MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 11 - 12, 2016 BOSTON CONVENTION AND EXHIBITION CENTER

Assessing the Predictability of an NFL Offense

Paper Track: Other Sports
Paper ID: 1529

December 14, 2015

1 Introduction

Pre-game preparation is paramount in the National Football League (NFL). Countless hours are spent reviewing film, strategizing, and understanding opponents. As the League has gotten increasingly competitive, even the smallest of advantages can make the difference between a win and a loss. With this as motivation, we used a machine-learning approach and NFL play-by-play data to predict whether an opposing offense executes a pass play or a run play. Using this prediction, we fine-tuned the expected play call: run left, run middle, run right, or long pass, short pass, screen pass.

In addition to shedding light on how well pass/run play selection can be modeled using readily available data and advanced analytics, our results provide a sense of individual team tendencies, *viz.*, to what degree does a team resemble or deviate from "average" behavior. While the label of average is somewhat arbitrary, the fact that our pass/run prediction model is developed using a season's worth of data across all teams and all games (including the postseason) implies that our results reflect an aggregate realization of an NFL offense. Using this realization as a baseline, we are able to identify teams that deviate from expectations. Similarly, we can identify teams that could be thought of as acting like a typical NFL team.

These insights, which can be derived ahead of a game, could help both players and coaches as they attempt to quickly anticipate their opponent's next move. In addition, these results would allow teams to better, and more efficiently, prepare for games by shifting some of the burden of characterizing an opponent from human to machine. The ability to identify potential game situations that warrant greater consideration given an opponent's expected actions would be a valuable advantage.

1.1 Background

While statistics and analytics have almost always played a role in sports, only relatively recently (the last several decades) have athletes, teams, and coaches fully embraced the more rigorous application of analytics to their sports. Certainly, Bill James and his pioneering work with sabermetrics for baseball brought the field of in-game analysis to the forefront. Building on the work of James and others, sports analytics has steadily evolved as more data has become available and the analytics have become more sophisticated, not to mention the explosive growth of computing power.

A large portion of the analytical work directed at football has either focused on specific scenarios (*e.g.*, [3, 8]) or considered aspects of the game from an optimization point of view (*e.g.*, [4]) — or both (*e.g.*, [1]). These analyses have been helpful in revealing ways in which analytics can be used to better understand the game of football; however, because they treat a singular aspect of the game (and are consequently limited in that regard), or because their results are aimed at optimal decision-making and may be difficult to realize in practice, many of these studies are more academic than practical.

The true power of advanced sports analytics comes from prediction — predicting the outcome of a game, or perhaps more importantly, beginning to understand what an opponent will do in a certain situation. This is no easy feat. One of the more important contributions in this area was the 2006 work of Kvam and Sokol [6], who developed a Markov Chain model that relied on win probabilities to predict NCAA Division 1 Basketball Tournament outcomes. In that same year, Joseph *et al.* [5] explored the use of various machine learning approaches to predicting the outcomes of English Premier League soccer matches.

More recently, prediction in the NFL has been aimed at pass/run prediction. At the 2015 MIT Sloan Sports Analytics Conference, a team from Booz Allen Hamilton showcased a play prediction tool [7] running on a handheld tablet. Similarly, at the 2015 Joint Statistical Meetings, Burton and Dickey [2] (hereafter, "BD15") presented their work on predicting the play-calling tendency (pass or run) of NFL teams. In their study, play-by-play data from the 2011–2014 seasons was analyzed and used to train a set of six models, depending on the quarter: quarter 1, quarter 2, quarter 3, and quarter 4 (winning, losing, and tied). The features used to develop each model included: yards-to-go, time remaining, point differential, offensive points, defensive points, interaction between yards-to-go and down, cumulative number of fumbles, cumulative number of interceptions, field position, timeouts remaining, and yards gained on previous plays. Both logistic regression models and random forest models were explored by BD15. To assess accuracy, BD15 tested their models on 20 randomly selected games between 2011 and 2014. While the results were impressive (mean accuracy near 75% and best-game accuracy near 90%), measuring accuracy using a subset of the training data is poor practice, and (1) typically over-estimates true accuracy and (2) results in a model that may fail to handle new data (*e.g.*, the 2015 season).

