36-460/660 Final Project Report

Alex Cheng, Melody Wang, Liz Chu, Kevin Ren

April 23rd, 2024

Contents

Run, Run, then Run Again

Introduction

The problem of playcalling is among the greatest challenges in sports coaching, particularly in the sport of American football. Particularly poor calls – e.g. the Seahawk's decision to pass rather than run with 1 yard to go in the closing seconds of Super Bowl XLIX – represent snap decisions that can cause franchises and players alike to lose critical games and ultimately millions of dollars. As such, in order to avoid such outcomes coaches must be very smart about what kinds of plays to run in different scenarios, depending on the previous plays that they have called along with the context of the game. In this project we hope to answer the age-old question of what plays a football coach should call, given the types of the previous three plays in the drive, along with contextual information about the field position such as yards-to-go until a touchdown, the strength of the offensive team in terms of their passing and rushing abilities, and the difference in score in the game. We ultimately find that our fitted models on the given covariates perform strongly on both in-distribution and out-of-distribution data on their respective predictive tasks, compared to a naive baseline model.

Data

Describes the data you're using in detail, where you accessed it, along with relevant exploratory data analysis (EDA). You should also include descriptions of any major pre-processing steps.

We obtained our data from the R packages `nflreadr` and `espnscrapeR`. The main feature of the `nflreadr` package that we used in this project was its wealth in NFL play-by-play data, which would give us detail on what was happening at each step in a NFL game. We primarily used play-by-play data from the most recent 2023 NFL season for our analysis, and while the dataset featured 372 features for each play, wed only used a subset of what we thought were the relevant features to our research question, such as yard to go and point differential. The primary purpose of using `espnscrapeR` was to obtain the rushing and passing rankings of each team, which we determined by using the total number of rushing and passing yards each team ahd at the end of the 2023 seasson.

With these two datasets, we created our own dataset through four major pre-processing steps.

(1) In `nflreadr`, we would redefine a successful play as having a positive Expected Points Added (EPA) for the first and second down, and covering the yards needed to achieve a first down for the third and fourth down.

(2) With the dataset from `nflreadr`, we would take the k previous rows before a play and parse out the play type. We would then augment the dataset by adding k columns, with the ith added column corresponding to the ith previous play.

(3) We would do a left-join operation with the `espnscrapeR` dataset - for the team with the possession, we would add its pass and rush ranking derived from the `espnscrapeR` dataset, resulting in two more additional columns to our dataset.

(4) We would filter out all plays that were not successes due to difficulties with counterfactuals, as well as all plays that weren't a rush or a pass. Thus, our dataset would only contain successful rush or pass plays.

This would result in the main dataset that we would work with training and building our models.

However, it would be unwise to blindly fit regressions without taking a look at possible patterns in our data, and our exploratory data analysis revealed some things that were intuitive, and some that were interesting.

