NFL Defensive Performance Analysis

Daniel O'Neel and Reed Johnson {doneel, rjj} @ stanford.edu
CS 229, Stanford University
December, 2014

***Abstract*—We seek to predict eight characteristics of defensive performance in a given NFL game based on previous games played by both that defense and the upcoming opposing offense. We choose a method of featureset generation that generalizes to any game in the season and then apply a variety of multivariate regression techniques in order to predict defensive performance before the game occurs. We compare the accuracy of our predictions to the actual results of their respective games, and also against existing benchmarks who publish predictions of "fantasy points", which are a known function of our eight predicted defensive statistcs.**

I. Introduction

Player performance in the NFL is a highly analyzed and debated subject, as evidenced by the rise of dedicated TV channels to "fantasy analysis" which seek to predict a player's performance for the sake of viewers, and of course by the massive gambling industry. The state of Nevada reported $1.34 billion of legal NFL betting in 2011 alone, which the National Gambling Impact Study Commission believes comprises less than 1% of annual football betting. [5][6]

Despite the size of the industry surrounding performance predictions, relatively little attention is paid to defensive unit performance, likely because it is generally considered less exciting and more team oriented than offense. Defense remains half of the football game, however, and due to this relative lack of attention, accurate predictions about the performance of defensive units offer a potentially large competitive advantage.

We look to predict a team's defensive performance as measured by eight statistics and base our predictions off of 36 statistics that represent the performance both the teams defense and the opposing teams offense. For those familiar with football, accuracy results offer simple interpretation, but to benchmark our performance, we also measure our predictions performance using a standard fantasy football scoring system used by the majority of leading "fantasy football" websites. This metric gives us a simple way to compare the efficacy of our analysis to the industry standard.

ID	Defensive	Offensive	Misc.
Game ID	Tackles for loss	**Passing yards**	Penalties
Team	**Sacks**	**Rushing yards**	Penalty yards
Week	Sack yards	First Downs	# of plays
	Interceptions	3 Down Conv. %	Posses. time
	Forced fumbles	4 Down Conv %	
	Defensive TDs	Avg. field position	
	Safeties	**Points**	

Fig. 1: Variables

II. Data

A. Source

We used data from the publicly available *NFLDB* database, which can be accessed at https://github.com/BurntSushi/nfldb. This dataset contains play-by-play statistics for every preseason, regular season, and post-season NFL play since 2009. We have narrowed our analysis to regular season games. Though our algorithms would apply to any season, all statistics given throughout this paper are from analysis on the 2013 season for the sake of consistency.

B. Data Attributes

Though *NFLDB* provides a massive number of statistics, we extracted a subset of 18 statistics for both teams in every game played in the 2013 NFL season, shown in table 1.

In bold are the eight output variables we seek to predict. Notice that for a given defense, we predict our own sacks, interceptions, etc. but we predict the *opposing* offense's passing yards, rushing yards, and points (these are referred to as yards allowed and points allowed).

C. Feature generation

Though our 8 output variables are clear, our features are not obvious. It would not be useful to predict the defensive performance of a team with the features being the offense performance in the same game. The value in these predictions is making the predictions before the game is played. Of the many possible ways to attack this problem, we chose to compute "resumes" for both the offense and defense of a team, which are composed of all statistics in the offensive and defensive column respectively from the table above as well as the miscellaneous column. We average these statistics for each game a team has played in up to the week we wish to predict for.

III. Models

Given the featureset described above and the output variables explained above, we explored various multivariate regression techniques. We originally also tried models comprised of univariate models for every output statistic, but after equivalent performance in some models (indeed, in some models the multivariate form is exactly the combination of independent univariate models) and underperforming in others during initial analysis, we decided to limit ourselves to multivariate models.

Throughout this paper, we measure our model's overall error by *normalized error*, which is defined by

$$\left|\left|\left(\frac{\hat{y}-y}{\sigma_y}\right)^2\right|\right|_1$$

A good prediction of yards allowed may have a residual of 20, but a bad estimate of sacks may have a residual of 2. This normalizes all our errors by the standard deviation of the actual values so we value our predictions in all eight categories equally.

