Predicting Playoff Results in the NFL

Arlena Tolmasoff, Anthony Paolillo, Matthew Schneider, Shubham Pandey

Problem Overview and Data:

Each season, 32 NFL teams vie for the ultimate goal of winning the Super Bowl. After duking it out for 18 weeks in the regular season, 14 teams split across two conferences advance to playoffs, with the top team in each conference earning a first round bye. The playoffs consist of four rounds (Wild Card, Divisional, Conference Championship, and the Super Bowl) and are single elimination. In the end, one team hoists the Lombardi Trophy and returns home for a Super Bowl Parade.

With the growing fanbase of the NFL, as well as the legalization of sports betting across many states in the U.S., more and more people are searching for ways to successfully predict the Super Bowl Champion. This serves as the inspiration for our final project. By using the statistics of a team accumulated throughout the regular season, our goal is to predict the Super Bowl winner of a given NFL season. With our model, we hope that we can identify which of the playoff teams of a given season will come out on top using the teams' current season statistics. This would allow us to best understand the landscape of the NFL and what types of teams tend to succeed in the playoffs.

We obtained our data for this project from https://www.pro-football-reference.com/ over the 2000-2022 NFL seasons. We scraped data tables from this website for each year, containing team information and performance measurements such as number of wins and losses, point differentials, and offensive and defensive statistics. From these tables, we obtained the performance statistics of interest for each team that made it to the playoffs during the years 2000-2022. As a result, we produced a dataframe containing 282 rows (corresponding to the 282 playoff teams) and 73 columns containing team information and performance measurements.

In order to measure the success of a team, we created a categorical response variable, 'Playoff_Result' to indicate how a team performed in the playoffs for a given year. This variable has levels 1-5. A value of 1 indicates this team lost in the Wild Card round, a value of 2 indicates this team lost in the Divisional round, a value of 3 indicates this team lost in the Conference Championship, a value of 4 indicates this team lost the Super Bowl, and a value of 5 indicates this team won the Super Bowl.

Exploratory Data Analysis:

When it comes to predicting the Super Bowl winner, it first is essential to understand which variables best determine a team's success. Given that our data had approximately 70 statistics from 282 playoff teams, we needed to perform some exploratory data analysis in

order to best select our predictor variables for our final model in an effort to avoid overfitting and correct for collinearity between our predictors. To accomplish this, we split these statistics into three categories (General, Offensive, Defensive) and from there, determined which had a significant effect on a team's playoff result.

The first part of our analysis worked with general team statistics such as Wins, Losses, Point Differential, Strength of Schedule, and other non-playing statistics. Grouping by each of the 5 responses to our variable "Playoff_Result" allowed us to compare the distributions of each of these statistics over the twenty year period. Figure I below shows the boxplots across the playoff outcomes for each stat.

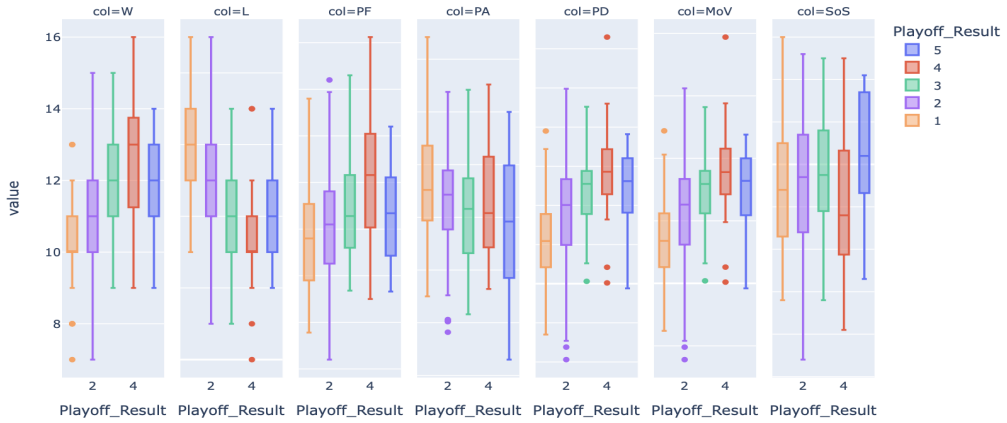

Figure I: Distributions of Seven "Non-Playing" Statistics based on Playoff Result

