Win-Loss Percentage in the NFL: Understanding Variables Within the NFL

Morgan Thomas, Micah Holmquist, Brayden Tyler

College of Business, Northern Michigan University

CIS 422: Data Mining

Prof. Jaeseung Baek

April 13, 2024

Introduction

The National Football League (NFL) is the most viewed sport within America. Few institutions throughout the world hold as much cultural significance and widespread appeal as the National Football League. The NFL was created in 1920, previously known as the Professional Football Association. Since 1920, the NFL has evolved into a juggernaut of entertainment. Florio (2023), states that the NFL brought in an approximate national revenue of 11.98 billion dollars. This in itself shows that the NFL stands as the pinnacle of professional football, and allows its athletes to showcase their skills in front of millions of people worldwide.

Throughout the years, teams have looked for every advantage possible in order to increase their chances of winning against their opponent. In order for a team to be successful in the NFL, a team must understand the skillset(s) that is needed to best fit the current and new players of a team. For example, if a team is having tremendous success passing the ball, but lacks performance when running the ball, would it be more beneficial in drafting or trading for another top wide-receiver to capitalize on a teams passing strength, or would it be more beneficial to draft or trade for players that would enhance a teams rushing game? In order to determine this, analysts must understand the statistical data that goes into achieving a win in an NFL game.

A study done in September of 2023 shows a predictive analysis model that forecasts outcomes within the NFL using decision tree and logistic regression. Within the article, Gifford (2023), states that turnovers are correlated to winning. Gifford goes on to explain that 44% of the variation in a team's win percentage was determined by the team's turnover differential. This data shows that turnovers play a crucial part when it comes to winning within the NFL, and also shows that NFL analysts likely take this into account when determining who should be traded and picked for each team.

Additionally, another study was also conducted in March of 2022 which also analyzed several other variables which impacted the prediction of an NFL team's win/loss ratio. It was concluded that ("Using Data and Analytics…," 2022) a team who wins the coin toss in overtime, has a higher chance of winning the game by roughly 15%. This is due to rules in overtime that prevents the opponent from obtaining the ball if the team that won the coin toss in overtime elects to receive the ball first, and then scores a touchdown. The NFL saw a huge backlash of this in 2022 in an intense NFL divisional round playoff game between the Buffalo Bills and the Kansas City Chiefs. This specific game was tied at the end of regulation, and therefore was forced to go into overtime. Due to rules in place at the time, whoever received the ball first had a higher chance of winning due to the fact all they needed to do was score a touchdown, and the opposing team wouldn't get the same chance. Kansas City won the coin toss and elected to receive the ball first, and ended up scoring a touchdown, thus resulting in an automatic win for the Chiefs in the playoffs. The NFL has since changed the rule for playoff games only, which gives both teams an opportunity with the ball.

Furthermore, with the advancements within technology and analytics, teams today are now able to access a wide variety of data that encompasses player performance metrics, game situations, and opponent tendencies. For example, ("2023 NFL Offense…," 2024) the Tampa Bay Buccaneers had the worst rushing offense in the 2023 NFL season. The Buccaneers were only able to rush for an average of 88.8 yards per game, compared to the Baltimore Ravens who had the best rushing offense, averaging 156.5 rushing yards per game. Analyzing data such as this allows a team like the Tampa Bay Buccaneers to properly analyze players on their team, and determine what changes need to be made. Vise versa for the Baltimore Ravens, they can also determine what rushing plays are the most successful for them, and what teams rushing works

best against. Being able to understand and leverage these complex datasets, coaches and analysts can identify strategic insights and patterns that can then be used when determining how to play against a certain opponent.

The win/loss ratio also plays a significant role in the realm of sports betting. For both professional sports bettors, and casual fans alike, analyzing the win/loss ratio and the variables that go into a teams win or loss, is essential for making an informed betting decision and predicting the outcome of a game. By examining historical statistics on a team's tendencies and patterns, sports bettors can identify trends that may influence the outcome of an NFL game, and can gain an advantage when determining what bet to place.

In conclusion, the ability to utilize statistical data and properly analyze the data can drastically improve not only a team's win/loss ratio, but can also prove to be useful for sports bettors. When analyzing these datasets, NFL teams can gain valuable insights on not only their team, but also their opponents, and sports bettors can aim to capitalize on their earnings. The aim for this project is to dive deeper into understanding these variables and extract meaningful information that can help accurately predict a team's win/loss ratio.

Data Description

Data can be either numerical or categorical. Categorical data is data in which the output is not numbers, but rather words or categories. Numerical data is data that has an output containing numbers. All variables used in this study are numerical. Numerical variables can then be further categorized as continuous or discrete. Continuous variables are variables in which the output is measured, and the outcome can be a decimal or an integer. Discrete variables are variables in which the output can be counted. The outcome of these variables is always an integer. All

variables analyzed in this study are categorized as continuous variables, as all variable outputs are either measurements or averages. The data collected also only includes regular season games, not playoff games. The dependent variable in this study is win-loss percentage. This is the percentage of individual games won over the regular season and is categorized as continuous. This study focuses on which variables allow a team to get the highest win-loss percentage possible.

There are 22 independent variables of interest in this study, which will determine how to earn the highest chance of winning a football game. The average number of points scored per game is the calculation of the average number of points a team scored in the normal season. Points opp, or points against, is a variable that measures the average number of points a team allowed another team to score against them in the regular season. Total yards per game is a variable that shows the average number of offensive yards completed per game by a team in the regular season. Offensive plays per game measures the average number of plays ran while the team was on offense. Offensive yards per play calculates the average number of yards gained per game while on offense. These calculations include plays in which the ball was passed or ran. Turnovers per game calculates the average number of times the ball was turned over in a game and the team lost control of the ball. Fumbles lost per game shows the average number of times the team fumbled the ball and did not recover it, losing possession. First downs per game is a variable that shows the average number of times a team successfully converted to first down in a game. Passes completed shows the average number of successful passes completed on offense in the normal season. Pass attempts per game shows how many times the quarterback attempts to pass the ball. These pass attempts are recorded whether or not the pass was completed or not. This also includes passes that result in turnovers. Passing yards per game shows the average

number of passing yards the team completed per game. Passing touchdowns shows the average number of touchdowns a team scored in a game as a result of a completed pass. This variable does not include passes that were intercepted that resulted in a touchdown. Passing net yards per game shows the average number of yards gained per pass attempt. Passing first downs per game is a variable that shows the average number of completed passes made that resulted in a first down. Rushing attempts is a variable that shows the average number of times a team attempted to rush, or run the ball. While rushing usually means a team ran the ball, any play that does not involve a forward pass is classified as a rushing play. Rushing touchdowns shows the average number of touchdowns scored after a rushing play. Rushing yards per attempt shows the average number of yards gained during rushing plays. Any plays in which the quarterback is sacked do not contribute in a negative way towards the total number of rushing yards. Rushing first downs shows the average number of times that a rushing play resulted in a first down. Penalties per game is a variable that shows the average number of penalties against a team in a game. Penalty yards per game shows the average number of yards lost by a team as a result of a penalty per game. Score percentage is the percentage of total runs that result in a score, whether this be by touchdown or field goal in a game. Turnover percentage shows the percentage of plays that result in a turnover, or the offensive team losing possession of the ball. The figure below shows an example of what the data set of interest includes, and an example of some of the data used in this study. The second figure shows the remaining columns of the data set.

year	team	win_loss_perc	points scored per game	points_opp_pe_rgame	total_yards_per_game	plays_offense_per_game	yds_per_play_offense	turnover_per_game	fumbles_lost_per_game	first_downs_per_game	pass_cmp_per_game	pass_att_per_game	pass_yds_per_game
2003	New England Patriots	0.875	21.75	14.875	314.938	65.125	4.8	1.5	0.6875	18.375	20	33.5625	214.5
2003	Miami Dolphins	0.625	19.4375	16.3125	288.063	60.5	4.8	2.125	0.9375	16.625	16.0625	28.125	174.5
2003	Buffalo Bills	0.375	15.1875	17.4375	271.75	61.25	4.4	2.125	1.0625	16.75	18.3125	31.375	167.75
2003	New York Jets	0.375	17.6875	18.6875	309.438	58.5	5.3	1.25	0.375	17.125	19.5	31	207.25
2003	Baltimore Ravens	0.625	24.4375	17.5625	308.063	63.0625	4.9	2.375	1.1875	16.1875	13.5625	25.9375	140.938
2003	Cincinnati Bengals	0.5	21.625	24	333.063	64.875	5.1	1.375	0.4375	19.5625	20.25	32.5	208.875
2003	Pittsburgh	0.375	18.75	20.4375	299.5	63.75	4.7	1.75	0.6875	17.1875	19.125	33.25	206.5
2003	Cleveland Browns	0.313	15.875	20.125	281.5	60.0625	4.7	2.0625	0.9375	17.25	19.5625	31.8125	177.125
2003	Indianapolis Colts	0.75	27.9375	21	367.125	65.0625	5.6	1.25	0.625	21.75	23.8125	35.5625	261.188