1.2 Overview of Our Approach

In this paper, we present an approach to NFL play prediction that builds off these most recent works by adding greater analytical depth, including prediction of run direction and pass type, and offering new insights. Using a similar source of play-by-play data, we analyzed the 2013-season dataset and developed a set of features that shares some overlap with BD15, but also includes new features that we developed explicitly for offensive play prediction. Furthermore, we considered both down and quarter to be critical parameters and created models for each combination, along with models for the final 2 minutes in each half. In essence, this is equivalent to conditioning the models on each of these parameters. To train the various models, we used a novel stacking approach that combines the predictions of five separate learning algorithms (neural network, support vector machine, random forest, naïve Bayes classifier, and gradient boosting machine). This approach yields performance that is better than any of the individual algorithms alone and can improve the model's ability to handle new data. Finally, using our prediction results as a representation of a typical NFL offense's behavior, we can identify teams that deviate from expected play calling. Similarly, we can identify teams that could be thought of as acting in a predictable way.

2 Modeling NFL Offensive Play Selection

2.1 Data, Features, and High-Level Methodology

Play-by-play data from two complete NFL seasons (2013 and 2014) were available to us for the purpose of developing an offensive play prediction algorithm. We focused on a pure machine-learning approach and used the 2013 data for training (33,722 observations) and the 2014 data (33,310 observations) for testing. A number of the features we used came directly from variables in the raw data, while others were derived. The final set of all features is as follows:

- Yards-to-go
- Field position
- Offense timeouts left
- Defense lead (negative for trailing)
- Time remaining in quarter
- Cumulative interceptions, fumbles, hurries, and sacks (per game)
- Pass:run ratio (cumulative for the season, for the game, and for the current drive)
- Passing effectiveness (cumulative for the game, and for the current drive)
- Running effectiveness (cumulative for the game and for the current drive)
- Interactions between defense lead and time remaining
- Interactions between time remaining and field position

Passing effectiveness was developed as a way to measure the benefit of choosing to pass. It was defined as:

$$\text{effectiveness} = \frac{1}{P}\sum_{p=1}^{P}\begin{cases} 1, & \text{if gain} \geq 3 \text{ or first down} \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where the sum is taken over all passing plays. A similar expression was used for run effectiveness.

The high-level methodology we adopted for play prediction was to train a set of models to first predict pass or run and then use that outcome to predict run direction (left/middle/right) or pass type (short/long/screen). As a way to improve performance, we used a stacking approach [9] that combined the predictions of five separate learning algorithms. This approach has been shown to yield performance that is better than any of the individual algorithms alone and can improve model generalization.

2.2 Model Building Details

Figure 1 illustrates the overall modeling pipeline. Twenty models were developed depending on the combination of quarter, down, and time remaining in a half. Depending on the outcome of those models (pass or run) a second bank of models is used to fine-tune the prediction. This second bank of models, which only conditions on down, produces one of the six final outcomes.

While a number of machine-learning algorithms were explored, our final approach was to use a stacking technique that blends the predictive outputs of several individual algorithms. We chose to implement five algorithms: neural network, support vector machine, random forest, naïve Bayes classifier, and gradient boosting machine. While the choice of these five was somewhat subjective, it was