To supplement this graph, we conducted an ANOVA test in order to determine which means were significantly different. We were able to conclude that all of the variables, outside of PA, had a significant effect on playoff results. When analyzing these variables, it is clear that some will be highly correlated with each other (Ex. Wins and Losses, Margin of Victory and Point Differential). Therefore, we subset based on each of the variables' relationship with playoff result and concluded that Wins, SoS and Point Differential were the best predictors of playoff result given our constraints.

Using our in-game offensive statistics to determine which predictors to use for our final model required a similar process to the one outlined above. Taking into account passing, rushing, conversion, and penalty statistics and adjusting for multicollinearity, we were able to select the following for our model that best predict playoff result: score%, expected points, third down conversion, total offensive yards, passing yards, yards per pass, total first downs, rushing touchdown, and yards per play. With these offensive statistics, we now had a better understanding of what gameplay measures significantly impact a team's playoff success and use these, along with others, to predict a Super Bowl winner given a field of playoff teams.

We also wanted to explore the relationship between defensive metrics and playoff results to find interesting trends. Similar to the EDA for offensive statistics, we used box plots to explore the distribution of each metric compared to the ordinal playoff result. The goal was to find metrics that had a large discrepancy between different playoff results, as those metrics might be effective predictors for our model. We created box plots for each metric and then selected the ones that looked significant. The variables that we deemed significant from the exploratory data analysis were third down conversion rate, fourth down conversion rate, net yards per pass attempt, red-zone touchdown percentage, and yards per play. For each of these metrics, the lowest median for each playoff result was by the superbowl winner, indicating that the better defenses usually make it further. These plots can be seen below in Figure II.

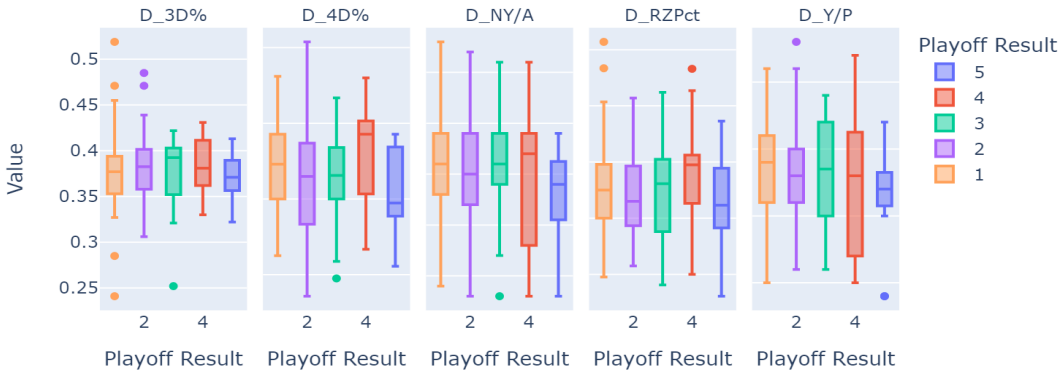

Figure II: Distribution of five defensive metrics based on Playoff Results

Another approach that we attempted was normalizing metrics by year. The NFL is an ever evolving league and stats that might have been impressive in some years, might be mediocre in other years. For example, the top 10 single seasons in passing yards all occurred in the last 12 seasons. The playstyle of the league changes and so we thought it might increase the predictive power of our model if we normalized certain metrics by the year. We normalized the metrics by ranking every playoff team for each metric. For example, the team with the least amount of points scored would have a 1 rather than their total number of points. The defensive metrics that we found interesting were fourth down conversion rate, defensive expected points, net yards per pass attempt, yards per rushing attempt, yards per play, and strength of schedule. Some of these metrics were significant in both the ranked and raw form. Figure III below shows the boxplots for these ranked metrics.

