# Data Analysis Report - NFL Playoff Entrant Analysis

Bhavin Bohre, Samhit Kasichainula, Leon Chang, Jessica Li

2025-04-06

## Abstract

Predicting who will make the playoffs each year is difficult. Performance before the playoffs could provide valuable insights into a team's strength that year and whether they can make the playoffs. We are interested in analyzing how different metrics and standings of NFL teams relate to whether they make the playoffs. Our research focused on answering what metrics contributed the most to a team making the playoffs, whether non-game-related metrics increase the odds of a team making the playoffs, and what metrics are more important relative to others for an NFL team to make the playoffs. The data we are using to conduct the analysis contains 638 entries of 32 different teams performances from 2000 to 2019 every NFL season.

Our outcome would follow a Bernoulli distribution, relating to whether the team made the playoffs or not. The co-variates that we focused on exploring were wins, point differential, margin of victory, strength of schedule, simple rating, and total yearly attendance. From our experimentation, we discovered that the only statistically significant predictor was wins for all three Bernoulli models, with strength of schedule being marginally significant. We additionally delved into a Poisson model to model the rate at which teams are making the playoffs using the years $2000 - 2019$ as the offset for the model. Instead of using the base values for the co-variates, we took the average of each and modeled this rate. However, from our results, we saw that none of these predictors were significant and that there was a lot of multicolinearity between the co-variates. Overall, we can claim that wins are the most consistent and significant predictor of whether a team makes the playoffs.

## Introduction

Every year, 32 football teams in the National Football League (NFL) compete in a single elimination tournament of 16 games for a chance to compete in the playoffs and the Super Bowl. Only 14 teams are able to compete in the playoffs and only one will ultimately win the Super Bowl. It is difficult to predict who will win the Super Bowl or make the playoffs each year. Performance before and during the playoffs could provide valuable insights into a team's strength that year and whether they can advance and win. Over 200 million fans tune into the Super Bowl each year, and they spend a total of over 17 billion on tickets, food, apparel, and other game-related merchandise (Cerullo, 2024). The city that hosts the final game can

expect to generate approximately 500 million for their economy with the additional visitors. The sports betting industry is also growing, with a record-breaking $1.39 billion expected to be spent on this year's Super Bowl (Duster, 2025). With such a lucrative industry that is only increasing in popularity each year, it's only natural to want to predict which teams have a better chance of making it to the playoffs and why. We are interested in analyzing how different metrics and standings of NFL teams relate to if they make the playoffs or not. We hope to better understand the sports industry and the factors contributing to an NFL team's success. For example, does performance during the game, such as the margin of victory, significantly impact the probability that a team makes the playoffs, or does an external factor like game attendance play a role?

## Data and Methods

Our group is utilizing data on NFL stadium attendance, team standings, and game statistics spanning from 2000 to 2019 from the TidyTuesday Project GitHub repository. The data in the data set comes from the Pro Football Reference team standings. For our project, we will only be focusing on the standings and attendance data sets. The data itself was already very clean and we did not have much trouble preparing the data due to no NULL or missing values being present. In order to process our data, we first joined the attendance.csv and standings.csv file to form one csv file which we joined on team, team name, and year. We then had to manipulate the attendance count to total yearly attendance for every team in order for it to fit in our 638 row table. Lastly, we had to code Playoffs to be 1 and No Playoffs to be 0 to be able to use a binary response variable for our models.

For all Bernoulli models our group explored, we define the outcome as whether or not an NFL team reaches the playoffs. The observed values are denoted as $y_i$ from $i = 1, ...n$ where $n = 638$. We let $y_i = 1$ when a team makes the playoffs and 0 if they have failed to reach the playoffs. We also assume each $y_i$ is an independent realization from $Y_i \sim Bernoulli(p_i)$ where $p_i$ represents the probability of making the playoffs.