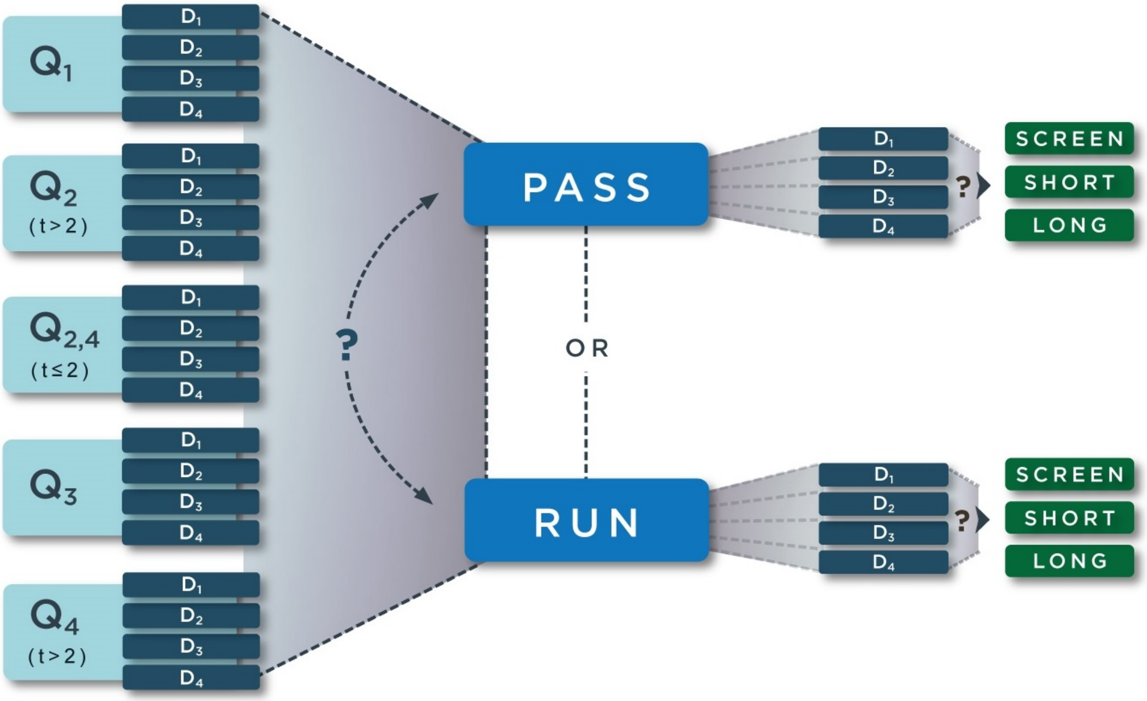

Figure 1: Model pipeline showing first bank of 20 models (left) used to predict pass or run based on quarter ($Q_i, i = 1 \dots 4$) and down ($D_j, j = 1 \dots 4$), and second bank of 8 models (right) used to predict play detail based on down.

based on previous experience and intuition with these algorithms, as well as their fundamentally different underpinnings. Some of the key details of each algorithm are summarized in Table reftab:algs.

To implement the stacking approach, the 2013 training set was divided into two subsets: one subset to train the five algorithms (level-0 models) and a second subset to train the blending model (level-1 model). We chose to train the level-0 models using 2/3 of the base training set and 1/3 of the base training set to train the level-1 model. We experimented with several approaches to the blending model (linear regression, simple averaging), ultimately deciding on a regression neural network (16 nodes in a single hidden layer). To be clear, each of the 28 models in Figure 1 is a stacked model. As is typically done, we preprocessed each input feature by demeaning and scaling it by its standard deviation. Finally, we used 10-fold cross-validation to minimize overfitting.

2.3 Model Performance Assessment

To quantify the performance of our approach, we tested on the 267 games of the 2014 season, which included playoffs and the Super Bowl. If N is the total number of all cases, TP is the number of true positives (a correctly predicted pass[1], TN is the number of true negatives (a correctly predicted run), accuracy was computed as follows:

$$\text{accuracy} = \frac{TP + TN}{N}. \tag{2}$$

[1] The choice of pass as a positive outcome and run as a negative outcome is arbitrary; the choices could easily be swapped.

Algorithm	Details		
Neural Network (NN)	Single hidden layer with 7, 14, or 21 nodes	Decay weights: $\in$ (0.005, 0.01)	$\leq$ 1000 iterations
Support Vector Machine (SVM)	Radial basis function kernel	Cost parameter: $\in$ (4, 8, 12); sigma parameter: $\in$ (0.05, 0.1)	$\leq$ 1000 iterations
Random Forest (RF)	Splitting parameter: $\in$ (4, 6)		$\leq$ 500 iterations
Naïve Bayes (NB)	Kernel density estimation enabled	Laplace correction: $\in$ (0, 0.33)	
Gradient Boosting Machine (GBM)	Interaction depth: $\in$ (3, 5, 7); shrinkage: $\in$ (0.01, 0.05);	Minimum number of terminal observation nodes: $\in$ (5, 10, 15)	1000 trees

