not relevant here. To adapt the match importance factor to the nuances of college soccer, we will have a hierarchy of conference tournament games, in-conference games, and out-of-conference games, in order of decreasing importance. The rationale is that in-conference games determine the regular season conference champion, so teams tend to play harder in those games than out-of-conference games. Moreover, a team’s rivals are almost always in their own conference. Conference tournament games are given even higher importance because they are particularly crucial for the NCAA tournament. The champion of each conference tournament receives an automatic bid to the NCAA tournament, and NCAA tournament seeding is very sensitive to conference tournament results. Thus, all teams have an incentive to play to their highest potential in conference tournament games.

Ranking college soccer teams necessarily entails adjusting a team’s performance by the strength of its opponent. FIFA’s model accomplishes this by using a team’s and its opposition’s ratings to create a probability of victory for each team prior to a match. In this way, a team is rewarded less for winning games that they were expected to win and more for games they were not expected to win. Furthermore, a team’s rating is updated according to a comparison between how a team was expected to perform and how it actually performed. By updating ratings in this way, 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’s team rating is always increasing as the season progresses, making comparison less intuitive. Moreover, accumulation-based systems do not easily allow for a differing number of games for each team, which will be the case for DI college soccer due to weather and postseason play.

To improve on FIFA’s model, we will make four adjustments to their model. First, their model does not consider goal differential. We would expect goal differential to be a significant indicator of team strength. Teams that win with greater margins of victory tend to be better than teams that scratch out wins. Second, their model is not very interpretable. The current ratings 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. Third, the function for computing the a priori win probability is not interpretable. Their model uses base-10 logistic regression on the prior ratings of both teams, which is not accessible to the average person. Lastly, 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.

To address these shortcomings, we will initialize each team’s rating at 0 and update the rating by the goal differential of the match in addition to the difference in actual versus expected outcome. To prevent lopsided scores from creating too large of updates, we will use the square root of goal differential divided by 2. Moreover, the win probabilities in our model will be calculated using the cumulative distribution function (CDF) for a standard normal random variable. The average person is much more familiar with this distribution and what it implies in terms of probabilities. Specifically, we will make updates according to

where is the CDF of a N(0,1) random variable, HE is the home-advantage effect, and is a scaling hyperparameter. A nice property that follows from this model specification is that ratings updates are symmetric, i.e., every team’s gain in rating corresponds to another team’s loss of rating. This property implies that an average rating of 0 is preserved, so an average team always has a rating of 0. By initializing at 0 and updating each team’s rating according to its goal differential then, we can interpret each team’s rating as the number of goals by which it would win (if positive) or lose (if negative) if it were to play an average DI college soccer team. Ratings updates are simple, computationally efficient, and done after every game, allowing rankings to be determined at any point in the season. Notice that this model does not include a factor for match importance. To keep our model simple and because all games have a significant effect toward a team’s inclusion in the NCAA tournament, we have discarded this factor.

Though the model above uses goal differential, this measure does not always provide an accurate depiction of the outcome of a match. Since soccer is a low-scoring sport, outcomes can have high variance. A team may have possession for much of the game and outshoot its opponent but lose due to an otherworldly performance from the opposing goalkeeper. Goal differential would clearly punish this team’s loss too harshly. To better reflect this team’s performance, we could adjust its goal differential. This adjustment can be made by using in-game statistics, such as shots, corners, and saves, to predict what this team’s 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. To confront this concern of goal differential being a bad measure of success, we will test three additional models, each with a different method of adjusting goal differential.

One other concern we may have with our model is the initialization of each team’s rating at the same rating. This implies for our model that every team has close to a 50% (not exactly 50% due to the home-advantage effect) chance of winning its first game. The model may need a substantial number of games before the win probabilities are well calibrated and accurately reflect each team’s strength. We need to be cognizant of this concern and analyze our results over time so that we can understand the effects of this assumption. To mitigate its effect, for each season after the first, we initialize each team’s rating at its final rating from the previous season.

We should mention that our hyperparameters, and HE, cannot be tuned through a rigorous method such as cross-validation. The win probabilities and ratings updates are created 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 used a grid of values for these hyperparameters and iterated through each possible combination, ultimately selecting the combination that corresponded to the best predictive performance. The same applies to the match importance factors for the baseline FIFA model.

