7-2011

A Statistical Data Mining Approach to Determining the Factors that Distinguish Championship Caliber Teams in the National Football League

Ernest Fokoué
Rochester Institute of Technology

Dan Foehrenbach
Rochester Institute of Technology

Journal of xxx xxxxx xxxxx x (xxxx) x-xx Submitted xx/xx; Published xx/xx

A Statistical Data Mining Approach to Determining the Factors that Distinguish Championship Caliber Teams in the National Football League

Dan Foehrenbach DEF7501@RIT.EDU
Center for Quality and Applied Statistics
Rochester Institute of Technology
Rochester, NY 14623, USA

Ernest Fokoué EPFEQA@RIT.EDU
Center for Quality and Applied Statistics
Rochester Institute of Technology
Rochester, NY 14623, USA

Editor: xxxxxx xxxxx

Abstract

This paper uses machine learning and data mining techniques to explore most of the performance measurements used in American football. The main goal is to determine/extract those factors that are most responsible for the success of the so-called great NFL teams. We consider a very large number of commonly used performance statistics and variables along with success indicators like winning percentage, playoff appearance, and championship wins. It is held by many football analysts/experts that defense wins championships. In this paper, we seek to establish if indeed there is ample evidence that the so called dominant teams are based on more defense than offense. Other football analysts strongly believe and declare that high third down conversion percentage is a very strong indicator of playoff/championship caliber teams. Using five years worth of data from 2006 to 2010, our application of techniques such as cluster analysis, principal component analysis, factor analysis, support vector machine and traditional logistic regression reveal compellingly interesting and consistent (over the years) elements of NFL greatness.

Keywords: *Statistical Data Mining, NFL teams, Performance Indicators, Playoff, Championship, Cluster Analysis, Principal Component Analysis, Logistic Regression Analysis, Support Vector Machine, Offense, Defense, Third Down.*

1. Introduction

Football enthusiasts all over America are all hugely fascinated by debates that seek to determine what makes a particular football team better than another. With the tremendous growth of fantasy football in recent years, it is even more exciting to dig deep into the numbers that characterize team performances, with the hope of finding ways to get an advantage over fantasy opponents. Building models to predict outcomes of NFL games has been a subject of great interest to a good number of top class statisticians and professors of statistics throughout the world. The quite recent article by Abbey et al. (2010) proposes a model of the NFL centered around the fundamental differences between predominantly pass oriented and predominantly rush oriented teams, and seeks to make the most accurate predictions of the outcomes of NFL games. With the ever increasing popularity of power rankings, experts and analysts from a variety of sources come up regularly with both ad hoc and scientific ways to create ranking of teams. For instance, the recent article by Govan et al. (2009) proposes an offense-defense approach to ranking team in sports, and as one would expect uses the NFL as a case. Indeed, beyond the NFL, fans and experts are just as busy churning and crunching numbers in an attempt to extract the indicators or factors that distinguish good from bad teams. In basketball for instance, Koh et al. (2011) explore data from youth basketball in Singapopre to determine what distinguishes the successful from the unsuccessful teams. Clearly, these authors do not build a model per se, but instead look at some of the variables that measure performances in Basketball and try to find out if there is a statistically (and practically for that matter) significant difference between good and bad teams. In a similar spirit, this paper considers five years (2006-2010) worth of NFL end of season statistics, and seek to use data mining and machine learning techniques to find out if teams can be automatically classified as good or bad based on those statistics, and also identify as much as possible those factors that seem to discriminate between the good and the bad teams. For our purposes, we define a good team to be one of the 12 that qualify for the playoffs for that given season. The rest of this paper is organized as follows: section 2 presents various of the data and performs both exploratory data analysis and some simple inferences on individual performance measures for a single season; in section 3, we merge all the five years worth of data into a single data file and perform a variety of data mining exploration and statistical analyses. Section 4 provides our conclusion and discussion and some pointers to our future work in this field.