pass_td_per_game	pass_int_per_game	pass_net_yds_per_att	pass_fd_per_game	rush_att_per_game	rush_yards_per_game	rush_td_per_game	rush_yds_per_att	rush_fd_per_game	peanalties_per_game	penalties_yds_per_game	pen_fd_per_game	score_pct	turnover_pct
1.4375	0.8125	6	11.0625	29.5625	100.438	0.5625	3.4	5.6875	6.9375	62.375	1.625	27.9	11.3
1.0625	1.1875	5.8	9.0625	30.4375	113.563	0.875	3.7	6.1875	6.4375	57.0625	1.375	28.1	17.2
0.6875	1.0625	4.9	9.375	26.6875	104	0.8125	3.9	6	6.625	55.6875	1.375	21.9	17.6
1.25	0.875	6.3	11.3125	25.5625	102.188	0.5	4	4.875	4.3125	34.375	0.9375	32.4	11.8
1	1.1875	4.9	7.5625	34.5	167.125	1.125	4.8	7.1875	7.875	60.625	1.4375	31.8	16.6
1.625	0.9375	6	11.3125	30.0625	124.188	0.75	4.1	6.3125	6.6875	52.875	1.9375	33.3	11.1
1.1875	1.0625	5.8	10.875	27.875	93	0.625	3.3	4.8125	6.9375	62.8125	1.5	27.1	13.5
1.0625	1.125	5.2	9.5625	25.75	104.375	0.5	4.1	5.6875	6.125	47.9375	2	26.7	17.2
1.8125	0.625	7.1	13.25	28.3125	105.938	1	3.7	6.5	5.75	41.375	2	46.3	10.2

Pre-Analysis

This project focuses on team statistics that are correlated with winning percentages. For this dataset, it is likely the most interesting variables to pay attention to are offensive plays per game (plays_offense_per_game), turnovers per game (turnovers_per_game), first downs per games (first_downs_per_game), average number of penalties per game (penalties_per_game), average number of yards per game that a team is penalized for (penalties_yds_per_game),and the percentage of a team's offensive drives that end in a turnover (turnover_pct).

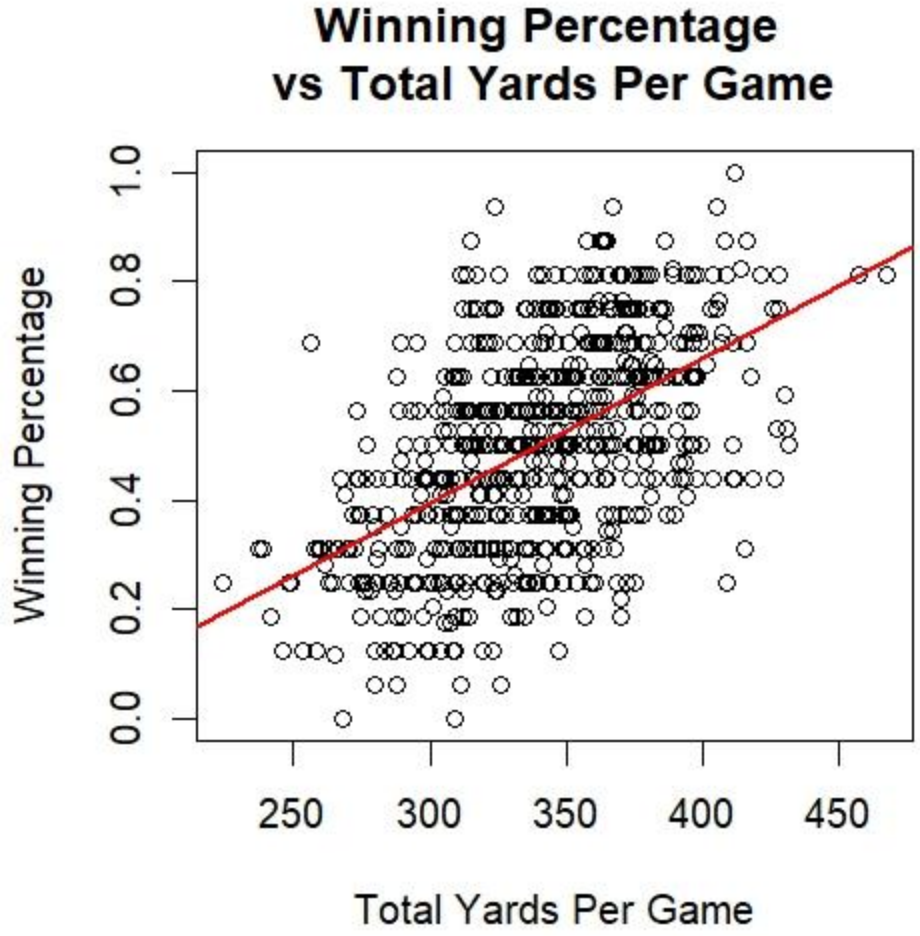


Fig. 1. Total Yards Per Game.

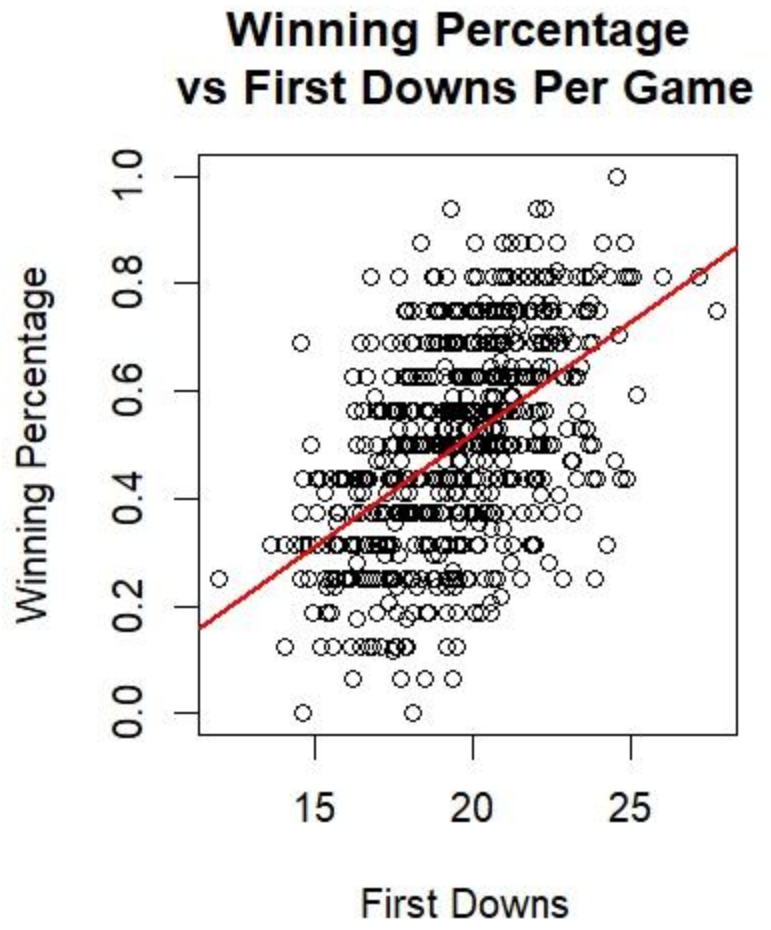


Fig. 2. First Downs.

There is a strong correlation between a team's average number of offensive yards per game and their winning percentage (Figure 1). Based on the single variable regression, each additional average yard per game is correlated with an increase in their winning percentage of 0.3% and this is significant at the less than 2e-16 level, which indicates that there is virtually no chance the relationship occurred by chance. (All significance levels discussed will be based on T-tests unless otherwise noted.) Something similar occurs with the average number of first downs per game (Figure 2), where the coefficient is 0.042 and the significance is less than 2e-16.

Nothing discussed so far is counterintuitive, as first downs allow teams to have additional offensive plays and additional offensive plays mean the other team is unable to score, save for situations involving turnovers. And so, it is notable that the results of a regression with both independent variables show that only total yards per game variable is statistically significant. (In this situation, total yards per game has a coefficient of 0.011 and is significant at the 5.88e-07 level.) From these results, it is safe to predict that yards per game will also play a more sizable role in the larger regression.

Still, despite these levels of statistical significance, it is notable that the R Squared values are only 0.2876 for the total yard per game regression, 0.2639 for the first downs per game

regression, and 0.2898 for the bivariable regression.

