Understanding Pass/Run Tendencies in the NFL

Jordan Butler

DS7240: Applied Data Mining

December 11, 2025

Predicting whether a team is going to call a pass or a run on a given play is a classic classification problem that combines football strategy with modern day analytics and is something that many coaches have been able to make a whole career out of in the NFL. Play-calling is influenced by a wide range of contextual factors: down/distance, score differential, time remaining, field position, personnel groupings, team tendencies, and more. With all of these variables being taken into account by coaches attempting to give themselves a competitive edge, it only stands to reason that, along with seemingly everything else in the modern world, feeding these numbers into a computer would be able to optimize and fine tune the process as a whole. With analytics and data continuing to dominate the NFL more and more every day, it only stands to reason that this particular aspect of the game could also be an ideal setting to apply statistical analysis and machine-learning techniques. By leveraging historical play-by-play data and relevant situational variables, predictive models that estimate the probability of a pass versus a run before the snap could provide a huge competitive edge for one team versus another. Such models would not only benefit the teams using them but could also help us understand offensive behavior and coaching patterns at a level never seen before. With this being said, such models would require a vast database of numerous variables for hundreds of games. Luckily, the nflfastR dataset exists providing over 300 variables of information for every single play that has been run in the NFL dating back to 1999. As part of the larger R-based nflverse, this data can be joined with many other packages to provide even further context such as personnel packages and individual player information. This data, combined with the power of predictive modeling, should serve as a powerful foundation towards the ultimate goal of being able to predict whether a team is passing or running the ball at a more accurate level than what is possible solely through chance. With the true percentage of passing plays in the NFL being 61% (2024), that number can serve as a benchmark for predictive accuracy assuming that, with consideration to other factors, a human being would be able to predict play type over a large period of games/plays at an accuracy only slightly above 61%. All that remains to be seen is whether or not a machine-learning model would be able to beat that number and prove what we all already know to be true. Anything humans can do; computers can do better and faster.

When it comes to using computer algorithms to build predictive models, there is no such thing as too much data. For this reason, narrowing down the 372 available variables in the nflfastR dataset is a relatively simple process, with the only step being taken to ensure realistic conditions would be to eliminate any variable that would not be available to an opposing team about a specific play before the ball is snapped. For example, anything pertaining to the outcome of the play will of course not be fed to the model (excluding whether the play was a pass or run which is needed for model training purposes however is not used in the test data). Every other variable, regardless of how unimportant it may seem, will be fair game for the models to pick apart and decipher exactly how much (or little) that variable contributes to the main outcome variable itself, play_type. The resulting 34 variables chosen to be predictors in the models can be seen below in Table 1.

Table 1

Variable Name	Description	Variable Name	Description
play_type	passing or running play	goal_to_go	& goal yes/no
year	year of game	last_play_type	last play pass or run
div_game	divisional game yes/no	no_huddle	play ran from no huddle yes/no
postseason_game	postseason game yes/no	offense_formation	formation of offense (shotgun, pistol, wildcat, etc.)
week	week # of game	num_rb	# running backs on field
rain	raining yes/no	num_fb	# fullbacks on field
snow	snowing yes/no	num_wr	# wide receivers on field
temp	temperature (F)	num_te	# tight ends on field
pos_team_home	home or away game for possession team	defenders_in_box	# defenders in the box (area near line of scrimmage)
quarter_seconds_remaining	seconds left in quarter	num_db	# defensive backs on field
half_seconds_remaining	seconds left in half	posteam_timeouts_remaining	# timeouts for possession team
game_seconds_remaining	seconds left in game	defteam_timeouts_remaining	# timeouts for defense team
game_qtr	quarter #	score_differential	current difference in score for possession team (+ if leading, - if losing)
game_half	half #	game_pass_pct	% of total plays that were passes so far in game
drive_play_num	# play in drive	game_pass_epa_per_play	Average expected points added per pass so far in game
game_play_num	# play in game	game_rush_epa_per_play	Average expected points added per rush so far in game
down	down		
ydstogo	years to first down marker (yards to goal line if & goal)		
yardline_100	# yards away from opposite endzone		

While the variable selection process was relatively straightforward, there were a handful of filters that needed to be applied. Firstly, the data used could only be from 2016 onwards as this was when the offensive formation variable were initially added to the dataset. Additionally, all plays ran in overtime were filtered out due to missing/strangely classified variables for certain observations. Neither N/As not duplicates were an issue, however only plays that were either a pass or run were fed into the models (excluding play types like special teams, 2 point

conversions, etc.). Some variables had to be manually created such as ones that could be converted into a binary variable or ones that required classification groupings such as weather which originally was a string of words simply describing the weather. After all pre-processing was completed though, the data was comprised of 36 total variables for 300,921 individual plays.

