Keys to Victory: 
Predicting NFL Outcomes via Machine Learning

Wayne Zhu 
Advisor: Drew McDermott 
May 1, 2013

Abstract

The goal of this project is to construct a robust model for predicting National Football League (NFL) games through the use of machine learning techniques. Achieving this objective requires access to large sets of historical data as well as a broad comprehension of the current literature on the football analytics. The applications to computer science include substantial programming to scrape, clean, and merge data from multiple sources as well as understanding of several approaches in machine learning. The ability to predict the outcomes of games at a high rate is of interest to both casual fans and professional teams. The insights offer greater appreciation of the game alongside advantages in team organization. We begin this report with a brief history of advanced statistics in professional sports, leading into an extensive description of the methodology behind our research and the resulting model. We explored machine learning algorithms such as random forest and support vector machines in order to create a model that ultimately predicts the winner of individual games with a greater accuracy than that of leading ESPN analysts. The final model consisted of 29 features for each team, ranging from basic box score statistics to highly advanced aggregate statistics developed by the online sources of Pro Football Reference and Football Outsiders. The model was able to predict at a 67% accuracy across all games in the past two years and up to 77% for its most confident predictions. Finally, we offer a discussion of the implications of our work and fruitful areas for additional research that could impact how professional teams understand strategy and organization.

1 Introduction

The origin of statistics applied to sports beyond basic box scores is often traced back to Sabermetrics, a term coined by Bill James of the Society for American Baseball Research (SABR). Though James began publishing booklets as early as 1987, it was not until the late 1990’s that his ideas began to take hold in baseball communities. As detailed in the book,
Moneyball: The Art of Winning an Unfair Game, the Oakland Athletics were the first sports organization to seriously adopt the statistical findings of Bill James. The general manager at the time, Billy Beane, took on the challenge of leading an organization that had cut its payroll in half. Despite being amongst the lowest spending baseball teams in the league and having less than a third of the budget of the wealthiest team, Oakland was able to achieve one of the best records in the league along with moderate playoff success by finding deep bargains by way of James’s statistical scouting methods. This stark divergence from audience expectation legitimized statistics in baseball and sparked the beginning of modern analysis.

Recently, other sports, including football, have caught on to the statistical craze that swept baseball. Naturally it took longer to be accepted in football because while actions are very focused and discretized in baseball, there is a flurry of unpredictable movement between 22 players at a time in football. Not only does this make it more difficult to devise useful statistics, but it also causes results to be less provable and straightforward. Furthermore, statistical analysis in football has to face a highly risk-averse coaching culture that is slow to innovate. Tradition has a significant hold on the National Football League (NFL), and coaches often err toward established conventions and practices over new ideas.

As a result, the existence of rigorous statistics in football is quite young (less than a decade old), and it has only recently begun to turn heads in the industry. All franchises now have at least a few statisticians on the payroll. Many online communities and publications have emerged with truly inventive ideas. Even journalists are now beginning to reference what were once obscure metrics in their articles and news stories.

Through this project, we aim to utilize publicly available NFL statistics, in combination with machine learning techniques, to predict the outcomes of individual NFL games. The results serve as commentary on the relative importance of the different facets football; furthermore, they offer a way to measure the level of penetration of technology and data science in the league. The study also has implications on the value of professional NFL analysts who must publicly announce their own picks as part of their job. The model can serve as a basis for an objective lens through which to evaluate the contributions of such analysts who are respected for their subjective expertise.

2 Methodology and intermediary findings

For this project, we collected raw data NFL data from Pro Football Reference, a massive online database of NFL statistics. The data includes a plethora of statistics describing
individual game performance for all 32 teams in the past eleven seasons, 2002-12. Though
the current incarnation of the NFL has existed since 1970, there are several reasons for
restricting the data to the last eleven seasons. Firstly, the year 2002 marks the latest
major restructuring of the league, which included the reorganization of conferences, new
standardizations of scheduling, and the introduction of new teams. Additionally, there has
been a steady shift in the balance of power in how football is played, with the passing game
becoming increasingly more important relative to the run game. Though there is no specific
year in which this effect has increased dramatically, drawing a line at the 2002 season serves
as a reasonable boundary for delineating the “modern era” of the NFL. Since our restricted
set has plenty of data, these reasons suggest that going farther back than 2002 may actually
hurt rather than help the predictive power of the model.

