NFL Play Predictor - Statistical Analysis

Viraj P. Modak

To recap, we started with following problem statement; predicting the type of play (pass/run) in an NFL game based on the game situation. We can represent the "game situation" in the form of multiple discreet and continuous variables. For this purpose, we cleaned up the raw play-by-play data available on the web for all NFL games for the past 10 years. We then performed Exploratory Data Analysis to identify trends in the variables. In this write-up we will report on the statistical analysis on the pertinent variables. The variables had diverse types - continuous as well as discreet and/or categorical. We were indeed able to explore the relationship between some of these variables and the play type as shown in the data wrangling report. For example, we saw a strong influence of Quarter, Down, Yards to go and Timeouts remaining. Somewhat of an influence of score differential and a weak influence of distance to goal, previous run/pass performance and the dual threat rating of the team QB.

Although the cleaned dataset is consistent and hence analyzed as a whole, it can be divided into subgroups - "regular season games" and "post season games". This may lead to differences in trends noticed because teams can behave differently in a "must-win" play-off scenario compared to a regular season game. But the regular season games outnumber the play-off games by a factor of 25. Hence any comparison between the two sub-groups may be heavily biased.

With this background we can now proceed to perform statistical analysis on our dataset. However, not all variables may lead to meaningful results. For example, we explored the effect of quarter/down as shown in figure S.1 in the supplementary material. Given that both these variables are categorical, comparing frequencies rather than means or variances makes sense. Same is the case with timeouts.

When it comes to continuous variables such as score differential, yards to go or distance to goal, we can indeed perform a more robust analysis along with hypothesis testing. We can use these results to support our conclusions from exploratory data analysis. In multiple cases we will see that the distributions of the variables are not normal. However, according to the central limit theorem, we can still perform statistical tests designed for normal distributions.

1. Yards to go

Based on the plot shown in figure S.2 of the supplementary material, 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 would be that 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.

2. Score differential

We can perform a similar analysis for the score differential, in this case though we can look at quarters 2, 3 and 4 because the distributions resemble those of a continuous variable. Once again the null hypothesis in this case would be 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, we can see 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 is a discreet 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.

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

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 S.3 - 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.

