Gerardo Ferrigno

# AI Project Interim Results: Bayes Network for NFL Playoff Prediction

This folder contains all the materials utilized for the first project intermediate results. All the datasets used for compiling information are in the Data Folder. The following datasets were used:

- NFL_Merged_Data.csv: This is our most important dataset, alongside Binary_Dataset.csv. This is a merged dataset consisting of clean_nfl_1999_stats.csv combined with nfl_first.csv.
- nfl_first.csv: This file contains data for all NFL teams with their wins, losses, Points Against, Points For, Strength of Schedule, Simple Rating System, Win Loss Percentage, and other data that was not used for this analysis. The playoff column was calculated by myself alongside the Strong_Start variable, which means whether a team started 3-0 or not. Although the data dates back to 1990, measures since 1999 were used. The data was gathered from https://www.pro-football-reference.com/
- nfl-team-statistics.csv: This file contains relevant offensive and defensive data from each team dating back to 1999. However, not all of these variables were used for the Bayes Network, as it was found that some of them do not contribute significantly.
- clean_nfl_1999_stats.csv: Clean version of **nfl-team-statistics.csv.** Teams had to be renamed to achieve same naming convention as other datasets. This dataset was merged with **nfl_first.csv**
- result_matrix.csv: Correlation matrix that contains relevant relationships between variables given their thresholds.
- annotated_result_matrix.xlsx: Excel file that contains highlighted relevant relationships. Some relationships were ignored due to similar metrics. Relationships highlighted in orange are strong, while yellow are moderate. Green indicates strong relationship with the playoff variable, while blue indicates a moderate relationship.

Unused data

- NFL_Conversion_Data.csv: Contains metrics like third down percentage, fourth down percentage, red zone efficiency. No variable from this dataset was added since there was no significant relationship between these metrics and a team making the playoffs.
- NFL_Full_Conversion_Data.csv: Merged **nfl_first.csv** with **NFL_Conversion_Data.csv.**

Notebooks:

- data_1999.ipynb: data cleanup to generate clean_nfl_1999_stats.csv
- merging_cleanup.ipynb: notebook used to merge datasets and create NFL_Merged_Data.csv, which is our relevant dataset.
- Binary_dataset.ipynb: Converts NFL_Merged_Data.csv to a Binary Dataset called Binary_Dataset.csv, which will be used for calculating the probabilities in the Bayes Network.

Gerardo Ferrigno

# Methodologies for Establishing Thresholds

There were various methods for establishing thresholds, some based on domain knowledge, others by using averages for playoff teams.

- SoS metric: Using domain knowledge, it was concluded that an above strength of schedule is above 0.449.
- SRS (Simple Rating System): This metric is the sum of OSRS (Offensive Simple Rating System) and DSRS (Defensive Simple Rating System). Using domain knowledge, it was concluded that an above average SRS is greater than 0.
- OSRS: Above average OSRS is greater than 2.
- DSRS: An above average DSRS is above 0.1
- MoV: means Margin of Victory. An above average MoV is when a team wins by a spread of 6.5.
- PD (Point Differential): A strong team has a point differential of more than 50.
- Average Points Allowed: A strong team allows less than 21 points per game.
- Average Points Scored: A strong team scores more than 24 points on average.
- Offensive Completion Percentage: A strong offense completes their passes more than 55% of the time.
- Average Passing Yards per Play: More than 6.6 passing yards per play indicates a strong offense.
- Average Rushing Yards per Play: More than 4.5 rushing yards per play indicates a strong offense.
- Defense Completion Percentage: A defense with more than a 55% completion percentage is considered weaker.
- Passing Yards Allowed: A defense that allows more than 5.9 passing yards per play is considered weak.
- Rushing Yards Allowed: A defense that allows more than 4.5 rushing yards per play is considered weak.
- Defense Passing Success Rate: A passing success rate above 43% is considered weak.
- Defense Rushing Success Rate: A rushing success rate above 40% is considered weak.
- Offensive Passing Success Rate: A passing success rate above 46% is considered strong for an offense.
- Offensive Rushing Success Rate: A rushing success rate above 46% is considered strong for an offense.
- Offense Average EPA Pass Rate: A rate above 0.15 is considered strong for an offense.
- Offense Average EPA Rush Rate: A rate above 0 is considered strong for an offense.
- Defense Average EPA Pass Rate: A rate below -0.02 is considered strong for a defense.
- Defense Average EPA Rush Rate: A rate below -0.06 is considered strong for a defense.

- Interceptions: Above 16 are considered above average, as the league average has been around 15.
- Average Offensive WPA Pass Rate: Above 0.002 is considered above average. In this context, a stronger offense contributes on average 0.002 to their win probability per play.
- Average Offensive WPA Rush Rate: Above 0 is considered above average.
- Average Defensive WPA Pass Rate: A defense with a rate above 0 for passing is considered weaker.
- Average Defensive WPA Rush Rate: A defense above a -0.0018 rate is considered weak.
- Average Offensive Success Rate: The average between an offense's passing success rate and rushing success rate. An average success rate above 0.44 means a stronger, more balanced offense.

Additional Notes for variables:

- Success rate depends on play context. A successful play gains at least 40% on first down yardage, 60% on second down, and 100% for third and fourth down. For reference, visit: https://www.sports-reference.com/blog/2023/09/success-rate-comes-to-pro-football-reference/#:~:text=Instead%2C%20Success%20Rate%20shows%20us,on%203rd%20or%204th%20down.
- EPA (Expected Points Added) is an advanced statistic used in the NFL. It considers the points expected to be added by offensive play. The inverse is true for defense. This means that plays closer to opponent territory have higher values. For reference, visit: https://bestballstats.com/expected-points-added-a-full-explanation/
- WPA (Winning Probability Added) is another advanced metric used that considers how a play adds to the win probability of a certain team. For reference, visit: https://www.advancedfootballanalytics.com/2010/01/win-probability-added-wpa-explained.html

Limitations:

- As there are more variables present in this analysis, there is a heightened probability for errors in identifying thresholds, potentially overestimating or underestimating relationships.
- There has been evolution in NFL playstyles over the years, meaning some variables were more relevant in the past while others are more relevant today.

Gerardo Ferrigno

## Final Variables Chosen for Bayes Network

Upon analyzing each variable and their relationship using domain knowledge and statistical analysis like developing a correlation matrix, the chosen variables that contribute most to predicting whether a team advances to the playoffs are the following:

- Average Offense Passing Yards Per Play
- Offensive Simple Rating System
- Average Points Scored
- Defense Average EPA Pass Rate
- Defense Simple Rating System
- Average Points Allowed
- Point Differential
- Average Offensive WPA Pass Rate
- Defense Average Passing Yards Allowed
- Simple Rating System
- Margin of Victory
- Win-Loss Percentage

## Findings

For now, this section will not contain much, as the probabilities have not been generated yet. However, there were several discoveries when researching and analyzing the data gathered, like the following:

- Conversion data like efficiencies on redzone, third, and fourth down have a weaker relationship compared to other data found on nfl-team-statistics.csv. Offensive and Defensive efficiency, alongside average yards per play, and success rates contribute more to a team making the playoffs. That is not to say that this is the case in practice; limitations have been identified above.
- A strong (3-0) start has a weaker relationship to making the playoffs compared to the variables listed above for the creation of the Bayes Network.
-