A. Linear Regression

Our linear regression model can be expressed as

$$Y = X\beta$$

In this model, our X matrix is composed of the raw resumes of the two teams. Our Y is an $n \times 8$ matrix with n equal to the number of games in the season up to the previous week. The performance of this model was ultimately weak.

B. Weighted Linear Regression

The weighted linear model is very similar to the linear model, except when fitting the model, we seek to minimize a *weighted* sum of residuals, defined by

$$\sum_{i=1}^{n} W_i(\hat{y} - \vec{X}_i\beta)$$

This models allows us to give more *weight* or influence to particular observations which we train our model on. We weighted games, which are rows $X_i \forall i \in [1, n]$ in our feature matrix, according to the following equation

$$(w-1)(1_T(X_i)) + 1$$

where $1_T(X_i)$ is the indicator function that is true if team T (for which we are predicting) played in that game, and w is a parameter that specifies the relative importance of such a game. When predicting an upcoming game for team T, this has the effect of making past games in which T has played w times as important in our model as games in which T did not play. We found the best weighting to be $w = 4$. This offered significant performance improvement over the linear model.

Note, however, that because the weightings depend on the team for which we want to predict, this requires re-fitting a new model for every prediction.

C. Partial Least Squares Regression

PLS is a dimensionality reduction or shrinkage technique we thought may be of use given the relatively large amount of predictors we have. Especially early in the season, there are only a limited number of games played, and we worried that our variance would be very high. PLS offers a technique for reducing the dimensionality to help combat overfitting. Indeed, as can be seen in Fig. 2, PLS does offer significantly better performance over linear and weighted linear regressions in the first half of the season.

D. Weighted Random Forests

Multivariate random forests are an extension of decision trees which use bagging and decorrelation to reduce overfitting. We modified Leo Breiman's original algorithm to do *weighted* random forests as described by Chen, Liaw and Breiman [1] [2]. The modification causes points to be drawn by weighted rather than equal

Fig. 2

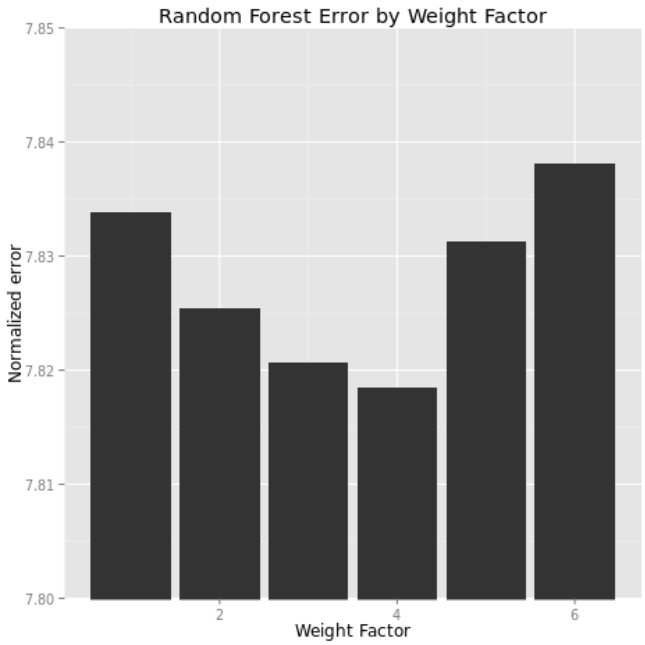

Fig. 3

probabilities when fitting each individual tree. We use the same weighting scheme above:

$$(w-1)(1_T(X_i))+1$$

Which leads the algorithm, when randomly selecting data points to fit in each individual tree, to select games in which team T has played with probability

$$\frac{w}{(n-g)+wg}$$

where n is the number of total datapoints and g is the number of games that team T has already played, as opposed to $\frac{1}{(n-g)+wg}$ for all other points. This has the effect of exaggerating the importance of games in which team T has played. As with weighted linear regression, this requires fitting a new model for every team.