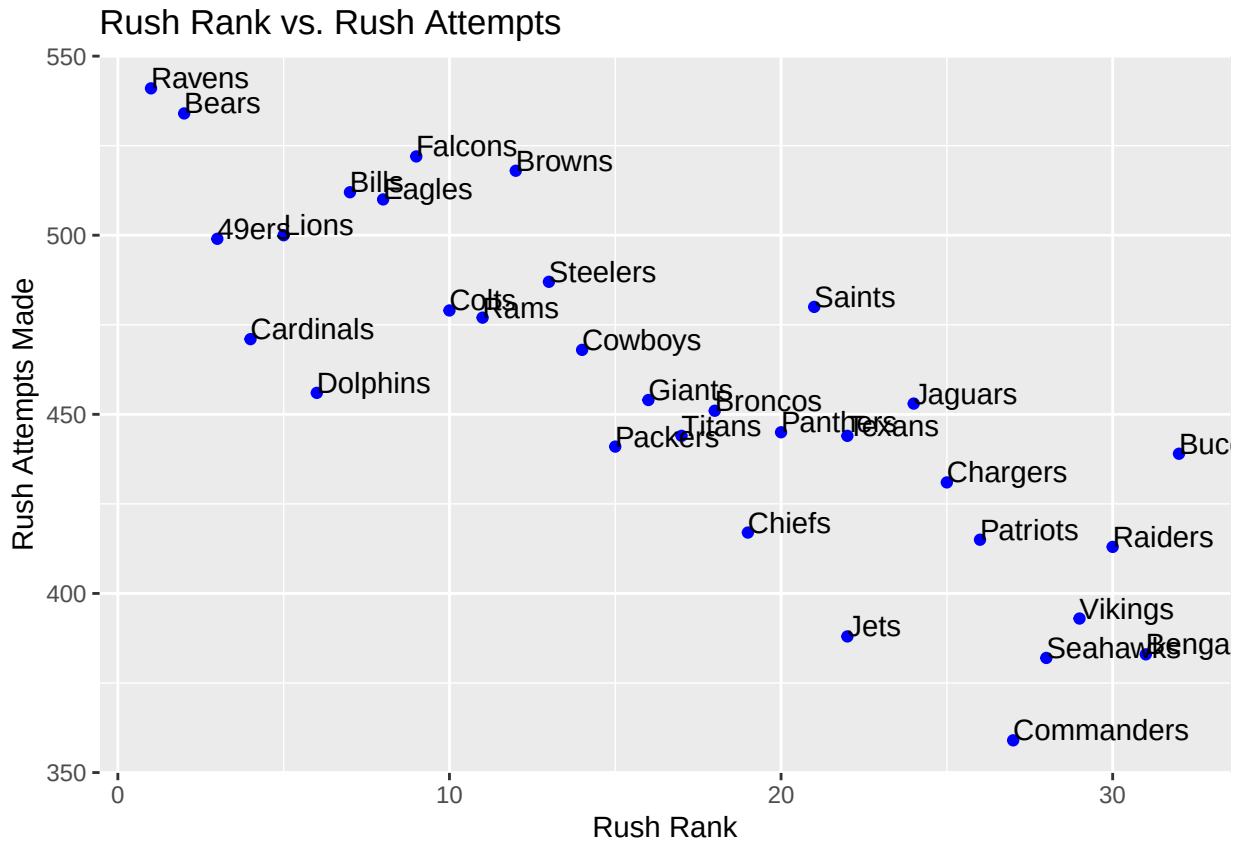
```
## * 13:52:11 | Cleaning up play-by-play...
```

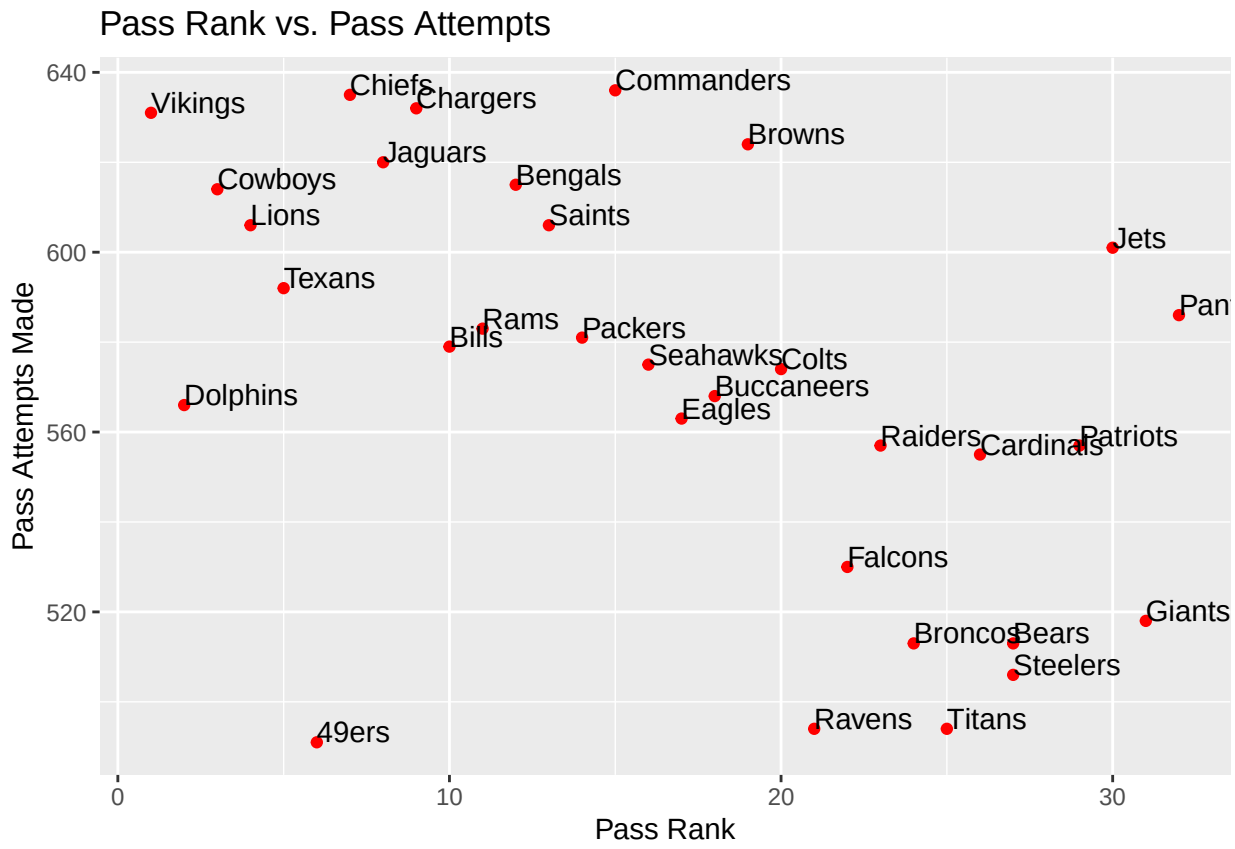
```
## v 13:52:16 | Cleaning completed
```