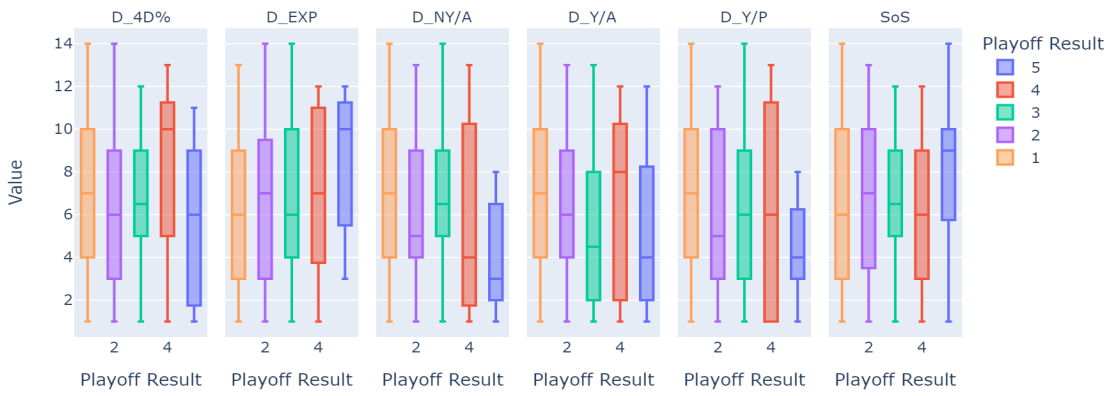

Figure III: Distribution of ranked metrics based on Playoff Results

Methodology and Analysis:

After performing exploratory data analysis, we selected a subset of 17 performance statistics which we thought would be most helpful in predicting playoff results. We fit a series of Random Forest Classification models to the subsetted data. We first fit these models using the raw data alone, and then in an attempt to standardize over each year due to the changing landscape of the league, we also calculated the ranks of each team's statistics within each year. We used 5-fold cross validation for all models to tune the 'ccp_alpha', 'max_depth', and 'max_features' hyperparameters. For each model, we selected the values of these hyperparameters which produced the optimal 'roc_auc_ovr', which is an adaptation to the ROC AUC metric that can be used for multiclass classification.

We evaluated the performance of each model as follows: we held out one year of data to use as a test set and used data for the remaining years as a training set. We then fit a Random Forest Classification model to the training data and calculated the test error and AUC score (using the previously mentioned 'One-vs-Rest' method). We repeated this process for each of the years, and calculated the average test error and average AUC score over all years. After comparing the performance of models fit on the raw data alone, the descending ranks of the performance statistics alone, and both the raw data and the ranks, we found that the best performing model used both the raw data and the descending ranks of the performance statistics within each year. This model used hyperparameters 'max_depth' of 9, 'max_features' equal to the base-2 log of the number of features, and a 'ccp_alpha' of 0.0035.

As an alternative approach, we also simplified our problem from an ordinal classification problem to the binary classification problem of predicting whether or not a team will win the SuperBowl for a given year. Using our previous 'Playoff_Result' response variable, we created a new binary response variable with a value of 1 if a team won the SuperBowl and a value of 0 if a team did not win the SuperBowl. We then fit a series of

Random Forest Classification models in a similar manner as before, using the raw data alone, the ranks of the performance statistics within each year alone, and also the raw data combined with the ranks of the performance statistics within each year. 5-fold cross validation was used once again to tune the 'ccp_alpha', 'max_depth', and 'max_features' hyperparameters. We selected the hyperparameters which produced the highest AUC score for each model.