Table 1: Key Machine Learning Algorithm Details

In addition to accuracy, we computed the F-score, which takes into account the precision and recall of the prediction, where

$$\text{precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{recall} = \frac{TP}{TP + FN}, \tag{4}$$

and FP the number of false positives (a run incorrectly predicted as a pass) and FN is the number of false negatives (a pass incorrectly predicted as a run). The F-score is defined as

$$F\text{-score} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{5}$$

Relevant performance metrics are provided in Table 2. With a mean pass/run prediction accuracy near 73% and a best-game accuracy near 83%, our approach proves to be effective. Although these numbers are slightly lower than BD15, our performance assessment considers a much more challenging, and realisitic, scenario — we train on a single season of data and test on a completely different season. Compare this to BD15, where training was done on four seasons, and evaluation was done on 20 games taken from those four seasons. Beyond pass/run prediction, our approach also predicts the offensive play details. Over all games, we correctly predict the detailed outcome (run direction or pass type) almost 58% of the time, which is remarkable.

Table 3 shows the prediction results for a three notable 2014 games, including Super Bowl XLIX. Because the play-calling of the Seahawks in the final minute of the Super Bowl received quite a lot of scrutiny and criticism, we will revisit it here. Trailing by 4 points with 26 seconds left in the game,

Metric	Minimum	Mean	Maximum
Pass/Run Accuracy	61.3%	72.7%	82.8%
Pass/Run F-score	0.60	0.75	0.85
Play Detail Accuracy	40.0%	57.6%	78.9%
Play Detail F-score	0	0.47	0.80
NN Accuracy (Pass/Run)	52.8%	64.4%	77.4%
SVM Accuracy (Pass/Run)	52.4%	66.9%	78.6%
RF Accuracy (Pass/Run)	**54.8**%	68.7%	79.7%
NB Accuracy (Pass/Run)	48.4%	66.6%	79.5%
GBM Accuracy (Pass/Run)	53.2%	**69.3**%	**80.3**%

Table 2: Performance Results (Stacked and Individual)

Label	Week	Match-up	Pass/Run Accuracy
Highest Accuracy	2	Falcons v. Bengals	82.8%
Lowest Accuracy	3	Colts v. Jaguars	61.3%
Super Bowl XLIX	–	Patriots v. Seahawks	74.4%

Table 3: Performance Results for Notable 2014 Games

the Seahawks had the ball on second down at the Patriots' 1-yard line. To the surprise of many, the Seahawks called a pass play and the ball was intercepted, sealing the win for the Patriots. Interestingly, when this game is run through our prediction code, we actually predict "Pass," and specifically, "Short Pass." While this is not particularly significant, it is noteworthy, nonetheless.

The results of Table 2 provide a sense of the benefit of the stacking approach. For pass/run prediction, stacking improves overall accuracy by 3.4% over the best individual algorithm, GBM, and by 8.3% over the weakest, NN. It also appears that stacking mitigates poor predictive performance, with a minimum accuracy of 61.3% compared to minimum accuracies for the individual algorithms near 50%. Among the individual algorithms, GBM and RF perform the best, SVM and NB are intermediate, and NN is the least accurate. These results should not be considered indicative of any particular algorithm's merits. Quite often, certain algorithms are better-suited to certain types of problems. Furthermore, while each algorithm was tuned, fine-tuning was not a focus of this analysis and as a result, we may not realize optimal performance across all algorithms.

3 Characterizing Pass/Run Tendencies

Having the ability to accurately predict offensive play calling in the NFL would be a potentially valuable asset to a defensive coach; however, NFL rules currently prohibit the use of any kind of technological aids (beyond specialized tablets used for in-game photo review) on the sidelines and in press-box coaching booths. Nonetheless, there is value to advanced predictive tools such as the one described in this paper — in particular, the ability to gain additional insight into an opponent during game preparation. For instance, if we consider the output of our predictive model to represent the anticipated average behavior of an NFL offense (since the model was built as an aggregate of all games and all teams), then poor predictive results for a specific team could imply that team deviates from anticipated average behavior. Conversely, a team that is well-predicted could be considered to act in a manner that is consistent with the behavior of an average NFL offense. While we recognize that these are broad generalizations and that real value would come from studying tendencies in specific situations, this thinking can create the foundation for more detailed analyses.