The first thing we took a look at was the relationship between pass ranking and number of pass attempts, as well as rush ranking and number of rush attempts, based on data from the `espnscrapeR` dataset. At a glance, these scatterplots seem pretty intuitive - the higher the ranking for a particular aspect of the game, the more attempts that a team would have with it. This is particulary true for rushing, where the scatterplot seems extremely linear, and from a qualitative perspective also makes sense. Teams like the Ravens, 49ers, and Dolphins have fast runners in Lamar Jackson, Christian McCaffrey, and Raheem Mostert, and it would make sense to just let these players do the work in creating their own yardage. Things get a bit more interesting when you start looking at the relationship between pass attempts and pass rankings. Though this scatterplot is still relatively linear, it is significantly more scattered than the rushing plot, which may indicate that the quality of your passing offense doesn't dictate how much you actually pass as much as rushing would. What's particularly interesting is that the top rushing teams we just mentioned - Ravens, 49ers, and Dolphins - all seem to have lower pass attempts than expected, even though the rankings for the Dolphins and 49ers are actually quite high. The 49ers and the Dolphins to have plenty of lethal receiver options (Brandon Aiyuk, Tyreek Hill), but it seems like teams just prefer to rush if they're good at it, regardless of their passing abilities, which may be attributed to ball safety in these two types of plays.

