NFL demonstration system

This document describes how I build an NFL betting system with a target 55% win rate and no maintenance for three years. The point was to prove that it's simple to do. I made a lot of arbitrary choices, there are a huge number of equally good systems that could be built using similar methods, as well as many much better systems that could be built either by devoting more effort using these methods or using other methods. But no one, to my knowledge, is successful at sports betting using the methods taught in most statistics courses.

The system is built with an additive binary factor design. Instead of looking for a single complex indicator with a 55% win rate, we'll look to combine less accurate indicators. The reason for this is 55% indicators are too hard to find, or rather, they're too easy to find, so they are eliminated. Any simple rule based on easily available data that made money against the spread is unlikely to survive long. However a 52% indicator is not only much less obvious, it does not by itself represent a profit opportunity for a customer paying full juice, so it is slower to be corrected.

Another advantage of factors is each one can be monitored independently. With a complex model it's hard to tell if it's not working, except by losing money. With a factor model, we can adjust or replace factors that stop working, hopefully while we're still operating with an overall profit. Binary factors in particular are easy to monitor. The more values a factor can take, the more complicated it is to evaluate its performance. With a binary factor, we look only at the rate teams with the factor cover the spread.

For similar reasons, we're just going to add our factors together rather than using more complex logic. Here the advantage is we can monitor each factor individually over all games, not just the games we bet. Another advantage of additive binary factor models is the scope for overfitting is less than with most schemes.

Factor one: over-reaction

For our first factor we will look at one of the most generally reliable predictors, overreaction. We will bet on the team the line moved against versus what we would have expected the prior week. For example, suppose Green Bay plays Chicago in week 7. Before the games of week 6, we would have expected Green Bay to be a 3.5 point favorite, but after the week six results, Green Bay is only a 1.5 point favorite. We're going to guess that this is an overreaction, and bet on Green Bay.

There are three reasons we think this should work. First is the mathematical concept of shrinkage, a very powerful real world tool that is usually taught as a minor mathematical curiosity in statistics texts, if it is mentioned at all. The line move is a combination of signal plus noise. In the example above, we know the line move was 2, so we can call it the sum of S + N, where S is the move that would have made the game exactly a 50% proposition.

If N is positive, the move was too big, so we want to bet on the team the line moved against. If N is negative, the move was too small, and we want to get on the team the line moved in favor of. The magical thing is that knowing nothing at all about the facts, we know that if N is positive, S + N is more likely to be positive than it would if N were negative. So since S + N is positive, we conclude that N is more likely than not to be positive.

One defect of the academic treatment of shrinkage is it tends to assume the volatility of N is constant, so the shrinkage effect is most powerful for line moves near zero and negligible for large line moves. Empirically I have found that the opposite is more common, the bigger the line move, the bigger the volatility of noise relative to signal; and not just in football line moves, in general changes. So I trust shrinkage more for the big line moves than the small ones.

The second reason is behavioral. If you want to win a majority of bets, it's usually wise to bet that less will happen than people expect (if you want to win a minority of bets—but win by a lot—take the other side, but in football betting we only care if a team covers the spread or not, not by how much). People systematically underestimate the probability of nothing happening, and also underestimate the likely size of what happens if something does happen.

Finally a feature of price-setting systems like football betting is over-reaction to new information. Most bookies take lines from others, but the decision maker bookies show their lines to insiders and experts before announcing them to the public (those insiders and experts are allowed to make fixed-size bets that cost the bookies money on average, but gives information worth more to the bookies than the cost of paying off). If the money favors one side, the bookie will adjust the line accordingly. However the cost of underadjusting is less than the cost of overadjusting, so she will tend to move the line too much.

Market makers in financial markets do the same thing when their book becomes unbalanced. Moving too much at least gets rid of the imbalance, and means that if you have to make a second move, it's more likely to be in the opposite direction than the same direction. Moving too little makes the imbalance worse and increases the probability that you have to make a subsequent move in the same direction.

Every reaction has a predictable over-reaction

To bet on overreaction, we need to measure the reaction. That means we need to estimate what the line in a week seven game would have been before the week six games were played. I chose a simple model. Each team has a power rating. In a game at a neutral site, a team is favored by the difference between its power rating and the other team's power rating; a negative difference means a team is an disfavored by that many points. I assumed a home field advantage of three points, so we expect the home team to be favored by three points more than the difference in power, and the visiting team to be favored by three points less.