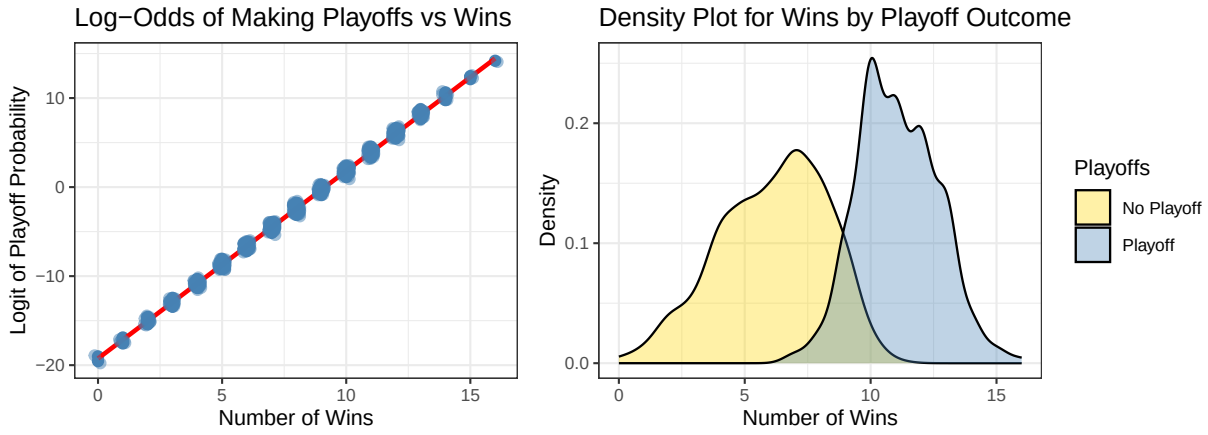


Figure 1: The plot on the left displays a scatterplot with the log odds of a team making the playoffs versus the number of wins they had that season. A line of best fit was added to show correlation, linearity, and trend. The plot on the right is a density plot that shows the

concentration of the number of wins for teams who made the playoffs (in blue) versus those who did not (in yellow).

The Bernoulli model including wins, point differential, and margin of victory is:

$$\begin{aligned}\text{logit}(p_i) = \beta_0 + \\ \beta_1 \times \text{wins}_i + \\ \beta_2 \times \text{points\_differential}_i + \\ \beta_3 \times \text{margin\_of\_victory}_i\end{aligned}$$

where $\text{wins}_i$, $\text{points\_differential}_i$, and $\text{margin\_of\_victory}_i$ are all continuous predictors.