As before, we evaluate performance by iteratively holding out one year of data as a test set and using data for the remaining years as a training set. However, to account for the fact that there can realistically only be one Super Bowl winner per year, rather than using the predict() method to assign class labels we instead use the predict_proba() method to compute the probability of each team winning the SuperBowl according to our model. We then assign a '1' as the response for the team with the highest such probability, and a '0' as the response for all other teams. Using each year's data as the test set, we store the test error and AUC score and compute the average of these over all years for each model. We found that the Random Forest Classification model fit on just the descending ranks of the performance statistics within each year produced the highest AUC score on average. This model used hyperparameters 'max_depth' of 5, 'max_features' equal to the base-2 log of the number of features, and a 'ccp_alpha' of 0.

Results and Interpretation:

The best performing ordinal Random Forest Classification model fit on the raw performance statistics combined with the descending ranks of the performance statistics within each year produced an average test error of 0.5595 and an average AUC score of 0.6974. The test errors and AUC scores using each year of data as the test set for this model are plotted in Figure IV. From Figure IV, we see that the performance of this model is very inconsistent from year to year. For some years this model produces a test error of as low as 0.33, but for others the test error is as high as 1, indicating that the model did not predict a single team's playoff result correctly.

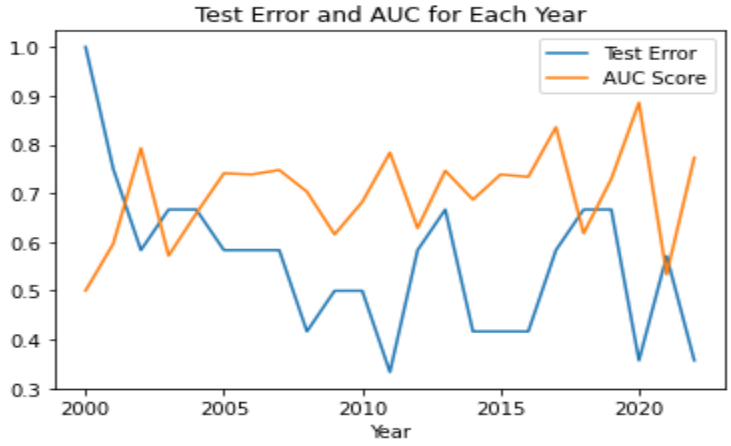

Figure IV: Test Error and AUC Scores for the Ordinal Random Forest Classification Model

The best performing binary Random Forest Classification model fit on just the ranks of the performance statistics within each year produced an average test error of 0.1128 and an average AUC score of 0.7182. The test errors and AUC scores using each year of data as the test set for this model are plotted in Figure V. We see that the test errors for this model are somewhat more stable than in the ordinal classification model. The AUC scores, however, are still quite inconsistent from year to year. For some years, this model produces an AUC score as high as 1, but for others the AUC score is as low as 0.09.

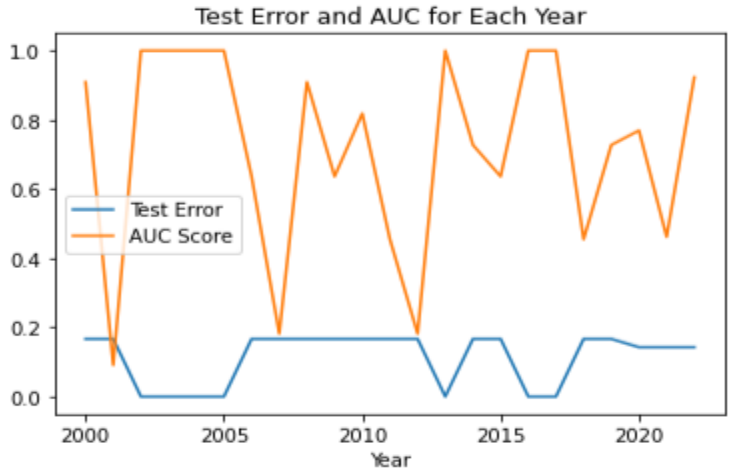

Figure V: Test Error and AUC Scores for the Binary Random Forest Classification Model