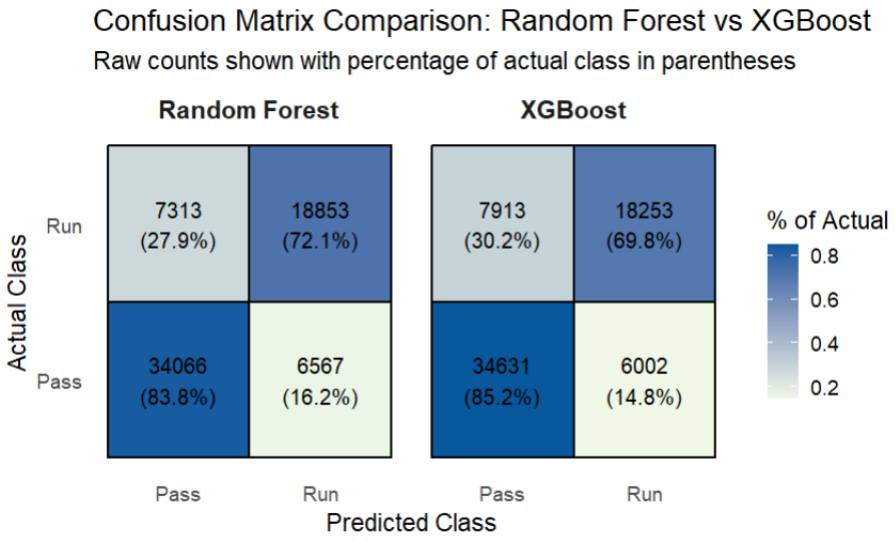
In order to capture the trends in the data as accurately as possible, both a RandomForest model and XGBoost model were created to cover the possibilities of both underfitting and overfitting. The RandomForest model will serve as the more conservative of the 2 hoping to be more stable, whereas the XGBoost model serves as the aggressive one, with its boosting algorithm framework allowing it capture more subtle trends. The data was one-hot encoded and then split with the 2016-2022 seasons serving as the training dataset and the 2023-2024 seasons serving as the testing dataset (77.78%/22.22% train/test split). The year variable was included with a numeric data type to accurately capture any trends existing on a chronological basis (with the idea in mind that the NFL has slowly but surely been becoming more pass heavy over the past couple of decades). In order to test the differences in predictive capabilities between the models, both accuracy and AUC will be measured. The accuracy will serve as a raw depiction of predictive capability simply indicating what percentage of plays in the testing dataset were predicted correctly. The AUC score will serve as a more context-based metric taking into account the confidence of the models for each prediction that will weight observations of the same prediction class differently depending on the model's confidence (predicted probability) in that prediction. The resulting accuracy and AUC scores can be seen below in Figure 1.

Figure 1

Model	Accuracy	AUC
Random Forest	0.7922125	0.8697485
XGBoost	0.7916885	0.8741405

Surprisingly, both the RandomForest and XGBoost models perfect almost exactly the same by both standards, with accuracies of 0.7922 and .7917 respectively both blowing the target accuracy of .061 out of the water. Furthermore, the AUC scores of 0.8697 and 0.8741 show that the models class separation abilities performed equally well. Predictive accuracy can be drilled down to even further to by looking at the differences in the classes themselves for each model. These results are shown below in Figure 2.