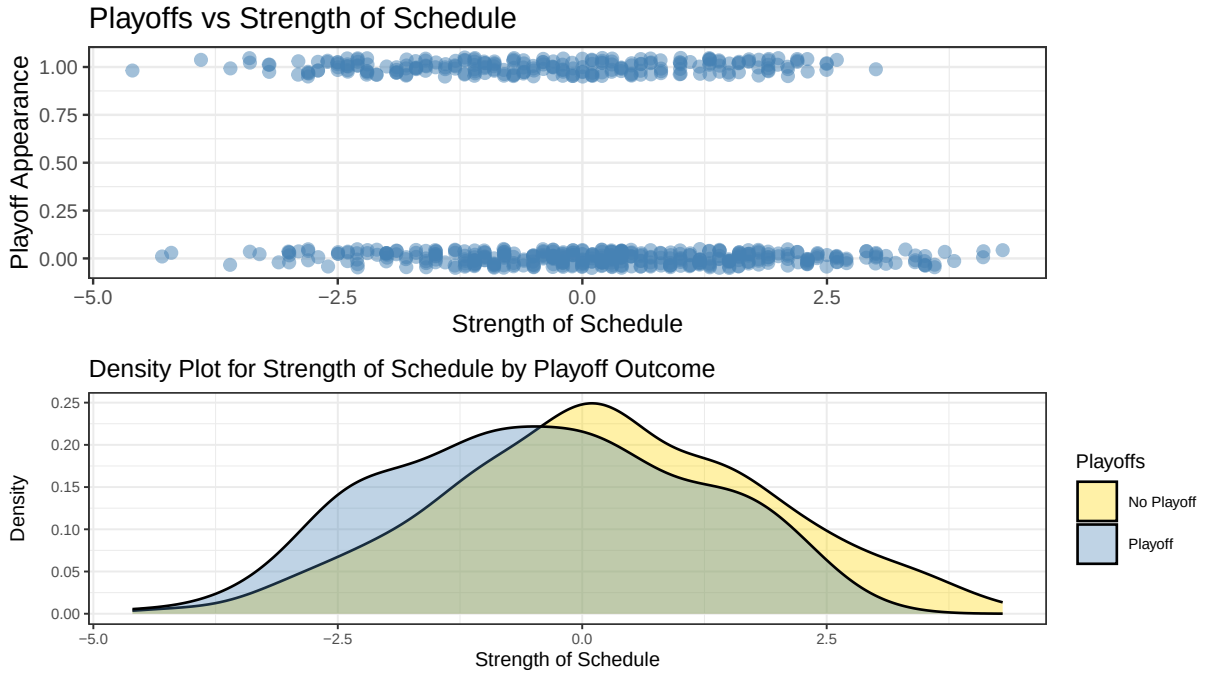


Figure 2: The top plot displays the relationship between an NFL teams strength of schedule (schedule difficulty or opponent strength) versus whether or not they make the playoffs, 1 representing successful playoff entry and 0 meaning a team failed to reach the playoffs. The bottom plot is a density plot that compares the distribution for teams that made the playoffs and teams that did not make the playoffs over 19 years. The x-axis is the strength of schedule and the y-axis represents the density or concentration of teams at each strength of schedule value. The yellow curve depicts teams that have not made the playoffs while the blue curve depicts teams that have made playoffs.

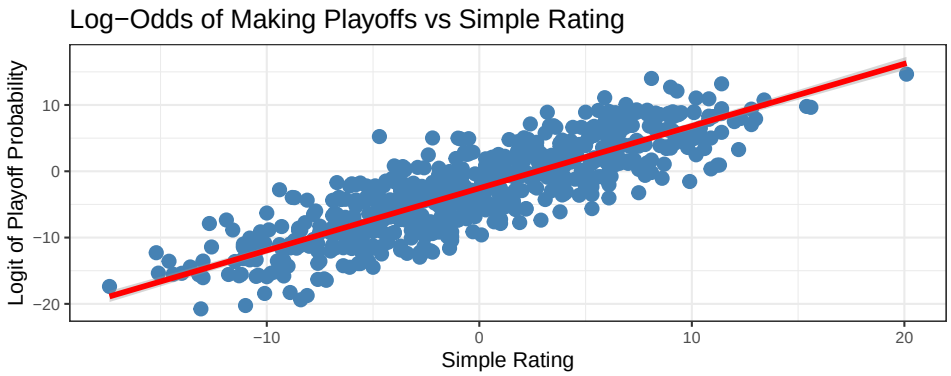


Figure 3: This plot is a scatter plot showing the log odds of making the playoffs versus simple rating (offensive rating + defensive rating) of an NFL team. A red line of best fit is displayed to represent correlation, trend, and linearity.

The Bernoulli model including strength of schedule and simple rating is:

$$\begin{aligned}\text{logit}(p_i) = \beta_0 + \\ \beta_1 \times \text{wins}_i + \\ \beta_2 \times \text{points\_differential}_i + \\ \beta_3 \times \text{margin\_of\_victory}_i + \\ \beta_4 \times \text{strength\_of\_schedule}_i + \\ \beta_5 \times \text{simple\_rating}_i\end{aligned}$$

where $\text{wins}_i$, $\text{points\_differential}_i$, $\text{margin\_of\_victory}_i$, $\text{strength\_of\_schedule}_i$, and $\text{simple\_rating}_i$ are all continuous predictors.