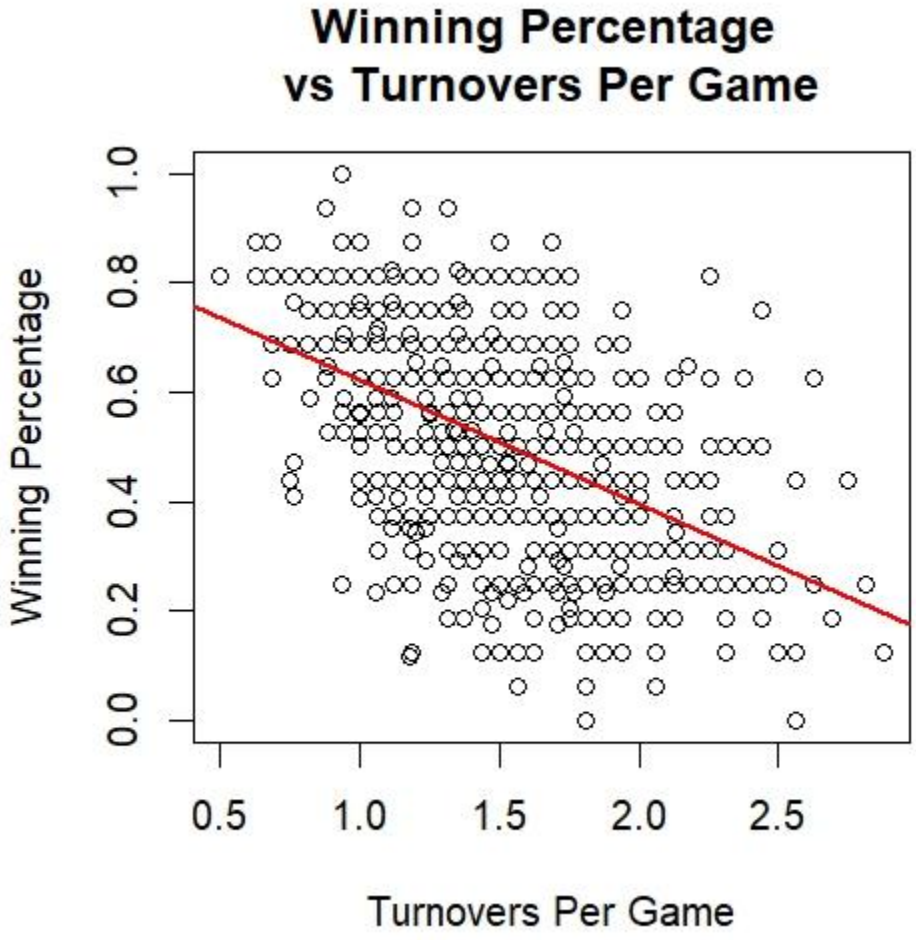


Figure 3

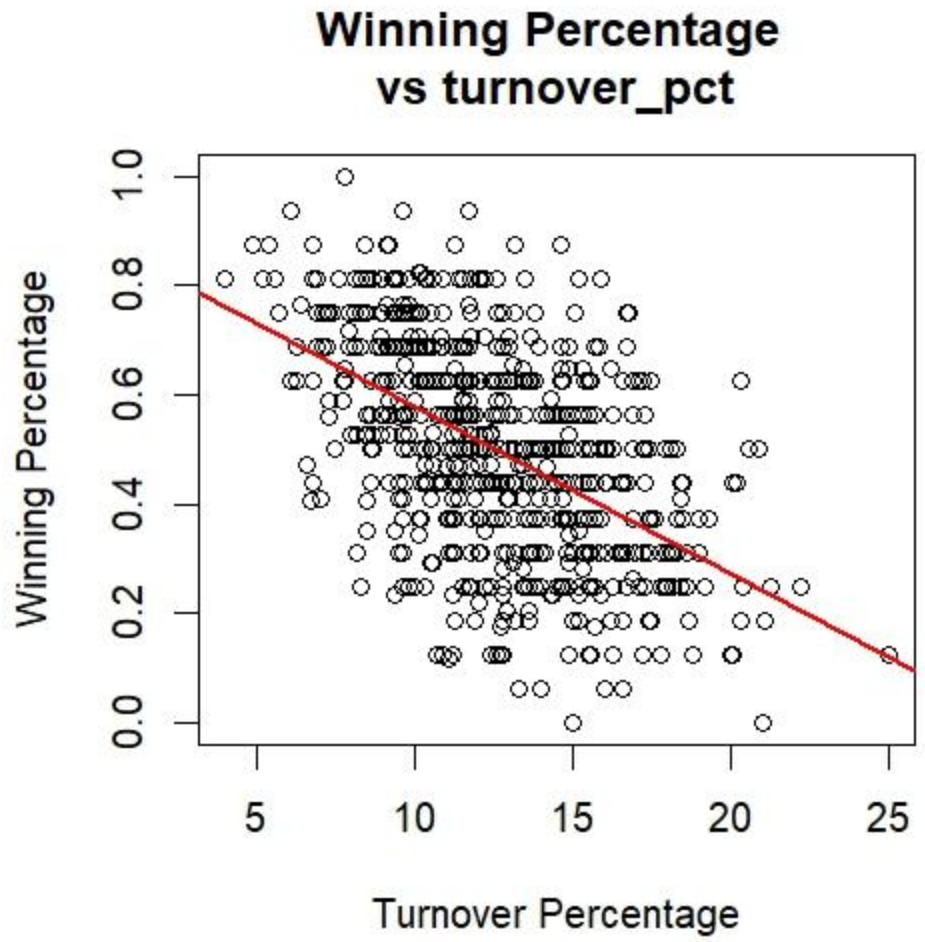


Figure 4

As seen in Figures 3 and 4, there is, when doing a single variable regression, there is a negative correlation between both the average number of turnovers a team commits in a game and the percentage of offensive drives that end in a turnover. The coefficient for the average number of turnovers per game is -0.227 with a significance of less than 2e-16. The coefficient for turnover_pct is -0.031 with a significance of, again, less than 2e-16.

When a bivariate regression is run, the coefficient for the average number of turnovers per game, -0.014, is statistically insignificant, while the coefficient for the percentage of

offensive drives that end in a turnover, -0.029, is significant at the 4.87e-05 level. The R Squared value does not rise above 0.2604 in any of these models.