Examining our pass/run prediction output for all teams, we identified the top three best- and worst-performing predictions for 2014 (Table 4). We next chose to look at the Chargers, Cowboys, Packers, and Patriots in greater detail. Figures 2 and 3 show the relative pass/run prediction accuracy by down for these four teams as a function of quarter. Values are relative to the average accuracies across all teams. A positive value in these figures indicates our pass/run model prediction accuracy is greater than the average of our prediction accuracy over all teams in the League. A negative value indicates our predictions are worse. In the former situation, we would infer that such a team behaves more like the average NFL team, and in the latter, we would infer the team behaves less like an average team. Absent bars for fourth down indicate a team did not execute a play in that quarter on fourth down during the 2014 season. Without exploring all aspects of the data in these figures, we offer the following select observations related to pass/run play calling:

Team	Pass/Run Accuracy	Season Record
Best-Predicted:		
Cowboys	77.0%	13 – 5
Chargers	75.2%	9 – 7
Broncos	75.0%	12 – 5
Worst-Predicted:		
Packers	67.9%	13 – 5
Eagles	68.0%	10 – 6
Patriots	70.2%	15 – 4

Table 4: Best- and Worst-Predicted Teams for 2014 (Pass/Run only)

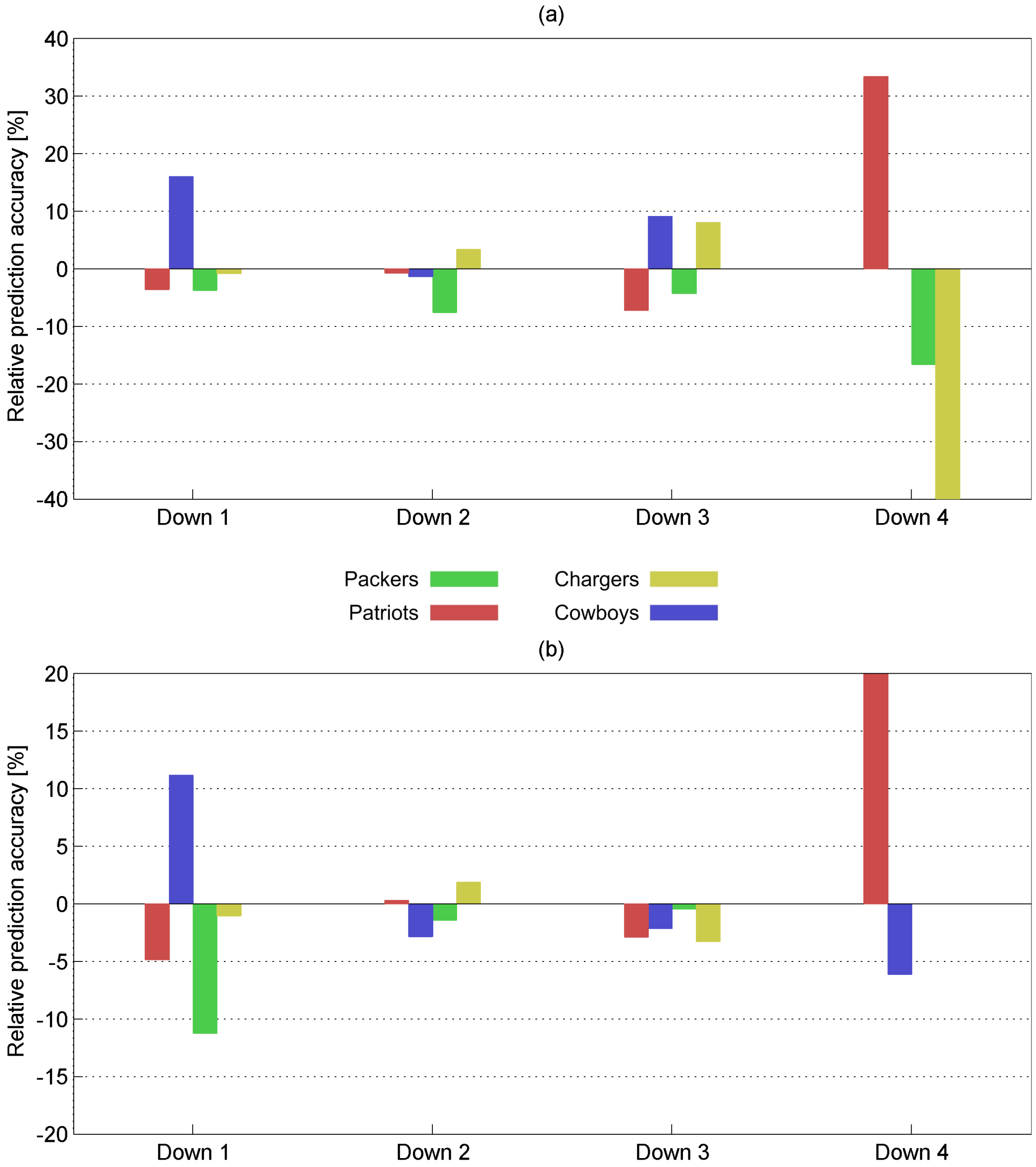

Figure 2: Relative pass/run prediction accuracy by down for four teams in the (a) first quarter and (b) second quarter. Positive values indicate a team is "more predictable," negative values, less.