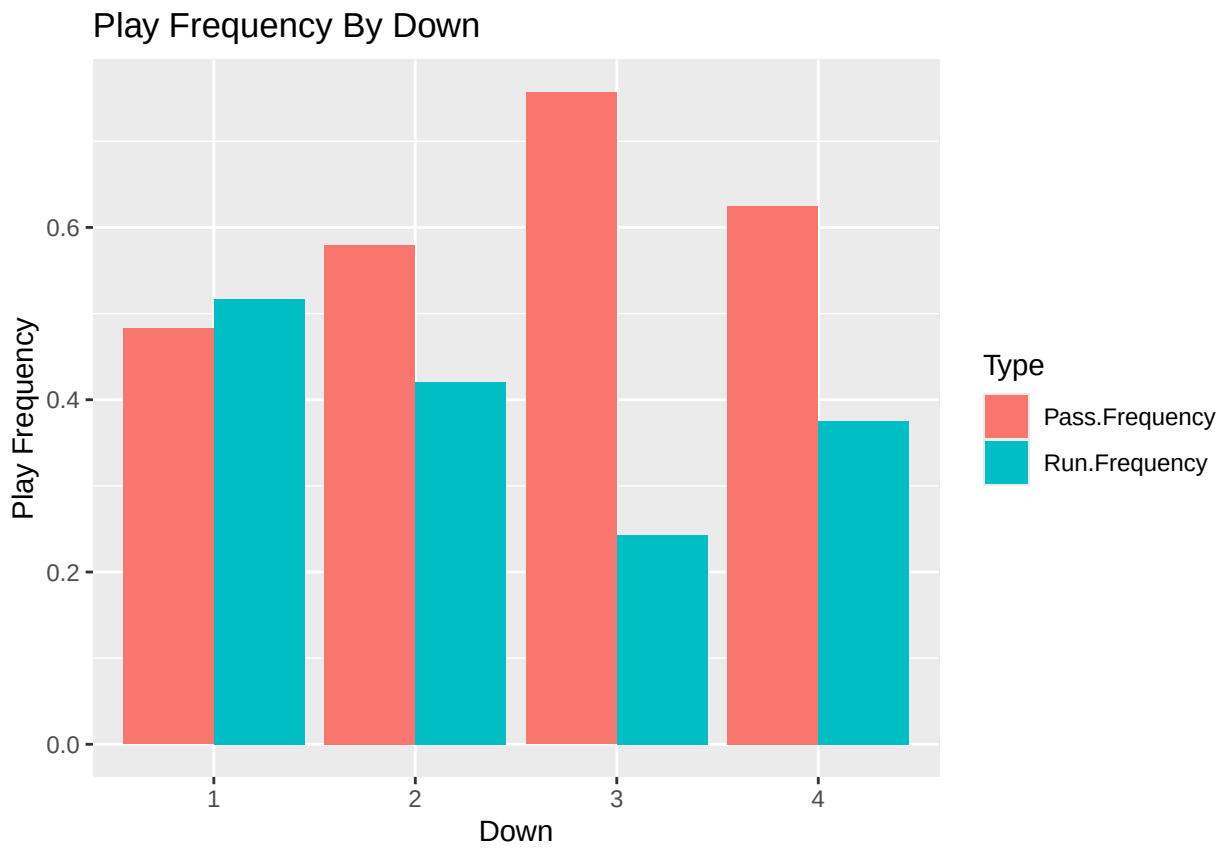
```
## Scraping rushing for offense from 2023!

## Scraping passing for offense from 2023!
```