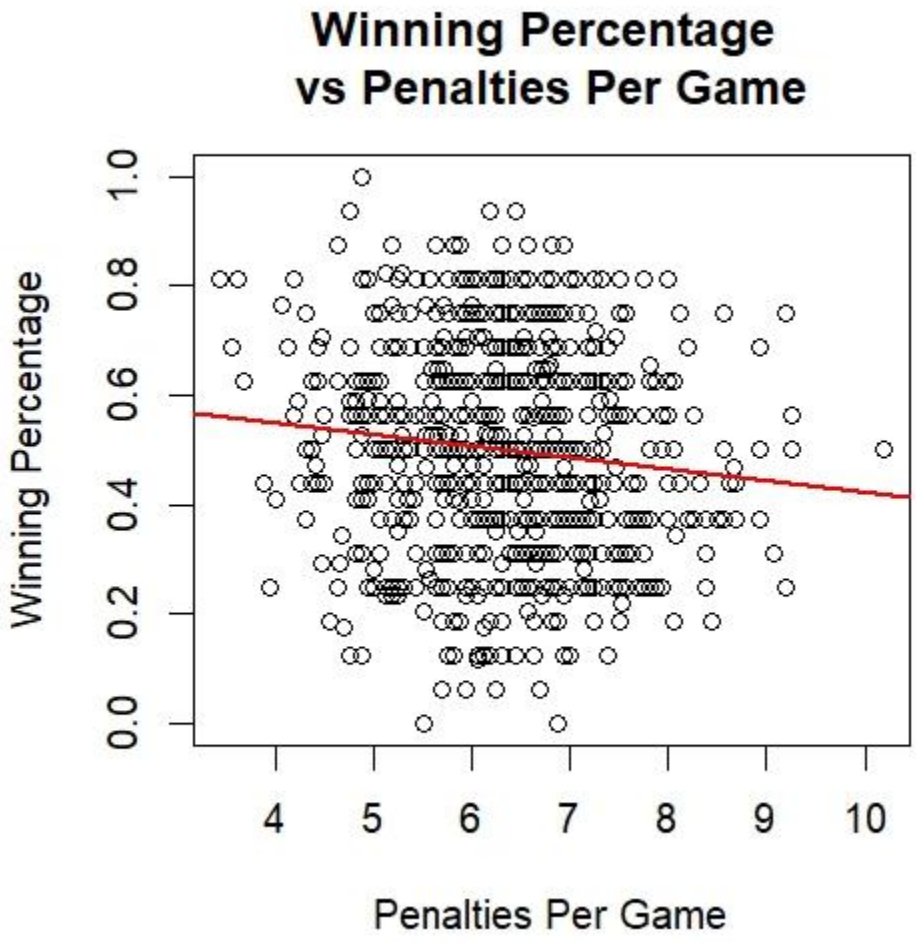


Figure 5

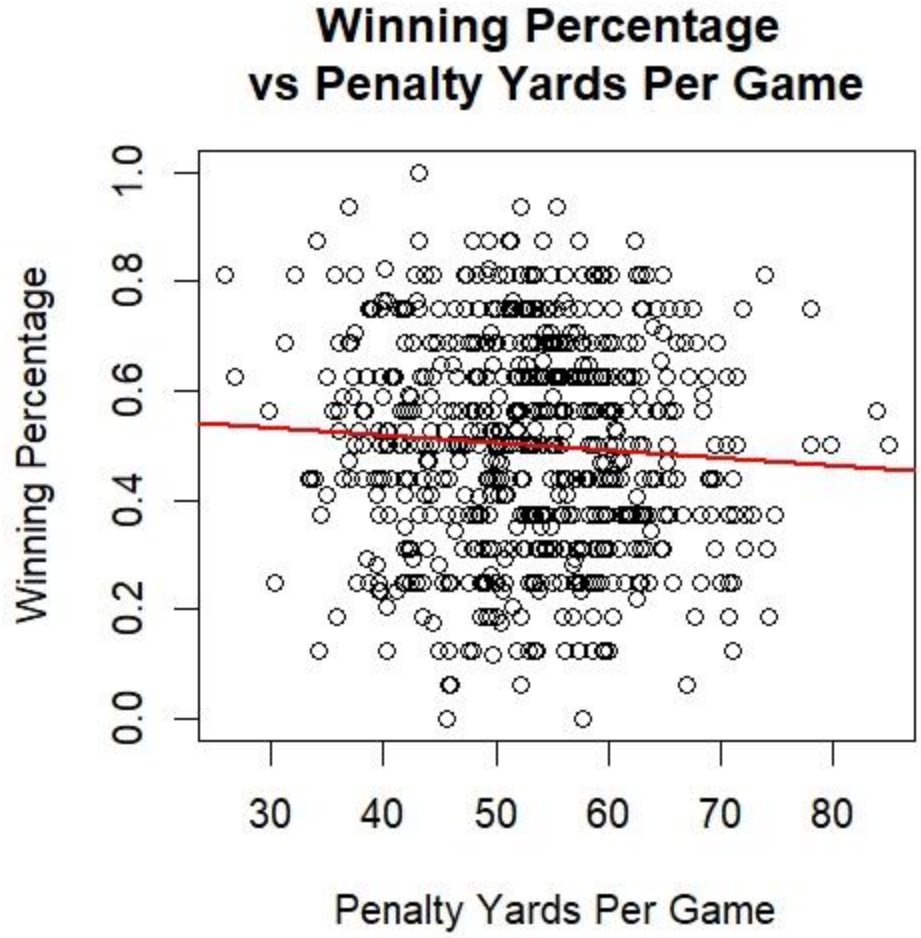


Figure 6

It turns out, intuitively enough since committing penalties hurts football teams, that there is a negative correlation between both the average number of penalties committed per game (Figure 5) and the average number of yards a team is penalized for per game and the team's winning percentage (Figure 6). Using a simple regression, the average number of penalties per game has a coefficient of -0.0212 and that is significant at the 0.0032 level. For the average number of yards a team is penalized for per game, these numbers are -0.0014 and 0.0861, respectively.

When a multivariate regression is done, the coefficient for the average number of penalties per game is -0.0579 and the significance level is 0.0008. The coefficient is 0.0045 and the significant level is 0.0195 for the average number of yards a team is penalized for. The positive correlation between penalty yards and winning percentage has no immediately clear explanation. Of course, the R squared value for this regression is only 0.02, and in none of the models involving penalties discussed here does R Squared rise above 0.02. This suggests that, on their own, penalties are not particularly helpful in predicting winning percentages.

Methodologies for Data Analysis

Multiple Linear Regression

The first model chosen is a Multiple Linear Regression (MLR). MLR is a method of statistics that looks at the relationship between one or more dependent variables and at least two independent variables in order to determine if there is a relationship between the independent variables and the dependent variable(s). The result is an equation where the individual impact of each independent variable on the dependent variable is represented as a coefficient. Once the model is produced, both the certainty of the impact of individual independent variables and the certainty of the model as a whole can be evaluated.

In this case, the single dependent variable of winning percentage winning percentage will be used and the independent variables are the various other statistical categories that have been described earlier. The strength of this model for the question of NFL winning percentages is that the individual statistics only describe part of a team's performance and thus cannot explain how a team will do in terms of wins and losses of their multifaceted games where every position -i.e. quarterback and cornerback- is highly specialized and many starting players will never be in the

game at the same time as another starter. An MLR, however, does have the ability to account for a variety of factors that go into winning football games and weight their impact. In all cases, the null hypothesis is that both the independent variables or the model as a whole do not have an impact on winning percentage while the alternative hypothesis is that they do.

This project will also evaluate the accuracy of the MLR model to predict a team's winning percentage by testing its accuracy against similar data that was not used to create the model. It will do this by using a Median Absolute Deviation (MAD) score. The process of getting a MAD score starts with taking the equation produced by the MLR and calculating what the model predicts, in this case, the winning percentage of team's in the testing data would be. Then the absolute difference between the predicted winning percentage and the actual winning percentage for each team is calculated. Finally, the MAD score is calculated by taking the median of those values. The lower the MAD score -i.e. the smaller the difference between the median predicted winning percentage and the actual winning wining percentage- the more accurate the model is considered to be.

The first step in this process is actually to prepare for the last step and separate the training data that will be used to create the model and the testing data upon which the model will be tested. In this case, 80% of the data was made into training data while the remainder will be used for testing.