2. Basic Statistical Analysis of the 2009-2010 NFL season statistics

To gain insights into what a single season can reveal in the way of factors that distinguish dominant teams from weak ones, we first consider the data from the 2009-2010 season, and focus on the offense. After removing some of the variables like fourth down attempts that clearly showed not apparent discriminating power, we remain with 13 variables. We also add the indicator variable `Playoffs`, which we later use to check how well unsupervised learning techniques succeed at partitioning the teams into good and bad. Below is a simple partial view of the data with some variable omitted to ease the display.

	Pts.G	Yds.G	Yds.P	X1st.G	X3rd.Pct	Pen	Pen.Yds	FUM	Lost	TO	Playoffs
Arizona	23.4	344.4	5.6	19.8	36	108	886	32	18	-7	1
Atlanta	22.7	340.4	5.2	20.6	42	78	664	19	8	3	0
Baltimore	24.4	351.2	5.5	20.0	42	115	1094	19	9	10	1
Buffalo	16.1	273.9	4.8	14.6	26	107	855	24	11	3	0
Carolina	19.7	331.1	5.2	18.1	37	88	698	23	11	6	0
Chicago	20.4	310.3	5.1	16.4	37	100	836	26	7	-6	0

From a pure informal perspective, a natural and plausible thing to do here is to consider each variable in turn, finding both graphically and numerically if indeed there is an indication of significant differences between the successful and the unsuccessful teams for each of the variable. The comparative boxplots of Figure (1) seem to reveal for the most part that playoff teams do indeed differ from non-playoff teams in ways that anyone who knows football would expect. For instance, it is clear from the plots that playoff teams on average fumble less than non playoff teams (at least for this season), and that playoff teams have a higher third down conversion percentage than non play off teams. Despite these apparent differences, one still has to establish formally which of the differences are actually statistically significant. Besides, it is important to emphasize that we are considering a single season, and only looking at the performance of the offense.