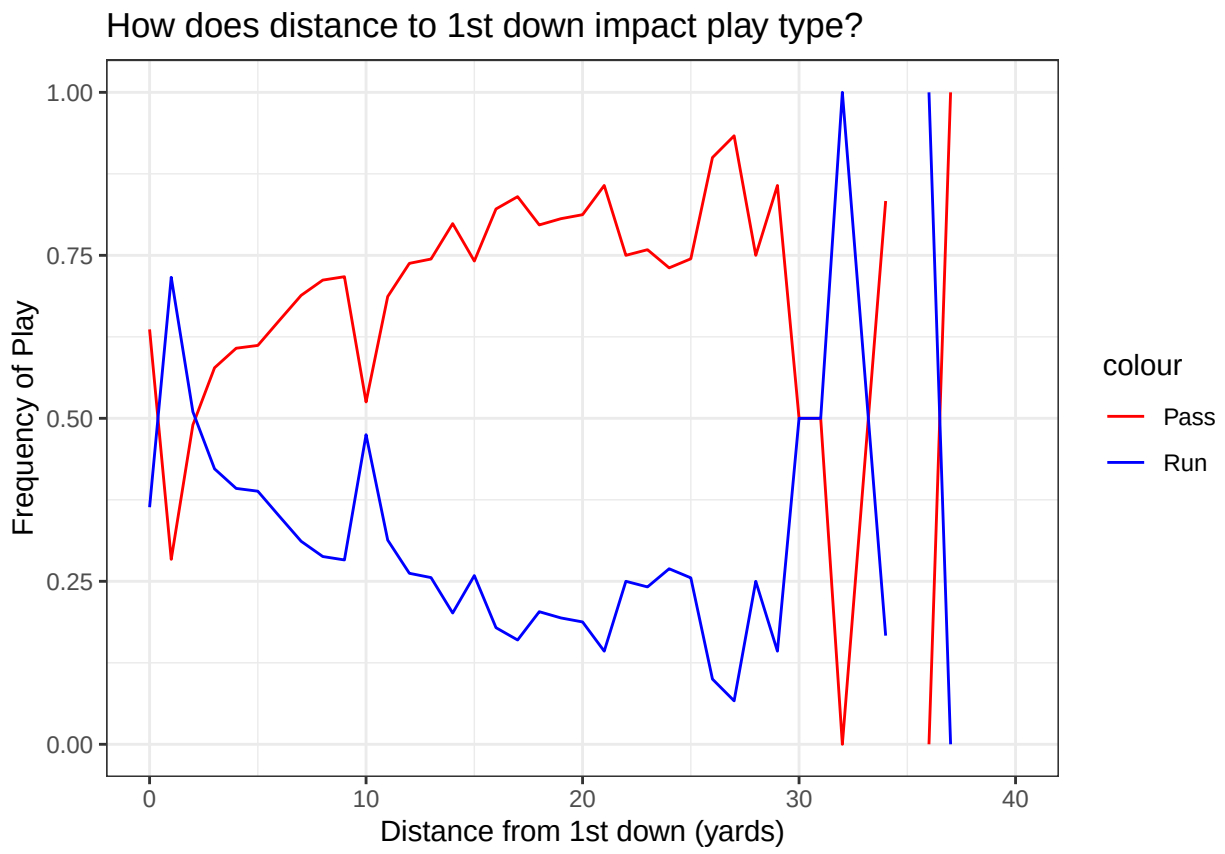




We then took a look at the types of plays that were occcuring at each down. Recall that the goal of each down is to either obtain the first down or to score the ball, and after the fourth down, teams are required to turn the ball over to the opponent. Thus, from first to third down, it makes sense that the number of pass plays increase significantly, as teams are more desperate to cover the amount of yards needed to get to the first down. Fourth down is special in that there is some selection bias in these plays - NFL teams only really go for fourth down when (1) not doing so would essentially mean the forfeiture of the game, or (2) the number of yards is small enough that the team is willing to risk turning the ball over to the other team for a chance to extend the play. Nevertheless, it is still interesting that teams are more inclined to pass rather than rush on fourth down. Since yards needed is usually small, it would make intuitive sense that teams would actually prefer to *rush*, so this may be a part of NFL play calling that depends on the previous plays, which is something we want to explore.



Lastly, we took a look at the relationship between the number of yards to go and the pass/rush attempts at each point. What we see is that under 3 yards the chance of a rush or a pass is generally a coin toss, but as the number of yards increases up to 30, teams are much more likely to pass rather than rush, due to the fact that they are desperate to cover more yards in their plays. Past 30 yards however, and it seems like it a coin toss once again, since the number of yards needed is so high that teams may either opt to make safer plays that won't result in turnovers (rushes), or still try to cover the amount of yards needed and extend the drive (passes).



Methods

Describes the modeling techniques you chose, their assumptions, justifications for why they are appropriate for the problem, and your plan for comparison/evaluation approaches. Additionally, you will need to describe how you will quantify uncertainty for your estimates of interest, with sufficient descriptions of the approach and justification for why it's appropriate for your data and problem of interest.

To link our data to our curiosity of playcalling strategy, we focused on 4 methods: Logistic Regression, Generalized Additive Modeling, Multimonial Modeling, and Multilevel Modeling.

Logistic Regression

As we assume a linear relationship among our covariates, a logistic regression was a relatively straightforward choice given that the nature of our question involves classifying the next best play, which is a classification problem.

$y_i \ f(y_i|\eta_i, ...), \eta_i = \beta_0 + \beta_1(p_{i-1}) + \beta_2(p_{i-2}) + \beta_3(p_{i-3}) + \beta_4(ydstogo) + \beta_5(scorediff) + \beta_6(passrank) + \beta_7(rushrank)$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9143	0.07784	11.75	7.423e-32
as.factor(prev_play_1)pass	-0.294	0.06234	-4.716	2.403e-06
as.factor(prev_play_1)run	-0.3718	0.06271	-5.929	3.039e-09
as.factor(prev_play_2)pass	-0.5452	0.06517	-8.365	6.017e-17

	Estimate	Std. Error	z value	Pr(>|z|)
as.factor(prev_play_2)run	-0.4015	0.06431	-6.243	4.299e-10
as.factor(prev_play_3)pass	0.3159	0.05615	5.626	1.846e-08
as.factor(prev_play_3)run	0.5934	0.0576	10.3	6.885e-25
ydstogo	-0.1193	0.005033	-23.7	3.763e-124
score_differential	0.01975	0.00177	11.16	6.217e-29
pass_rank	0.01432	0.00193	7.421	1.163e-13
rush_rank	-0.01443	0.001882	-7.669	1.737e-14