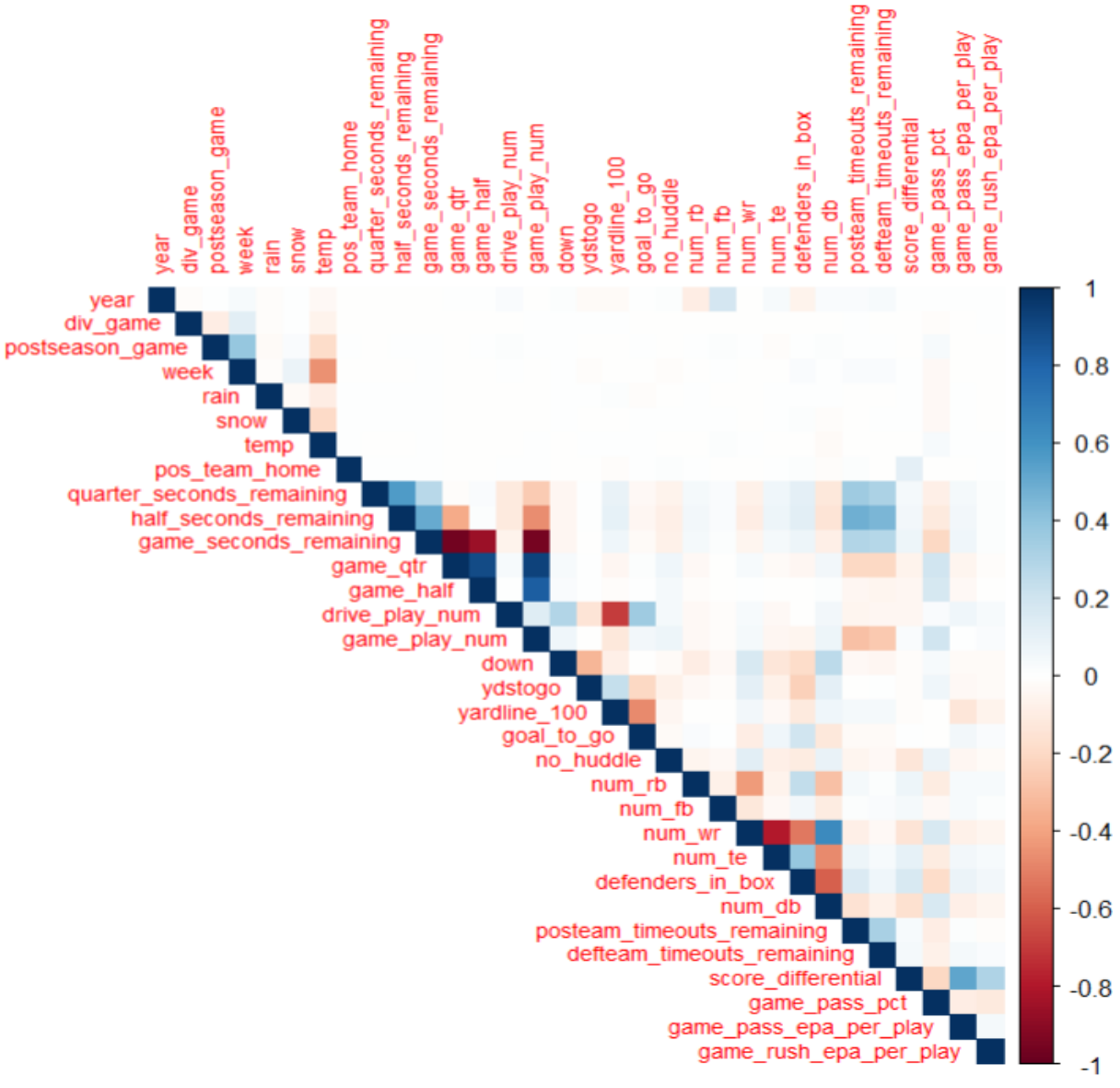
Figure 2



It is clear that both models were able to predict passes correctly at a much higher rate than runs. It is also interesting that the conservative model (RandomForest) was able to predicted runs at a higher percentage than the aggressive model (XGBoost) and vice versa. This difference highlights how the overall structural formatting of the trees in each model perceives runs and passes differently depending on the model's aggressiveness. However, while the main goal of this study is to focus on raw predictive capability, being able to highlight the actual trends and patterns in the dataset according to the models is something that could also prove something of interest. Though XGBoost model did have a higher AUC, the RandomForest model will be the main focus for all post-hoc analysis with accuracy still being the main metric of interest for the study as a whole and there not being enough of a difference in class-specific predictive capability.

Checking for multicollinearity is a good way to show any potential direct correlation between 2 variables. While tree-based models are typically quite good at filtering out any multicollinearity, the test can still give valuable insight into showing correlated variables that one might not think would have any correlation at all. A correlation plot of all numeric variables can be seen below in Figure 3.