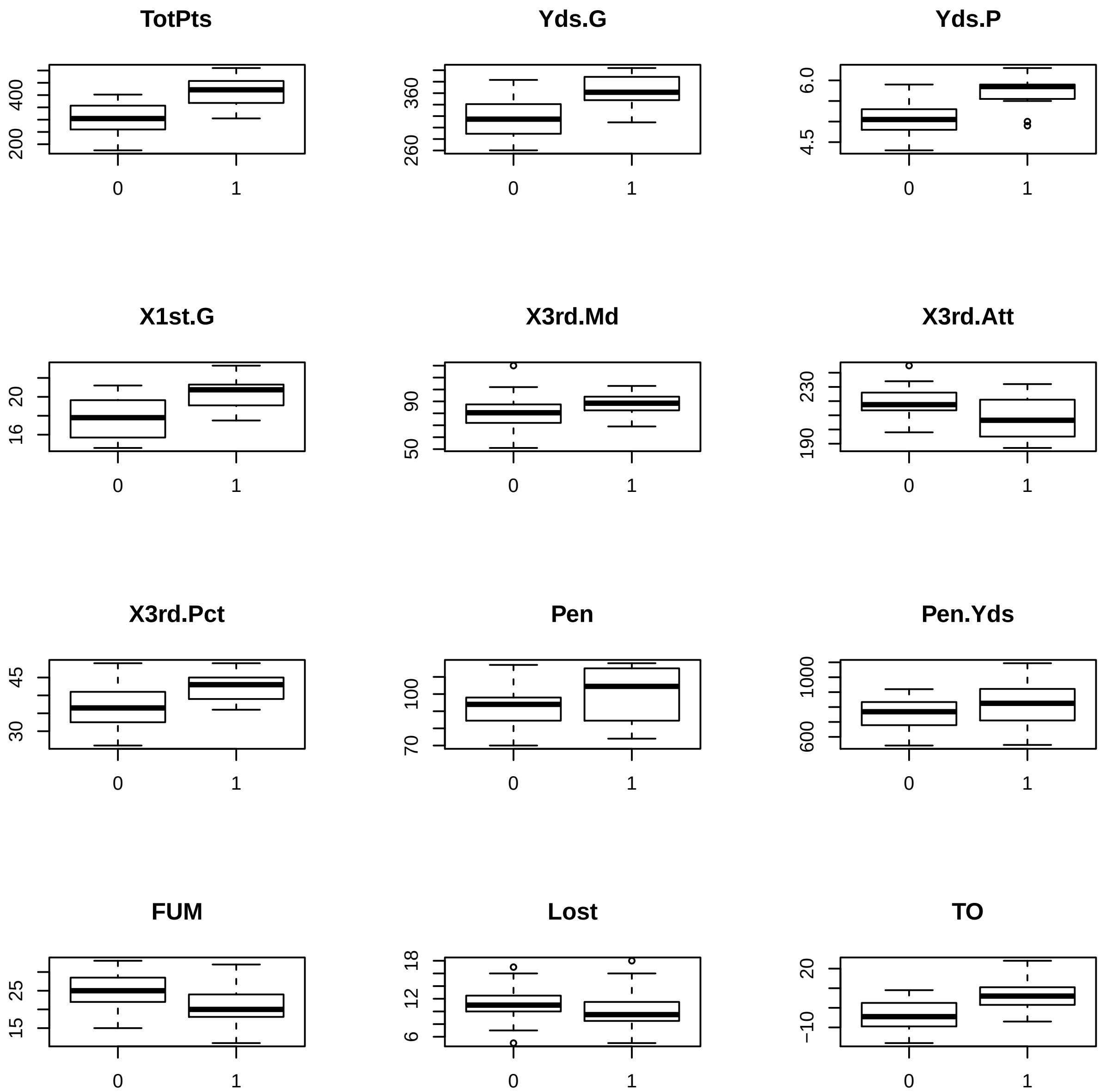

Figure 1: Comparative boxplots of the performance of NFL teams for the 2009-2010 season.

By way of a formal analysis of the difference, we consider each variable in turn, and perform a simple two sample t-test. It is fair to remark that only 12 teams make it to the playoffs against the remaining 20 that do not. Despite this apparent (imbalance) difference in sample size, we believe that each sample has enough observations to help make a plausible inference based on this season. It is refreshing to note that for most of the variables, there are no outliers, and the bulk of the data follows an approximately normal distribution (not really surprising here, since most are averages).

Variable	Observed Difference	Interval	P-value	Significance
Points/game	4.35	$(1.51, 7.23)$	0.0040	Yes
Fumble	-2.50	$(-6.33, 1.34)$	0.1896	No
Lost Fumble	-0.28	$(-2.75, 2.19)$	0.8420	No
Yards/game	29.00	$(5.74, 52.23)$	0.0164	Yes
3rd Down Att	-7.45	$(-17.95, 3.03)$	0.1530	No
Total Points	69.8	$(23.97, 115.62)$	0.0041	Yes
3rd Down%	3.51	$(0.13, 6.89)$	0.0424	Yes
1st Down/game	1.67	$(0.34, 3.00)$	0.01550	Yes
Penalty Yards	43.2	$(-73.65, 160.00)$	0.4449	No
Turnovers	6.27	$(0.38, 12.17)$	0.0379	Yes

Table 1: *Assessing the significance the statistical significance of the differences between playoff teams and non playoff teams on various variables.*

Now, a quick look at the following partial view of the sample correlation matrix reveals that some of the variables either plain redundant or strongly correlated as expected.