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	20195 on 14959 degrees of freedom
Residual deviance:	19222 on 14949 degrees of freedom

While all of our coefficients estimates are significant, it seems like the play types of the 3 previous plays brought the most influence on outcome in terms of magnitude. In specific, the 3rd previous play yielded the most change in log-odds of a successful game outcome - the change in the game outcome when the 3rd previous play was a pass compared to when it was the base condition (first_play) was 0.3159 while it was 0.5934 for a run. Since the log odds for the 1st and 2nd previous play was negative, we convert them to an exponential scale and discover that for the 1st previous play, a pass had a higher odds ratio of 0.7452765 than that of the a run. On the second play, however, a pass produced a lower odds ratio of 0.5797258 than the 0.6693153 yielded by a run. Interestingly, it seems like the odds ratio for both types of plays does not necessarily decrease or increase as k changes. More so, yards-to-go, score differential, play_rank and pass_rank seem to have more of an influence on outcome than the previous 2nd and 3rd plays. This may signify that we need to work with more intricate relationships and compare these results with the results of more complex models.

Table 3: Table continues below

(Intercept)	as.factor(prev_play_1)pass	as.factor(prev_play_1)run
2.495	0.7453	0.6895

Table 4: Table continues below

as.factor(prev_play_2)pass	as.factor(prev_play_2)run
0.5797	0.6693

Table 5: Table continues below

as.factor(prev_play_3)pass	as.factor(prev_play_3)run	ydstogo
1.372	1.81	0.8876

score_differential	pass_rank	rush_rank
1.02	1.014	0.9857

General Additive Model

Our logistic regression developed promising results, but did not necessarily account for the possibility that we had a non-linear relationship between our predictors and game outcome. By allowing us to smooth over predictors that may have non-linear relationships, we can take advantage of general additive models to flexibly model these intricate relationships.

$g(E[Y_i]) = \beta_0 + f(p_{i-1}) + f(p_{i-2}) + f(p_{i-3}) + f(ydstogo) + f(scorediff) + f(passrank) + f(rushrank)$

Table 7: Table continues below

(Intercept)	as.factor(prev_play_1)pass	as.factor(prev_play_1)run
-0.1945	-0.08644	-0.06765

Table 8: Table continues below

as.factor(prev_play_2)pass	as.factor(prev_play_2)run
-0.649	-0.5013

Table 9: Table continues below

as.factor(prev_play_3)pass	as.factor(prev_play_3)run	pass_rank	rush_rank
0.3263	0.5968	0.01483	-0.01441

Table 10: Table continues below

s(ydstogo).1	s(ydstogo).2	s(ydstogo).3	s(ydstogo).4	s(ydstogo).5
-2.489	-3.508	-2.481	2.841	-2.101

Table 11: Table continues below

s(ydstogo).6	s(ydstogo).7	s(ydstogo).8	s(ydstogo).9
-1.89	1.665	-7.651	-0.4053

Table 12: Table continues below

s(score_differential).1	s(score_differential).2	s(score_differential).3
0.365	-0.05332	0.08052

Table 13: Table continues below

s(score_differential).4	s(score_differential).5	s(score_differential).6
-0.144	-0.1041	0.005754

s(score_differential).7	s(score_differential).8	s(score_differential).9
-0.0008692	0.2834	-0.06723

Surprisingly, after smoothing for score_differential and yds_to_go, the previous 1st play type lost its significance while the ratio for both play types increased substantially. Additionally, the run had a higher odds ratio than that of a pass for the first previous play - the exact opposite as our previous model. However, we developed similar results as logistic regression for the previous second and third play types. We hypothesize that the significant change in the first previous play stems from ydstogo and score differential having a sizable interaction effect on the first previous play, which makes sense as ydstogo and score differential measure current conditions of the game, and the most recent previous play contributes more change to current situations than further plays in time.

