The Run-Pass Conundrum: What's The Better Play?

Arthur Jakobsson, Elin Jang, Prathik Guduri

2024-04-30

Introduction

The NFL, the biggest stage in America's favorite sport, is viewed by tens of millions of people per game. With its immense popularity, the NFL has become a crucial part of American society, shaping meaningful conversations amongst friends, family, and even strangers. In the center of this craze are the coaches, whose strategy and decision-making can make the difference between victory and loss. These coaches are paid multi-million dollar salaries as a testament to the immense pressure placed on them and a lot is at stake in every single play, with the fate of the players, other coaches, and entire organizations in the palm of their hands.

With this context, this report dives into the data covering the tens of thousands of plays across the NFL season, aiming to uncover the underlying trends that determine the effectiveness of different offensive strategies, particularly the ongoing debate between running and passing plays. One of the most interesting concepts about American football is the strategic chess match between passing and running plays in the NFL, with game theory in optimizing offensive decisions amidst defensive ones. We analyze the optimal conditions and situations where certain types of plays can maximize influence on the outcome of a given game.

To provide a preview of the results, our final conclusion based on the analysis is that _____. Further detail is discussed in the latter sections.

Data

The data used in this report is sourced from the play by play data of the 2023 NFL season provided by the NFL readr R package. This dataset is quite extensive, consisting of over

300 variables for each play in every game of the season. The following are the most relevant variables of this dataset that we will be using for the bulk of our analysis:

play_type — keeps track of the play type (run or pass) ydstogo — yards needed for a first down yards_gained — yards gained on the play down — down number for the play posteam_type - home/away team (offensive team)

To get a better understanding of the dataset and preliminary information on the variables, we performed the following exploratory data analysis.

Play by Play EDA:

```
# R chunk with EDA
```

We see

Drive by Drive EDA:

```
# R chunk with EDA
```

We also wanted to perform some trends to analyze how the amount of run and pass plays affect success on a drive level. To do this we wanted a dataset containing drive by drive data which we created from the play by play dataset. We merged data points with the same game id and drive number into one drive data point. For this we counted interesting variables in the play by play data set when merging them into a singular drive. The important variables formed in this process include.

- total_yards_gained - total yards gained over the whole drive
- posteam_type - home/away team (that is on offense for the drive)
- drivePoints - number of points that the drive resulted in
- runPercent - percent of the drive plays that were run plays
- passPercent - percent of the drive plays that were pass plays

Methods

Linear Model

We started our models with one of the simplest models, a linear regression model. For this model we used some of the variables we deemed the most important in the play by play data. This included the playtype (run/pass), yards for first down, and down number. For the purpose of analyzing run vs pass plays, the data was cleaned to include only run and pass type plays and plays that were downs one, two or three.

Zero-inflation model: predicting yards

We noticed many plays resulted in zero yards gained or even negative yards progressed. We were interested in what factors most significantly contributed to making a positive number of yards gained but also given that a positive number of yards were gained we were also curious which factors most significantly impacted the number of yards gained.

To study this, we decided to build a zero-inflation model with a binary logistic regression model to predict whether or not a team would gain zero yards. Following this, we used a Poisson function to model and predict the number of positive yards gained (given that the play had positive yards gained). In order to perform this, we changed all negative values for yards gained in the dataset to zero so that we could differentiate forward moving plays and those without progression.

It is important to note that in football success can not always be determined by gaining yards vs losing yards in a play. There are many plays that gain yards that would be considered failures and considering them as an equal success to a 90 yard touchdown pass would not be the most accurate representation of football. Also though it is more rare, there are times when losing yards isn't really a failure, for example when a team is kneeling to run down the clock at the end of a game. Due to these issues, the use of this model is most useful for coaches when trying to understand how many yards they need to progress. We further elaborate on the implications of this in the results and conclusion sections.

Predicting Drive Success Based on Ratio of Play Types

Since we wanted to measure more definitively the success of a play, we seek to analyze on a basis of drive success and a series of plays. To determine this, we create a drive-by-drive dataset and create two separate binary logistic general linear models. We created a binary variable which marked whether a drive resulted in a positive number of scored points and regressed this with whether more runs were made or more passes and the home and away team variables. This helped us understand the importance of incorporating more or fewer runs into a drive and how it impacts success.