	TotPts	Yds.G	X1st.G	X3rd.Md	X3rd.Pct	Pen	Pen.Yds	FUM	Lost	TO
TotPts	1.0	0.9	0.8	0.5	0.7	0.0	0.1	-0.4	-0.1	0.7
Yds.G	0.9	1.0	0.9	0.5	0.7	0.0	0.1	-0.4	0.0	0.5
X1st.G	0.8	0.9	1.0	0.7	0.8	-0.2	0.0	-0.4	0.0	0.4
X3rd.Md	0.5	0.5	0.7	1.0	0.9	-0.3	-0.2	-0.3	-0.1	0.2
X3rd.Pct	0.7	0.7	0.8	0.9	1.0	-0.3	-0.2	-0.4	-0.1	0.3
Pen	0.0	0.0	-0.2	-0.3	-0.3	1.0	0.9	0.2	0.2	0.2
Pen.Yds	0.1	0.1	0.0	-0.2	-0.2	0.9	1.0	0.1	0.2	0.2
FUM	-0.4	-0.4	-0.4	-0.3	-0.4	0.2	0.1	1.0	0.6	-0.5
Lost	-0.1	0.0	0.0	-0.1	-0.1	0.2	0.2	0.6	1.0	-0.3
TO	0.7	0.5	0.4	0.2	0.3	0.2	0.2	-0.5	-0.3	1.0

We obviously restrict ourselves to those variables that when taken alone are significant in discriminating between good and bad teams, and we ignore any variable that is redundant. Having said that, it turns that Points per game is unsurprisingly the main factor of a healthy and successful offense, with a P-value of 0.004. In other words, successful teams score significantly more points per game than their counterparts. With the average number of yards per game following right after in terms of significance with a p-value of 0.016, it seems clear again that great teams do indeed end up being the ones that pile up the number of yards per game, quite unsurprising again. Then, comes the Turnover variable with a significance captured by a p-value of 0.0379. This variable is hard to explain here, but it seems to point to risk/reward aspect of a great team. Finally, the famous third down conversion percentage often mentioned by analysts and experts sneaks through with a p-value of 0.0449. To us this might be the indicator of the tactical savvy of the coaching staff and the sheer tenacity of the team as a whole, namely their ability to force their will in tough situation, a mark indeed of winners.

3. Statistical Data Mining of the 2009-2010 NFL season

It is clear that attempting to perform formal inference with only $n = 32$ observations when each has observed dimensionality of $p = 13$, exposes our analysis to all sorts of challenges. At the very least, we first consider some of the techniques of unsupervised learning that are less vulnerable to short fat data. A look at the hierarchical clustering of the 32 teams yields the dendrogram in Figure (2). The cluster on the right side of the dendrogram contains all

