Predicting the 2024 College Football Playoff

Jesse Davila, Collin Travasos, Molly, McCann, Josh Garzaniti, Tyler Seroczynski

Executive Summary

With the 12 team playoff there is a wider range of teams and with that significantly greater room for variability. These 12 teams have had wildly different stories through the season and have a wide range of roster construction and play styles. Their journeys are all unique but nonetheless they, according to the College Football Playoff (CFP) committee at least, are the 12 best teams in football.

Our goal was to create a machine learning algorithm that accurately predicts the winner for each matchup ultimately leading to the national champion. This model needed to be significantly more complex than models for the past CFPs. The four team set up was much simpler than the new format which required a simpler model. There was far less at stake, no campus games, no bye weeks, long periods to rest and prepare, and typically more similar teams. The added variables included in the expanded playoff require a more complex model than had been able to be used in past years.

In our model we factored in metrics such as EPA, WPA, and a Gini coefficient index. With the expanded playoff and the variables it brings we also brought in win probability and a health specific factor for each team. We ran an XGBoost model using a variety of fine tuning that is addressed in our Analysis and Feature Selection.

Our model was able to predict the results of the games at an 81.23% accuracy level. We predicted that the eventual national champion will be Oregon. While there are a few upsets along the way we ended up with one of the favored teams taking home the national championship trophy.

Background

The inaugural 12-team college football playoff has people talking about football more than ever. Teams from the SEC, Big Ten, ACC, Big 12, and Mountain West all have a shot at winning a title. There are teams like Texas and Oregon who can throw the ball ridiculously well. Teams like Notre Dame and Boise State are dominant in the run game and need success on the ground to get going in order to be efficient. Additionally there are teams like Tennessee, Ohio State, and IU who didn't win nor play for a conference championship, and here they are competing for a national title. At the end of the day, all 12 teams as a whole are different, and each of their roads to getting to the playoff have been different. Every week in college football there are upsets few expect to happen, so why should it be different when lights are the brightest? It is our goal to push aside how we feel when it comes to predicting the outcomes of the games and compose a model that we are confident with.

Problem Statement

There has never been a 12 team playoff in the history of College Football, and this has brought up the questions like whether this is the right format to use. We have no idea what to expect. Are all the heavy favorites going to dominate, or will we see tournament magic comparable to that of march madness? There are countless factors when it comes to having success on the field including some things people would consider unquantifiable. As analysts, we do what we can to make variables like home field advantage, bye weeks, strength of schedule, and injuries, among other things, quantifiable. The beauty of sports and specifically a postseason bracket is that the tournament plays out differently every single year. Some years all the favorites win, and other years an underdog will get hot and everything in between. This is College Football, and there has never been a 12 team playoff before so this is uncharted territory. It is our goal to predict the winners of each game and the eventual champion. Will Oregon be the first 16-0 team, or will a season-long Cinderella story like Arizona State continue it's magical run?

Analysis

We ran an *XGBoost* model to answer these questions. First we took the 2024 play-by-play data and partitioned it into an 80/20 training-test split for the model selected. This project's *XGBoost* model was built off of the features previously mentioned and then trained over the course of the last 15 weeks of regular season college football data for the 2024. Then the model was tuned through the use of creating a list of parameters: number of rounds, tree depth (in terms of where branches were split), column and row subsamples, amount of features used, the learning rate of the model, the model's gamma, and child-weight (how many samples needed to be left in each tree's leaf before splitting it again). Each of these parameters were expanded out into a grid and then tested vigorously in a 5-fold cross validation test to select the best possible parameters for the final *XGBoost* model.

This model was used to predict each of the matchups by pitting the home playoff teams against the away opponents for each round of the 2024 CFP playoffs. A health factor was hardcoded into each round of the playoffs for each unique team based on how healthy the group collectively believed each team was. One specific example of this was *Georgia* being given a health factor of .8, while most of the other teams hovered around 1 since *Georgia's* primary quarterback, *Carson Beck* was hit hard during the SEC championships back in week 15, and diagnosed with a UCL injury, which sidelined him for not only the remainder of the game, but put his playoff appearance in question since (Meyer). These health factors were directly mutated to multiply each team's win probability, thereby giving *Georgia* 75% of their original win probability, since quarterback injuries immediately impact both the team's morale and offensive capabilities.