**Data**

To create our ranking system, we have obtained team-level and player-level data for DI men’s and women’s college soccer teams for the 2020-2022 seasons. Though the data has been cleaned and relatively few values are missing, simple exploration of the data yielded concerns that measures were given zeroes if the data was not recorded. This is particularly true of the player-level data. For this reason, player-level stats were discarded. Otherwise, we are fortunate to have team-level stats and results for every game for every DI team for the seasons listed. For this paper, we will focus only on the women’s data.

**Preprocessing**

To create our models to adjust goal differential, we first had to preprocess our data. In particular, we fixed one game in which the goal differential was calculated incorrectly and removed another game in which the only information available was the number of goals scored by each team. For one game in which the number of shots was below the number of goals scored by that team, the shots were imputed as the number of goals plus the average difference in shots and goals for all games. 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, all but 70 games were a case of the shots off target not being recorded, so this statistic could be inferred as the difference in shots and shots on target. For the other 70 games, which is about 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 along with 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.

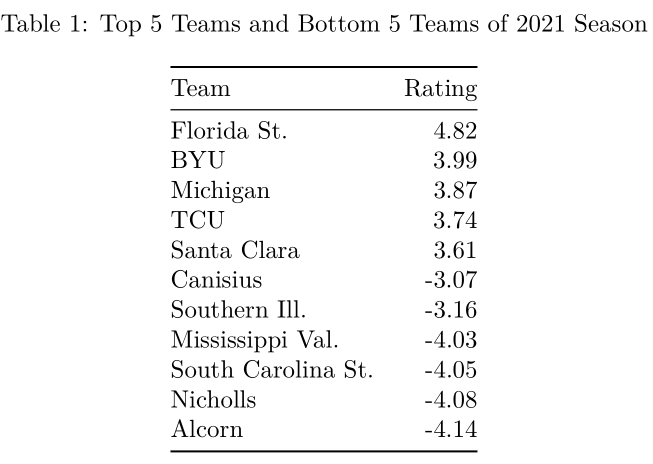
**Variable Selection and Modeling**

After conducting preprocessing, we moved to determining 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 70%: shots, shots on target percentage, corners, offsides, goal kicks, saves, and save percentage, where each variable is for both the team and its opponent.

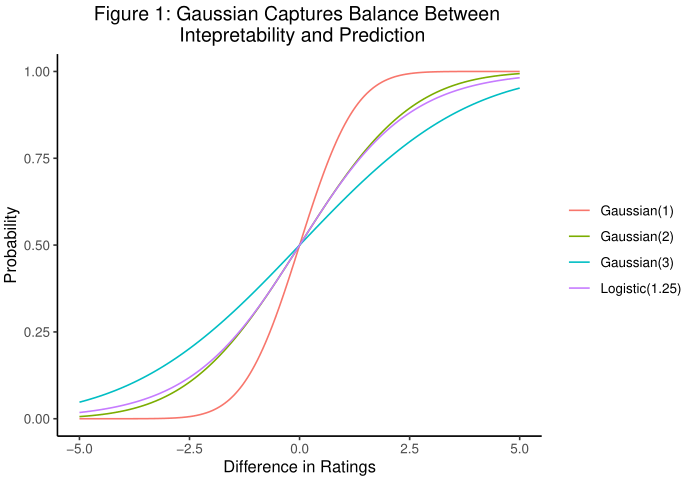
For our models of goal differential, we considered random forests, gradient-boosted trees, and multi-layer perceptrons. For each model, its hyperparameters were tuned via ten-fold cross-validation on the data for the prior seasons. For the 2021 season, the 2020 data was used to select the hyperparameters; for the 2022 season, the 2020 and 2021 data were used. Once our hyperparameters were selected, we used the data from the prior seasons as the training data for each season. The predicted goal differentials from these models were then used to update the ratings as previously described. To allow for sufficient data to be gathered to adequately predict goal differential, we used the true goal differentials for the 2020 season. 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 values to all be above 0.9.

To convert the win probabilities for each game into match predictions for three-way outcomes, any win probability below 48% corresponded to a loss, any win probability between 48% and 52% corresponded to a draw, and any win probability above 52% corresponded to a win. Two-way outcomes were determined in the usual way with a threshold of 50%. With respect to tournament projections, in reality, the tournament field is created from the winners of each conference tournament with the rest of the spots filled by at-large bids. To replicate this for tournament projections, the highest-rated team in each conference received a spot in our projected tournament field. Each successive spot was then filled by the next highest rated team not already in the projected field. We will not consider score projections in this paper due to time constraints, but it will be addressed in future work.