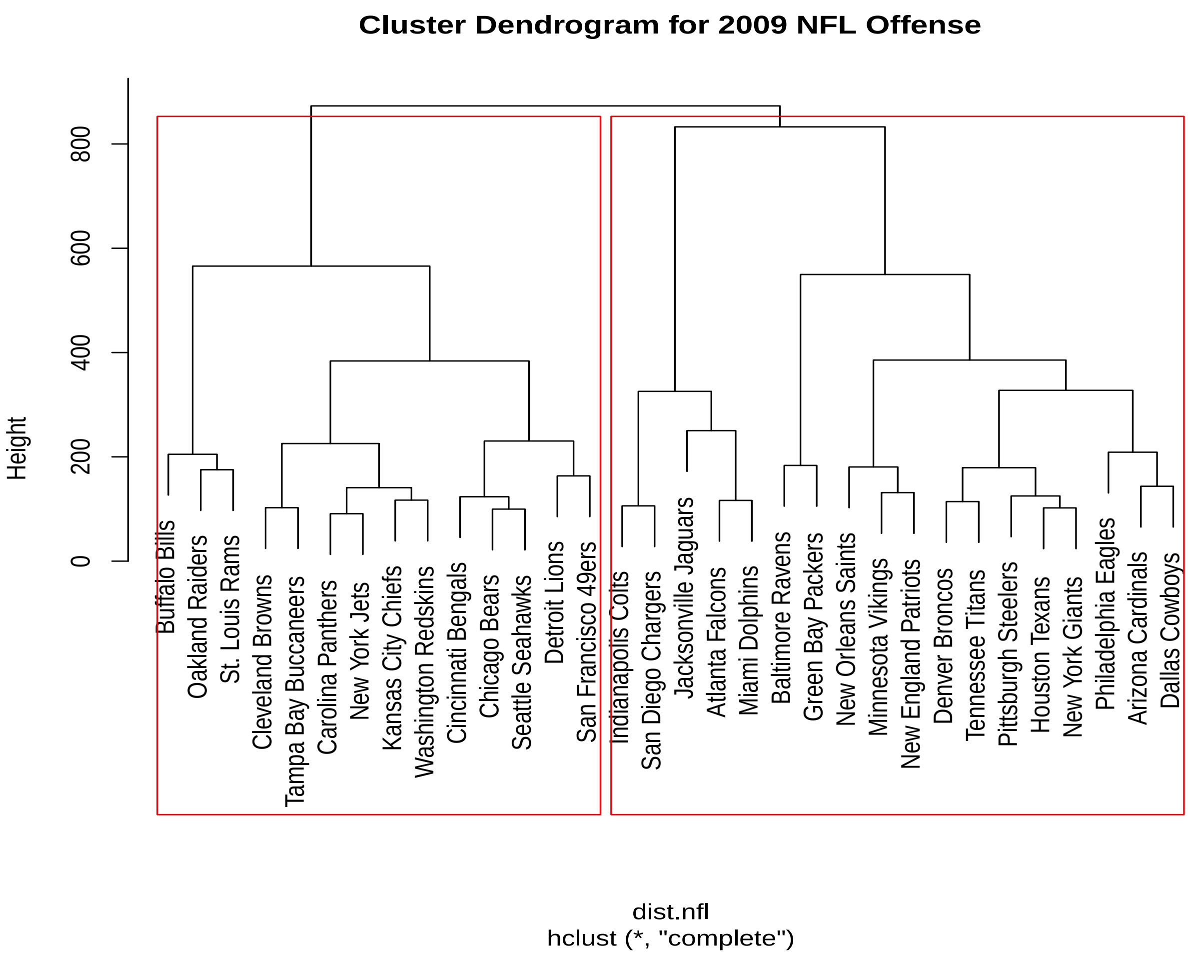

Figure 2: On this dendrogam, the two hypothetical clusters are marked by the rectangles.

the playoff teams except for two of them, namely the Cincinnati Bengals and the New York Jets. For the record, these two teams barely made it to the play offs that season. Indeed, a quick plot of the first two principal components shows that those two teams exhibit more of the characteristics of the teams that failed to make the playoffs (See Figure (3)). Even with only two principal components, playoffs teams are clearly separated from non playoff teams by a rather simple decision boundary. Although they made the playoffs, the Jets and Bengals were not offensive powerhouses during the 2009-2010 season, hence their presence in the midst of teams that did not make the playoffs.

