SYBIL - the NFL play predictor

Milestone Report

Viraj P Modak

December 08 2019

CONTENTS

1. Problem Statement:

Imagine you are the defensive coordinator (DC) of an NFL team, devising a strategy for a must win or a play-off game. During the game, based on the situation the opposition offense is going to call and execute plays. Your goal is to disrupt those plays or minimize their impact. That defines the success criteria of your job as DC. In this situation, imagine that you have a tool to accurately predict what the opposition play is going to be. That would be a huge tactical advantage on the field. Sybil - the NFL play predictor is that tool.

To add more context to the problem, in an NFL game, there are multiple factors which can decide the offensive play call. They include the play clock, down, yards to go, yards to goal among others. In obvious situations, an experienced practitioner of the game can predict the play call. However, in situations with a significant degree of uncertainty, a human brain may not be able to process all the information and that's where Sybil's data driven decision making and use of ML techniques can help in the decision making.

The target customers for Sybil would primarily be the NFL coaches, in particular the defensive coordinators and their support staff. However, NFL analysts, and enthusiasts such as Fantasy Football players are also potential users of Sybil.

2. Data to be used:

To train Sybil, the **"nflscrap-R play-by-play"** data will be used. It is available freely on GitHub. Data is recorded for every NFL game (regular and post-season) from 2009 through current. Analysis will be restricted to the end of 2018-2019 season. Data can be found at the following link. No code files or data files are uploaded separately with this milestone report. But the relevant links are provided in the Appendix.

https://github.com/ryurko/nflscrapR-data/tree/master/games_data/regular_season

3. Problem solving approach:

Sybil is a predictive model which can predict a Run/Pass play based on the game situation. The game situation can be summarized using multiple variables including:

1. Down
2. Yards to go
3. Field position (Yards to goal)
4. Play clock
5. Score line
6. Play formation
7. Team pass performance in the game
8. Team run performance in the game
9. Dual-threat rating of the quarterback (QB)
10. Timeouts remaining

The data clean-up process is describes in Section 4. Data was cleaned up to a format which lists out the individual variables and the final outcome. Some of the factors e.g. QB dual threat rating were calculated independently. The Exploratory Data Analysis (EDA) and Statistical Analysis results are reported in Section 5 and Section 6 respectively.

The data will eventually be used to train Sybil using one or more ML techniques. Sybil's performance will then be evaluated using standard ML performance metrics. These two aspects are not included in this milestone report.

4. Data Wrangling/Clean-up

The data wrangling approach and techniques used to prep the data is described in the following steps. The prepped data will then be used to train Sybil.

a. CSV files were read into a single data frame. These include the player roster and the actual play by play data

b. The dataframe was then cleaned up to remove entries which are not plays. These include Timeouts, Two-minute warnings, End of game/QTR, game suspensions and resumptions. The str.startswith function was used for this purpose This also includes plays which do NOT have a time clock associated with them. The data is read as string, which caused the missing values to be read as empty strings. These values were replaced by np.nan and then removed using the dropna method of a data frame. Additional features such as play clock, run/pass performance and the dual threat rating were calculated using individual functions

c. The play_clock was represented as either MM:SS or MM:SS:00 and was read as string in the original data frame. This was then split into minutes and seconds using the str.split function and the minutes and the seconds were then converted to float using astype. The final play clock in minutes was then calculated using a simple arithmetic operation over the floats: Minutes = Minutes+Seconds/60

d. Whether a play is a pass/run can also depend on how the teams pass/run performance has been until that point in the game. To calculate the pass yardage, the string value was first converted to numeric using pd.numeric. Following this, groupby and apply functions were used to calculate the cumulative pass completion % and the cumulative pass yardage - grouped by team and game. There were stray non-numeric or 'NA' values for cases such as incomplete passes. These were addressed using errors='coerce' wherever applicable. This operation was repeated to calculate the cumulative run yards grouped by team and game.

e. To calculate the dual threat rating of a team, the dual threat performance of individual quarterbacks was first calculated. For this, total yardage and run attempts were calculated for QBs over the entire season using the groupby and agg functions. Now the dual threat factor was then calculated using a simple arithmetic calculation as a product of yardage and run

attempts, each normalized by the games played. The final dual threat rating was a percentile count for all QBs going from 0 through 1.

f. The dual threat rating calculation for a team in a game proved quite tricky because of the possibility of multiple QBs playing in the game. The QB-specific dual threat ratings were first converted to a default dictionary using zip and defaultdict which was then called using the map function. The contribution of each QB for a team in the game was estimated using the pass count, once again using groupby and agg. The team's dual threat rating was then calculated as a weighted average of the QBs' dual threat rating over the contribution. A separate local function was written for this purpose and then called using apply.