As seen in Fig. 3, the minimum normalized error occurs with $w = 4$, which curiously is the same weighting factor that acheived minimal error in weighted linear regression, though we suspect coincidence rather than hidden relationship.

E. Multivariate Adaptive Regression Spline

Seeking a model that allowed nonlinearity with the possibility variable interaction, we fit a multivariate adaptive regression spline (mars) model to good success. Using the earth software implementation, we allowed 2nd level interactions. MARS offers us a particular advantage: though we suspect there are nonlinearities and interactions, we did not know which. MARS infers these itself through the fitting process. [4]

	Linear	Weighted	PLSR	MARS	WRF
Sacks	0.9022	0.8942	0.8942	0.8862	0.8825
Interceptions	0.9405	0.9310	0.9238	0.9196	0.9057
Forced fumbles	0.9447	0.9388	0.9357	0.9255	0.9209
Defensive TDs	0.9200	0.9100	0.9137	0.9103	0.9108
Safeties	0.9825	0.9724	0.9639	0.9501	0.9492
Points allowed	0.8774	0.8609	0.8431	0.8139	0.8089
Pass allowed	0.8864	0.8435	0.8285	0.8103	0.7910
Rush allowed	0.8513	0.8454	0.8280	0.8067	0.7948
Mean	0.9131	0.8995	0.8910	0.8776	0.8702

Fig. 4: Normalized error by category and model

IV. Results

Table 4 shows the normalized error for each algorithm in each category.

First, notice the perfrmance of MARS and weighted random forests models consistenly outperform other models in nearly every category. Algorithms which produce the overall lowest error tend to offer small improvement in all categories rather than significant improvement in one category.

Second, notice that some categories have consistently

Fig. 5: Fantasy point prediction performance with mean error and varance band

lower error than others. Particularly, predictions for yards and points allowed are consistently better predictions than turnovers and safeties. For those who know football, this result should make sense: events such as turnovers are single, inherently unpredictable events, whereas points and yards allowed are less affected by individual plays.

Also notice that the numbers in this table imply that that our models explain only a low percentage of the variance in these statistics. In the best case, we explain slightly over 20% of variance. Compared to many machine learning applications, our model appears to have exceptionally low predictive power. In the context of football predictions, however, our model appears very competitive.

We've also compared our weekly error to predictions from another well known fantasy prediction provider: Fantasy Football Freaks, a USA Today Sports Subsidiary. [3] FFF, unlike every other projection provider we've checked, privides historical predictions for every team and every week of past years, which means we can directly compare our predictions to theirs. Fig. 5 shows the average error of both their fantasy predictions and ours. Note that we do not directly predict fantasy points, but rather use our 8 predicted statistics as inputs of the fantasy points function.

Category	Average error
Sacks	1.2312
Interceptions	0.9185
Forced fumbles	0.6149
Defensive TDs	0.3166
Safeties	0.0314
Points allowed	6.7637
Pass yards allowed	57.0780
Rush yards allowed	36.9433

Fig. 6: Average Errors

V. Conclusion

Our predictions demonstrate consistently better accuracy than those of Fantasy Football Freaks. In both a week by week and a team by team comparison, our prediction model has lower error. Interestingly, our errors and theirs show some correlation (0.2870). This would seem to imply our algorithms both fail to predict some "unpredictable" events. In the context of football, we would indeed expect that some events are truly unpredictable, and that all predictions (whether by humans or algorithms) would be similarly affected.

Our average errors also provide some insight into the different categories. Table 6 shows our unnormalized errors. The residual for our points prediction is on average 6.76 points - slightly less than a touchdown. This is signficant especially given that betting lines are commonly 3 or 7 points (for winning by a field goal or touchdown, respectively). Our predictions are accurate within 7 points 58.25% of the time.

