National Football Leauge(NFL) Analysis

(COMP3125 Individual Project)

Nicholas McGourthy

# Introduction

The increasing integration of analytics in professional sports has transformed how teams approach strategy and decision-making. The National Football League's (NFL) partnership with Amazon Web Services (AWS) has particularly advanced this trend. Providing sophisticated metrics such as touchdown probability and expected yards at the play level, there is a proven opportunity for analysis to give team’s leverage. My research aims to analyze play-by-play data to identify statistical trends that correlate with team success both in the current game and season long.

Research questions:

1. Does early down efficiency lead to team wins?
   1. Early downs are 1st and 2nd down plays.
   2. ‘Efficient’ 1st down plays gain 3+ yards.
   3. ‘Efficient’ 2nd down plays gain 50% + the yards to gain for a first down.
2. Does 1st quarter explosiveness translate to offensive success in the remaining quarters of the game?
   1. Explosive plays are plays gaining 10+ yards.
   2. To determine success in remaining quarters, I will look at the total yards gained in those quarters.
3. Does a previous drive influence scoring probability of the following drive?
   1. A drive is the consecutive plays a team runs before giving the ball to the other team.
   2. For each drive, the team has 4 plays to gain 10 yards and if they gain those 10 yards they get another 4 plays, this is called a first down.
   3. Factors that could influence the following drive is: number of first downs in a drive, time of possession(length of time in seconds), play count(number of plays ran in the drive) and if the drive ended in a score.
4. How does a team’s efficiency look in the 4th quarter and how does this correlate to winning.
   1. 4th quarter is the final 15 minutes of a game, we are trying to determine if teams finish strong and this correlation to winning.
   2. Efficiency will be a single metric combining early down efficiency we used in question 1 along with 3rd and 4th down efficiency which is simply if the play gains enough yards for a first down.
5. Teams will ‘script’ there first 15 plays of a game meaning they know exactly what plays they will run instead of choosing a play depending on the game, how does the first 15 plays compare to the rest of the game?
   1. I will compare the efficiency of plays using the same efficiency characteristics as questions 1 and 4.
   2. I will also compare mean yards gained for scripted vs non scripted plays.

# Datasets

## Source of dataset

The dataset used for analysis is from the GitHub repository NFL-Verse [1].

## Character of the datasets

The dataset contains play by play data for NFL seasons, my analysis utilized the 2023 and the current 2024 season data. Both datasets are in CSV format and contain 374 columns.

## Feature Engineering

Preliminary feature engineering utilized Pandas[2] included calculating a team record, I used the columns home\_team, away\_team, home\_score and away\_score. I created a dictionary where the key’s are teams and the value is a array indicating [wins, losses, ties]. I then iterated through each game, determined the home and away team’s then compared home\_score vs away\_score which is the final scores of the game to determine the winning side.

# Methodology

## Early Down Efficiency

To begin I needed to calculate efficiency, I created a dictionary for teams where the key is team name and value is an array. The array is in the format [total 1st and 2nd down plays, efficient 1st and 2nd down plays]. I iterated through each team’s offensive plays and updated the array accordingly.

Using the array I then calculated a efficiency metric diving index 1 by index 0 for each array and multiplying by 100 to get a %.

The model I used was a linear regression model to calculate the value. I used Numpy[3] to do so. I choose because can explain variance of data. Essentially this model will explain the variance/correlation of wins based on early down efficiency.

## Early Game Explosivness

Like early down efficiency, there is some preliminary calculation needed to be done. I utilized a dictionary again but this time in a different structure. The key is the number of explosive plays (10+ yards gained) and the value is an array containing the yards gained in the remaining quarters of the games.

To do this I iterated through each team’s games and created two dataframes for each game.

Dataframe 1: Plays in the first quarter. [col ‘qtr’ == 1].

Dataframe 2: Plays in the remaining quarters. [col ‘qtr’ != 1].

With these data frames I found the sum of plays gaining 10+ yards in the first quarter dataframe using the yards\_gained column. I then found the sum of yards gained in Dataframe 2.

The dictionary was updated accordingly appending the yards gained in the remaining quarters where the key is the number of explosive plays in the first quarter.

I utilized a linear regression model with Numpy[2] again to calculate the to explain if there is any correlation between first quarter explosiveness and yards gained in the remaining parts of the game.

## Strong Drive

Much of the columns already exist in the dataset with the columns: drive\_ended\_with\_score, drive\_first\_downs, drive\_time\_of\_possession, and fixed\_drive(drive #).