As an additional component of the prediction model created, a Gini index was produced. This Gini index incorporates data from all 12 playoff teams' games from week 10 to week 15. The index measures how unevenly the WPA and EPA for the favored team in each matchup varies amongst each other. Because the original dataset did not have a home and away team specific WPA, a derived variable dependent on whether the favored team was at home or away was made. For this analysis, a low index may have a more balanced play style whether it be offensively or defensively, while a high Gini index could suggest heavy reliance on star players or key plays. A high Gini value for WPA means that only a few players or plays have significant impact on game outcomes. A high Gini index for EPA indicates that most of the point-scoring contributions come from a few explosive plays or players. The results of the Gini index were then added to the XGBoost model as a feature for predicting the overall win probability of the teams. Before doing so, the results were expanded to all 15 of the weeks and teams, thus, it was necessary to assign a Gini value for the teams not competing in the playoffs. This value was 0.5, as it is the middle value for the Gini index which is on a scale from 0 to 1.

Feature Selection

When creating this 2024 CFP prediction model, there are a few variables baked into the overall win probability being used to pit teams over each other. The first of these features is EPA or expected points added. For each team and their respective 2024 games, EPA was taken both from the offensive and defensive perspective and mutated to create a differential between the two (EPA diff). This differential essentially represents the difference between a team's ability to score and allow points marginally. So when taking two teams, their EPA differential can be directly compared against each other to create a scaled version of how good they are individually respective to one another. EPA differential is the basis for the comparative analysis used in the predictive XGBoost model. Built off of that are several other layers of derived variables added to predict wins. The largest after EPA is WPA or each team's calculated win probability, which was simply taken as a mean of each team's WPA on a play by play analysis. Win probability was calculated as the sum of those means for each team Injuries were once again added into calculations, but after the model was already trained on the college football dataset. Injuries were created as a list of individual teams' health index ratings on a scale of 0 to 1 inside of each round of the bracket for the respective teams remaining. As such values were hardcoded to remain consistent through each round.

Results

The final results of the cross validation tuning of the XGBoost model indicated that the best parameters to use in an XGBoost model were: *nrounds 1*50, *eta* of 0.2, *max* *depth* of 3, *gamma* of 1, *colsample* of 0.8, *child-weight* of 5, and *subsample* of 0.8. The final training accuracy of this final model was 81.23%.

The win probabilities created regarding health factors taken into consideration were then directly paired up for each CFP matchup for each unique round and the winners were determined by taking whichever team's win_prob_with_injury was greater than the threshold of 0.5 or 50%. That team would then advance to the next round and the process would continue until a championship team had been produced over the remaining 4 rounds of matchups.

Outlined below are the following model results:

1st Round

- [ ] Texas beats Clemson with a pW of of 93%
- [ ] Ohio State beats Tennessee with a pW of 79%
- [ ] Penn State beats SMU with a pW of 58%
- [ ] Notre Dame beats Indiana with a pW of 88%

2nd Round

- [ ] Boise State beats Penn State with a pW of 72%
- [ ] Notre Dame beats Georgia with a pW of 71%
- [ ] Arizona State beats Texas with a pW of 76%
- [ ] Oregon beats Ohio State with a pW of 98%