Figure 3



Certain obvious correlations show almost perfect collinearity such as the number of seconds left in the game being negatively correlated with the quarter and the number of wide receivers being negatively correlated with the number of tight ends. Listed below are some of the more interesting correlations that can be noted from the correlation plot.

- Perfect summarization of the demise of fullbacks as shown by the negative correlation between year and num_fb complimented by the inversely positive correlation between year and num_rb, also shows that defenses can been putting more people in the box as the years have gone on
- A slightly more humorous relationship is showing just how severe the negative correlation is between week and temperature as it obviously gets much colder into the month like December and January
    - This likely explains the slight negative correlation between week and game_pass_pct
        - In addition to temperature, both the presence of rain and snow show a slight decrease in game_pass_pct, however not as much as would be expected (likely due to their not being an indicator showing the severity of rain/snow as a slight drizzle almost certainly would not affect the pass % as much as a downpour)
- The slight positive correlation between pos_team_home and score_differential confirms the notion of home field advantage and perfectly captures why teams are given a 3 point boost in spread lines solely for being the home team
- The percentage of passing plays shows a negative correlation with the number of seconds left in the quarter, half, and game, showing that teams tend to run the ball more earlier in the game and pass the ball more as the game progresses
- The amount of plays run from no huddle offense increase as time goes on as expected (teams hurrying when they are losing), however this is likely something that wouldn't exist in college football where more teams run no huddle hurry up offense randomly throughout the game just to try to throw the other team off
- The existence of a negative correlation between number of defenders in the box and game pass percentage shows that a level of play type predictiveness already exists, as stacking the box is a common strategy when you suspect a run is coming
- Average epa per pass shows basically the exact same correlations as average epa per run except more severe, highlighting that proficient passing will always lead to more expected points than proficient running
    - Interestingly enough though, no relationship exists between year and percentage of passing plays like one might expect. If the dataset were to be expanded to include all plays going back to 1999 though, it would likely tell a different story

Overall, the correlation plot provides a great deal of insight into the passing tendencies of the NFL and how they relate to various game situations which in part offers a glance into the thought process of the RandomForest model on an elementary level.

Diving further into the thought process of the model, a look at exactly which variables the model found most important when deciding whether a play was a run or a pass adds yet another layer of information on top of showing how the variables themselves were correlated with one another. The importance metric is determined by randomly shuffling the values of that column

within the dataset and then retesting the accuracy. The less accurate a model is after the reshuffling, the more important the variable is. Additionally, while the importance numbers themselves are simply relative magnitudes and don't have any practical meaning, you can compare the numbers amongst themselves to determine relative importance between variables. The list of all 34 predictor variables and their importance is shown below in Figure 4.

Figure 4

	Variable	Importance
1	offense_formation	6.677322e-02
2	game_pass_pct	3.566746e-02
3	down	2.624663e-02
4	game_seconds_remaining	2.327976e-02
5	half_seconds_remaining	2.136176e-02
6	game_play_num	1.580868e-02
7	last_play_type	1.462213e-02
8	ydstogo	1.384421e-02
9	quarter_seconds_remaining	1.270118e-02
10	defenders_in_box	1.249511e-02
11	score_differential	1.247754e-02
12	game_pass_epa_per_play	1.105598e-02
13	game_qtr	8.567223e-03
14	drive_play_num	7.411141e-03
15	num_wr	6.710003e-03
16	num_db	5.809030e-03
17	game_rush_epa_per_play	5.121656e-03

	Variable	Importance
18	yardline_100	5.096058e-03
19	game_half	4.287750e-03
20	posteam_timeouts_remaining	2.213231e-03
21	num_te	1.781895e-03
22	num_rb	1.066530e-03
23	defteam_timeouts_remaining	9.117908e-04
24	temp	7.520354e-04
25	week	6.417868e-04
26	year	6.310094e-04
27	no_huddle	6.005816e-04
28	goal_to_go	4.584438e-04
29	div_game	7.868018e-05
30	pos_team_home	7.818915e-05
31	rain	4.703747e-05
32	postseason_game	4.683324e-05
33	num_fb	0.000000e+00
34	snow	-4.411533e-07