The feature engineering needed to be done was adding a column for next drive score(next\_drive\_score). The value for this will be Boolean. Another piece of feature engineering was adding a unique drive\_id with is a concatenation of game\_id, offense team, and drive number.

I will use a logistic regression utilizing Scikit-learn[4] where the output is next\_drive\_score and I will be using the columns drive\_first\_downs, drive\_time\_of\_possession, drive\_ended\_with\_score and drive\_play\_count.

To prevent overfitting, I will utilize the unique drive\_id to ensure only one instance of each drive is used in the prediction.

## Fourth Quarter Efficiency

For this much of the code from question 1 could be reused. The only addition was to include 3rd and 4th downplays and compare the yards\_gained column and if the value is greater than or equal to yds\_to\_go column.

For this I will also use linear regression with Numpy[2] like question 1 to calculate a to determine if there is correlation between a teams wins and there fourth quarter efficiency.

The idea behind this is in football if teams are efficient late in games, that means if they are wining they will continue to make the correct plays and gain yards to seal the game and win. If they are not efficient down the stretch then these leads can quickly be erased and there record will show this.

## Scripted Play Performance

For feature engineering, there is a column in the dataset called play\_id, but this is not what I needed.   
  
For feature engineering I created a dictionary for each game where the keys are the two teams, and the values are the play number. I then iterated through the plays and if the offensive team runs a valid play, I add the column game\_play\_num and increment the value of the corresponding dictionary. A valid play is a run or pass play without a penalty.

Once the game\_play\_num was added, I could then query game plays by this and be able to identify the first 15 plays versus the rest.   
  
I utilized similar strategies to efficiency where I calculated scripted efficiency vs rest of game efficiency. I also calculated the mean yards gained for both scripted and non scripted plays.

I will use t-tests to compare the two samples and utilize the p-value to determine if the difference occurred by chance. I utilize Skikit-Learn[4] to do this.

# Results

## Early Down Efficiency

Reference charts [1.1 & 1.2] in the GitHub repository.

2023 Season : 0.488

2024 Season : 0.221

The of the 2023 season describes a strong correlation between wins and early down efficiency. The value of 0.488 means approximately 48.8% of the variance in wins can be explained by early down efficiency.

In football, this is expected as teams who get positive plays early can choose a variety of plays as the yards to gain gets smaller. This offensive flexibility makes them more unpredictable and difficult for defenses to stop, leading to more points and wins.

The drop in the 2024 season to a of 0.221 means about only 22% of the variance in wins can be described by early down efficiency. This could be explained as the 2024 season is still ongoing meaning the sample size is smaller as opposed to the 2023 season.

Furthermore, there could be a shift in ideology in the NFL putting more emphasis on defenses stopping early downs.

## Early Game Explosiveness

Reference charts [2.1 & 2.2] in the GitHub repository.

2023 Season = 0.062

2024 Season = 0.064

Both are extremely weak and give reason to believe there is not a strong correlation in early game success projecting game long performance.

This can be explained for some reasons.

Reason 1: Explosive plays often result from defensive breakdowns such as missed tackles or coverage errors. These specific weaknesses are quick to be addressed by coaching staffs to ensure the correct adjustments are made to fix these issues.

Reason 2: Linear regression models are sensitive to outliers and since explosive plays are relatively rare events, games with high numbers can disproportionately influence the correlation creating this difficulty in the to explain variance.

Reason 3: There is a term called game script in football that influences offensive strategy which may explain outliers. Teams with many explosive plays often get out to points lead and be more conservative with play calling resulting in less yards. On the flip side a team with few explosive plays may be very aggressive and attempt to gain many yards to catch up to a lead.

## Strong Drive

2023

|  |  |  |
| --- | --- | --- |
| COLUMN | Coefficient | Odds Ratio |
| Drive\_first\_downs | -0.046 | 0.955 |
| drive\_time\_seconds | 0.002 | 1.002 |
| drive\_play\_count | -0.074 | 0.928 |
| Drive\_ended\_with\_score | 0.154 | 1.116 |

2024

|  |  |  |
| --- | --- | --- |
| COLUMN | Coefficient | Odds Ratio |
| Drive\_first\_downs | 0.175 | 1.191 |
| drive\_time\_seconds | 0.004 | 1.004 |
| drive\_play\_count | -0.205 | 0.814 |
| Drive\_ended\_with\_score | 0.131 | 1.140 |

Many of the coefficients are very small making them hard to interpret but the odds ratio is more interpretable.