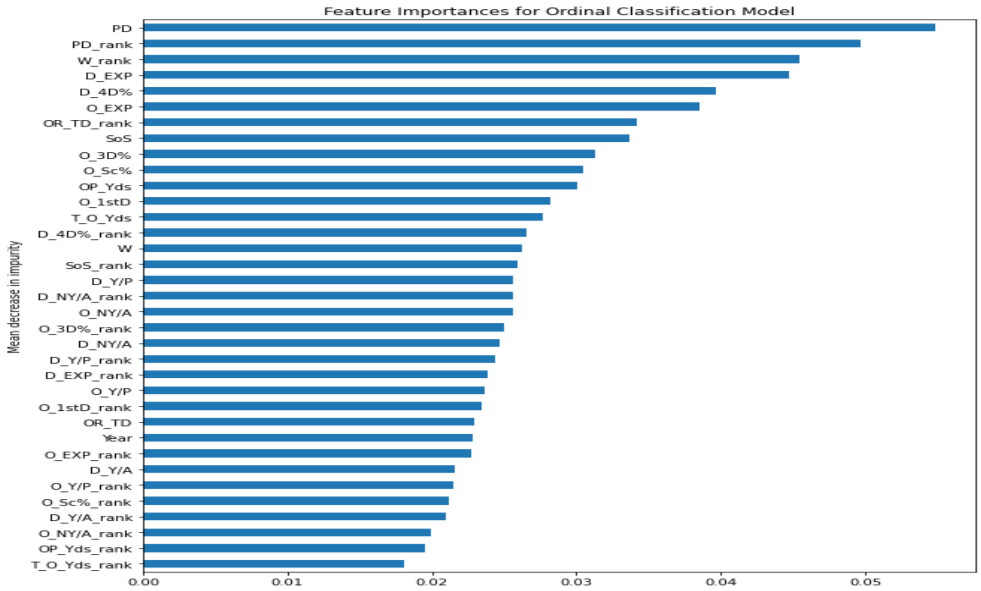

Figure VI : Most important features selected by the best Ordinal Classification model

Looking at the most important features in our best ordinal classification model, we see that most weightage is given to stats like point differential, win_rank and offensive stats like rushing touchdowns, third_down conversion % and so on. Offensive stats in general are more important here than defensive stats, and ranked stats are more significant than their

unranked counterparts, as expected. Some stats like net yards per pass attempt, which have commonly been highly associated with winning, are not ranked as high, probably because this is ordinal classification, where there is granularity in outcome, instead of just winning it all, hence other advanced stats take precedence over simpler ones.

Comparing that to the binary classification model, we see that strength of schedule is the most significant for ultimately winning it all, which makes sense. Defense-based stats dominate the top features here compared to the ordinal classification model, which is probably due to the fact that our data goes back till 2000, and traditionally defensive juggernauts like Patriots dominated the early 2000s, which has influenced our model. Offensive stats don't fall behind too much, with stats like Offensive expected points featuring in top 5 features. Net yards per pass attempt doesn't fall behind too much in this case, falling in the top half of features.

We can also interpret the win probabilities predicted on our test set after adjusting for opponent win probabilities (all scores should sum up to 1). For instance, for the 2021 season, the predicted Super Bowl winner had a win probability of 18.6%, while a large proportion of teams had less than 5% chance of winning. There are other competing teams in the mix too, with two other teams having around a 12% chance of winning. Comparatively, in the early 2000s, the strongest team had a 26% chance of winning, while several teams with less than 4% probability of winning, our model indicates that the overall parity in the league has improved.