The next step was running an MLR on the training data. The results appear in figure 7. The model will be known as the primary model.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.1097875  0.5438191   0.202   0.8401
points_scored_per_game   0.0260789  0.0035325   7.383 6.26e-13 ***
points_opp_per_game     -0.0239024  0.0011634 -20.546  < 2e-16 ***
total_yards_per_game    -0.0016098  0.0025638  -0.628   0.5303
plays_offense_per_game  -0.0113566  0.0103649  -1.096   0.2737
yds_per_play_offense     0.0518813  0.0971817   0.534   0.5937
turnovers_per_game       0.0630891  0.0459168   1.374   0.1700
fumbles_lost_per_game   -0.0071763  0.0235240  -0.305   0.7604
first_downs_per_game     0.0163798  0.0084487   1.939   0.0531 .
pass_cmp_per_game        0.0047769  0.0039506   1.209   0.2272
pass_att_per_game        0.0146055  0.0063360   2.305   0.0216 *
pass_yds_per_game       -0.0004475  0.0023033  -0.194   0.8460
pass_td_per_game        -0.0154067  0.0225656  -0.683   0.4951
pass_net_yds_per_att     0.0392008  0.0426809   0.918   0.3588
pass_fd_per_game        -0.0054855  0.0101855  -0.539   0.5904
rush_att_per_game        0.0195893  0.0124278   1.576   0.1156
rush_td_per_game        -0.0228189  0.0253634  -0.900   0.3687
rush_yds_per_att         0.0020322  0.0545408   0.037   0.9703
rush_fd_per_game        -0.0115109  0.0104794  -1.098   0.2725
penalties_per_game       0.0111739  0.0082166   1.360   0.1744
penalties_yds_per_game  -0.0017052  0.0009156  -1.862   0.0631 .
score_pct               -0.0005566  0.0017177  -0.324   0.7460
turnover_pct            -0.0118577  0.0051874  -2.286   0.0227 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07704 on 516 degrees of freedom
Multiple R-squared:  0.842,     Adjusted R-squared:  0.8353
F-statistic:   125 on 22 and 516 DF,  p-value: < 2.2e-16
```

Figure 7. MLR Primary Model Regression Table.

Taking a team’s results in each of the listed independent variables and plugging them into the equation will result in the model’s prediction of a team’s winning percentage.

The “Adjusted R-squared” score found near the bottom right corner of Figure **7** indicates the percentage of variation in team’s winning percentages that can be explained by the model. The Adjusted R-squared score of roughly 0.835 here means that the overall input model can predict nearly 84% of the variation in the winning percentage of team’s in the testing data. The

p-value score seen in the bottom right is a measure of how likely it is that these results could be achieved if the null hypothesis was true and there was no relationship between the overall model and team winning percentages. A score of nearly 0, such as the one found here, indicates that there is nearly zero chance of these results happening at random.

The statistics on the far right under "Pr(>|t|)") indicate the results of a t-test of a t-test that is used to evaluate how likely the observed results are if the corresponding game statistic does not have an impact on the winning percentage of an NFL team, which is the null hypothesis. The larger the absolute value of the t-value, the less likely it is that the results are by chance. Generally speaking, any probability below 0.05 is considered statistically significant, although the lower the number the more confident one can be.

Four independent variables are statistically significant. Both points scored per game ("points_scored_per_game") and points allowed per game ("points_opp_per_game") have p-values of nearly zero. The coefficient for points scored per game is roughly 0.026, which means that for each additional average point per game a team scores, this model predicts an increase in winning percentage by 2.6%. In the context of a 17 game regular season schedule, this means that the model predicts that scoring 17 additional points over the course of a season results in winning roughly 0.442 games (The product of 17 and 0.026 is 0.442.) Of course, the experience of individual teams is likely to vary. It would, for example, result in two additional victories if a team scored an additional 7 points in a game lost by 3 points and an additional 10 points in a game that was lost by 8. On the other hand, scoring an additional 17 points in a game that a team lost by 28 points would result in no additional victories.

The coefficient for points allowed per game ("points_opp_per_game") is roughly -0.024. For each additional average point per game allowed, the model predicts a drop in the winning percentage by around 2.4%. For a 17 game season, the model predicts that allowing 17 additional points will result in winning roughly 0.408 games. (The product of 0.024 and 17 is 0.408).

Average pass attempts per game ("pass_att_per_game") and the percentage of offense drives that result in a turnover ("turnover_pct ") are also statistically significant, but at the 0.05 level. The coefficient for the average number of pass attempts per game is approximately 0.015, meaning that the model predicts that 17 additional passing attempts over the course of 17 game season will result in approximately 0.255 additional wins per season. (The product of 0.015 and 17 is 0.255). The coefficient for the percentage of offensive possessions that result in a turnover is -0.012. The primary model thus predicts that, over the course of a 17 game season, a 1% rise in the percentage of offensive drives resulting in a turnover will lead to about 0.20 fewer wins. (The product of -0.012 and 17 is 0.204.)

None of the other independent variables are statistically significant at the 0.05 level, although two, average first downs per game ("first_downs_per_game") and average penalty yards per game ("penalties_yds_per_game") do come close. That such independent variables, average total yards per game ("total_yards_per_game" and average turnovers per game (turnovers_per_game), are not significantly significant -the t-score for average total yards per game is 0.5303- seems odd, but is probably the result of including the average number of points scored per game and the average number of points allowed per game in the model. Turnovers reduce points scored and, generally speaking, it is by moving the ball that teams score points.

The MAD score for this model is roughly 0.063. This amounts to a 6.3% difference or, over the course of a 17 game season, about 1.07 games. (The product of 0.063 and 17 is 1.071.)

In order to help determine the impact of removing the average points scored per game and average points allowed per game variables, a MLR was run without those independent variables. The results appear in Figure 8. The model will be known as the no points model.

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.6984499  0.7700049  -0.907 0.364790
total_yards_per_game     -0.0035911  0.0036370  -0.987 0.323929
plays_offense_per_game   -0.0298473  0.0146831  -2.033 0.042585 *
yds_per_play_offense      0.0919699  0.1380279   0.666 0.505507
turnovers_per_game        0.1661078  0.0626599   2.651 0.008272 **
fumbles_lost_per_game    -0.0004194  0.0332927  -0.013 0.989953
first_downs_per_game      0.0041942  0.0119152   0.352 0.724975
pass_cmp_per_game         0.0083557  0.0055688   1.500 0.134109
pass_att_per_game         0.0331827  0.0089274   3.717 0.000224 ***
pass_yds_per_game         0.0005980  0.0032725   0.183 0.855089
pass_td_per_game          0.1035350  0.0216690   4.778 2.31e-06 ***
pass_net_yds_per_att      0.0413983  0.0605504   0.684 0.494470
pass_fd_per_game          0.0014267  0.0144685   0.099 0.921486
rush_att_per_game         0.0637699  0.0174484   3.655 0.000284 ***
rush_td_per_game          0.0959650  0.0267409   3.589 0.000364 ***
rush_yds_per_att         -0.0060184  0.0774968  -0.078 0.938128
rush_fd_per_game         -0.0230018  0.0148447  -1.549 0.121874
penalties_per_game        0.0229506  0.0116309   1.973 0.048999 *
penalties_yds_per_game   -0.0028623  0.0012986  -2.204 0.027957 *
score_pct                 0.0100165  0.0018266   5.484 6.52e-08 ***
turnover_pct             -0.0290426  0.0071162  -4.081 5.19e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1095 on 518 degrees of freedom
Multiple R-squared:  0.6797,    Adjusted R-squared:  0.6673
F-statistic: 54.96 on 20 and 518 DF,  p-value: < 2.2e-16
```

Figure 8

As was the case with the previous model, the P-value for the model as a whole is nearly 0. The Adjusted R-Squared score, however, is lower at 0.6673. And, as expected, a greater number of independent variables are statistically significant in this new model.

The average number of turnovers per game, the average number of pass attempts per game, the average number of passing touchdowns per game, the average number of rushing attempts per game, the average number of rushing touchdowns per game, the percentage of offensive possessions that a team scores on, and the percentage of offensive possessions that result in a turnover are all statistically significant at the 0.01 level. Additionally, the average number of offensive plays per game, the average number of penalties per game, and the average number of penalty yards per game) are all statistically significant at the 0.05 level.

Some of the coefficient results are to be expected, at least in terms of direction. The coefficient for average passing attempts per game is roughly 0.033. The model thus predicts that increasing passing attempts by only 1 per game, on average, would result in roughly a 3.3% increase in winning percentage or, over the course of a 17 game season, winning over 0.5 additional games. (the product of 0.033 and 17 is 0.561.) Scoring, on average, one additional passing touchdown per game increases, according to this model, a team's winning percentage by roughly 10.4%. Over the course of a 17 game season, that would result in over 1.75 additional wins, according to this model. (The product of 0.104 and 17 is 1.768.).

Similarly, the coefficient for the average number of rushing attempts per game is 0.064. Over the course of a 17 game season, 17 additional rushing attempts would be expected to result in nearly 1.1 additional victories. (The product of 0.064 and 17 is 1.088.) The coefficient for average number of rushing touchdowns is approximately 0.096, which means the model predicts

that over the course of a 17 game season an additional 17 rushing touchdowns would result in roughly 1.6 additional wins. (The product of 0.096 and 17 is 1.632.)

The percentage of offensive possessions that a team scores on is positively correlated with victories as the coefficient is roughly 0.01. That means that in a 17 game season the no points model predicts that a 1% increase in the number of offensive possessions that team scores on will result in an additional 0.17 victories. (The product of 0.01 and 17 is 0,17.) The percentage of offensive plays that result in a turnover has a coefficient of roughly -0,029. This means that an increase of 1% in the percentage of offensive plays that result in a turnover would, according to this model, result in losing nearly 0.5 fewer games over the course of a 17 game season. (The product of -0.029 and 17 is -0.493.)

The average number of turnovers per game, however, has a positive coefficient of approximately 0.166 in the no points model. The p-value for the percentage of offensive possessions that result in a turnover (≈5.19e-05) is lower than the corresponding value for the average number of turnovers per game (≈0.008), but both are significant at the 0.01 level. A possible explanation for his seeming paradox could be that more aggressive offensive teams win more games, but also have higher turnovers due to greater risk taking and also have more possessions due to their aggressive manner, which allows for still having a lower percentage of possessions resulting in turnovers.

It is also viscerally puzzling that the coefficient for the average number of offensive plays per game in the no points model is roughly -0.03. In a 17 game season, the model thus predicts that 17 additional offensive plays will result in winning about 0.5 fewer games. (The product of -

0.03 and 17 is -0.51.) Seemingly more offensive plays would result in more points, but as the website Rookie Road explains:

> If a team is ineffective in scoring points while running lots of time off the clock, they have little chance to come back at the end of a game if they are down. Also, if they are used to running the ball heavily, switching to a more up-tempo style of offense to score points in a comeback attempt will be harder than it would be for a team who implements this style all the time. (2022)

Although no statistics are cited by Rookie Road, this could be an explanation for the negative correlation between a team's winning percentage and the average number of offensive plays a team has in a game.