Supplementary material

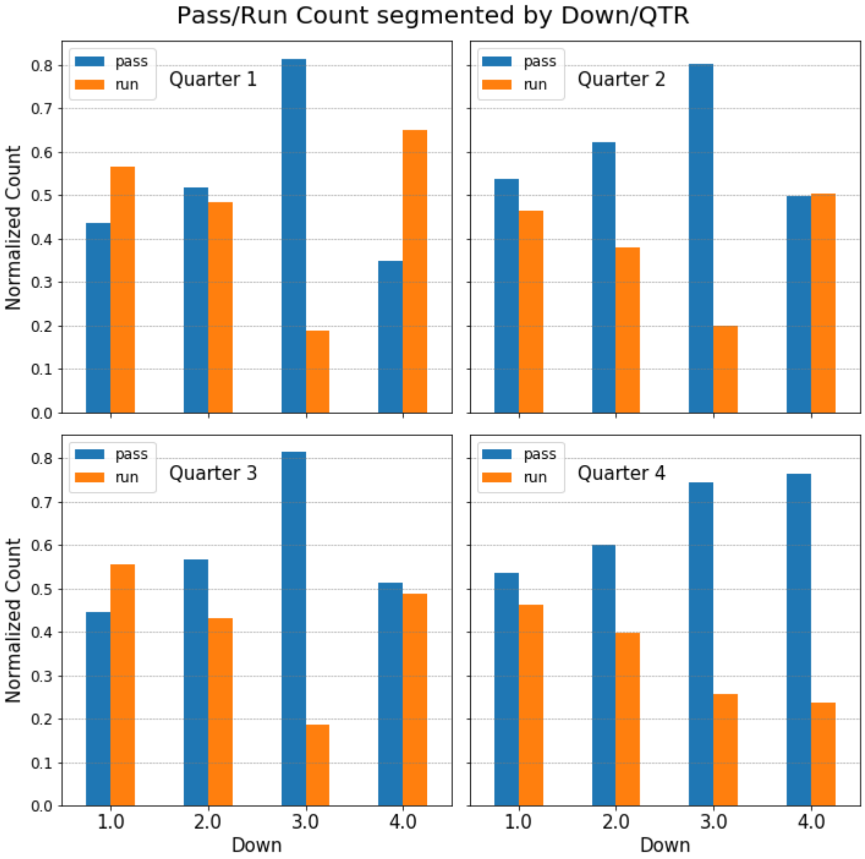

Figure S.1: Count of pass/run plays segmented by play down/qtr

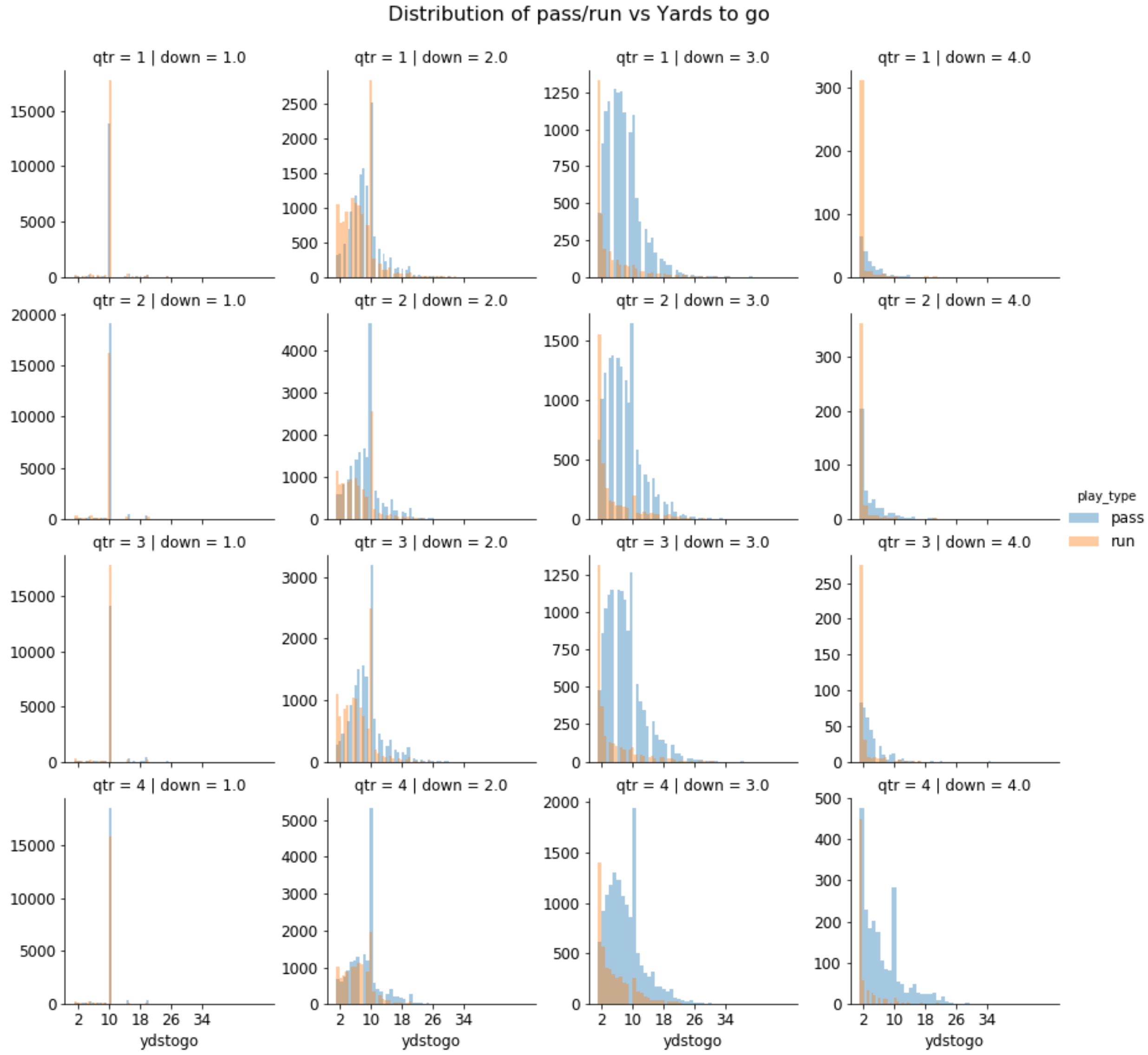

Figure S.2: Count of pass/run plays segmented by play down/qtr and yards to go

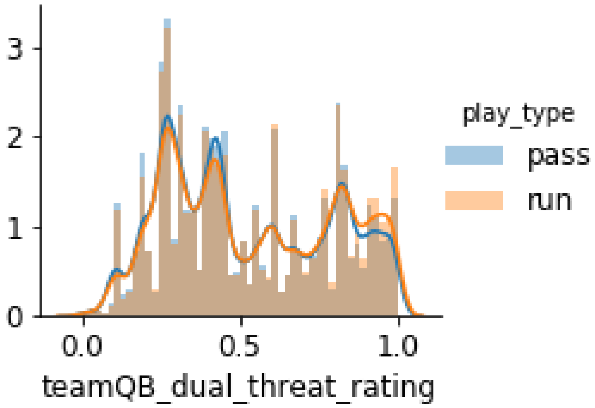

Figure S.3: Distribution of pass/run plays segmented by team dual threat rating