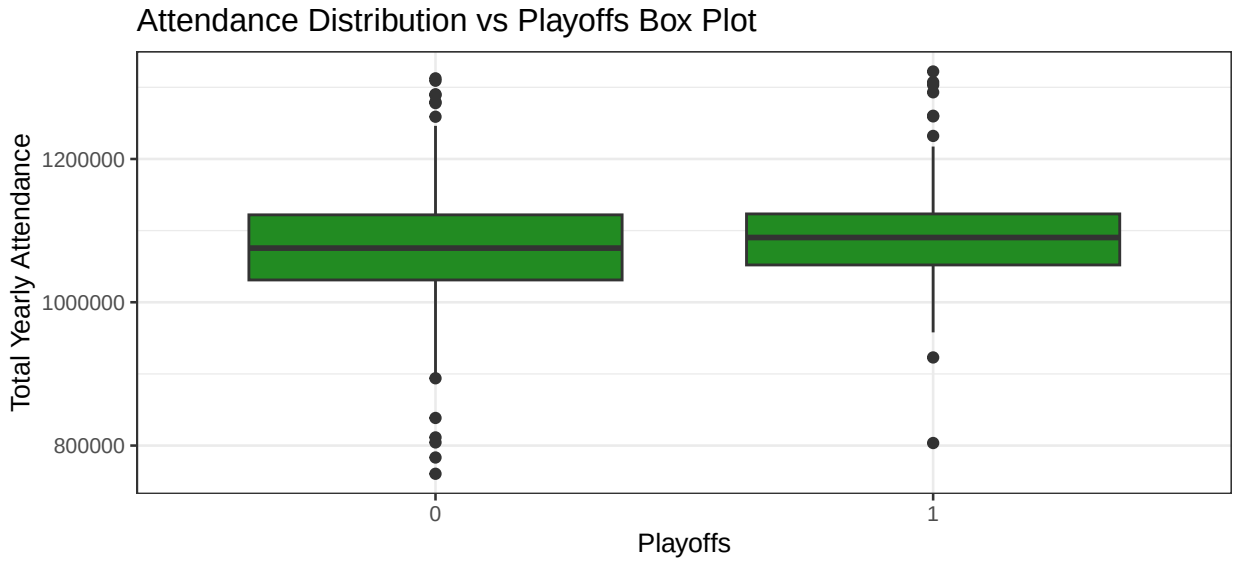


Figure 4: The above box plot displays the relationship between making the playoffs (1) and not making the playoffs (0) against total yearly attendance. The box shows the interquantile range and median of total yearly attendance with the points representing outliers in the

data. The lines coming out from the box (whiskers) represent the variability of total yearly attendance from the interquantile range.

The Bernoulli model additionally accounting for total yearly attendance is:

$$\begin{aligned}\text{logit}(p_i) = \beta_0 + \\ &\beta_1 \times \text{wins}_i + \\ &\beta_2 \times \text{points\_differential}_i + \\ &\beta_3 \times \text{margin\_of\_victory}_i + \\ &\beta_4 \times \text{strength\_of\_schedule}_i + \\ &\beta_5 \times \text{simple\_rating}_i + \\ &\beta_6 \times \text{total\_yearly\_attendance}_i\end{aligned}$$

where $\text{wins}_i$, $\text{points\_differential}_i$, $\text{margin\_of\_victory}_i$, $\text{strength\_of\_schedule}_i$, $\text{simple\_rating}_i$, and $\text{total\_yearly\_attendance}_i$ are all continuous predictors.

In addition to using a Bernoulli model to model whether or not a team makes the playoffs, we were also interested in looking at the rate at which teams enter the playoffs between $2000 - 2019$. The observed values are denoted as $y_i$ from $i = 1, ...n$ where $n = 638$. We let $y_i$ be the number of times a team has made the playoffs between $2000 - 2019$ and $N_i$ be the number of seasons a team has played between $2000 - 2019$. The rate at which a team makes the playoffs is represented by $y_i/N_i$. We also assume each $y_i$ is an independent realization from $Y_i \sim Poisson(\lambda_i)$ where $\lambda_i$ represents the expected count team $i$ makes the playoffs between $2000 - 2019$.

The Poisson model including wins, point differential, margin of victory, strength of schedule, and simple rating is:

$$\begin{aligned}\log(\lambda_i) = log(\text{years}_i) + \\ &\beta_0 + \\ &\beta_1 \times \text{avg\_wins}_i + \\ &\beta_2 \times \text{avg\_points\_differential}_i + \\ &\beta_3 \times \text{avg\_margin\_of\_victory}_i + \\ &\beta_4 \times \text{avg\_strength\_of\_schedule}_i + \\ &\beta_5 \times \text{avg\_simple\_rating}_i\end{aligned}$$

where $\text{avg\_wins}_i$, $\text{avg\_points\_differential}_i$, $\text{avg\_margin\_of\_victory}_i$, $\text{avg\_strength\_of\_schedule}_i$, and $\text{avg\_simple\_rating}_i$ are all continuous predictors and $log(time)$ is the offset used to account for season played disparities between the years $2000 - 2019$ for NFL teams.