The variable importance numbers show that the most important thing in determining a run or a pass is the offensive formation, and that it is almost twice as important as the second most important variable, percentage of passing plays so far in the game. If level-specific analysis were completed, it would likely show that passes are more likely to occur from the shotgun formation and runs are more likely to occur from the pistol (which is common knowledge but still nice to see it statistically confirmed). Some other very important variables are how often a team has already passed so far in the game, what down it is, how much time is left in the game, and whether the last play ran was a pass or a run. On the other end of the spectrum are the least important variables, most of which are to be expected such as the number of fullbacks (which no one uses anymore, sorry Kyle Juszczyk), being the home vs away team, and whether it is a regular season or postseason game. There are certain variables that are surprisingly low in importance, however. As mentioned earlier, the likely reason why rain and snow are now as important is due to the lack of a rain/snow intensity metric which makes sense. However, whether or not the team is in an & goal situation not being labeled as important, as common knowledge might suggest that a team is more likely to run the ball when they are that close to the goal line, but the variable importance metrics shows that is not the case. Additionally, metrics

like game_pass_epa_per_play (passing efficiency) and game_rush_epa_per_play (rushing efficiency) being listed below variables like down, yards to go, and time remaining show that the game situation present for a play is more important than things like how well a team has been passing or running the ball so far that game, which is somewhat surprising, but not totally unexpected. Overall, the variable importance numbers help highlight the decision making process of the RandomForest model and quantifying it in such a way that would other be impossible to ascertain.

As a final step of post-hoc analysis, one can expand on the variable correlation by looking into exactly how the model grouped variables together by performing a Principle Component Analysis and seeing which variables the model put into the principle components that explain the majority of the variance within the data. This serves as a sort of combination of the previous post-hoc analysis techniques by not only "grouping" related variables together, but also showing how much of the variability in the outcome variable can be explained by that group of predictors. This concept of dimensionality reduction that PCA accomplishes is perfectly highlighted by the added interpretability it accomplishes through clearly showing which groups of variables are the most important as a whole rather than individually. A scree plot showing the percentage of explained variance by principle component, along with a list of the variables grouped within each of the top PCs, are shown below in Figures 5 and 6.

