Gerardo Ferrigno

AI Project Final 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 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.
- NFL_Bayes_Net.png: Visualization of the Bayes Network.
- CPTs.xlsx: Excel Spreadsheet with calculated CPTs. These CPTs are the ones obtained from **probabilities.ipynb**

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.
- probabilities.ipynb: This notebook was created to calculate the CPTs in our Bayes Network. It defines the thresholds for each metric and takes in **Binary_Dataset.csv.** The calculations were made in order of breadth of the tree. The final values were stored in CPTs.xlsx
- main.ipynb: This is the main notebook where the Bayes Network is created and answers important research questions regarding the established variables.

Libraries

The following libraries were used for the creation of this project:

- Pandas: For creating, manipulating datasets
- NumPy: For calculating CPTs
- probability4e: From aima-python, this module contains the Bayes Network class implementation used for creating the NFL Bayes Network
- sys: This module was used for importing probability4e. It modifies the current path to probability4e's path. If running the code, replace the path argument from, sys.path.insert(1, "/Users/user/Desktop/aima-python") to the path where aima-python is located.

Code Explanation

The main notebook for this code is main.ipynb. As discussed, this notebook creates the Bayes Network and contains answers to various research questions. The notebook runs as is and does not ask for any input. If desired, a user can calculate their own probabilities by using enumeration_ask. Various notebooks were used but the most important one is main.ipynb. Information about the other notebook is found above. The code works simply by running main.ipynb and installing the required dependencies.

Methodologies for Establishing Thresholds

There were various methods for establishing thresholds, some based on domain knowledge, others by using averages for playoff teams. These were the established thresholds for each metric:

- 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.

Gerardo Ferrigno

- 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.

- Win-Loss Percentage: A strong team wins more than 59% of the time.

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

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/Test Results

The Interim results for this project contained the following findings:

- 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.

After successfully creating the Bayes Network and answering various research questions, the most important findings were:

- A strong defense, particularly against the pass, significantly improves a team's likelihood of success. Teams with a low Defense Average EPA Pass Rate are more likely to allow fewer than 21 points on average. This highlights the importance of defending the pass in today's game.
- Teams that allow fewer than 5.9 yards per play or fewer than 21 points per game have an 80% chance of making the playoffs.
- Average Passing Yards Per Play play a crucial role in a team's ability to score. Teams with at least 6.6 passing yards per play have a 57% chance of scoring at least 24 points per game on average. Conversely, teams that average less than 6.6 passing yards per play have only a 26% chance of scoring at least 24 points.
- Margin of Victory plays a crucial role in making the playoffs, as teams that secure wins by more points are less likely to lose.
- Offensive ratings provide accurate metrics for measuring the effectiveness of an offense. It was found that a team, given they made it to the playoffs and had a strong OSRS, were 82% likely to score at least 24 points on average.
- While stopping the run is crucial for defenses, in today's league a pass defense is more important. A strong passing defense has a heightened probability of allowing less than 21 points on average. However, as discussed, a balance in both types of defenses is critical.
- SoS (Strength of Schedule) has no significant relationship to a team making the playoffs.

Final Limitations of the Network

Although this model is perceived as accurate, there are various limitations that need to be discussed:

- 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.
- The model relies on historical performance metrics and team statistics but does not account for structural factors unique to the NFL. Division winners are guaranteed a playoff spot regardless of their overall record, as demonstrated by outliers like the 2010 Seattle Seahawks. This playoff seeding rule is not directly captured in the data and can lead to unexpected results. The model does not incorporate information about divisional competitiveness within a season. Teams in weaker divisions have inflated playoff probabilities due to lower

competition, while strong teams in tougher divisions may face more challenging paths to the playoffs.

- A fundamental assumption of Bayesian Networks is conditional independence among parent nodes. While this simplifies calculations, it may not fully capture the intricate interdependence among parent nodes. For example, team strength, measured by SRS, may indirectly influence multiple downstream factors in ways the model cannot represent accurately.
- The CPTs in the network are derived from observed frequencies and may not generalize well to unseen data.
- The model does not consider offensive and defensive data regarding rushing. Although this data is crucial and should be incorporated, there were no significant relationships found between these metrics and the variables incorporated in the network.
- The model does not consider down efficiencies. Although these are critical for a team winning, they were not incorporated into the model because no significant relationships were found.
- Although important, special teams data was not considered either, as they were difficult to incorporate.

Conclusion

The elaboration of this Bayes Network has demonstrated the predictive power of Bayesian Networks in the field of sport analytics. The insights acquired from this study provide actionable takeaways for how different variables influence scoring, winning, and eventually making the playoffs. As the NFL continues to evolve with more sophisticated play styles, these tools will become vital in understanding the complexities of the game. Future work could focus on addressing these limitations and potentially making this model more accurate.