**DEPARTMENT OF THE NAVY**

**NAVAL** POSTGRADUATE SCHOOL

1 UNIVERSITY **OR**

MONTEREY, **CA93943-500**

From: Gridiron Success Team

To: Commander, OA3801: Comp Methods II

Via:  (1) Capt. Rashad Brown

(2) Capt. Ryan Harris

(3) LT Noah Richwine

(4) LTJG Mark Talvacchia

Subj: EXECUTIVE SUMMARY AND KEY FINDINGS OF “GRIDIRON SUCCESS: UNRAVELING THE WINNING FORMULA”

Ref: (a) Pro-Football Focus

(b) Pro-Football Reference

* + - 1. As lifelong, die-hard fans, our group decided to dive into the data-rich environment that is professional sports. As Howie Long said in 2000, "Football is (America's) passion," and this is the complex sport we chose. The basic goal of our group was to look at team data to really find out what it takes to be a successful team. However, we quickly realized that success is difficult to define as a real-world concept. Financial, marketable, regular season (IE winning percentage), or playoff success were all areas we could have focused on. Though every aspect deserves rigorous examination, our team decided to focus on season outcome in terms of win percentage and playoff performance. Even narrowing down ‘success’ to these metrics posed challenges though, since there is no single statistic to measure this kind of success. Which would a General Manager rather have, a team that goes 17-0 in the regular season and loses their first playoff game or a team that goes 9-8 in the regular season but puts together a playoff performance worthy of Super Bowl victory? To answer this question, we decided to create our own "Season Performance Factor" (SPF) to allow us to fairly compare regular and postseason play in one model.
      2. For data collection, both Reference (a) and (b) were utilized. Both websites have a reputation for being the industry standard. This allowed us to have minimal data cleaning responsibilities from website formatting or pulling from many different sources. For our web-scraping algorithm, the python package “Beautiful Soup” was used. Our Algorithm created CSV files for each year we selected (2010-2023) categorized by team stats, team games, rankings and coaches. These years were used since it was a range long enough to sufficiently provide trend analysis, while also minimizing factors such as changes in the way the game is played and significantly increasing code run time. To combine the data, we used the glob function while ensuring each file added a new column with the year so that we would not lose track of our date for future plotting. Cleaning the data became a lengthy challenge for the group due to very small but significant differences. For example, the team faced challenges from one specific game being canceled and several teams undergoing name and/or city changes. These differences caused several duplication or base ten errors in early models.
      3. Once our data was sufficiently cleaned and organized, it was time to start the actual data analysis. As stated earlier, the team decided to define our own metric of success, being the Season Performance Factor. To elaborate, the performance factor takes the win percentage from the season uses a multiplier based on how far the team made it into the playoffs. The following table shows the multiplier used in our Season Performance Factor.

|  |  |
| --- | --- |
| Week | Record Multiplier |
| Super Bowl | 1 |
| Conf. Champ | 0.9 |
| Divisional Round | 0.8 |
| Wild Card | 0.7 |
| No Playoff Game | 0.65 |

Table 1: Season Performance Factor Multipliers

The playoff position was determined by assigning a week number to each round of the playoffs. For example, the Wild Card Round took place during week 19 and Super Bowl during week 22. Each number of weeks played (18 if the team did not make the playoffs) was then assigned a multiplier based on Table 1 above. The product of this multiplier and the team’s season win percentage is our final Season Performance Factor for the given season.

4. Due to the wide range of data available to us, the team decided to utilize machine learning to help identify key metrics contributing to our Season Performance Factor. The ML algorithm was used to assign scalar values to every metric in the dataframe. Once the dataframe had scalar values, the ML performed a set of linear regressions to look for high correlation trends. We had the ML process train on 80% of the data, and test on the remaining 20%. The result of our ML algorithm are two bar graphs, Figure 1 and Figure 2 below.

A graph of a number of blue bars

Description automatically generated

Figure 1: Feature Importance Percentages by Metric Assigned by Machine Learning Algorithm

A graph of different teams

Description automatically generated

Figure 2: Feature Importance Percentages by Team Assigned by Machine Learning Algorithm

Figure 1 shows what team stats had the highest correlation to team success. Intuitively, forced\_points\_mean or more eloquently the average number of points scored was the largest contributor to team success. Interestingly, the next most important stats were defensive pass yards ratio (pass yards allowed / total yards allowed), rush yards allowed, and offensive passing yards. Figure 2 shows the top teams by correlation to our linear regression. As seen, the New England Patriots best fit our model. This means that the success (or lack thereof) of the New England Patriots fit what our ML algorithm defined as important more than any other team. In other words, we can conclude that points scored, defensive passing yard ratio, and rushing yards allowed contributed to the performance of the New England Patriots more than any other team. The teams that appeared in Figure 2 are largely teams that experienced either large amounts of success (such as the New England Patriots or Kansas City Chiefs) or failure (Cleveland Browns and Jacksonville Jaguars), which proved to be a nice sanity check for our team. It was no surprise that the teams that best fit the model were both annual Super Bowl favorites, and teams that often finished in last place. This shows that the model was successful in finding trends that contributed to both success and failure.