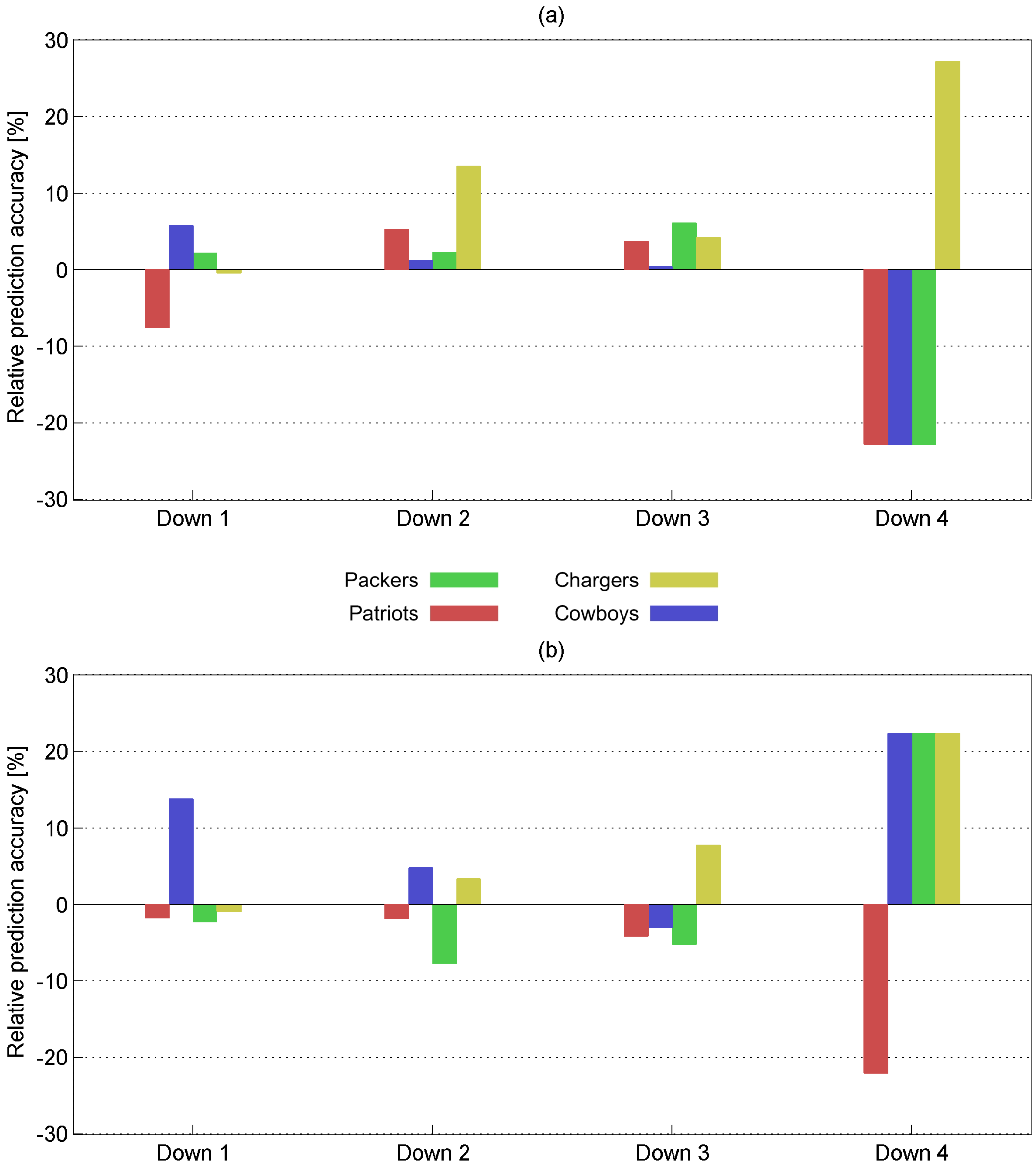

Figure 3: Relative pass/run prediction accuracy by down for four teams in the (a) third quarter and (b) fourth quarter. Positive values indicate a team is "more predictable," negative values, less.

- The Cowboys are very predictable on first down in all quarters
- Second and third downs appear easier to predict than first and fourth
- Our worst predictive accuracy occurs in the second quarter, and our best occurs in the third quarter
- On fourth down, the Patriots are predictable in the first half and unpredictable in the second half

As a final note, we looked to find correlations between how well a team is predicted and various measures of offensive "success," such as win/loss record, offensive yards per play, and other metrics. In all cases, we failed to find clear correlations.

4 Conclusion

In this paper, we have presented an advanced machine-learning approach to predict whether an opposing NFL offense chooses to pass or run. Additionally, our approach predicts the direction of run and the type of pass. Our pass/run prediction accuracy is, we would argue, the best reported to date. We have illustrated how such a predictive tool can be used to augment game preparation efforts by identifying teams and game situations that warrant greater scrutiny.

The results presented in this paper only scratch the surface of a deeper analysis of various teams, games, and plays. In particular, it would prove interesting to review the incorrect predictions in detail to determine if the deviations can be explained. Incorporating additional seasons of data would also be a worthwhile direction for future efforts. This would give rise to several new questions that could be explored: How do teams behave from year to year? Can an offensive coaching change be detected through predictive modeling? Does predictive performance improve with additional seasons of data?