Figure 5

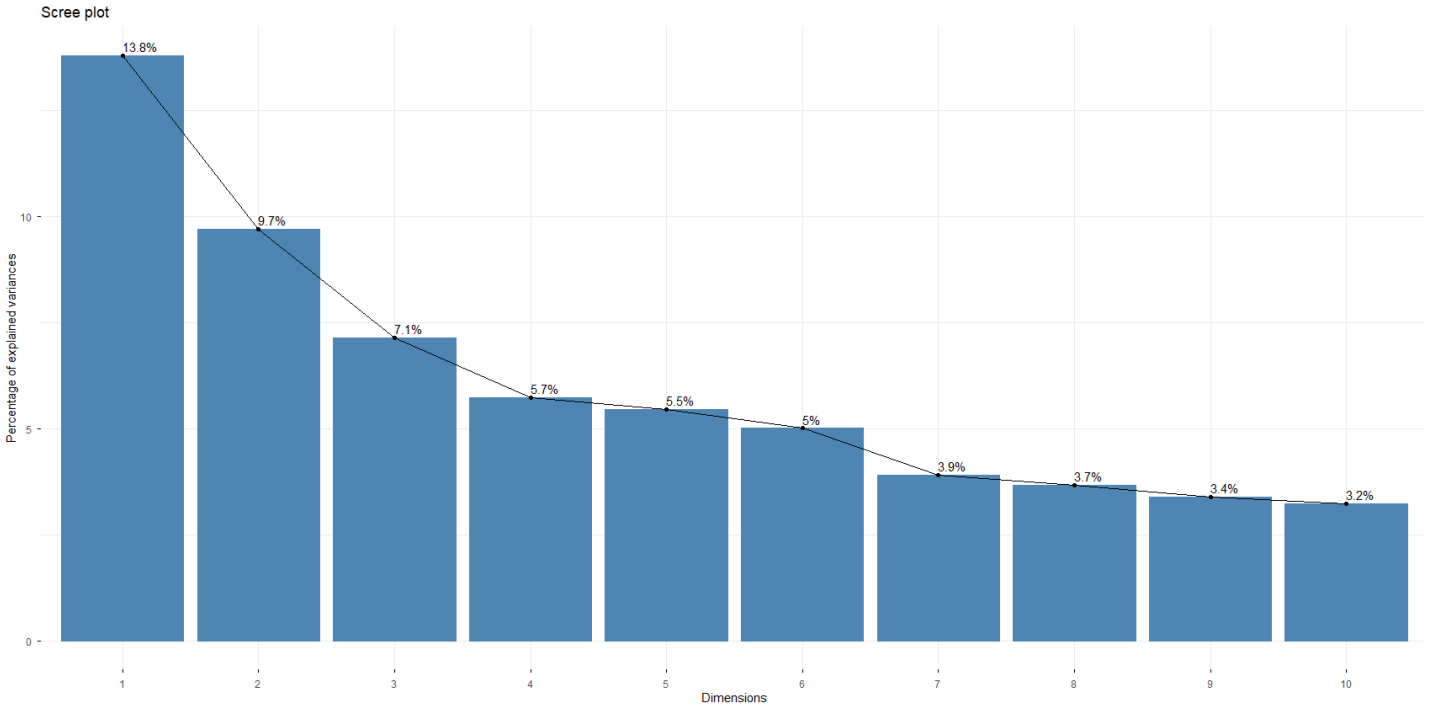


Figure 6

PC1

	Variable	Contribution
1	game_seconds_remaining	0.441199430
2	game_play_num	0.427862424
3	game_qtr	0.408964593
4	game_half	0.337494476
5	half_seconds_remaining	0.288903160
6	posteam_timeouts_remaining	0.212867765
7	defteam_timeouts_remaining	0.194818580
8	quarter_seconds_remaining	0.185067692
9	num_db	0.156913034
10	game_pass_pct	0.154572524

13.8%

PC2

	Variable	Contribution
1	num_wr	4.548574e-01
2	num_db	4.144015e-01
3	defenders_in_box	3.924568e-01
4	num_te	3.631380e-01
5	num_rb	2.001711e-01
6	game_half	1.968238e-01
7	game_qtr	1.874087e-01
8	game_play_num	1.855550e-01
9	game_seconds_remaining	1.696494e-01
10	score_differential	1.682164e-01

9.7%

PC3

	Variable	Contribution
1	yardline_100	0.497184701
2	drive_play_num	0.492569747
3	goal_to_go	0.366584535
4	ydstogo	0.272176406
5	game_half	0.242408274
6	down	0.238299476
7	quarter_seconds_remaining	0.211975881
8	game_qtr	0.178942761
9	half_seconds_remaining	0.176546091
10	posteam_timeouts_remaining	0.135640070

7.1%

A look at the scree plot in Figure 5 shows that the "elbow" or point in the bar graph in which there is a significant drop off in percentage of variance explained per additional PC, appears to be right after PC 3. These first 3 PCs combined explain just under 1/3rd of the total variance observed in the outcome variable at a total of 30.6%. Looking further into the variables listed in each of the PCs, a clearly defined "group name/theme" emerges from each of them. PC1 can be described as "game situation at time of play" consisting of variables such as how far into the game it is (combination of time remaining and number of plays run so far). PC2 can be described as "offensive/defensive formation" consisting of variables such as the number of each offensive skill position on the field and number of defensive backs/defenders in the box. The fact that the defensive numbers are contributing just as highly if not higher as a whole to the PC could be indicative of audibles being called at the line of scrimmage by the qb based on the defensive setup. Finally, PC3 can be described as "drive situation", which differs from PC1 by focusing more on things like down and yards, along with where they are on the field and how far into the drive they are. While ultimately, grouping these variables together like this likely isn't as effective as looking at the variables individually, the PCA results do provide an alternative method of detailing predictors at a grouped level rather than individually which is something that can still provide valuable insight in the context of this study.

In summary, the idea that anything humans can do machines can do better is ultimately confirmed with predictive accuracies almost 20% higher than what could be achieved through pure chance. The additional information gained through the post-hoc analysis consisting of a correlation plot, variable importance, and PCA provide a clear and concise description that, while maybe not contributing anything additional to the raw predictive capabilities of the model, do

provide valuable insight into the decision making process of the models. Additionally, these insights can be quite interesting for those not necessarily interested in the predictive capabilities of the models outlined in this study but rather looking for detailed information that could improve their overall knowledge of the game and enhance their viewing experience at home. While ultimately a relatively simple model in terms of peak machine-learning capabilities, this study as a whole serves as a stepping stone to understanding what is truly possible when combining the power of machine and man when it comes to the ever-growing presence of analytics in the NFL today.

References

nflfastR package: http://nflfastr.com

nflreadr universe: https://nflreadr.nflverse.com