Multimonial Regression

We proceed in experimenting with model performance with multimonial modeling, which allows the model to perform with more granularity while still maintaining the form of logistic regression. In specific, we have the presence of short/deep and left/middle/right passes and left/middle/right runs, so a multimonial model would be suitable for interpreting the effect on predictors of multiple categories on the outcome.We first filter out the NAs in the pass types, as there are more null values in our more granular predictors.

$Pr(Y = c|X) = \frac{e^{x}\beta_c}{\sum_j^c x\beta_j}$

Table 15: Table continues below

	(Intercept)	as.factor(prev_play_1)pass
pass deep middle	-0.8677	-0.07861
pass deep right	0.3305	-0.1952
pass short left	2.267	0.07894
pass short middle	1.417	0.2813
pass short right	2.072	0.01266
run left	3.036	-0.157
run middle	3.29	-0.4028
run right	2.838	-0.1908

Table 16: Table continues below

	as.factor(prev_play_1)run	as.factor(prev_play_2)pass
pass deep middle	-0.06128	-0.2089
pass deep right	-0.3435	0.1175
pass short left	0.02593	-0.5642
pass short middle	0.101	-0.4538
pass short right	-0.005989	-0.5757
run left	-0.3512	-0.9784
run middle	-0.5809	-0.9026
run right	-0.26	-1.115

Table 17: Table continues below

	as.factor(prev_play_2)run	as.factor(prev_play_3)pass
pass deep middle	0.1105	0.121
pass deep right	0.08538	0.2977
pass short left	-0.5918	0.5447
pass short middle	-0.5285	0.5005
pass short right	-0.559	0.653
run left	-0.8596	0.8214
run middle	-0.8461	0.8476
run right	-0.9037	0.8258

Table 18: Table continues below

	as.factor(prev_play_3)run	ydstogo
pass deep middle	0.09535	0.04201
pass deep right	0.1724	-0.01801
pass short left	0.4945	-0.06975
pass short middle	0.4066	-0.05108
pass short right	0.5998	-0.06537
run left	1.097	-0.1534
run middle	0.9353	-0.2508
run right	1.07	-0.1438

	score_differential	pass_rank	rush_rank
pass deep middle	-0.01007	-0.01845	0.005774
pass deep right	0.005032	-0.008449	0.003738
pass short left	-0.0005548	0.004021	-0.002277
pass short middle	-0.00449	0.002739	0.001306
pass short right	0.002082	0.004806	0.009402
run left	0.02132	0.0108	-0.01258
run middle	0.01489	0.01683	-0.002175
run right	0.02084	0.02261	-0.01667

Pass short middle was the best-outcome strategy in term of odds ratio for all the previous plays with the exception of the 3rd previous play. We do not see a lot of difference in Run directions for the previous plays or for the other predictors, but do note that a middle run has a lower log odds than other directions for score_differential and ydstogo. This is reasonable - while a middle run may bring you closer to the other side of the field, the linebackers and nose tackles on defense often clog up the middle lane, which minimizes score potential. Middle runs also have lower odds ratio more often than not in previous plays. Analysis from other football analytics gurus have alsos shown that [short passes to the middle generate the most EPA]{https://sumersports.com/the-zone/hitting-the-hard-shots-why-the-middle-of-the-field-is-the-most-effective-throw-in-football-despite-the-best-quarterbacks-succeeding-elsewhere/}, but are just generally difficult to execute due to the amount of time the ball is in the QBs hands to generate such a play, as well as defensive schemes often wandering near the middle.

Multilevel Regression

Since our effects may not be constant, we let go of the fixed effects assumption present in the previous models to create a multilevel model. Specifically, we take advantage of the model's suitability for nested

structures in accounting for dependencies among the same cluster. We also included random intercepts for the offensive and defensive teams and their interaction with each other. The idea behind this model is not only include the differences among each offensive team, but also the influences of which team is on the defense and how they interact with the the offensive team. For example, the Dolphins and the Jets have very different run tendencies, and would also exhibit different play-calling behaviors when playing the Browns versus the Seahawks.