During the course of this work, it quickly became clear that in-game prediction is complex and challenging. Ultimately, in-game prediction, in any sport, will always be challenging because it is driven by human behavior — behavior that cannot be captured by a computer model, no matter how sophisticated. And, "That's why we play the game — to see who'll win." [attrib. Adolph Rupp, 1965]

References

[1] J. Boronico and S. Newbert. An empirically driven mathematically modeling analysis for play calling strategy in American football. *European Sport Management Quarterly*, 1:21–38, 2001.

[2] W. Burton and M. Dickey. NFL play predictions. In *JSM Proceedings, Statistical Computing Section*, 2015.

[3] V. Carter and R. E. Machol. Optimal strategies on fourth down. *Management Science*, 24(16):1758–1762, 1978.

[4] J. D. Jordan, S. H. Melouk, and M. B. Perry. Optimizing football game play calling. *Journal of Quantitative Analysis in Sports*, 5(2):1–34, May 2009.

[5] A. Joseph, N. E. Fenton, and M. Neil. Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553, 2006.

[6] P. Kvam and J. S. Sokol. A logistic regression/markov chain model for NCAA basketball. *Naval Research Logistics (NRL)*, 53(8):788–803, 2006.

[7] D. Lariviere. Play prediction analytics could be coming soon to the NFL. *Forbes*, February 2015.

[8] S. Sahi and M. Shubik. A model of a sudden-death field-goal football game as a sequential duel. *Mathematical Social Sciences*, 15:205–215, 1988.

[9] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.