2.1 Passing and rushing yardage

With the data set in place, we began to delve further into the individual statistics in
order to form explanatory variables for the model. In order to gain a rough idea of a teams
offensive performance, one natural direction to pursue is the team’s total yardage output.
Since gaining yardage is the medium through which teams score, it intuitively has a direct
relationship with points scored per game. The yards gained can be further divided into
passing yards and rushing yards. The data shows that the correlation between passing
yards per game and points scored per game is .658 while the correlation between rushing
yards per game and points scored per game is .365. The fact that passing yards serve as a
better predictor of points scored is in line with the observation that passing has become an
increasingly important part of the game. The relative influence of the two types of offense
on points scored can be better understood with the plot in Figure 1.

The contour plot of course shows that points per game is increasing with both passing
yards and rushing yards. Additionally, it is difficult for a team to score a lot (25+ points
per game) without being able to gain yards both ways. However, one complicating factor is
that teams play at different paces. A team can play at a high pace by allowing less time to
run off the clock between plays, thereby garnering more possessions in a single game. One
can easily imagine a team with this playing style that has a mediocre offense; it manages
to gain a great amount of yardage through its sheer number of offensive plays but it cannot
convert its drives into touchdowns. To better understand the quality of an offense, we can
examine the effectiveness of each individual possession. Therefore, we should look at the
same plot instead with passing yards per attempt (passing play) versus rushing yards per
carry (rushing play). Their respective correlations with points per game are .753 and .291.
The resulting contour plot can be seen in Figure 2.

Though it is similar to Figure 1, this plot allows us to better see the uneven weights of the two facets of offense. The most effective offenses are largely driven by a highly efficient passing game.

2.2 Home-field advantage

Next, one factor that is widely known throughout sports to have a strong effect on the outcome of a game is home-field advantage. Though the way schedules are structured during the regular season does not suggest the home team would be an inherently better team, there are certainly intuitive factors as to why teams would overperform when at home. The differences in playing at home and away are vast. For one, the crowd makes noise to disrupt the away team and break concentration. The away team has to face the stress of travelling, including flying several hours and adjusting to an unknown city, while the home players have the luxury and comfort of their own houses. The home team is able to practice
on the same field and roughly the same conditions for days leading up to the game. All of these reasons and more could contribute to the very real effect of home-field advantage.

Regardless of what causes it, we can measure the extent of the effect by looking at how it influences the outcome of the game. Figures 3 and 4 show the win percentage and average point differential for the home team across the years 2002-12. Overall, the home team wins games roughly 57% of the time and achieves an average point differential of 2.6 points.

2.3 Defense-adjusted Value Over Average

As mentioned in the introduction, advanced statistics are beginning to take hold in the world of football, and there are people pursuing football analytics as a full time career. One notable publication at the head of the business is Football Outsiders (FO), an online magazine that applies data science to football. Because of the vast resources of FO, they are able to create a team evaluation system that looks not at aggregate game statistics but rather each individual play by itself. Their proprietary evaluation formula called Defense-adjusted Value
Figure 3: Home team win percentage by year

Figure 4: Home team point differential by year
Over Average (DVOA), assigns value to every single offensive, defensive, and special teams play, producing a single score for each team that is adjusted for the strength of opponents. By running analysis on a scale as small as each individual play, FO is able to accurately capture a team’s true rate of producing successful plays. The authors understand that even a team with a low rate of success can string together a few successful plays resulting in a score, while there is predictive value in successful plays that do not lead to a score. Figure 5 displays full season data of average margin of victory versus cumulative DVOA, showing that even though DVOA looks at the value of individual plays rather than scores, it correlates very well with actual scoring. The correlation between the two for seasons 2002-10 is .928.