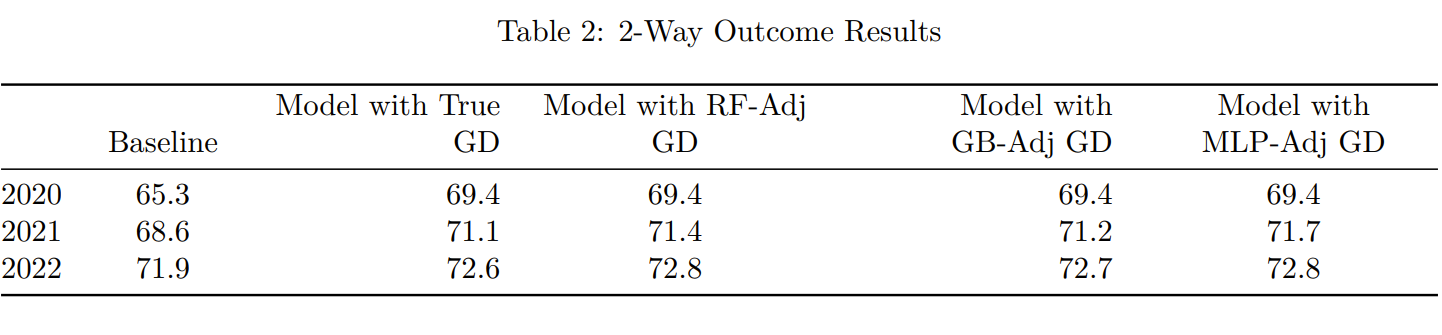
**Results**

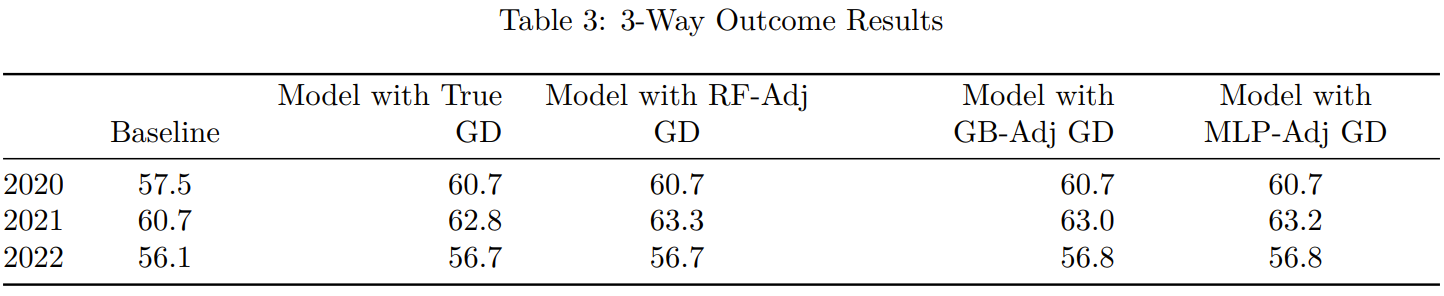


To examine our results, we will address the issue of interpretability first. In Table 1, we have the five highest rated and five lowest rated teams from the 2021 season for the model with the true goal differentials. We can see that the ratings generally lie in the interval [-4,4]. With 350 women’s DI college soccer teams, this implies that the best teams in the country would beat the 175th best team in the country by about four goals and the worst teams in the country by about eight goals. This matches our intuition and provides a very easy comparison of teams, both to average and each other.



