**Case Study: NFL Super Bowl Aspirations**

It’s January, and so it is one of my favorite times of the year – it’s NFL playoff time! For my analysis, I wanted to take a look if there was any way of predicting playoff (and Super Bowl) performance based on their performance in years past. In order to do so, I created my dataset using information from <https://www.pro-football-reference.com/>. I couldn’t find a dataset that had the information I needed. Per their wishes, I didn’t scrape the information using a spider but did so over an afternoon watching playoff football. Because I can be arbitrary, I only used information for the Super Bowl area (1966 season on, although 1970 was the first truly merged season).

Are records, point differential, or place in division any indicator of Super Bowl success? We will find out!

**Graph Analysis Portion:**

1. Load the data from the “nflstats.csv” file into a DataFrame.
2. Display the dimensions of the file.
3. Display the first 5 rows.
   1. In the initial dataset, an asterisk next to team name meant that they made the playoffs. I have created a new column that represents whether they made playoffs as a 1. Not making the playoffs is a zero.
   2. If they didn’t make the playoffs, their playoff finish is NaN.
   3. To make analysis easier, I took the text field DivTotal which would say “4th of 4” etc, I created DivPlace (Nth) of DivMax (of X). That makes it numerical for us.
   4. Playoffs Result is our target.
4. Some questions for our dataset:
   1. Is record the best indicator of Super Bowl hopes, or is it strength of Offense/Defense/PD?
   2. Can we work with our categorical variables to make them numerical?
5. Look at summary information about your data (total, mean, min, max, freq, unique, etc.) Does this present any more questions for you? Does it lead you to a conclusion yet?
   1. No conclusions yet, but a good look at our data.
6. Make some histograms of your data.
   1. We can certainly conclude that we’ve got some normal data. WL differential, points differential are both zero sum values (someone wins, another person has to lose. Ditto points) so that makes sense! The vast vast majority of teams don’t make the playoffs.
7. Make some bar charts for variables with only a few options.
   1. Conference and Playoff result are the only two we have. The other categorical variables have hundreds of unique values.
8. To see if the data is correlated, make some Pearson Ranking charts
   1. Definitely some correlations. GamesOverEven (W – L) is very correlated… but that doesn’t tell us much. Teams that score more points than their opponents win more games. I’m having John Madden flashbacks. But we do see some correlation between wins, points, and playoff results.
9. Use Parallel Coordinates visualization to compare the distributions of numerical variables between team records and playoff success.
   1. Done. Very spaghettiesque.
10. Use Stack Bar Charts to compare teams who made it far in the playoffs to teams who didn’t make it based on the other variables.
11. Now it’s time to reduce some of the features so we can concentrate on the things that matter!
    1. The features we will get rid of are: "DivTotal" (it's redundant), "Coaches", "AV", "Passer", "Rusher", "Receiver". (Names don't really tell us performance and there are hundreds of unique values.
    2. We can also fill in missing values if there were any. In this case, we don't, but it would have been something to try. Maybe we can fill in the NAs for playoff results if we hadn't already.
12. If you go back and look at the histograms of MadePlayoffs, you’ll see that it is very skewed. Most teams don't make the playoffs or even make it very far. Let's try a Log Transformation: it is a good method to use on highly skewed data.
13. Convert your categorical data into numbers (TeamName, Playoffs Result)
14. Training - Split your data into two sets: Training and Testing.
15. Evaluation – We are trying to predict Super Bowl success. There are many algorithms that could be used but we’re going to use logistic regression.
    1. Metrics for the evaluation:
       1. Confusion Matrix (2 false negatives, 15 false positives)
       2. Precision, Recall & F1 score (.968 for yes, .333 for no for precision, recall .966, F1 .982 for yes.)
       3. ROC curve Initially, PlayOff Results was included which made our prediction a little too good. When removed, we get much more realistic predictions. According to our ROC curve, our model does better than random guessing.
16. When taken all together, it does look like our model does slightly better predicting than guessing. There were a few false positives and more false negatives, so we could refine even more but all in all not a bad start. Sadly, my own prediction failed as the 49ers did not win the Super Bowl, which was a shame!