Figure 5: DVOA vs. average margin of victory

2.4 Special teams

A component of professional football that is often overlooked is the special teams plays, perhaps because it constitutes such a small proportion of the time of play. Special teams can cause drastic shifts in just one play, and they include kickoffs, punts, field goals, and kick returns. To fully understand what drives the outcome of games, we need to scrutinize this part of the game as well. While there has historically been fewer available statistics on special teams, Football Outsiders has created SPDVOA, a subset of DVOA limited to special teams.
that quantizes its impact on the game. However, upon carefully exploring the SPDVOA data, we learned why it is so often overlooked analyses: The variance in performance is so great from game to game that it is nearly impossible to predict the contribution of special teams in a single game. Intuitively, this is true because such a high proportion of special teams value comes from low percentage, idiosyncratic events such as blocked kicks and returns for touchdowns. We found that past special teams performance is a very poor predictor for current special teams production, and that it is simply not very useful for predicting the outcomes of games. As we see from Figure 6, there is only a mild relationship past special teams ability and margin of victory. The correlation between the two variables is only .1338, and the team with a better special teams unit wins just over 53% of the time. As a result, special teams performance is not a significant variable in our quest for predicting games.

Figure 6: Past special teams performance vs. margin of victory

2.5 Previous season adjustment

One major hurdle to overcome is lack of data for the beginning of the season. By the end of the season, a team has data from several games worth to form a robust prediction. However, early in the season there is much less data and more importantly none for the case of the first game. As a result, we had to carry over data from the previous season. Doing so
involved careful analysis because much is left up to discretion in the process. We considered using the previous season’s data not only to fill in for the first game but also to influence the data of several initial games in a decreasing manner. The result is a decreasing vector of adjustment factors between 0 and 1 that we use to interpolate between previous season averages and current season averages. To understand this better, let us use turnovers per game as an example. Suppose a team averaged 1 turnover per game in the previous year and 5 per game so far this year. If the adjustment factor is .25, then the predictive value for the next game will be \((.25) \times 1 + (1 - .25) \times 5 = 4\). In order to determine the best way to form this adjustment, we must first resolve two things. First is the length of time that the previous season should affect the current season’s predictions: anywhere between 1 to 16 games. Second, it requires us to determine the manner with which this influence decreases. We experimented with different linear, harmonic, and exponential sequences. Ultimately, the most effective result we found was a linear sequence of length 8 from 1 to .125. In other words, we allowed the previous season’s data to influence the predictive model for the first 8 games of the season, with the adjustment decreasing by .125 per game.

### 2.6 Other features

The previous sections have shown several methods of analysis and data inspection for understanding whether a variable is suited for inclusion in our model. We examined dozens of possible features with varying degrees of success, and a breakdown of each one would be repetitive without being additionally informative. Ultimately, the model includes 29 features derived from the data to describe each team. Comparing these features across two teams allows us to predict the outright winner, the point spread winner, the margin of victory, and the result of the over-under. The techniques used to make predictions include linear regression, logistic regression, support vector machines, and random forest learning. The performance and comparison with leading NFL analysts are detailed in the results section.

### 3 Results

The following results were compiled from predicting the 2011 and 2012 regular seasons, and the respective training sets used were 2002-10 and 2002-11. Across the two seasons tested, there was only one tied game and it was not included for the purposes of calculating winning and prediction percentages. The data was compiled in CSV format and all of the data scraping, analysis, and modeling were done in the statistical programming language, R.
3.1 Game winner

The best model for predicting the outright winner for the games was formed through random forest learning. Random forest is a technique that constructs a “forest” of decision trees, and prediction works by returning the median (for classification) or mean (for regression) response by these decision trees. For our model, we ran a 5000 tree random forest against the margin of victory, predicting wins for values over 0 and losses otherwise. The results are compared to fourteen ESPN predictors: one prediction algorithm “Accuscore,” and one crowd-sourced poll “Pick’em,” and twelve expert human analysts. The prediction percentages shown are calculated from cumulative records from up to the past five years. These are the prediction rates when the model is forced to select a winner for each game. When we can choose to only declare a winner for the most confident predictions (limited to a quarter of observations), a linear regression model can result in prediction accuracies of up to 80.8% for 2011 and 77.1% overall.