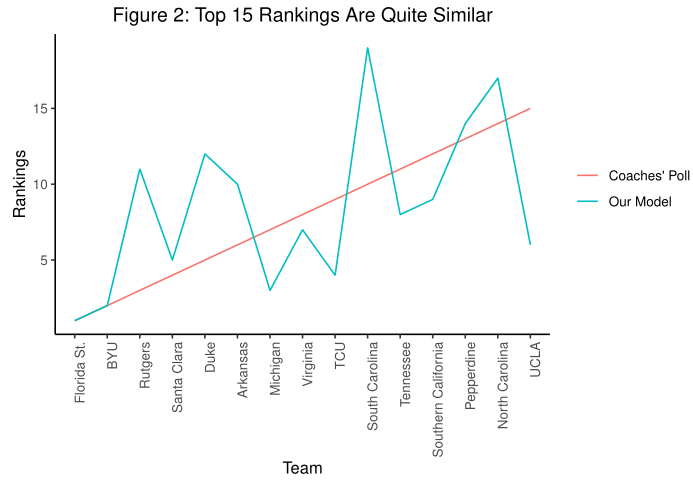
In choosing the scaling hyperparameter of the win probability function, we noticed a trade-off as this hyperparameter changed. When was too large, i.e., the teal curve in Figure 1, the model gave too much credit to teams for winning games. The win probabilities were too modest, resulting in the ratings becoming too large in absolute value (upwards of nine in some cases) and rendering their interpretation useless. Conversely, with too small, i.e., the red curve in Figure 1, the ratings were constrained to too narrow of a neighborhood around 0, which hurt predictive performance by a few percentage points and did not allow for as easy comparison of teams. In the end, was chosen to be 2, which corresponds to the yellow curve in Figure 1. We can see that for an aptly chosen value of the logistic function, the normal CDF and logistic function provide almost identical probabilities. This is just to say that the normal CDF is not as much an improvement in prediction as it is in interpretability. With respect to the home-advantage effect, the model was found to be perform best when this hyperparameter was set at 0.35. This implies that teams receive an extra third of a goal in goal differential by playing at home. For the baseline FIFA model, the match importance factors were chosen to be 0.25, 0.5, and 0.75 for out-of-conference games, in-conference games, and conference tournament games, respectively.





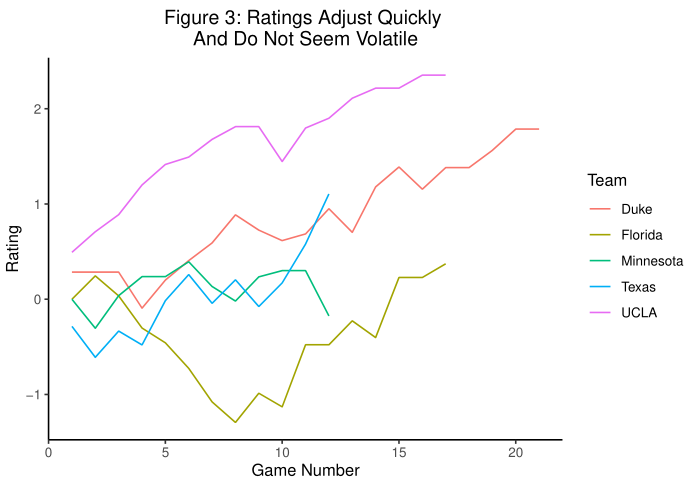
Having demonstrated that our ratings are interpretable and discussed hyperparameter selection, we can move to applying our rankings towards match outcome predictions. The most critical component of our model is ensuring that the win probabilities are well-calibrated. For our model to make the correct updates, the win probabilities need to provide a true representation of each team’s chance to win their next game. The only way for us to measure whether these probabilities are well-calibrated is by analyzing what they imply as far as predicted outcomes. In Table 2 and Table 3, we have the accuracy for each model tested along with the baseline FIFA model for 2-way and 3-way match outcomes. Each of our proposed models is meeting our performance benchmarks in addition to outperforming the baseline model. Moreover, the model’s performance is improving each season, assuring us that the model can better learn each team’s strength with more and more games. To provide such high accuracies with such a simple, interpretable model is almost astonishing. For comparison, Ulmer et al. achieve an accuracy of just over 50% for 3-way outcomes using English Premier League data. With these results, we can be confident that our win probabilities are well-calibrated, and our model is making accurate ratings updates.

Of the proposed models, we selected the model that uses the true goal differentials. Even though the models with adjusted goal differential offer a slight improvement in terms of prediction, this improvement was not deemed significant enough to justify the extra computational time. The advantage of our model is simple, interpretable, efficient updates, but this advantage is lost when a complex machine learning algorithm is employed to create these updates. Updates become ambiguous because it is not obvious how in-game features have been combined to adjust goal differential. This loss of interpretability is not worth a marginal improvement in predictive performance.