Similar to the average number of turnovers per game and the percentage of offensive possessions that end in a turnover, the coefficient results in the no points model for the average number of penalties per game and the average number of penalty yards per game go in different directions. Both are statistically significant at the 0.05 level with the average number of penalties per game ($p-value: \approx 0.049$) has a coefficient of roughly 0.023 while the average number of penalty yards ($p-value: \approx 0.028$) has a coefficient of roughly -0.003. The model consequently predicts that 17 additional penalties over the course of a 17 game season would increase a team's win total by nearly 0.4 games (the product of 0.023 and 17 is 0.39389) while 17 additional penalties yards over the course of a season of identical

`length` would decrease a teams win total by about 0.05 games (the product of -0.003 and 17 is 0.051). The average number of yards per penalty for all teams in the used data set is roughly 8.373.Over the course of a 17 game season, if a team committed 17 additional penalties with average yardage of 8.373, the result would, according to the model, would be winning nearly 0.04 few games. (The product of 8.373 and -0.003 is -0.025119. The sum of 0.023 and -0.025119 is -0.002119. The product of -0.002119 and 17 is -0.036023.) If a team committed, on average, 1 additional penalty per game with an average penalty length of 5 yards, this model says the net impact would be positive by roughly 0.14 games. (The product 0f 5 and -0.003 is -0.015. The sum of 0.023 and -0.015 is 0.008. The product of 0.008 and 17 is 0.136.) The overall impact of penalties appears small according to the model. The positive correlation between penalties in total and winning percentage and the negative correlation between penalty yards and winning percentage might be explained by team's benefiting from aggression as a penalty like offsides - generally sanctioned with a 5 yard penalty (National Football League, "2023 NFL Rulebook")- would help a team if a player gets ways with it. (Teams would have to figure that the cost of getting away with any such penalties is sometimes being caught.)

The MAD score for the no points model is roughly 0.093, which amounts to a 9.3% median difference or a median difference of, over the course of 17 games, about 1.58 games. (The product of 0.93 and 17 is 1.581.)

Backwards and Forwards Stepwise Regression

Another method for creating a model is to use stepwise regression. Backwards stepwise regression starts with all of the variables in the dataset and tests eliminating independent variables until it is not possible to eliminate additional independent variables without resulting in a statistically significant loss of explanatory power in the model. Conversely, forward stepwise

regression starts with 0 variables and adds independent variables until it is not possible to add more independent variables without reducing the explanatory power of the model. Just as with MLR, the models created by backwards and forwards regression can be used with training data to create MAD scores that can be compared with other MAD scores.

The regression table for the model produced by backwards stepwise regression is in figure 9.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.3135845  0.1139073   2.753 0.006108 **
points_scored_per_game   0.0254872  0.0015582  16.356  < 2e-16 ***
points_opp_per_game     -0.0241328  0.0011377 -21.212  < 2e-16 ***
total_yards_per_game    -0.0009214  0.0003175  -2.902 0.003861 **
plays_offense_per_game  -0.0099818  0.0063096  -1.582 0.114245
pass_att_per_game        0.0145196  0.0059885   2.425 0.015661 *
pass_net_yds_per_att     0.0400760  0.0125713   3.188 0.001518 **
rush_att_per_game        0.0162001  0.0059974   2.701 0.007131 **
penalties_yds_per_game  -0.0005123  0.0003632  -1.411 0.158894
turnover_pct            -0.0044503  0.0012063  -3.689 0.000248 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07677 on 529 degrees of freedom
Multiple R-squared:  0.8391,    Adjusted R-squared:  0.8364
F-statistic: 306.6 on 9 and 529 DF,  p-value: < 2.2e-16
```

Figure 9. Backwards Stepwise Regression Table.

The adjusted R-squared for this model, which will be referred to as backwards model, is 0.8364, while the average number of points scored per game, the average number of points opponents score per game, the average total yards per game, the average number of passing yards per game, the average number of rushing attempts per game, and the percentage of offensive possessions that result in a turnover are all statistically significant at the 0.01 level. The average number of pass attempts per game is statistically significant at the 0.05 level.

The coefficient for the average number of points scored per game is roughly 0.025 per game. This means that the backwards model predicts that 17 additional points over the course of a 17 game season would result in about 0.43 additional wins. (The product of 0.025 and 17 is 0.425). The coefficient for the average number of points scored by opponents per game is roughly -0.024. This means that the backwards model predicts that allowing 17 additional points over the course of a 17 game season would result in winning about 0.41 more games. (The product is of -0.024 and 17 is -0.408.) The fact that absolute value of these two coefficients are very close together is interesting because it indicates that the backwards model is saying that any additional points scored is worth approximately the same number of additional victories as a reduction in the same number of points allowed would be worth.

The coefficient for the average number of total yards per game is interesting because it is roughly -0.0009. The backwards model thus predicts that, over the course of a 17 game season, that 1 additional yard per game would result in a decrease in wins by about -0.0009. The key here is not the direction of the coefficient so much as how close it is to 0, which is the result of the backwards model utilizing only a very small correlation between a team's winning percentage and the team's total number of individual yards. Since the backwards model does utilize the average number of points scored and points allowed, the message is that yards without scoring does not impact a team's success very much or that small numbers of additional yards do make much difference.

The coefficient for the average number of pass attempts per game is roughly 0.015. An increase of 17 pass attempts over the course of a 17 game season would result, according to the backwards model, in about 0.25 more wins. (The product of 0.015 and 17 is 0.255.) The coefficient for the average number of passing yards per game is roughly 0.04. The model thus

predicts that 17 additional passing yards over the course of a 17 game season would result in 0.04 additional wins. The coefficient for the average number of rushing attempts per game is approximately 0.016. The backwards model predicts that 17 additional rushing attempts over a 17 game season would result in a 0.016 additional wind. The coefficient for the percentage of offensive possessions that result in a turnover is roughly -0.004, which means that backwards model predicts that an increase of 1 in the percentage of offensive possessions that result in a turnover would result is -0.004 fewer wins in a 17 game season.

The MAD score for the backwards model is approximately 0.065. In a 17 game season, this means the mean deviation is about 1.1 games. (The product of 0.065 and 17 is 1.105.)

When forward stepwise regression is used, the result is the regression table seen in Figure 10.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.3135845  0.1139073   2.753 0.006108 **
points_scored_per_game   0.0254872  0.0015582  16.356  < 2e-16 ***
points_opp_per_game     -0.0241328  0.0011377 -21.212  < 2e-16 ***
turnover_pct            -0.0044503  0.0012063  -3.689 0.000248 ***
pass_net_yds_per_att     0.0400760  0.0125713   3.188 0.001518 **
penalties_yds_per_game -0.0005123  0.0003632  -1.411 0.158894
rush_att_per_game        0.0162001  0.0059974   2.701 0.007131 **
total_yards_per_game    -0.0009214  0.0003175  -2.902 0.003861 **
pass_att_per_game        0.0145196  0.0059885   2.425 0.015661 *
plays_offense_per_game  -0.0099818  0.0063096  -1.582 0.114245
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07677 on 529 degrees of freedom
Multiple R-squared:  0.8391,    Adjusted R-squared:  0.8364
F-statistic: 306.6 on 9 and 529 DF,  p-value: < 2.2e-16
```

Figure 12. Forward Stepwise Regression Table.

The model created by this table will be referred to as the forward model from this point on. The forward model has an adjusted R-squared value that is identically to the backward model. In fact, all of the independent variables and their coefficients are also identical. The MAD score is also identical.

Because the backwards and forward models both use the average number of points scored per game and the average number of points allowed per game, for the reason described above in the MLR section, it makes sense to eliminate those two independent variables from consideration and then use backwards and forward regression.