The final data frame was then saved as a CSV file. This will be used to perform EDA, Statistical Analysis and eventually train Sybil. The results from the EDA and statistical analysis are reported in Section 5 and Section 6 respectively.

5. Exploratory Data Analysis Results

The EDA approach is described in a fully executed notebook and the link for the same is provided in the Appendix. Key highlights and plots are shown here.

1. When segmented by quarters and downs Q1 is when teams are most conservative and prefer to run the ball the most. Q2 and Q4 is when teams are more likely to pass the ball and 3rd down is when the teams are most likely to pass. This is shown in Figure 1.

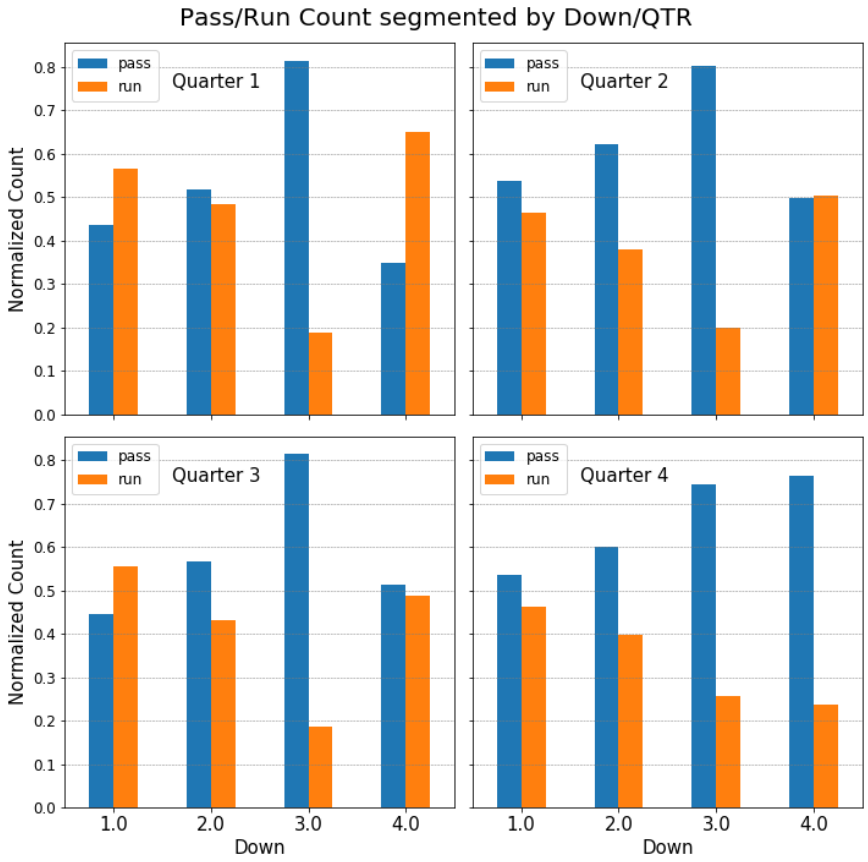


Figure 1: Count of pass/run plays segmented by play down/qtr

2. Moving to yardage - in case of long yardage situations as expected, teams prefer to pass on 3rd and 4th downs. Run plays on 3rd and 4th down do happen but only in short yardage situations. On 1st and 2nd downs, visually it is hard to distinguish between pass and run preference. This behavior is shown in Figure 2.