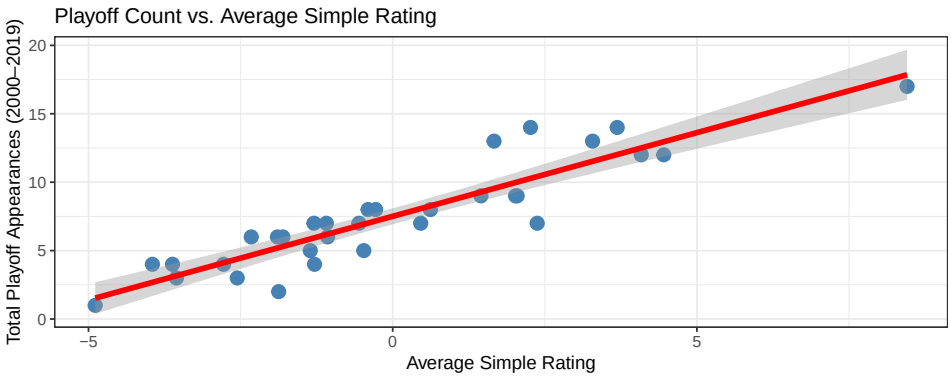


Figure 5: The scatter plot shows the relationship between average simple rating and total playoff appearances from the years $2000 - 2019$. A red line of best fit along with a grey confidence interval is displayed to display the relationship.

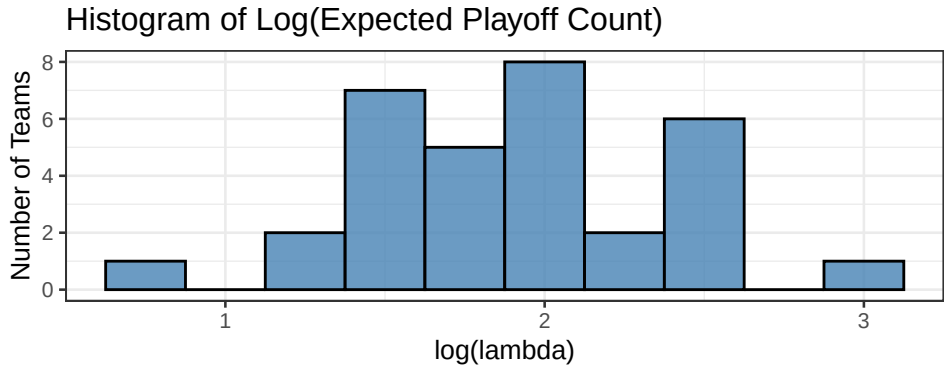


Figure 6: The histogram represents the distribution of the log expected playoff counts for the 32 NFL teams where $log(\lambda)$ represents the natural log of playoff counts.

## Results

In our first Bernoulli model, we can see that in the left scatter plot in Figure 1, there is a strong positive linear relationship between the number of wins a team has throughout the season and the log odds probability of them making the playoffs. From the density plot we can also see that teams who make the playoffs tend to have more wins than teams who do not. Teams with at least 10 wins are much more likely to make the playoffs.

Table 1: Bernoulli Model 1 Logistic Regression Results

| | Est. | S.E. | z val. | p |
|---|---|---|---|---|
| (Intercept) | -20.08 | 2.29 | -8.78 | 0.00 |
| wins | 2.21 | 0.26 | 8.51 | 0.00 |
| points_differential | 0.51 | 0.37 | 1.39 | 0.16 |
| margin_of_victory | -8.19 | 5.84 | -1.40 | 0.16 |

From Table 1 we can see that wins is the only statistically significant variable in predicting whether a team makes the playoffs with a p-value below the significance level of 0.05. The points differential and margin of victory are both less significant with a p-value of 0.16 each.