We also built a generalized additive model (gam) to further understand the relationship between the percent of plays being runs and whether that play was a scoring play. This model calculates the relative impact of each of the variables included in the model on the output and generates predictions based on the sum of these relationships. Gams are non-linear and we explore the interesting non-linear relationship observed in our regression between run percent and play success in the results section.

Results

Linear Model

The linear model is very limited in the patterns it can capture in the data and thus it is limited in the results it can provide. However, we do find some interesting trends. These trends can be seen in figure X which provides the predictions of the linear model given if it is a run or pass play, the down number, and the yards to go. Firstly we see that 1st and 2nd down are very similar to each other while 3rd down has a significantly lower EPA. In the context of football this does make sense, as 3rd downs are considered the riskiest out of the three downs as often not making a 3rd down results in a loss of possession through punting and thus 3rd downs have the most potential to lose expected points. Another trend we see is that pass plays seem to have a significantly higher EPA then run plays. This is a little more interesting, likely due to the nature of pass plays and their higher potential for significantly large gains the run plays which most of the time are only for a couple yards.

```
# R code for linear model graph
```

Zero-inflation model: predicting yards

```
Call:
zeroinfl(formula = yards_gained ~ ydstogo + play_type + as.factor(posteam_type),
    data = pbp, dist = "poisson", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-2.0898 -1.0697 -0.4792  0.4765 30.4224

Count model coefficients (poisson with log link):
                             Estimate Std. Error  z value Pr(>|z|)
(Intercept)                 2.3274908  0.0027434  848.411  < 2e-16 ***
ydstogo                     0.0136557  0.0002446   55.818  < 2e-16 ***
play_typerun               -0.7113955  0.0021253 -334.724  < 2e-16 ***
as.factor(posteam_type)home  0.0067440  0.0019826    3.402  0.00067 ***

Zero-inflation model coefficients (binomial with logit link):
                             Estimate Std. Error  z value Pr(>|z|)
(Intercept)                -0.075992   0.014567   -5.217 1.82e-07 ***
ydstogo                    -0.024289   0.001363  -17.821  < 2e-16 ***
play_typerun               -1.313447   0.011974 -109.689  < 2e-16 ***
```

```
as.factor(posteam_type)home -0.035488   0.010722   -3.310 0.000933 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 11
Log-likelihood: -6.376e+05 on 8 Df
```

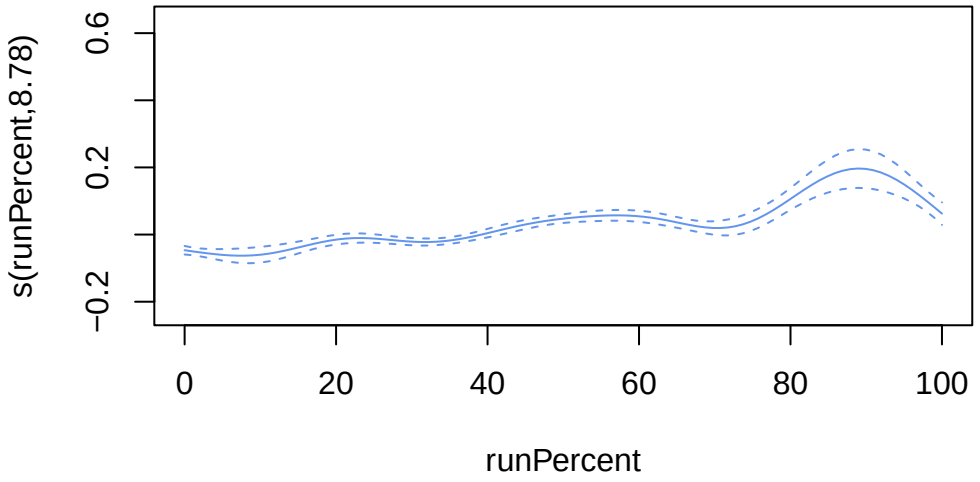
We find in our zero-inflation model that home field advantage yields an advantage on whether yards were gained but also on the number of yards gained. The logistic regression finds that the log odds of 0 points decreases by about 0.04. More notably having the play being a run decreases the log odds of zero yards gained being 0 more considerably by 1.31. This aligns with intuition about football because runs are more likely to make progress and are generally considered less risky. However, when we turn to our poisson model we find that running has a detrimental effect on the number of yards gained. This also aligns with intuition since passing plays generally cover a larger distance and are either successful or gain no yards. So when given that the play is successful it makes sense for passing plays to be predicted to be more beneficial to gaining more yards. Overall, these results imply that passing plays are riskier for making some measurable progress, but if the team needs to make considerable progress it is more worthwhile to throw a passing play.