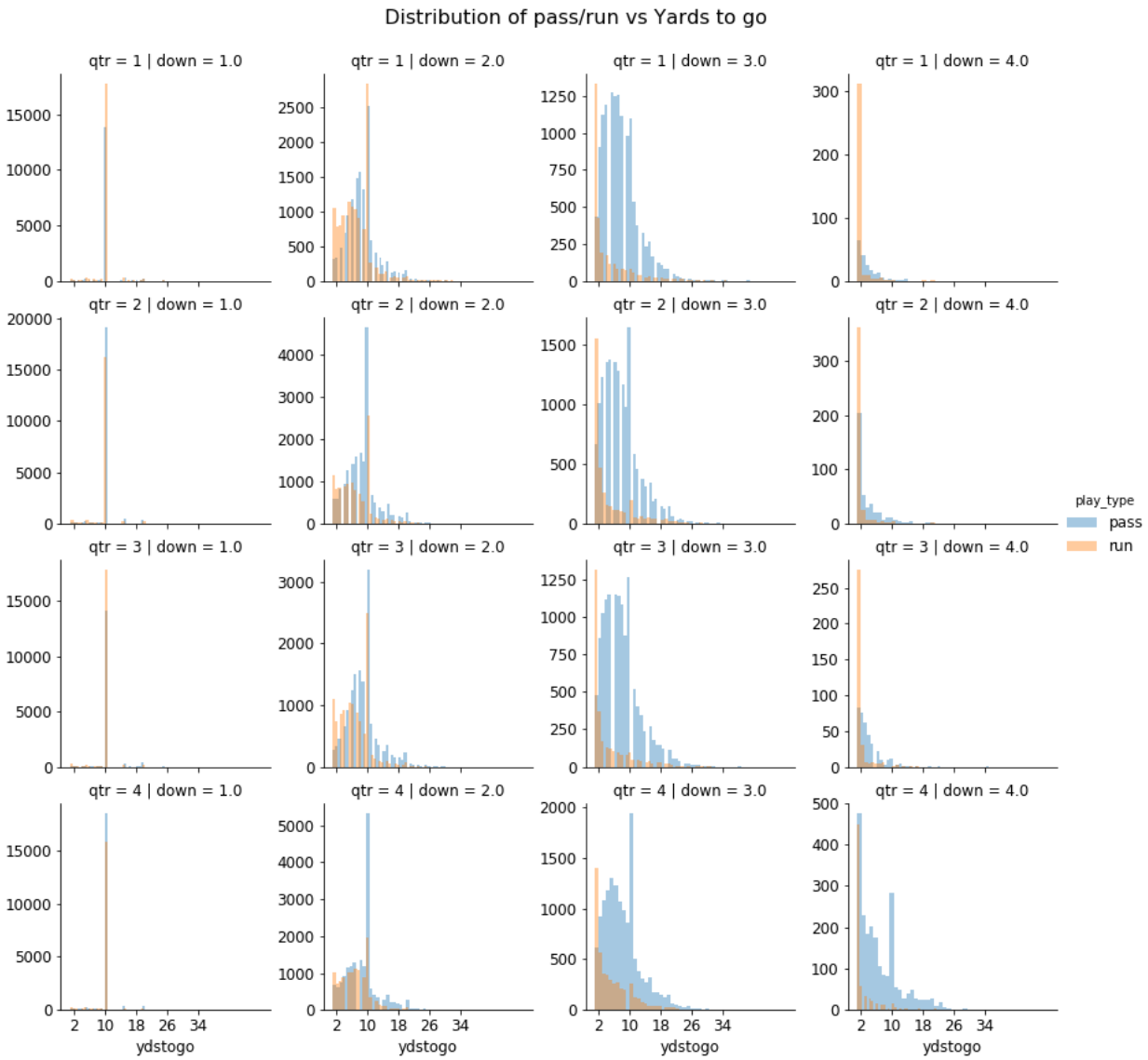


Figure 2: Count of pass/run plays segmented by play down/qtr and yards to go

3. Teams are more likely to run the ball if the score difference is positive and more likely to pass if it is negative. This is expected and which is shown in Figure 3.