To assign power ratings to all 32 NFL teams we have 31 unknowns (one team's rating is arbitrary since adding a constant to all ratings makes no difference). Each week gives us 16 or fewer lines to use as data, lower become teams have some bye weeks. So we need at least two weeks of games to get a fit, and usually three if there are bye weeks; but I chose to use four for stability. However because I want to get up-to-date estimates, I weighted the current week 1, the prior week 0.5, the week before that 0.25 and finally 0.125 for three weeks ago. I then solved for the set of power ratings that minimized the weighted sum of squares between predicted and actual lines. If I were really going to bet on this, as opposed to making a demonstration, I'd just download power ratings from one of the more reliable publications.

We bet on the teams favored by fewer points than the difference in our power ratings implies, or disfavored by more points. So if the difference in power ratings suggests Green Bay should be favored by 3.5 points, and it's actually favored by 1.5, we bet on it. If Green Bay were actually favored by 6 points, we'd bet on Chicago.

Before we test this idea on data, there is another important step. We always want at least two different ways to test a factor. One we'll bet on, one we'll use as protection against data mining during the fitting phase, and as a quality control mechanism afterwards. For real applications, I like at least three checks. Ideally these checks use different data and different methods to test the same theory, however they do not have have to be usable for betting, for example, they can use future data. We'll consider them in retrospect to see if our theories were sound.

Rematch

A simple check on the over-reaction theory is to compare the lines when two teams face each other twice in a season. If our theory is correct, whichever team is favored by more in the second game than the first (or disfavored by less, or favored in the second game but disfavored in the first game) should lose more than half the time. This is noisier than the indicator we're going to use for betting, because more weeks elapse between power estimation and game, and because we have fewer games to test. But it's more direct, it does not require fitting power ratings.

If the teams play each other at the same site, it's particularly easy, we can just compare the two lines. However this is the less common situation, most rematches are two teams in the same division playing one game in each of the two home stadiums. We could retain our assumption of a three point home field advantage to adjust the line, but I prefer to test that assumption as well. So instead I do a regression of the line change for all games of this type, with the dependent variable the amount the home team is favored in the second game and the independent variable the amount it was favored as a visitor in the first game. Positive residuals from this regression are treated as the line moving in favor of the home team in the second game, negative residuals are treated the opposite. This regression does not bias the in-sample results, because it uses no information about which team covered the spread, only the information in the line move itself.

Let the data begin

Now we finally get to some data. I did this in June 2006, and used data from the 1983 to the 2005 seasons. I used the odd years to fit the model, and held the even years out for out-of-sample validation. Over the test seasons, the team the line moved against covered the spread in 1,499 games and failed to cover in 1,302 (in 141 games either the line did not move, or the spread was matched exactly so all bets were returned). This is a 53.5% win probability, above our goal of 52% for a single factor.

How about our validation exercise? There were 40 examples in the eleven test sample seasons of two teams playing each other twice in the same season in the same stadium. In only 19 of them did the team the line moved against win. However in the 609 home-away matches, the team the line moved against won 320 times. Combining the two means this more direct version of our rule worked 52.2% of the time, giving us some confidence we are exploiting a real tendency of line moves to overreact rather than an accidental feature of the data.

As written here, the only use I made of the sample data was to fit the regression for the home-away rematch validation. However I did look at the data and played around with alternative formulations. The results above are definitely in-sample. So next I looked at performance in the twelve odd numbered seasons, which I had scrupulously avoided looking at while working on the model (this is not completely out of sample, because I have been a quantitative football bettor since the 1970s and was familiar with statistical properties and theories, but it was the best holdout I could do).

The power indicator was right in 1,398 games, wrong in 1,248 with 53 ties. This is a 52.8% winning percentage, below the test sample result but still above 52%. The rematch test worked in 22 of 40 same-stadium rematches and 327 of 651 home-away rematches. This combined to only a 51.0% success rate which is lower than I'd like to see, but not enough to knock this indicator out of the system.