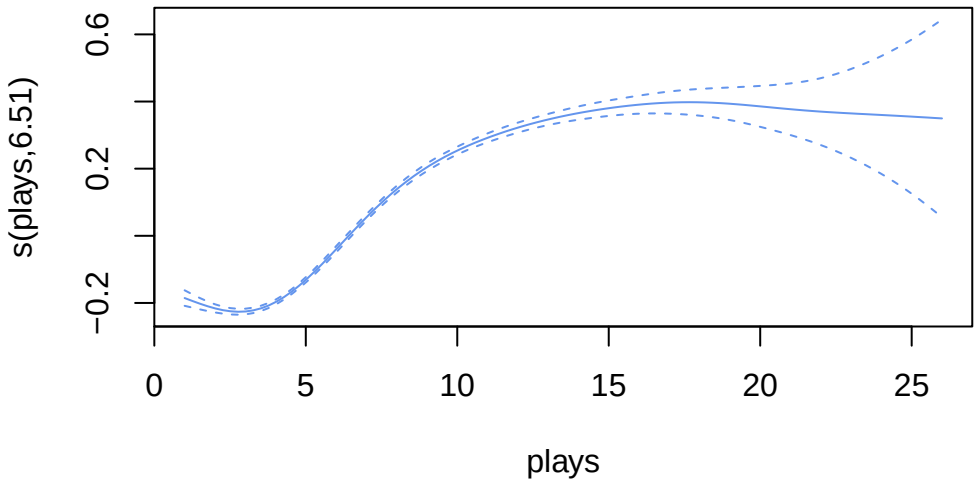
Predicting Drive Success Based on Ratio of Play Types

This model yielded the familiar result that teams that have home game advantage will likely outperform others in drive success. However, we find interesting results about the relationship between the ratio of run plays to pass plays and the success of the play. Fig X. visualizes this relationship. We find that the most successful drives are those that are around 90% runs. This is very interesting, as many of the professional coaches and even NFL fans will tell you that running 90% of plays will not result in successful drives. We think this may come from an issue in the low number of plays in a drive which means that a drive with 90% runs, has at least 10 plays in it due to the nature of whole numbers. A drive with over 10 plays is very likely to be a successful drive as it has lasted that long. This thus could create a miss representation of the variable. We also notice a peak at equal shares of passes and runs which seems to represent most of the data as we can see from our EDA. The cross-validated model shows that all of the coefficients are significantly non-zero and the standard errors for Fig X. are low.

In our glm model, we found that having strictly more runs than passes helped increase the odds that a drive would end in a touchdown. This likely can be attributed to having the need for many running plays and few passing, potentially risky, plays in every drive. We found that having more runs than passes increased the log odds of making a pass by 0.32 with a standard error of 0.03 and a p-value of less than 0.05 which suggests that this value is significantly not zero.

```
Call:
glm(formula = drivePoints > 0 ~ more_runs + as.factor(posteam_type) +
    plays, family = binomial, data = dbd_no_neg)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8948  -0.8658  -0.6644   1.0012   2.0286

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.18218    0.03292 -66.293  < 2e-16 ***
more_runsTRUE               0.32427    0.03158  10.268  < 2e-16 ***
as.factor(posteam_type)home 0.08820    0.02603   3.388 0.000704 ***
plays                       0.26120    0.00391  66.797  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 40500  on 29797  degrees of freedom
Residual deviance: 34672  on 29794  degrees of freedom
  (808 observations deleted due to missingness)
AIC: 34680

Number of Fisher Scoring iterations: 4
```





Discussion

Conclusion

Limitations

Although our model provides a good overview of comparing play types during certain situations including over different downs and over different distances needed for the first down, there are also many other variables present when considering play calling. For example a team like the lions who have Jahmyr Gibbs and David Montgomery running behind one of the best offensive lines in the league may want to run more than the average NFL team. Defense strength is another important factor, if you are playing against the top cornerbacks in the league, you might want to rely on passing plays less. The specific strengths of the offense and defense team which can be very different impact the play calls a lot. One further thing to consider is the predictability of play calls. Our data treats all of the data points as individual points for this analysis but in reality, there is a lot more depth to it. Offensive coordinators want to make play calls that the opposing defensive coordinator will not expect or is not ready for.

Future work

`Future work that would be very interesting would be looking into deeper models that consider`