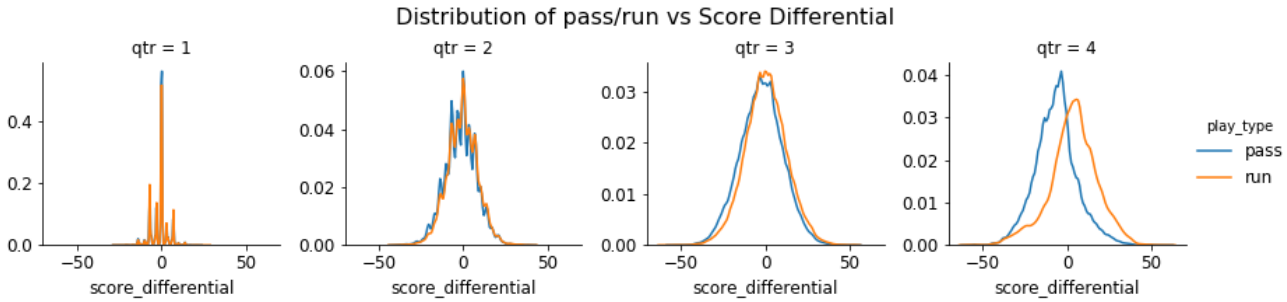


Figure 3: Count of pass/run plays segmented by play down/qtr and yards to go

4. The effect of timeouts is also explored and the tendency of teams to pass increases (and that to run decreases) consistently as they start using timeouts, which is shown in Figure 4

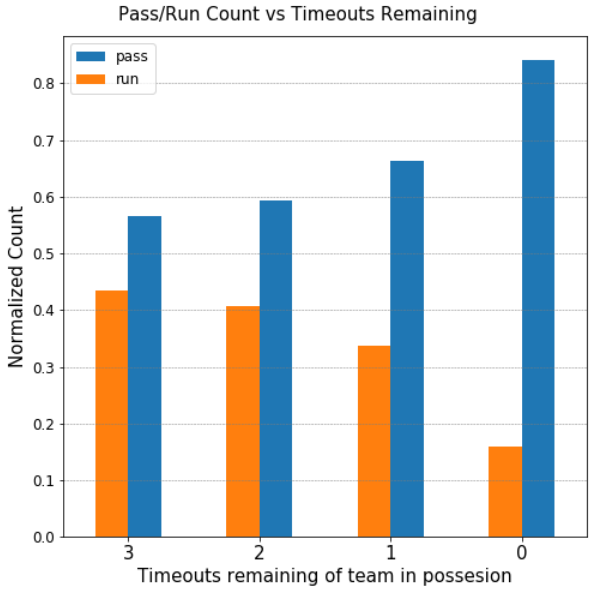


Figure 4: Count of pass/run plays segmented by play timeouts of team in possession

5. No strong visual correlation is seen between pass vs run with respect to distance from goal. However, 4th down plays happen if the spot is within specific areas of the field. And there is a very slight preference to pass on 4th down, in Q2 and Q4, if you are further away from goal

6. The effect of previous pass/run performance in the game was also tested. This is not a feature directly available in the raw dataset but had to be calculated independently. It will be seen in section 6, that the effect, although small, is statistically significant. The relevant data is shown in Figure 5.