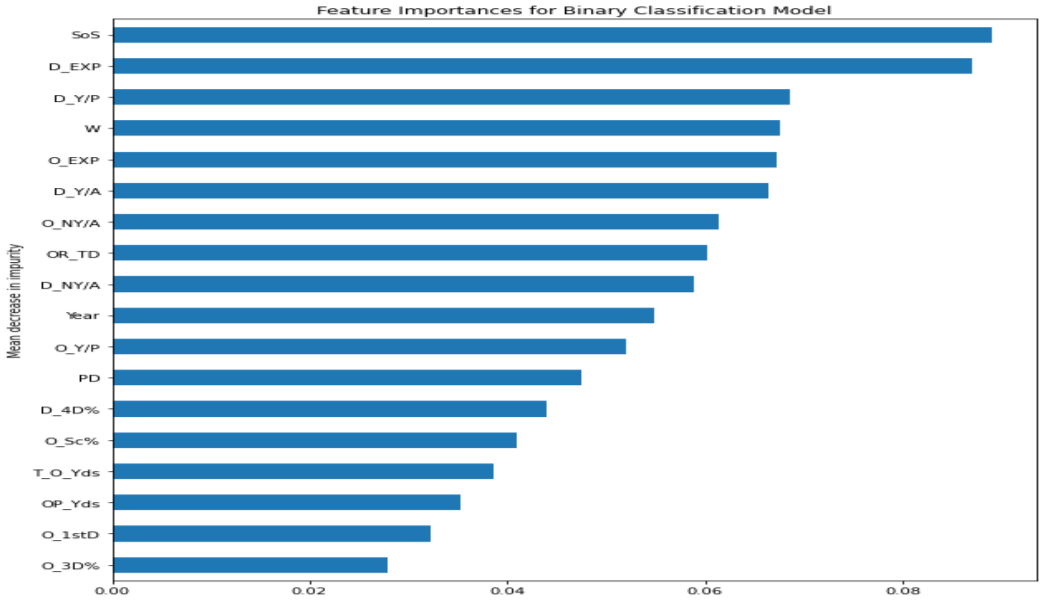

Figure VII : Most important features selected by the best Binary Classification model

Discussion and Conclusion:

When evaluating the results of our model, it is important to first remember the structure of the playoffs. Our statistics are collected over a span of 500+ plays during a season. However, being that the NFL playoffs is single elimination, a team's outcome can be decided on only a handful of plays. Thus, outcomes of games tend to have high amounts of variability. So, while our model is trained on statistics from an entire season, the outcome being highly dependent on very few plays can cause large fluctuations in our results, as well as noise in our dataset from the results of previous years.

Another factor that must be considered is the year to year change in the NFL. Every year, teams are constantly searching for the best strategies to win, and what may be very successful one year may be completely different in the next couple of years. Thus, since our data spans over the course of 20+ seasons, as the NFL evolves, it is easy to understand how the tail ends of our timeframe may have different play styles. While we attempted using different strategies, such as only the past 5 seasons as our training data, ultimately our model best succeeded with the inclusion of more data.

There can be various applications of Super Bowl winners, the most obvious being sports betting predictions. Using our adjusted win probabilities, we can conclude that the betting odds for recent Super Bowls shouldn't be too extreme, with at least 2 to 3 teams in the mix for winning it all. This does reflect the actual Vegas odds for the past few years, which is another indicator of an increase in parity between the teams. Our model can be particularly useful when these odds are significantly different, as a large gap could potentially lead to arbitrage opportunities in the betting market.

We believe that there is room for improvement that could be the focus of future work. The simplest approaches could include collecting data from more years. The problem with this approach is that as we go further into the past, the data becomes more unreliable. We could also try out different methods for normalizing the data. We used a yearly ranked normalization technique. It might be interesting to standard score normalize the data by year or by some defined time period. Another idea that might work is to collect more predictors for the data, such as EPA, which is a newer and more advanced statistical measure in the sport. Adding in data or playoff results from previous seasons might also improve our model. Lastly, since there are only a limited number of playoff games in general, a way to remedy the small sample size might be to create simulation studies/synthetic data to help the model learn better. All of the above are interesting avenues to explore to improve the results.

Overall, this project gave us the opportunity to apply many of the methods we have learned throughout the course in an area that we are highly interested in. We feel confident in what we have produced and look forward to potentially building on this in the future.