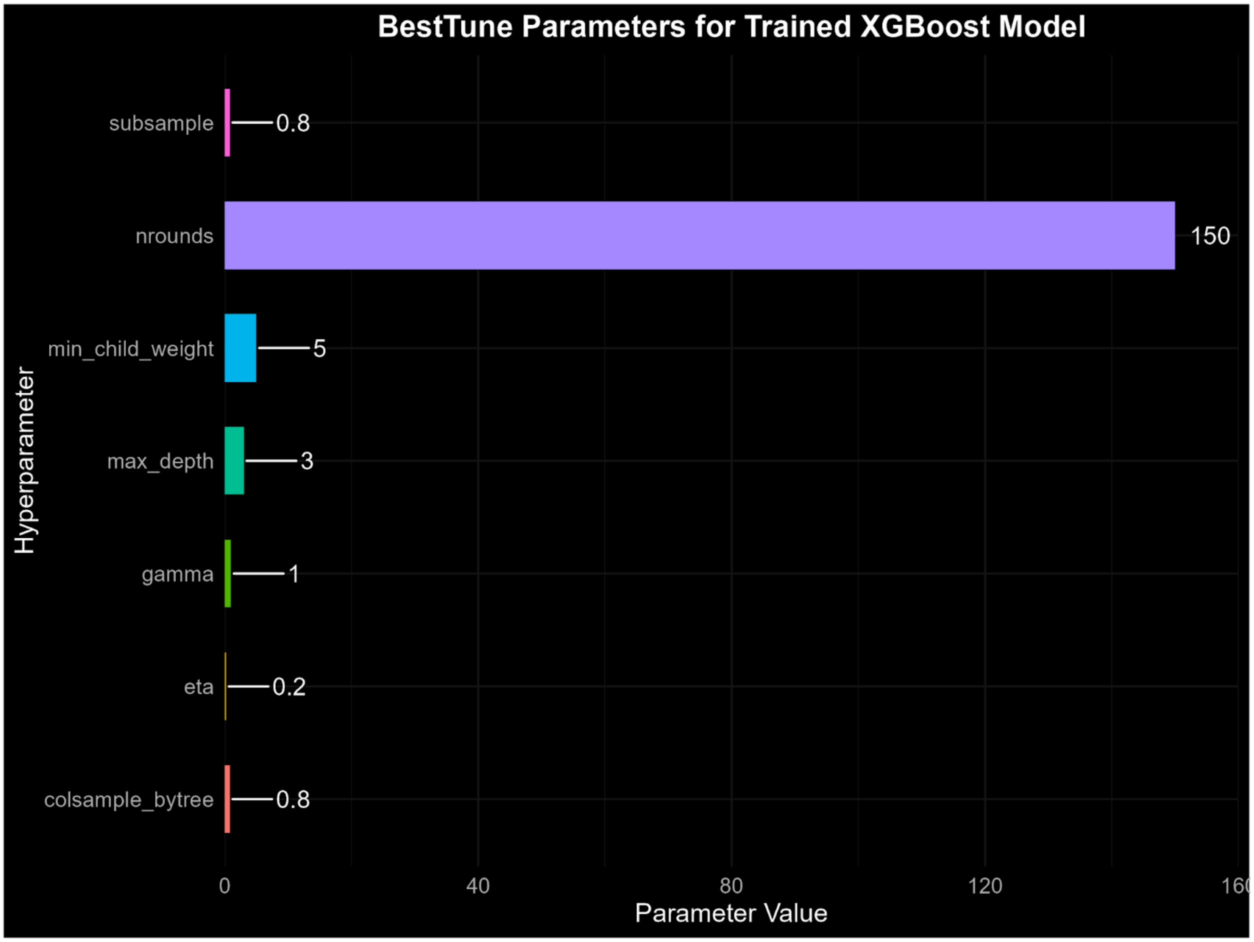
CFP Semi-Finals

- [ ] Notre Dame beats Boise State with a pW of 85%
- [ ] Oregon beats Arizona State with a pW of 78%

CFP National Championship

- [ ] Oregon beats Notre Dame with a pW of 52%

Appendix



Sources

"Evolution of Boosting." *AIML*, AIML, June 2024, https://aiml.com/wp-content/uploads/2024/06/evolution-of-boosting.png. Accessed 12 Dec 2024.

Gilani, Saiem, Akshay Easwaran, Jared Lee, and Eric Hess. *cfbfastR: The SportsDataverse's R Package for College Football Data*. 2021, https://cfbfastR.sportsdataverse.org/. Accessed 12 Dec. 2024.

Kimberly Fessel. “Measuring Statistical Dispersion with the Gini Coefficient.” *Kimberly Fessel’s Blog*, 5 June 2020, kimberlyfessel.com/mathematics/applications/gini-use-cases/.

Meyer, Craig. "Carson Beck Injury Update: Georgia QB’s Elbow Timetable and College Football Playoff Impact." *USA Today*, 9 Dec. 2024, https://www.usatoday.com/story/sports/ncaaf/2024/12/09/carson-beck-injury-update

-georgia-elbow-timetable-college-football-playoff/76868875007/. Accessed 12 Dec. 2024.

Notre Dame Fighting Irish Depth Chart. 2024 Notre Dame Fighting Irish Football Depth Chart | Ourlads.com. (n.d.). https://www.ourlads.com/ncaa-football-depth-charts/depth-chart/notre-dame/91487