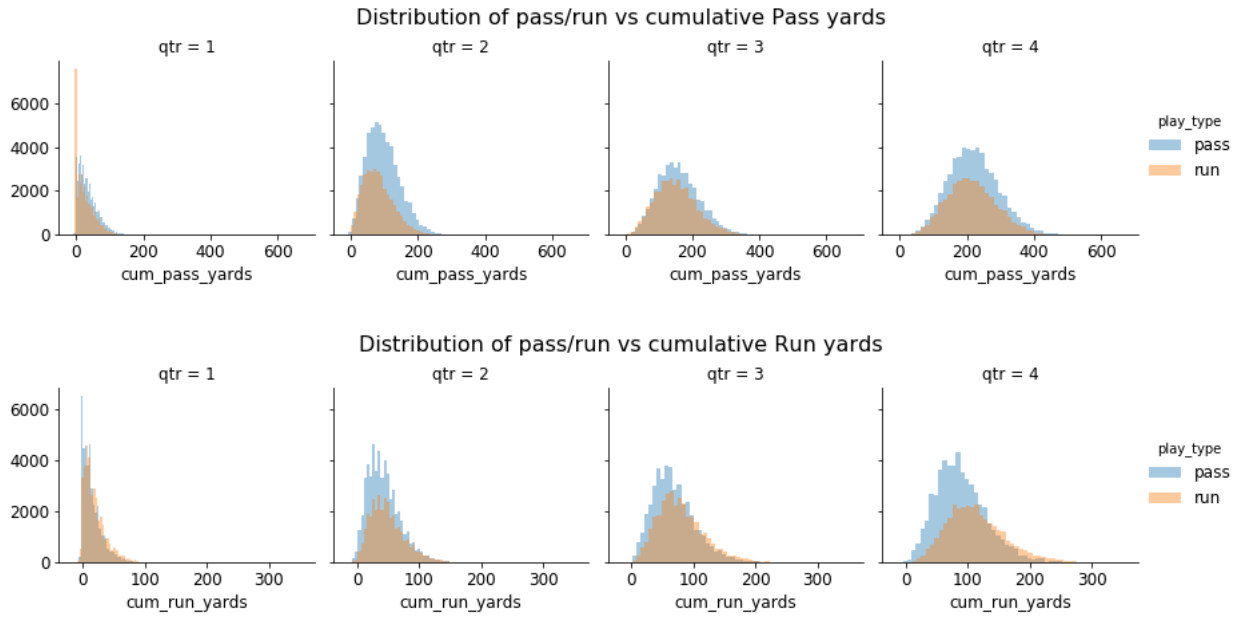


Figure 5: Count of pass/run plays segmented by quarter and previous pass/run performance

7. Another feature which was built and calculated independently, was the dual threat rating of the Team QB. Once again, the difference is not obvious on visual inspection - as seen in Figure 6, but is statistically significant as will be shown in Section 6.

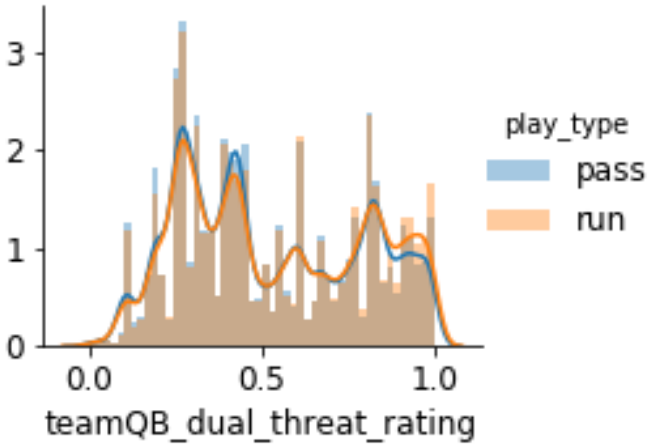
8. 

Figure 6: Distribution of pass/run plays segmented by team dual threat rating

6. Statistical Analysis

Statistical analysis was performed on continuous variables such as score differential, yards to go or distance to goal, along with hypothesis testing. The results from these report will support the conclusions drawn from the EDA in Section 5. In multiple cases it will be seen that the distributions of the variables are not normal. However, according to the central limit theorem, statistical tests and parameters designed for normal distributions can still be used.

6.1 Yards to go

Based on the plot shown in Figure 2, it is worthwhile to look at quarter 3 and possibly quarter 4. Given the distribution a t-test would be appropriate. The null hypothesis in this case is that the compared means are equal. The key statistics, p-values and the statistical inferences are shown in Table 1. To calculate the p-values we assumed that the variances were unequal

Table 1: Statistical analysis of yards to go vs qtr/down by play type

qtr	down	pass_mean	run_mean	t-test_pvalue
3	1	10.284213	9.828213	6.03E-83
3	2	8.936741	6.735528	0.00E+00
3	3	7.817511	4.561672	3.39E-180
3	4	3.868852	1.829971	1.25E-18
4	1	10.087728	9.72813	1.11E-58
4	2	8.674929	6.91863	2.56E-275
4	3	7.90534	5.397723	9.90E-186
4	4	7.016202	2.767442	7.01E-104

We can see that all p-values are very low, which means that indeed, the difference between yards to go for pass and run is statistically significant. A similar analysis was done for distance from goal, but as consistent trends/differences were not noted.

6.2 Score differential

To perform a statistical analysis on score differential, data from only quarters 2, 3 and 4 because the distributions resemble those of a continuous variable. Once again the null hypothesis is that compared means are equal. The key statistics, p-values and the statistical inferences are shown in Table 2 and once again, the variances are assumed to be unequal.