Table 1: Comparison of prediction accuracies

<table>
<thead>
<tr>
<th>Name</th>
<th>Accuracy</th>
<th>Name</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. Allen</td>
<td>61.4%</td>
<td>T. Jackson</td>
<td>62.9%</td>
</tr>
<tr>
<td>M. Golic</td>
<td>63.2%</td>
<td>K. Johnson</td>
<td>60.2%</td>
</tr>
<tr>
<td>M. Hoge</td>
<td>64.9%</td>
<td>M. Ditka</td>
<td>64.8%</td>
</tr>
<tr>
<td>R. Jaworski</td>
<td>65.9%</td>
<td>C. Carter</td>
<td>66.0%</td>
</tr>
<tr>
<td>C. Mortensen</td>
<td>63.7%</td>
<td>Accuscore</td>
<td>66.3%</td>
</tr>
<tr>
<td>A. Schefter</td>
<td>61.5%</td>
<td>Pick’em</td>
<td>65.7%</td>
</tr>
<tr>
<td>M. Schlereth</td>
<td>62.9%</td>
<td>Model (2011)</td>
<td>66.8%</td>
</tr>
<tr>
<td>S. Wickersham</td>
<td>64.7%</td>
<td>Model (2012)</td>
<td>67.8%</td>
</tr>
</tbody>
</table>

As Table 1 shows, the model bested all of the analysts’ career records in both years. ESPN’s Accuscore was its best predictor yet still fell short of both marks by our model. While the experts do have years of experience as players, announcers, and analysts, their subjective opinions ultimately cannot stem from the same amount of raw inputs that the random forest model can take in. Unless they watch each play of every game in the entire season, the raw amount of information the experts have to work with is dwarfed by the several seasons of statistics that serve as the base of our model. Considering that data scientists working within NFL franchises have far greater access to even more advanced statistics than the public, their ability to model data-heavy questions must far exceed that of individual experts. Any comparative advantage these experts have over statistical models must instead
lie in aspects of the game in which data is not easily quantized or compared.

3.2 Spread and over-under bets

The spread and the over-under are the two most common tools of the sports betting world. They are bets formulated such that both sides stand an equal chance of winning. While the spread bet is based on the difference between the two teams’ scores, the over-under is equivalently based on the sum. The spread bet is described as a “line” displayed in the following format: “Denver Broncos (-7.5) vs. Oakland Raiders.” This expression (usually with the home team listed first) indicates that Denver is a 7.5 point favorite against Oakland. Accordingly, participants may bet on which point total will be greater, Denver’s score minus 7.5 or Oakland’s score. Similarly, the over-under bet is simply described by a number equal to the expected sum of the scores, and participants can choose the “over” or the “under.” If the resulting point differential or sum of scores equals the spread or the over-under, respectively, a tie or “push” occurs, and money is returned as if no bets were made.

However, if a participant can model the outcomes of games better than the entity offering the bets, then he can choose the winning team more than half of the time. The reason for testing against these bets in this project is that doing so serves as a true measure of whether the model offers any new insight that is not priced into the market already. The following predictions are tested against lines from the Las Vegas Hotel (LVH), the standard in the sports betting industry. The best spread model resulted from linear regression, predicting the 2011 and 2012 seasons at rates of .595 and .541, respectively, for a combined .568. The overall accuracy is greater than .500 with 99% confidence. But, this result does not mean the model can serve as a money making machine. This is because in order for the bookkeeping entities to make money, the payouts for the bets are not even odds, meaning that the amount of money won from making a correct bet is less than the amount lost from an incorrect bet. Generally, participants must risk $1.10 to $1.15 to win $1.00. If we use conservative odds of 1.15:1, the break-even rate is .535. While the model’s accuracy does exceed this rate, it is not by a statistically significant margin.

A similar result can be seen with the over-under bet. Our best over-under model resulted from random forest learning, and it produced success rates of .578 (2011) and .541 (2012), for a total of .560. This accuracy also exceeds .500 with 99% confidence but does not exceed the break-even rate with statistical significance. The comparisons and p-values can be seen in Table 2.
<table>
<thead>
<tr>
<th>Bet type</th>
<th>Spread</th>
<th>Over-Under</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>LR (n = 498)</td>
<td>RF (n = 502)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>56.8%</td>
<td>56.0%</td>
</tr>
<tr>
<td>Better than even (.500) p-value</td>
<td>.002</td>
<td>.007</td>
</tr>
<tr>
<td>Profitability (.535) p-value</td>
<td>.134</td>
<td>.263</td>
</tr>
</tbody>
</table>

### 4 Code and deliverables

The data for this was collected entirely from *Pro Football Reference, Sunshine Forecast*, and *Football Outsiders*. It is all stored and organized in CSV formatted files and the data includes box scores, team statistics, betting lines, and expert predictions. Additionally, the project includes over 1000 lines of code in the *R* programming language, with purposes including the creation and testing of models, producing plots, cobbling together data sets, and analyzing the data in an understandable manner. Each script, primarily the prediction scripts, have abundant comments describing how to run and manipulate the code. Table 3 contains an extensive outline of the most important files and folders included in the project.