For drive\_first\_downs in 2024 the odds ratio of 1.191 represents an increase of odds of about 19% with every first down in the previous drive to score on the next. This is significant however goes against the column drive\_play\_count as more plays decrease the odds.

More plays likely mean more first downs however this decrease could be due to offense being on the field longer and getting tired.

We see a negligible increase for both years for drive\_time\_seconds meaning the longer a drive lasts really does not affect scoring on the next drive.

Finally with drive\_ended\_with\_score, we see a clear positive impact for both years and this makes sense, this creates momentum. If a offense scores they build confidence in themselves and teamamtes while defenses lose this confidence.

## 4th Quarter team efficiency.

Reference charts [4.1 & 4.2] in the GitHub repository.

2023 season = 0.52

2024 season = 0.308

For the 2023 season we can explain roughly 52% of wins by looking at teams efficiency in the fourth quarter. This is explainable as I mentioned in part II that if teams are efficient late in the game, they will seal victories where as inefficiency can cause leads to be erased and generate losses.

For 2024 the discrepancy drop to 30.8% of the variance being explained by late game efficiency is interesting. There are explanations and a reiterating theme could be the lack of sample size. As mentioned previously the 2024 season is still ongoing and there are less games that have been played meaning less plays have been run in the 4th quarter to really solidify these efficiency numbers.

Another reason for the discrepancy could be that in 2024, is looking at the chart we see that teams are inefficient but still have a high win total. This could be due to there defenses winning games for them down the stretch while the offense struggles to generate offense. Or furthermore the offenses could be very efficient in the beginning of the game and create a lead and they in a sense coast to a victory.

## Scripted Plays

Efficiency

2024 : T-stat = 4.323 and p-value = 0.000148

2023 : T-stat = 4.033 and p-value = 0.000333

Yards gained

2024 : T-stat = 0.887 and p-value = 0.381

2023 : T-stat = 1.739and p-value = 0.092

For efficiency both years show a very strong p-value < 0.05 with high t-stats indicating that efficiency is usually much higher with these plays as compared to the rest of the game.

However with average yards neither year has a significant p-value and the t-stat is much lower.

This can likely be explained as teams want to get going in both the passing and rushing attack but also just want positive plays, they do not intend to have massive explosive plays which is why we see the efficiency show strong correlation but average yards gained not so much.

# Discussion

For this analysis there is interesting analysis made however there is a wide array of outside factors that can and likely have affected the research.   
  
For sports in general the difficulty is that it is humans playing and human error plays a large factor in the game. There are factors like home fields and weather that can add onto this, take for example a home field where an away quarterback is having difficulty’s hearing his coach because the stadium is too loud, how do we think efficiency of an offense will compare at that situation then a quiet home game?

Future research in the NFL and sports across the globe are continuing to grow exponentially, take for example recent years Amazon signed a massive partnership with the NFL to stream all Thursday games on Amazon Prime. This is to not only show off prime deals but also to show there AWS Next Gen Stats which is a huge analytical breakthrough in the NFL.

# Conclusion

For my research it was a very interesting experience as I was able to take my passion and love for football and utilize my technical experience to see from the statistical level what makes a team good.

Efficiency is something measured in all areas of life and is a large prominent of my research and much of my findings relate with what coaches and players across the globe are trying to focus on to win more games.

If teams get ahead in early downs, the playbook opens up for what they want to do and really puts the defense on edge. This is shown by my linear regression model that provides a clear demonstration of the correlation between efficiency in early down with winning.

The most interesting part of this project to me was the strong drive section (question 3). What I took away was the column drive\_play\_count actually plays a negative effect in the chances of scoring on the next drive. This could be due to offensive players being tired, plays being run giving the defense an edge into the mindset of the coach, or really any reason to think of and hopefully future research can create some concrete definitions to why this is the case.

##### References

[1] “nflverse,” *GitHub*. https://github.com/nflverse

‌[2] Pandas, “Python Data Analysis Library — pandas: Python Data Analysis Library,” *Pydata.org*, 2019. https://pandas.pydata.org

‌[3] “NumPy — NumPy,” *Numpy.org*, 2009. https://numpy.org

‌ [4] Scikit-learn, “scikit-learn: Machine Learning in Python,” *Scikit-learn.org*. https://scikit-learn.org/stable/

‌