Prediction Perfection:

Paul Trusela

Executive Summary:

With annual bracket predictions, fans, sports outlets, and shows are all pitching to say who will win the CFB playoffs. In wake of the inauguration of the 12-team playoff, we devised one of the most comprehensive models to date, accounting for home and away team advantages, individual player talent and injuries, and head coaching prowess. Including these three key variables alongside common predicted scoring metrics, we can determine that Ohio State will successfully block out the outside noise to win it all and return to the top of the CFB mountain.

Background:

Every year fans, gamblers, and news broadcasters all over the country buckle down to predict the winners of the college football playoff. Some use their personal knowledge of the sport, some bet on their favorite teams or against their least favorite, others build and use machine learning models. Whether the purpose is financial benefit, bragging rights, or recreational interest, a model is a great way to predict the outcome of the playoffs. Major sites such as ESPN, CBS Sports, the New York Times, and Sports Illustrated all post predictions for fans to reference, with varying degrees of success. We believe we can improve upon all of these sites. As such, this project endeavors to accurately predict the outcome of each game in the 2025 college football playoff, and ultimately determine the champion.

Problem Statement:

Our analysis will be focused on predicting the College Football Playoff bracket as accurately as possible. We aim to build a model that will correctly predict the results, round by round, with minimal error. Our model uses playoff data from “cfbfastr” and other websites that include data on injuries, suspensions, player grades, etc. We were presented with the R package that included a lot of the stats

necessary to run our analysis, however we did manually collect player and coaching data to use in our model.

Analysis and Feature Selection:

We went with a Random Forest model for our college football predictions mainly because it's like the Swiss Army knife of models—pretty handy and adaptable to handle all the twists and turns in our input data. When you've got several decision trees working together, a Random Forest lowers the chances of overfitting disasters while still giving solid guesses even when faced with messy or not-so-straightforward patterns. It's cool how this model shows off which features really matter, almost like finding that needle in a haystack regarding what drives results.

Plus, since Random Forest handles both types of variables—categories or more number-focused ones—and doesn't sweat too much about scaling obstacles, we can throw in an eclectic bunch of stats from past team performances to player stats and other tidbits influencing outcomes. Basically, picking a Random Forest was about making sure our game prediction setup isn't just wobble-free but also easier to digest and ready to roll across different situations.

In building an initial Random Forest model we used several metrics from cfbfastR and college football play-by-play data. These included EPA, WPA, ELO, and SP+ Ratings. EPA and WPA were included in the play-by-play data, which we retrieved for the 2014-2024 college football seasons. We trained our model on these metrics from the 2014-2023 college football seasons, and tested it on the 2024 season.We used pre-developed functions to calculate the ELO and SP+ values, also for the last ten years. EPA and WPA enabled us to examine player and team performance and efficiency, while also taking into account time of the game (e.g. "crunch time"), and location of the ball on the field. SP+ and ELO enabled us to account for opponent strength and win margins.

We used a variety of different features to run our model from a wide array of sources. The first bit we used was the team PFF data from Pro Football Focus where we scraped every team's data from 2014-2024 to measure team success and different offensive and defensive grades. In order to further estimate a player's impact we multiplied their usage and their PPA. We also computed a team value by adding the player values for the top 15 used players on the team, hoping to narrow down their favored offensive roster. We also chose to get all of the season ending injury data from the past 10 years because it allows us to factor in player injuries which is very prominent this year as Georgia starting QB Carson Beck is likely to miss the entire playoffs with a UCL injury. For this we made player usage a 0 for that season. Multiplying by 0 ensured that that player will not have an impact on the team's playoff outcomes. This was essential information for the training data, as merely including it in the test data would throw off the model's response to the 0s.We acknowledged that this may skew our model in cases where players were injured mid-season, and if given more time and resources we would have preferred to break player contribution down to a weekly metric, using their previous week's metric to predict the current week's performance. Another issue with injuries was that consolidated college football injury data was only available for the 2024 season. Further work should include a more in-depth search for historical injury data. While our settled method had its flaws, we counterbalanced this issue by including the player's non-injured contribution in the model, as well as focusing on the top 15 used players in each team. Using the top 15 players significantly reduced instances where players injured midseason were considered in the model, and including the non-injured contribution allowed injured players to still be taken into account.

Our next metric of consideration was coaching, a very important part of team success late in the season when players get tired and hurt. We wanted to include something to mention the effect of coaches on team success and so our next step was to create a metric to calculate this. By multiplying the number

of years coached, career strength of schedule, and win percentage together we were able to come up with a method to rank the best coaches in college football and how much of an impact they have on their teams success.

Results:

Testing our model yielded an accuracy of 77.51%, meaning that in general, it correctly predicts the outcomes of games, win or loss, 77.51% of the time. Our sensitivity was 69.23%, meaning that it correctly predicted wins 69.23% of the time. While this number was lower than we would have preferred, our specificity was more encouraging, accurately predicting losses 83.55% of the time. However we believe our model is strong, and are confident in its overall predictive power.

Regarding the 2024 college football playoff, our results were found by simulating each round game by game. In the 5-12 matchup between Texas and Clemson, Texas advances to play Arizona State. In the 8-9 matchup the Ohio State Buckeyes beat Tennessee while Penn State beats SMU. Notre Dame beats Indiana and then Georgia in their first two games and they play Penn State in the semi-finals after the Nittany Lions take out Boise State. Texas beats Arizona State and Ohio State gets their week 7 loss to Oregon back and beats the ducks. The semifinal matchups pit the 5th seeded Longhorns against the 8th seeded Buckeyes and the 6th seeded Nittany Lions against the 7th seed Fighting Irish. We end up with a 7 vs 8 matchup in the national championship where Ohio State plays Notre Dame. The Irish do fall short though, losing to Ohio State in the national championship.

Conclusion:

As much as we wish we could say our model predicted a Notre Dame national championship, that is not how our model predicted the college football playoff. With an accuracy of 77.51%, and specificity of 83.55% we are confident in our model's predictions of losses in the series. While we would prefer a sensitivity higher than the 69.28% we achieved, we believe our derived variables

contribute a predictive strength other models do not have. For gamblers, we suggest betting on losses rather than wins, but we are confident that our model will strongly reflect the actual outcome of the 2024 college playoff. To further strengthen our model, we would require more detailed injury data from the last ten years, as well as an injury consideration on a weekly basis rather than a season basis.