Using backwards stepwise regression creates a model that will be referred to as the backwards no point model. The regression table for the backwards no points model is in figure 13.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.4519928  0.1676871  -2.695 0.007255 **
total_yards_per_game    -0.0022193  0.0005571  -3.984 7.75e-05 ***
plays_offense_per_game  -0.0314240  0.0093597  -3.357 0.000844 ***
turnovers_per_game       0.1557098  0.0516634   3.014 0.002704 **
pass_cmp_per_game        0.0088641  0.0052678   1.683 0.093031 .
pass_att_per_game        0.0330732  0.0086149   3.839 0.000139 ***
pass_td_per_game         0.1071972  0.0202735   5.288 1.82e-07 ***
pass_net_yds_per_att     0.0709091  0.0224521   3.158 0.001679 **
rush_att_per_game        0.0592617  0.0089778   6.601 1.00e-10 ***
rush_td_per_game         0.0979127  0.0255878   3.827 0.000146 ***
rush_fd_per_game        -0.0224107  0.0094684  -2.367 0.018300 *
penalties_per_game       0.0221108  0.0115118   1.921 0.055312 .
penalties_yds_per_game  -0.0028326  0.0012864  -2.202 0.028105 *
score_pct                0.0103333  0.0017580   5.878 7.40e-09 ***
turnover_pct            -0.0276426  0.0063170  -4.376 1.46e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1089 on 524 degrees of freedom
Multiple R-squared:  0.6792,    Adjusted R-squared:  0.6706
F-statistic: 79.23 on 14 and 524 DF,  p-value: < 2.2e-16
```

Fig. 13. Backwards No Points Stepwise Regression Table.

The backwards no points model has an adjusted R-squared value of roughly 0.671. The independent variables average total yards per game, average number of offensive plays per game, average number of turnovers per game, the average number pass attempts per game, the average number of passing touchdowns per game, the average number of passing yard per game, the average number of rushing attempts per game, the average number of rushing touchdowns per game, the percentage of offensive possessions that result in scoring, and the percentage of offensive possessions that results in a turnover are all statistically significant at the 0.01 level. The average number of rushing first downs per game and the average number of penalty yards

per game are both significant at the 0.05 level. The remaining independent variables are not statistically significant.

The coefficient for the average number of total yards per game is roughly -0.002, which is surprising since the average points allowed per game and average points scored per game variables have been removed. The coefficient for the number of offensive plays per game that a team has is -0.031, which translates, according to the backward no points model. It is worth noting that the coefficients for both the average number of pass attempts per game (roughly 0.033) and the average number number of rushing attempts per game (roughly 0.059) are both positive and greater than the absolute value of the coefficient for the average number of offensive plays per game, according to the backwards no points model. This means that, on average, any offensive play that results in either a rushing or a passing attempt does have a positive impact on the backwards no points model's estimation of winning percentage.

The coefficient for the average number of turnovers per game is roughly 0.156, which is interesting because it is a positive correlation. This gives some additional support to the idea that turnovers are a necessary part of more aggressive offensive plays.

The coefficients for the average number number of passing touchdowns (approximately 0.107), the average number of rushing touchdowns (approximately 0.098), he average number of passing yards per game (approximately 0.071), the percentage of offensive possessions that results in team scoring (approximately 0.01) are all positive in the backwards no points model, as you would expect. The coefficients are negative for the average number of penalty yards per game (approximately -0.003) and the percentage of offensive possessions that result in a turnover (approximately -0.028) in the backwards no points model, which again is what one would expect.

It does stand out that the coefficient in the backwards no points model for rushing first downs per game (approximately -0.022), although this might be related to the issue of how it is not advantageous to have a relatively slow-moving offense that has trouble scoring.

The MAD score for the backwards no point model is roughly 0.093, which is larger than any of the models discussed so far and amounts to a median adjusted deviation equivalent of about 1.58 games in a 17 game season. (the product of 0.093 and 17 is 1.581.)

Applying forward step regression to the no points model returns the regression table shown in figure 14. The resulting model will be referred to as the forward no points model.

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.4723433  0.1639561  -2.881 0.004127 **
score_pct              0.0105897  0.0017501   6.051 2.74e-09 ***
rush_att_per_game      0.0583502  0.0089811   6.497 1.90e-10 ***
rush_fd_per_game      -0.0223192  0.0094718  -2.356 0.018819 *
pass_td_per_game       0.1071093  0.0203018   5.276 1.93e-07 ***
rush_td_per_game       0.0982103  0.0255635   3.842 0.000137 ***
turnover_pct          -0.0274094  0.0062441  -4.390 1.37e-05 ***
total_yards_per_game  -0.0023539  0.0005478  -4.297 2.06e-05 ***
pass_net_yds_per_att   0.0745256  0.0224455   3.320 0.000962 ***
pass_att_per_game      0.0321631  0.0086257   3.729 0.000213 ***
plays_offense_per_game -0.0303112  0.0093695  -3.235 0.001293 **
turnovers_per_game     0.1595772  0.0511995   3.117 0.001928 **
pass_cmp_per_game      0.0088899  0.0052593   1.690 0.091559 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1092 on 526 degrees of freedom
Multiple R-squared:  0.6762,    Adjusted R-squared:  0.6688
F-statistic: 91.53 on 12 and 526 DF,  p-value: < 2.2e-16
```

Fig. 14. Forward Stepwise No Points Regression.

The forward no points model as an adjusted R-squared value of 0.6688. The percentage of of offensive possessions that the result in a team scoring (coefficient: 0.01), the average number of rushing attempts per game (coefficient: 0.058), the average number of passing touchdowns per game (coefficient: 0.107), the average number of rushing touchdowns per game (coefficient: 0.098), the percentage of offensive plays that result in a turnover (coefficient: -0.027), the average total number of yards per game (coefficient: -0.002), the average number of passing yards per game (coefficient: 0.075), the average number of passing attempts per game (coefficient: 0.032), the average number of offensive plays per game (coefficient: -0.03), and the average number of turnovers per game (coefficient: 0.16) are all statistically significant at the 0.01 level. Additionally, the number of rushing first downs per game (coefficient: -0.022) is statistically significant at the 0.05 level.

Although the coefficients for these statistically significant independent variables are not exactly the same in the forward no points model as they are in the backwards no point model, they are very similar, which explains the very similar adjusted R-squared scores. The MAD value for the forward no points model is roughly 0.092, or nearly identical to the MAD value for the backwards no points model.

K-Nearest Neighbors

One of the methods chosen for attempting to solve the problem is known as k-Nearest Neighbors regression, otherwise known as KNN regression. This method felt like a solid method to use compared to others because of its ability to handle non-linear relationships as well as messy data, which is exactly what this dataset is. KNN regression allows for a straightforward approach when predicting continuous outcomes based on what the values of the neighboring data points are. In order for KNN regression to be accurate, it was crucial that the ID columns in the

dataset (year, team) were removed in order to obtain accurate results. Once these columns are removed, the next step is to split the data into training, validation, and testing data, which was done by splitting it to 80% training data, 10% validation data, and 10% testing data. After splitting the data appropriately, the next step was to define the data for the training, validation, and testing data, as well as defining the data input and outputs.

```
library(FNN)
#Extract Input and Output Data
nfl_Data <- nfl_R[ , c(3:25)] # Remove ID columns
View()

#Split training, validation, and testing data
set.seed(1)      #make sure that we get the same result
idx <- sample(c(1,2,3), size = nrow(nfl_Data), replace = TRUE, prob = c(.8, .1, .1))

train_nfl <- nfl_Data[idx == 1,  ] #define training data for KNN
validation_nfl <- nfl_Data[idx == 2,  ] #define validation data for KNN
test_nfl <- nfl_Data[idx == 3,  ] #define testing data for KNN

train_nfl.X<-train_nfl[  ,c(1:23)] #define training input for KNN
train_nfl.Y<-train_nfl$win_loss_perc #define training output for KNN

validation_nfl.X<-validation_nfl[  ,c(1:23)] #define validation input for KNN
validation_nfl.Y<-validation_nfl$win_loss_perc #define validation output for KNN

test_nfl.X<- test_nfl[  ,c(1:23)] #define testing input for KNN
test_nfl.Y<-test_nfl$win_loss_perc #define testing output for KNN
```

The above code was used to obtain accurate data in order to proceed with obtaining the MAD score. Before obtaining the MAD score, the final step for KNN regression after splitting the data is to run the KNN regression function within RStudio. This requires a K value however, and since the dataset has 672 observations, a good starting point was to take the square root of 672, which comes out to ≈ 26.

```
MAD_nfl<-c()
for (i in c(1:50))
{
  #Run KNN regression for validation data
  knn.reg.model <- knn.reg(train_nfl.X, validation_nfl.X, train_nfl.Y, k = i)
  y_hat <- knn.reg.model$pred
  #Calculate error metrics
  MAD_nfl[i]<-mean(abs(validation_nfl.Y-y_hat)) #MAD
}
print(MAD_nfl)
#Based off above results, we chose K = 11
knn.reg.model <- knn.reg(train_nfl.X, validation_nfl.X, train_nfl.Y, k = 11)
y_hat <- knn.reg.model$pred
#Calculate error metrics
MAD_nfl<-mean(abs(validation_nfl.Y-y_hat)) #MAD
print(MAD_nfl)
#Mad = .09231
View(nfl_Data)
```

Running a “for” loop, as seen above, within RStudio with the K value ranging from 1-50, allows the ability to see the surrounding K values MAD scores in order to determine the best fit K value. K = 11 had the lowest MAD score (0.09231) after running the “for” loop, so when running KNN regression, this would be the best K value to use with this dataset.