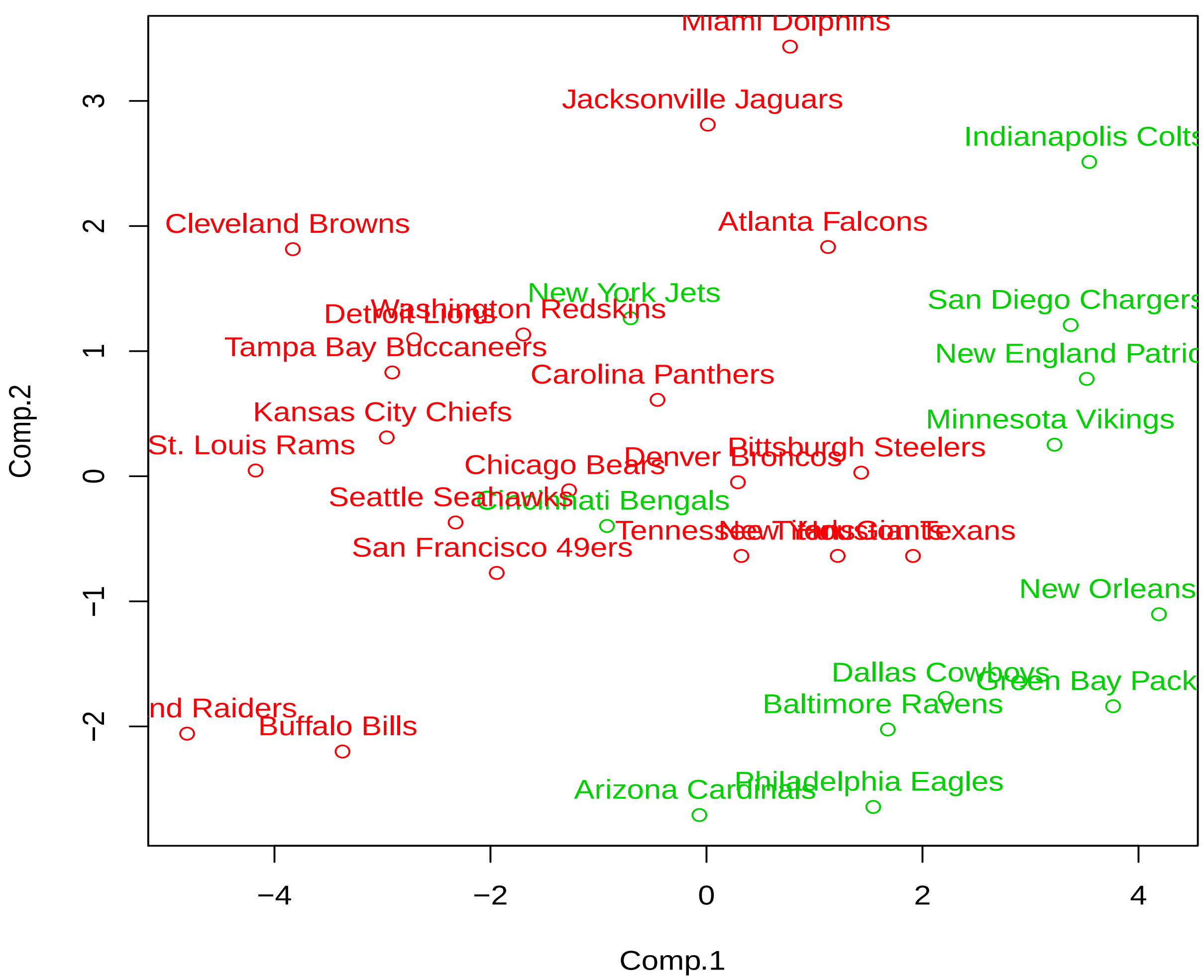

Figure 3: First two principal component scores with the label of the teams.

4. Conclusion and discussion

In this paper, we have focused solely on the data from 2009-2010 NFL season, and have used some very common statistical and data mining techniques to gain insights into the factors that might explain the differences between successful and unsuccessful NFL teams. We have considered offense and defense separately, which clearly does cause us to miss the subtle and obvious interplays between these two fundamental aspects of the game. In our future work, we plan on merge all the five years worth of data and providing a more thorough analysis with offense and defense considered together in the same analysis. We plan on applying various pattern recognition techniques to the whole data sets to gain even deeper insights in the workings of the performance measures used in the NFL.

References

R Abbey, J Holodnak, C May, C Meyer, and D Moeller. Rush versus pass: Modeling the nfl. *Journal of Quantitative Analysis in Sports*, 2010. Aritcle.

Anjela L. Govan, Amy N. Langville, and Carl D. Meyer. Offense-defense approach to ranking team sports. *Journal of Quantitative Analysis in Sports*, 5(1), 2009. Article 4.

Koon Teck Koh, Wang John, and Clifford Mallett. Discriminating factors between successful and unsuccessful teams: A case study in Elite Youth Olympic Basketball Games. *Journal of Quantitative Analysis in Sports*, 7(3), 2011. Article 21.