For the second Bernoulli model we decided to explore, we saw that top plot of Figure 2 shows that strength of schedule does not directly affect whether an NFL team makes the playoffs since teams with both easy and difficult schedules are seen in the playoffs between the $-2.5$ to 2.5 range. Teams that had a strength of schedule greater than 2.5 were much more likely to not make the playoffs. However, the density plot in Figure 2 shows that teams that made the playoffs tend to have slightly easier strength of schedules compared to teams that did not. Lastly, Figure 3 shows the strong linear positive relationship between simple rating and log odds of making the playoffs.

Table 2: Bernoulli Model 2 Logistic Regression Results

| | Est. | S.E. | z val. | p |
|---|---|---|---|---|
| (Intercept) | -20.81 | 2.38 | -8.74 | 0.00 |
| wins | 2.29 | 0.27 | 8.48 | 0.00 |
| points_differential | -0.01 | 0.48 | -0.03 | 0.98 |
| margin_of_victory | -7.18 | 6.01 | -1.19 | 0.23 |
| strength_of_schedule | -7.63 | 4.25 | -1.79 | 0.07 |
| simple_rating | 7.34 | 4.24 | 1.73 | 0.08 |

We see that wins is still the only significant predictor that has a statistically significant affect on modeling whether a team makes the playoffs or not. Adding strength of schedule and simple rating did have a statistically significant affect on whether or not team made the playoffs. The p-values for these 2 predictors of 0.07 and 0.08 are marginally significant to the 0.05 significance level but do not meet the threshold of being $\leq 0.05$.

In our last Bernoulli model factoring in total yearly attendance, we see in Figure 4 that median total yearly attendance between teams that made the playoffs versus teams that did not was not very different. Both playoff and non playoff have around the same median and variability with the interquantile range for playoff teams be slightly smaller. In addition, there seems to be more outliers for non-playoff teams than playoff teams. These results indicate that attendance does not factor much into whether a team makes the playoffs or not.

Table 3: ANOVA Table for Bernoulli Model

| | Df | Deviance | Resid. Df | Resid. Dev | Pr(>Chi) |
|---|---|---|---|---|---|
| NULL | NA | NA | 637 | 844.92 | NA |
| wins | 1 | 627.94 | 636 | 216.98 | 0.00 |
| points_differential | 1 | 0.66 | 635 | 216.32 | 0.42 |
| margin_of_victory | 1 | 2.00 | 634 | 214.32 | 0.16 |
| strength_of_schedule | 1 | 6.85 | 633 | 207.48 | 0.01 |

| | Df | Deviance | Resid. Df | Resid. Dev | Pr(>Chi) |
|---|---|---|---|---|---|
| simple_rating | 1 | 3.06 | 632 | 204.41 | 0.08 |
| total_yearly_attendance | 1 | 1.09 | 631 | 203.32 | 0.30 |

The model with wins, point differential, margin of victory, strength of schedule, simple rating, and total yearly attendance has the lowest deviance compared to the other models meaning fits to the data set the best. However, this does not mean it may be the best model to fit our data since only 1 predictor is statistically significant. Wins might be the only necessary predictors to predict whether a team is going to make the playoffs when factoring in total yearly attendance. The other predictors are not accounting for much of the deviance at all, and might be better off ignored. From the above deviance table as well, we see that the best model is the model that includes wins, point differential, margin of victory, and strength of schedule since its deviance is statistically significant with a p-value less than 0.05 significance level.

Looking into our poisson model, we see that in Figure 5 that there is a strong positive linear relationship between average simple rating and total playoff appearances between $2000 - 2019$. This means the higher the simple rating (offensive rating + defensive rating) the more the team appears in the playoffs. In Figure 6, we see that majority of the teams fell into between 1.5 and 2.5 for log number of playoff appearances with 2 having the most number of teams.