Table 2: Statistical analysis of score differential vs qtr by play type

qtr	pass_mean	run_mean	t-test_pvalue
2	-1.566904	-0.583737	2.98E-55
3	-2.27387	0.37842	7.41E-189
4	-6.564699	3.633725	0.00E+00

It is interesting because, to the human eye it is near impossible to spot a difference between the two means except for quarter 4. However, even for qtr 2 and 3 we do see that the differences between means is statistically significant. Furthermore, from a practical standpoint, it can be seen that the p-value approaches zero with each passing quarter and during qtr 4, the difference is beyond doubt as the p-value is zero. Play quarter (which is used here to segment the data) is a discrete quantity and it would be interesting to note how the difference behaves with respect to the play clock which is a continuous variable. That analysis will make for an impactful visual but for simplicity is not shown here.

6.3 Cumulative Run/Pass performance:

The next item we can investigate is how the play type depends on the teams pass performance and run performance. The hypothesis in this case can is a little tricky to frame but can be expressed as follows: difference between pass and run performance for pass plays is greater than difference between pass and run performance for run plays. It can be represented as follows:

$\mu_p > \mu_r$

where,

$$\mu_p = \langle cumulative\_pass\_yards - cumulative\_run\_yards \rangle_{pass\_plays}$$
$$\mu_r = \langle cumulative\_pass\_yards - cumulative\_run\_yards \rangle_{run\_plays}$$

The results for the 1-sided t-test are listed in Table 3. In this case the p-value is 1, which means that the likelihood of null hypothesis being true is certain.

Table 3: Comparing difference between pass and run performance for pass and run plays

diff_for_pass_play	diff_for_run_play	t-test_statistic	t-test_pvalue
75.859833	52.317236	87.260388	1

6.4 Dual threat rating:

The hypothesis in this case will be that the teams with higher dual threat rating prefer to run the ball more than teams with lower dual threat rating. It can be framed as follows:

$\mu_{run} > \mu_{pass}$

μ_{pass} = mean dual threat rating for all pass plays

μ_{run} = mean dual threat rating for all run plays

The results shown in Table 4 show that the difference in the means is really subtle at face value - also seen in Figure 6 - but statistically significant with the large sample size since the p-value is 1.

Table 4: Compare dual threat rating for pass plays and run plays

dual_threat_mean_run	dual_threat_mean_pass	t-test_pvalue
0.538057	0.520613	1

In summary we can say that, just by doing exploratory data analysis, it is possible to extract meaningful trends from subtle differences. However, given the large sample size, these differences are indeed statistically significant.

7. Summary

We started with the objective of designing a play predictor which can potentially be used by NFL coaches/defense coordinators or NFL enthusiasts, analysts and fantasy football players. Play-by-play data was obtained and cleaned up in the form of multiple independent variables - or features - and a response variable of play type, which can be pass or run. Some of these features had to be calculated independently. The effect of these variables on the play type was explored on the play type using visual plots as well as statistical tests.

The next step would be to use the data to train using multiple ML techniques. These will include Logistic Regression, Support Vector Machines, K Nearest Neighbors, Decision Trees and Random Forests. The model can be tuned by adjusting relevant hyperparameters and the performance can be tested using standard ML metrics and the best performing model will be used as the final product.

Appendix

Link to the cleaned-up dataset:

https://drive.google.com/open?id=1sL2ZLX1BU7o800UiPCSJ2n84D3Y5GUGJ

The raw dataset is not uploaded to a shared location for this project, but can be accessed freely on the link provided in Section 2

Link to the Python files for data wrangling/clean-up:

https://github.com/virajmodak16/NFL_Play_Predictor/blob/master/Python_Code.zip

Link to the Jupyter Notebook for Exploratory Data Analysis:

https://github.com/virajmodak16/NFL_Play_Predictor/blob/master/VirajModak_Capstone1_DataStorytelling.ipynb

Link to the Jupyter Notebook for Statistical Analysis:

https://github.com/virajmodak16/NFL_Play_Predictor/blob/master/VirajModak_StatisticalAnalysis_20191128.ipynb