Regression Tree Analysis

Additionally, another method chosen to potentially assist in solving this problem within the NFL is regression tree analysis. Opting to choose this method to address the problem was due to its ability to handle complex relationships and interactions with a dataset, thus making it a well-suited method for scenarios where traditional linear regression models may struggle to capture the non-linear patterns within a dataset. This is done by partitioning the data into smaller subsets based on the predictor variables, which then identify subgroups with similar response variable values; this alone allows for easy interpretation of the results.

```
#Regression Tree Analysis


#Split training and testing data
set.seed(1)      #make sure that we get the same result
idx <- sample(c(1,2), size = nrow(nfl_Data), replace = TRUE, prob = c(.8, .2))

train2_nfl <- nfl_Data[idx == 1, ] #define training data
test2_nfl <- nfl_Data[idx == 2, ] #define testing data

test2_nfl.X<- test2_nfl[  ,c(1:23)] #define testing input
test2_nfl.Y<-test2_nfl$win_loss_perc #define testing output
```

To start the process of using regression tree analysis, it is important to once again, split the data, but this time into only training and testing data. For this dataset, the data was broken into 80% training data, and 20% testing data, as seen above. Once again, after doing so, defining the testing input and output is also necessary.

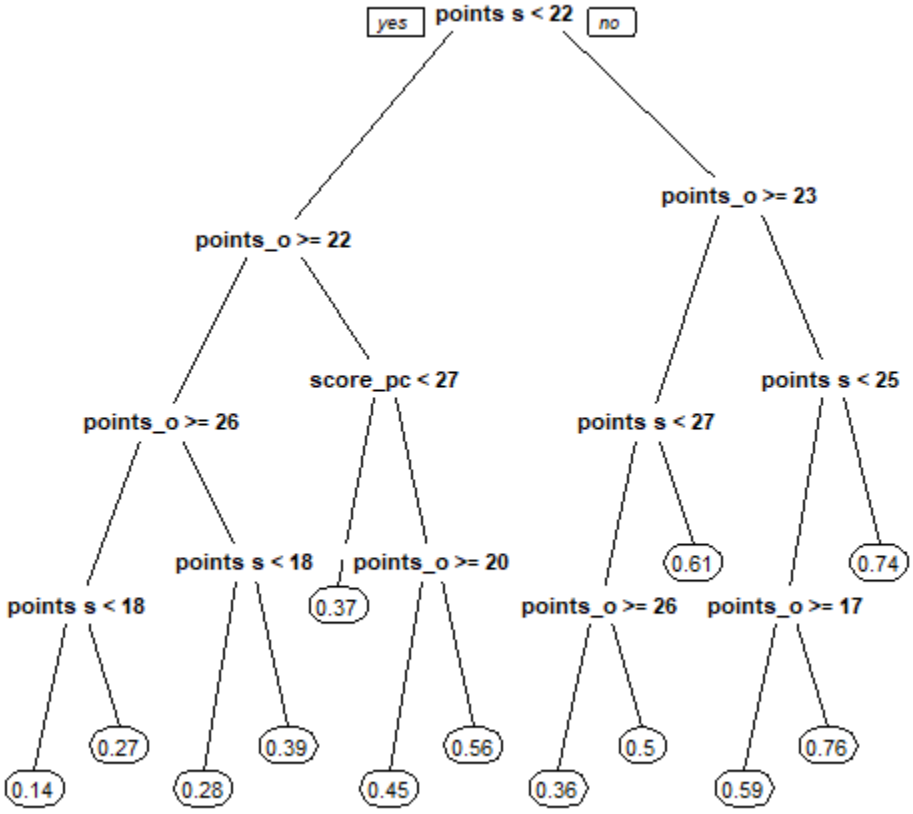
```
#Run regression tree
library(rpart) #this package contains decision tree
Reg_Tree = rpart(win_loss_perc~. , method="anova", data=train2_nfl, minbucket=1)
y_hat<-predict(Reg_Tree, test2_nfl.X) # predict the y_hat

#Calculate error metrics
MAD<-mean(abs(test2_nfl.Y-y_hat)) #MAD
print(MAD)

library(rpart.plot)
prp(Reg_Tree) #plot tree diagram
```

Once this is done, running the regression tree can then be started by using the “rpart” function, as seen above, within RStudio. When using the “rpart” function, for this dataset the method chosen was “anova”, and the minbucket was set to 1. Using the predict function within RStudio, the y_hat value can also be predicted, which is necessary for obtaining the MAD value. After all these steps are completed, the MAD value can be calculated, which comes out to 0.07972206.

Using the “prp” function within RStudio, allows the data to be plotted to a tree diagram as well.



“Points s” being the variable “points scored per game” within the dataset, and “points_o” being the variable “points_opp_per_game” otherwise known as points scored by the opponent per game. This diagram allows the ability to see which thresholds are used when determining a teams win/loss ratio.

Another method in determining the best model to analyze the dataset is the forward stepwise regression method. In the forward stepwise regression method, the set starts out empty. After determining the adjusted r^2 values for each variable, the variable with the highest adjusted r^2 is placed into the set. This process is continued until there are no further improvements in the

adjusted r^2 values for the variables not yet included in the regression model. After this model is implemented, the MAD value is computed. The MAD value for this dataset using the forward stepwise regression model is 0.06564451.

Another method to determine the best model to analyze the data is the backwards stepwise regression. In the backwards stepwise regression method, all 22 independent variables are analyzed. The relevant variables are determined by the p-value of each variable. The confidence level of this data set is 0.05, so if the p-value of a specific variable is higher than 0.05, that variable is removed from the model and the process is performed again. If more than one variable has a p-value greater than 0.05, then only the variable with the highest p-value is removed before repeating the process. After running the backward stepwise regression method and removing unnecessary variables, the final estimation equation is: $\hat{Y}=0.603+0.029x1-0.026x2-0.001x3-0.02x4+0.008x5+0.017x6+0.001x7-0.033x8+0.023x9-0.004x10$

where x1 is points scored per game, x2 is points against per game, x3 is the total amount of yards per game, x4 is offensive plays per game, x5 is first downs per game, x6 is pass attempts per game, x7 is passing yards per game, x8 is passing touchdowns per game, x9 is rush attempts per game, and x10 is turnover percentage. Upon completion of this method, the MAD value was computed using R Studio. The MAD value for the backwards stepwise regression method is 0.064611. Out of all methods tested for this dataset, the backward stepwise regression has the lowest MAD value, meaning that this method is most useful in determining the estimation equation.

Conclusion

In summary, an NFL team's win/loss ratio can be directly affected by various different variables as discussed in this paper. This is shown by the various regression models that were ran with these variables; these models include; forward stepwise regression, backward stepwise regression, KNN regression, as well as regression tree. From these models, the best one that was found to be used for running a regression analysis on this dataset was backward stepwise regression. After eliminating the unnecessary variables for this regression model, we were left with variables that were deemed to be significant in contributing to a team's win/loss ratio. These variables include; points scored per game, points against per game, total number of yards per game, offensive plays per game, first downs per game, pass attempts per game, passing yards per game, passing touchdowns per game, rush attempts per game, and turnover percentage. With a MAD value of 0.065, this was the lowest MAD value, and therefore was the best model to use. The second best regression method used was forward stepwise regression, with a MAD value of 0.066. Thirdly, the next best regression method used for this dataset was regression tree analysis with a MAD value of 0.07972206. Finally, the worst of the four regression methods that we used for this dataset was KNN regression with a MAD value of 0.09231. Each of these methods helped show the significance of the variables and how they play a role in predicting a team's win/loss ratio, while maintaining error to a minimum when doing so.

References

Cantalupa, N. (2024). NFL Team Data 2003-2023. Retrieved April 1, 2024 from https://www.kaggle.com/datasets/nickcantalupa/nfl-team-data-2003-2023.

Florio, M. (2023). NFL national revenue reaches $11.98 billion in 2022. Retrieved from https://www.nbcsports.com/nfl/profootballtalk/rumor-mill/news/nfl-national-revenue-reaches-11-98-billion-in-2022

2023 NFL Offense Rankings: Team Pass and Rush Stats. (2024). Retrieved from https://www.foxsports.com/articles/nfl/2023-nfl-offense-rankings-team-pass-and-rush-stats

Gifford, M. (2023). A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression. Retrieved from https://www.sciencedirect.com/science/article/pii/S2772662223001364

National Football League. (N.D.). 2023 NFL Rulebook. Retrieved on April 27, 2024 from https://operations.nfl.com/the-rules/nfl-rulebook/

Using data and analytics to evaluate the 2022 club proposals on overtime in the postseason: NFL Football Operations. (n.d.). Retrieved from https://operations.nfl.com/gameday/analytics/stats-

articles/using-data-and-analytics-to-evaluate-the-2022-club-proposals-on-overtime-in-the-postseason/

Football Ball Control. (2022, November 23.) Rookie Road. Retrieved April 27, 2024 from https://www.rookieroad.com/football/how-to/how-to-ball-control/