$log(\frac{p}{1-p}) = \alpha_0 + \alpha_1(p_{i-1}) + \alpha_2(p_{i-2}) + \alpha_3(p_{i-3}) + \alpha_4(ydstogo) + \alpha_5(scorediff) + u_p + u_d + u_{pd}$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.7073	0.09268	7.631	2.33e-14
as.factor(prev_play_1)pass	-0.2891	0.06274	-4.608	4.056e-06
as.factor(prev_play_1)run	-0.3614	0.06306	-5.731	9.99e-09
as.factor(prev_play_2)pass	-0.5473	0.06546	-8.361	6.21e-17
as.factor(prev_play_2)run	-0.3956	0.06471	-6.114	9.719e-10
as.factor(prev_play_3)pass	0.2939	0.05638	5.213	1.861e-07
as.factor(prev_play_3)run	0.5766	0.05803	9.937	2.864e-23
ydstogo	-0.1193	0.005069	-23.54	1.553e-122
pass_rank	0.01021	0.003472	2.942	0.003263

Term	Variance
posteam:defteam	0.05676
defteam	0.01376
posteam	0.01819

The multilevel model generally follows that of logistic regression in that the 3rd previous play had the highest effect and the 2nd previous play had the lowest. However, we do find that the log odds seem to increase slightly for the first previous play and decrease in magnitude for the 3rd one. In terms of random effects, the interaction between offensive and defensive teams had a variance of 0.05676, which suggests that there was moderate variability on the intercepts between different combinations of offensive and defensive teams. There seemed to be slightly more variability between different offensive teams than in different defensive teams. On the other hand, correlation between fixed effects seems to be more prominent in prev_plays and yards to go than the different k previous plays with each other.

In all its entirety, our four model each brought us varying insights on the underlying relationship between our data and our goal at predicting the best play given the k previous plays. We now aim to test the validity of each of these models.

Results

Describes your results. This can include tables and plots showing your results, as well as text describing how your models worked and the appropriate interpretations of the relevant output. I do not want to you to write out the textbook interpretations of all model coefficients! I only want you to interpret the output that is relevant for your question of interest that is framed in the introduction.

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.021988 (tol = 0.002, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00464698 (tol = 0.002, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00668153 (tol = 0.002, component 1)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0437912 (tol = 0.002, component 1)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0216843 (tol = 0.002, component 1)
```

	Misclassification Rate
Model 1	0.3626
Model 2	0.3489
Model 3	0.7796
Model 4	0.3646

With the four models we trained, we decided to evaluate them using misclassification rate, given that our problem is that of classification (labeling each play as either a run or a pass for three of our models, and classifying each play with its length and location for the multinomial model). In order to account for the uncertainty of our models' performance on out-of-sample data, we performed 5-fold cross-validation on each of our models and achieved average misclassification rates, shown above. Our second model, the GAM, performs the best with a 5-fold CV misclassification rate of 0.35, meaning that on average, we expect our GAM model to incorrectly predict the run/pass about 35% of the time. Our first model and last model (GLM and GLMER models) perform very similarly, but ever so slightly worse (predicting incorrectly about 36% of the time). Our third model, the multinomial model, performs the worst with a misclassification rate of about 77%. This is likely due to the fact that our classification is more granular, as we now have nine different classes we could predict instead of the two we previously had.

Discussion

Overall it was very interesting to see how similarly our chosen models performed given the breadth of the features we chose. This may suggest that our features were not distinct enough from each other or we did not select enough features. This could be plausible as play-calling is a clearly complex process representing the battle between offensive and defensive coaching minds along with the players on the field, not to mention the field position of the play itself. As such, it is plausible that one limitation of our approach is simply that we did not perform enough vetting of the features we chose. In the future we could potentially add more features such as the ranking of the defensive team on the field versus the pass or the run (currently no defensive statistics are considered, which could be troublesome) as well as including features such as which team is home versus away, or perhaps tuning our models on different number of prior plays that the model is able to see.

Secondly, the question of what our training data should look like was something that we could potentially change in the future as well. By nature our data fails to reveal counterfactual outcomes, e.g. the offensive team induces a treatment in the form of a play call, causing some outcome in the form of yards gained, which doesn't allow for seeing what *would* have happened should a different play have been run. As a result we chose not to include what we determined to be 'failed' plays in our training process, and so we modeled the probability that a given play occurred given that it was successful. This process therefore wastes lots of data: no information from the failed plays is reflected in our trained models. Expanding our analysis to be a causal inference question in which we attempt to do some modeling on what play calls could potentially reverse or alter the outcomes of failed plays could represent an interesting statistical question under which future research could be conducted.