5. The final step in our analysis was the visualization of our data. The organization of our data allows us to group any combination of teams desired, and plot against any given statistic. For example, if we wanted to validate our performance factor by checking it against the performance of NFC North teams, Figure 3 can be easily generated for analysis.

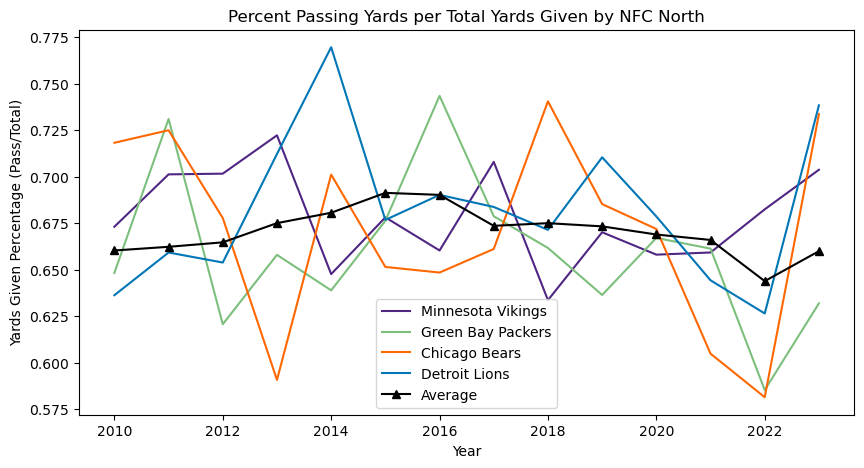


Figure 3: Passing Yards Ratio vs. Year for the NFC North

For further analysis, interactive plots were generated so that each available statistic can be viewed for each team. These plots were generated by the five most important statistics per Figure 1. For example, a plot of offensive pass yards for the Kansas City Chiefs would show the following:

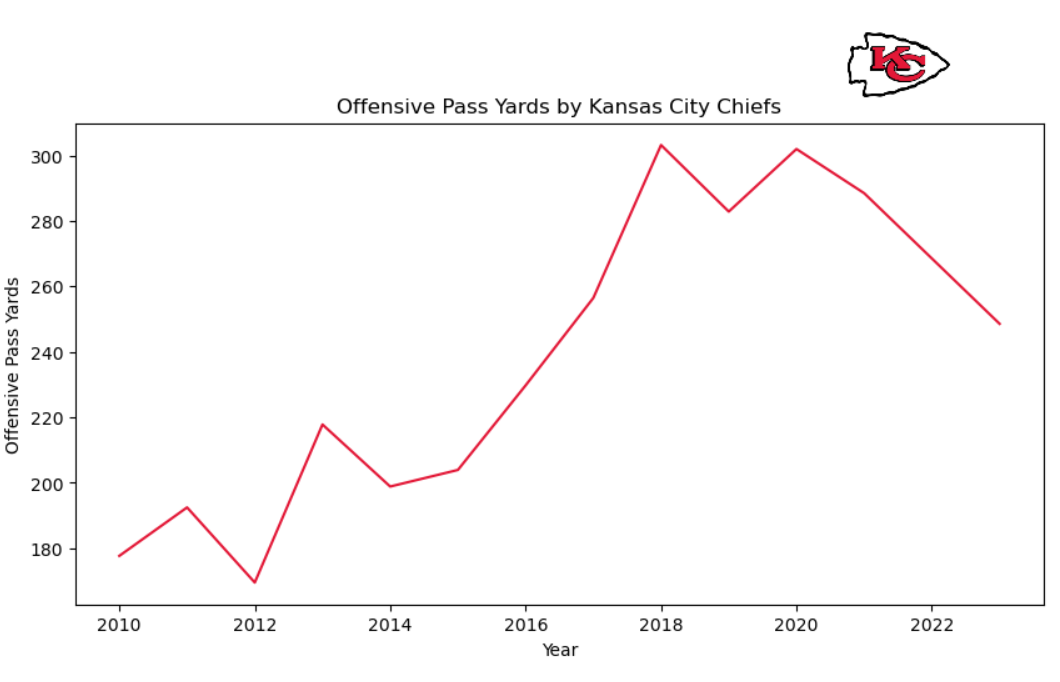


Figure 4: Example Output of Interactive Individual Plot

This allows the team to dive deeper into how the teams performed for the statistics identified and draw conclusions of what statistics were most important for a given team.

6. In summary, the team was able to gather significant quantities of data from reputable online sources, clean and sort the data, define performance metrics and apply these to machine learning algorithms to get results which can be analyzed visually through several bar or line plots. While this kind of analysis hardly scratches the surface of what is possible in the field of professional sports data, it was a useful exercise for the team to apply concepts learned throughout OA3801 and beyond.