Table 4: Poisson Model Logistic Regression Results

| | Est. | S.E. | z val. | p |
|---|---|---|---|---|
| (Intercept) | -3.63 | 1.77 | -2.05 | 0.04 |
| avg_wins | 0.32 | 0.22 | 1.44 | 0.15 |
| avg_point_diff | 0.10 | 0.92 | 0.11 | 0.91 |
| avg_margin_victory | 1.12 | 11.43 | 0.10 | 0.92 |
| avg_strength_of_schedule | 2.46 | 11.05 | 0.22 | 0.82 |
| avg_simple_rating | -2.72 | 11.00 | -0.25 | 0.80 |

From the above regression output, we see that when averaging all the co-variates and using a Poisson model that none of the co-variates are statistically significant. This indicates that this poisson model poorly represents the rate at which teams make the playoffs and that including all these predictors does not model this rate effectively.

Table 5: ANOVA Table for Poisson Model

| | Df | Deviance | Resid. Df | Resid. Dev | Pr(>Chi) |
|---|---|---|---|---|---|
| NULL | NA | NA | 31 | 62.57 | NA |
| avg_wins | 1 | 52.10 | 30 | 10.46 | 0.00 |
| avg_point_diff | 1 | 0.11 | 29 | 10.36 | 0.74 |

| | Df | Deviance | Resid. Df | Resid. Dev | Pr(>Chi) |
|---|---|---|---|---|---|
| avg_margin_victory | 1 | 0.01 | 28 | 10.35 | 0.93 |
| avg_strength_of_schedule | 1 | 1.45 | 27 | 8.90 | 0.23 |
| avg_simple_rating | 1 | 0.06 | 26 | 8.84 | 0.80 |

Using the above deviance table, we see that the model with average wins is only statistically significant compared to the others at 0.05 significance level. We see that the deviance is really close to 0 when including all 5 predictors meaning that the model is very saturated causing it to over fit to new data that it has not seen. Therefore, we can conclude the best Poisson model only includes average wins as a predictor.

From our analysis, we have clearly seen that wins was the only significant co-variate the contributed to whether a team made the playoffs. Tying back to our research question of whether other non-game related factors affected playoff entrance, we clearly see that it indeed did not play a part between differentiating the two groups.

The final model that our group has decided to choose to model whether or not a team has made the playoffs follows a Bernoulli distribution since our main focus is to figure out what model effectively predicts if a team makes the playoffs or not. From our analysis of deviance table, the model that we chose that best represents this is:

$$\begin{aligned}\text{logit}(p_i) = \beta_0 &+ \\ &\beta_1 \times \text{wins}_i + \\ &\beta_2 \times \text{points\_differential}_i + \\ &\beta_3 \times \text{margin\_of\_victory}_i + \\ &\beta_4 \times \text{strength\_of\_schedule}_i\end{aligned}$$

where $\text{wins}_i$, $\text{points\_differential}_i$, $\text{margin\_of\_victory}_i$, and $\text{strength\_of\_schedule}_i$ are all continuous predictors.

## Conclusion

To sum up our analysis, our main finding was the wins was the only significant co-variate for all Bernoulli models we tests. In the Poisson model, we saw that none of the co-variates were significant but the residual deviance for almost all the models was really low indicating very high saturation in the model. A key takeaway that we saw was that sometimes the simplest models fit the data the best. Due to our prior belief, we thought the predictors we chose to experiment with had a substantial effect on whether an NFL team made the playoffs. However, through our analysis we clearly saw otherwise. In the future, some potential further research that could be done is does the time and day games are played affect the number of wins a team has and whether this in turn effectively affects their chance at making the NFL playoffs.

**Refrences**

Cerullo, M. (2024, February 7). Americans expected to spend a record $17.3 billion on 2024 Super Bowl. CBS News. https://www.cbsnews.com/news/super-bowl-2024-las-vegas-economic-impact/#:~:text=Such%20numbers%20underline%20the%20event%27s,related%20merchandise%20and%20other%20items

Duster, C. (2025, February 8). 2 reasons why a record $1.39b is expected to be bet on the Super Bowl. NPR. https://www.npr.org/2025/02/08/nx-s1-5290099/one-billion-super-bowl-lix-bets#:~:text=A%20record%20$1.39%20billion%20is,American%20Gaming%20Association%20(AGA)