Our average errors for yards allowed is also interesting. While it is tempting to conclude that rushing offenses are more reliable, one must realize that the difference in accuracy only appears when the errors are unscaled. One can see in table 4 that our prediction for rush yards is no more "accurate", teams just tend to rush for fewer yards than they pass, leading to lower variance in number of rush yards than number of pass yards.

Another interesting trend shows up in Fig. 2: After a generally consistent decrease in error each week, there's a signficant a rise in prediction error in the final weeks of the season until the very last week, which again appears normal. This trend holds across 3 seasons we tested (the trend is not obvious in fantasy point error graph, as the fantasy point algorithm obscures much of the inaccuracy). We speculate the rise in error is due to the fact that these final weeks often feature games

which determine playoff spots, and often these games are against division opponents who have already played each other. Both of these factors may contribute to a propensity to "pull out all the stops" - to literally be less predictable to their familiar opponent (and unfortunately to us).

Why then, the drop to very low error in week 17? Week 17 notoriously features many games between teams who have their fate guaranteed one way or the other. Motivation is low and for the playoff bound teams, there is no incentive to risk injuries. Thus these games tend to be much less exciting, but wonderfully predictable.

It comes as no surprise that our predictions are "innaccurate" in the sense that we can only explain a small percentage of the variance in these statistics. We are rather surprised, however, that our ability to explain 10-20% of variance in many categories gives us noticeably more accurate predictions than a leading provider. We have chalked this up to both the inherent unpredictability and, again, the relatively lack of attention paid to the unglamorous matter of defensive performance.

VI. Next Steps

Though we consider our existing algorithm successful, there are some obvious opportunities for improvement. The most obvious problem we have is that we only generate high accuracy in the second half of the regular season. Using historical data from previous seasons as a prior will likely give us increased accuracy at the start of the season. Additionally, we can extend our data set to include all playoff games while still using our existing models.

One way to improve the accuracy of our models is to increase the number of significant features. For instance, if Peyton Manning is injured and cannot play in a game, the Broncos are likely to score much fewer points against a defense than if he was playing. Since we currently have no data available on injuries, our models will make the same prediction whether or not Manning is playing. Although it is hard to measure the impact each individual player has on a game, a simple count of the starting players missing on offense and defense could help improve our predictions.

Additionally, even at times when they are perfectly healthy, many starters will sit out games or parts of games when their team has already clinched a playoff berth. Adding a simple boolean value stating whether a team has already clinched their playoff spot could help account for some unusual values that may be throwing off our models.

Given that our there is an extremely high amount of unpredictability in performances, we think it would be useful in many cases to have confidence intervals for each prediction we make. A team likely to score between 20 and 26 points is significantly different from one likely to score between 10 and 36, even though our model will currently predict 23 for both. Including a confidence interval with our predictions will let users know the range of outcomes they should expect.

Finally, we would like to gather data from many prediction providers to better benchmark our own models. Currently we are confident that our predictions are superior to one provider, but to extend that claim to other providers such as ESPN, Yahoo, or the NFL itself, we need to log their predictions. This will likely have to wait until next season, when we will find out, from week one, whether our model is truly competitive with the best.

References

[1] L. Breiman, *Random Forests* Machine Learning Vol. 45 Issue 1. 2001

[2] C. Chen and A Liaw and L Breiman. *Using random forest to learn imbalanced data* University of California, Berkeley 2004.

[3] Weekly Projection Ratings. [Online]. Available: http://www.fantasyfootballfreaks.com/weekly-projections-ratings

[4] J. H. Friedman. *Multivariate Adaptive Regression Splines* The Annals of Statistics, Vol. 19, No. 1. 1991.

[5] Gillian Spear, *Think sports gambling isn't big money? Wanna bet?,* NBC News, July 15th 2013. [Online]. Available: http://www.nbcnews.com/news/other/think-sports-gambling-isnt-big-money-wanna-bet-f6C10634316

[6] National Gambling Impact Study Commission, *National Gambling Impact Study Commission Final Report*, 1999.