“Model.r” and “Predictor.r” are the main modeling scripts. The former contains all of the code for predicting 2011-12 outcomes of each type described in results. First, it takes in data scattered across the multitude of CSV files, standardizes it, and combines it into useable training and test data sets. Then, it runs the model depending on the prediction technique and the model type. The prediction techniques include logistic regression, linear regression, random forest, and SVM, while the model types include game outcome, spread pick, and over-under pick. The top of the script includes all of the available options including the choice of which variables to test. “Predictor.r” is a similar script except that it is meant to predict a game with data of the user’s choice. A user can enter data for two teams in “Test.csv” and run the predictor script to predict which team will win. All of the outcomes provided in the results section ultimately resulted from these scripts.

### 5 Further research

The analysis from this project lends itself directly to several broad, impactful paths to explore given more time and resources.

- Predictive abilities can be improved by using the more granular play-by-play data
Table 3: Descriptions of important files and folders

<table>
<thead>
<tr>
<th>File or folder</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model.r</td>
<td>Main R script that pieces together the data into training (2002-10) and test (2011-12) sets and makes predictions.</td>
</tr>
<tr>
<td>Predictor.r</td>
<td>R script that takes in the user-inputted contents of “Test.csv” and predicts the outcome of a single match between two teams.</td>
</tr>
<tr>
<td>Plot.r</td>
<td>Created several plots, both for visualizing trends in the data as well as those used for this report.</td>
</tr>
<tr>
<td>Fuse.r</td>
<td>Fuses together data from the TeamData folder and Lines folder in order to form a useable data set that is then saved into the Frames folder.</td>
</tr>
<tr>
<td>Frames (folder)</td>
<td>CSV formatted files that include full seasons of games with basic information such as the spread and over-under picks as well as the resulting scores and yardage for both teams.</td>
</tr>
<tr>
<td>TeamData (folder)</td>
<td>Several CSV files per year for the schedule, pass and rush offense, pass and rush defense, special teams, and miscellaneous statistics.</td>
</tr>
<tr>
<td>Lines (folder)</td>
<td>Raw betting information on the spread and over-under lines downloaded from Sunshine Forecast.</td>
</tr>
<tr>
<td>Raw (folder)</td>
<td>Original unmodified copies of the downloaded data.</td>
</tr>
<tr>
<td>Test.csv</td>
<td>Testing inputs for Predictor.r to predict single outcomes.</td>
</tr>
<tr>
<td>Variables.xlsx</td>
<td>Dictionary for abbreviations of all relevant variables.</td>
</tr>
<tr>
<td>Expert.xlsx</td>
<td>Record of expert prediction rates.</td>
</tr>
</tbody>
</table>

rather than aggregate game day. This venture presents more challenges, including language parsing, but it can lead to a greater understanding of which variables influence games and to what extent.

- Teams can use predictive models to better analyze which facets of the game they are strong or weak in. Given constraints such as salary caps and market rates at certain positions, managers can use models to discover what roster moves would bring them the greatest increase in their chances of winning. These results have implications on which players they should allot time to, draft over the summer, and sign or release during free agency.

- Analysis of how the weights of the variables in the model change over time can shed light on how the game itself has changed. Building an understanding of how the importance of certain aspects of the game have shifted over time can allow teams to predict future trends and stay on top of strategy and player management.
6 Acknowledgements

I want to thank Professor Drew McDermott for taking time this term to advise me on this project. His help in offering ideas and direction proved invaluable throughout the term. Additionally, I want to note my appreciation for all of the analysis done at *Football Outsiders*. The publication’s daily columns have greatly influenced my understanding of not only football but also methods and practices of data analysis.
References