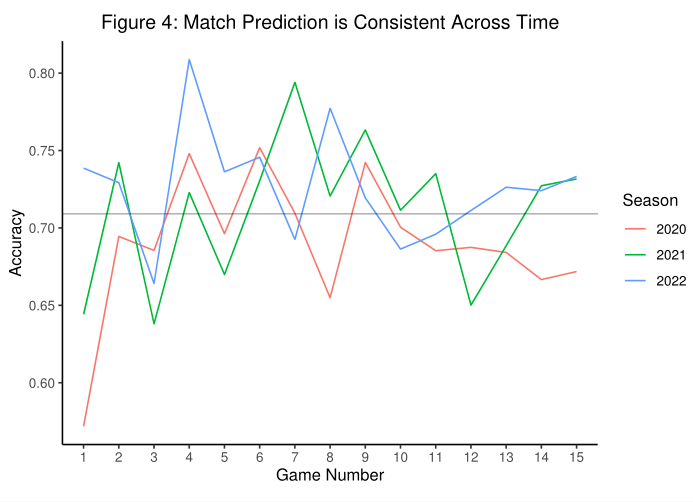
With our final model in hand, we can examine what it implies for the rankings of college soccer teams, our ultimate goal. In Figure 2, we compare the top 15 ranked teams in the United Soccer Coaches poll at the end of the 2021 season with the rankings from our model. Due to the subjective nature of the coaches’ poll, we do not want to match its rankings exactly, but we do want our rankings to lie somewhat close since these rankings incorporate expert knowledge. We can see that our rankings generally track those of the coaches’ poll; most teams lie within 2-3 spots of the coaches’ poll with a few teams ranked significantly higher or lower. The mean absolute deviation of our rankings from the top 15 in the coaches’ poll is 3.8. Overall, we believe that our model finds a nice balance between matching expert knowledge and making its own objective determination of a team’s strength.

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 while 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. We should mention though that our 2022 model has UCLA, the 2022 national champion, ranked as the best team in the country. As far as the 2021 NCAA tournament, we accurately predicted 29 out of 31 conference winners. Given the randomness of conference tournaments since they are only a few games in length, this is exemplary performance. The two that were not correctly predicted, Memphis and Santa Clara, were both ranked second in their respective conferences. As far as the at-large bids, we predicted 28 of the 33 teams, yielding an overall accuracy of just under 90% in predicting the women’s 2021 NCAA tournament field. The seven teams incorrectly included in our projected field were the seven lowest rated teams. Moreover, of the seven teams that made the tournament but were not included in our projected field, four were within ten spots of making the field and five were within 15. As with the polls, we believe that our projected field has a strong enough correspondence to the true one that we can be assured that our rankings resemble the true ones. At the same time, there is significant enough of a difference to suggest that our model is not biased like the tournament selection committee.



Two concerns we had with our model were that ratings updates may be large and volatile, and the model may need a substantial number of games to find each team’s strength. To address the volatility of ratings updates, we used the square root of goal differential. Nonetheless, to ensure that updates were not too volatile, we computed the mean update for each season, finding mean updates of 0.23, 0.21, and 0.17. These results are promising because not only are the updates relatively small, but they are also decreasing over time. Thus, we can be sure that the model is growing increasingly confident in its rating for each team. Figure 3 demonstrates that updates generally are no larger than 0.5.

With respect to our other concern, we have plotted our predictive performance over time for each of the three seasons in Figure 4. After only about 2 games, the model’s accuracy in predicting match outcomes is very close to the overall average, the black line.



In Figure 3 as well, each team’s rating generally remains steady after about 4-5 games. Therefore, we can be confident that our ratings learn each team’s strength relatively quickly.

**Conclusion**

We have created a ranking system for DI women’s college soccer that can be used to generate compelling narratives, predict match outcomes, and make tournament projections. The model is interpretable, efficient, and exhibits exemplary performance. Though the model has been applied only to women’s college soccer, we expect that the model will also perform well in men’s college soccer. The talent disparity in men’s soccer is much smaller than women’s soccer since there are almost half as many teams. As a result, we will likely not face any issues with large, volatile updates and uninterpretable ratings. To improve our model, we would like to automate the preprocessing and modeling pipeline for easy application in practice. Moreover, we would like to better understand where our model performs well and where it does not. We have examined trends over the course of a season, but there may be more subtle trends that we can identify and use to improve upon our model. Lastly, we would like to broaden the narratives that we can build. Currently, our ratings can be used to compare overall team performance, but we would like to create offensive and defensive ratings in order to predict match scores and identify teams’ strengths and weaknesses. One avenue to implement this would be factor analysis, using our created ratings to adjust for strength of schedule. Another direction to broaden our narratives would be to allow the home-advantage effect to differ across teams, allowing us to determine the toughest places to play across the country.

.

**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.