To the extent possible, I want the factors to be independent, to use entirely different data, and to rely on different types of theories. Overreaction used only line data, nothing else, and was based on mathematics and behavioral principles, as well as the organization of the betting industry. So the next one is based only on a statistic from game play, and is supported by an idea of how bettors, as opposed to bookies, interpret evidence.

Factors two and three: turnovers

One of the most important score determinants in football is turnovers—fumbles lost and interceptions—and they are pretty random, much more random than the other major aspects of game performance. Therefore the team that gave the ball away more in a game is probably better than the score suggests, and the team that received more turnovers than it gave, is probably worse than the score suggests. So we are going to bet on teams that had net giveaways in the last game, and against teams that had net takeaways.

Turnovers gives us two indicators, one for the visiting team and one for the home team. The indicator works when a team that gave the ball away more than it took the ball away last game wins, or when a team that took the ball away more than it gave the ball away last game loses.

In the test sample the indicator worked for home teams 1,188 times and failed 1,105 times, with 649 ties. For visiting teams the numbers were 1,229 wins, 1,088 losses and 625 ties. These are 51.8% and 53.0% success percentages respectively. In the holdout sample, the home indicator worked 1,067 times, failed 1,101 times with 531 ties; the visiting team version worked 1,092 times, failed 1,077 times with 530 ties. These are 49.2% and 50.3% success rates respectively.

The holdout failure of the turnover indicator is worrisome. It did work in the combined test-plus-holdout period, and it is a simple indicator with good intuition behind it. So we will hold it on probation for now.

Season-long validation

To validate the turnover indicator, we'll look at how season turnover differential predicts wins and covers. For each game we'll look at the net turnovers for each team in the prior games for the season and subsequent games. That is, we'll add up all the times the team gave the ball away through fumble lost or interception, and subtract all the times the team took the ball away. A positive number is bad for the team, it means it lost the ball by turnover more than it gained the ball. Then we'll subtract the visiting team's number from the home teams.

Common sense predicts a negative relation between the home-minus-away turnover number and the home team winning the game. A positive number means that the home team had more net turnover losses, or fewer net turnover gains, than the visiting team; thus it is probably a worse team; thus it will probably lose. The relation should be symmetrical, past turnovers and future turnovers should be equally good predictors of team quality.

This is borne out in the test sample by a regression of a home team win indicator on prior turnover differential and another regression on subsequent turnover differential. The coefficient on prior turnover differential is -0.017 (0.002 standard error, 8.6 t-stat) and on subsequent turnover differential it is -0.014 (0.002 standard error, 7.2 t-stat). So for each extra turnover the home team has given away over the season, whether in the past or the future, relative to the visiting team, the home team has about 1.5% less chance of winning the game (this is not precisely correct, we need to do more work to convert a regression coefficient to a probability, but it is reasonably close for small probabilities).

If our turnover indicator theory is correct, we will see the opposite relation for prior turnovers if we regress an indicator for the home team covering the spread instead of the home team winning. Although the team that gave the ball away more net is probably the worse team, we think the line overcorrects for that by ignoring the large random component to turnovers. So we expect the team with the worse turnover record to be more likely to cover.

On the other hand, the line cannot take into account subsequent turnover differential, so we would expect that to have the same negative sign in a cover regression as in a win regression. This is in fact what we see in the test sample. The coefficient of the regression of a home team cover indicator on prior season turnover differential is 0.002 (standard error 0.002, t-stat 0.9) and on subsequent season turnover differential it is -0.006 (standard error 0.002, t-stat 3.0). Although the t-statistic on the prior season regression is marginal, what gives us faith in the turnover indicator is the difference between the prior season statistic and the subsequent season statistic.

In the holdout sample, the validation test performs better. The prior season coefficient is 0.007 (standard error 0.002, t-stat 3.8) and the subsequent season one is -0.006 (standard error 0.002, t-stat 3.0).

The squeaky wheel gets the grease

Our final factor is based on betting economics. Many bettors bet consistently on their favorite team. Bookies prefer all of their customers to lose a little. If a customer loses too much, he might quit, or not pay, or cause trouble. If a customer wins too much, he might spend the money rather than saving it to give to the bookie later. Therefore it’s best for the industry if every team in the league covers in half its games and fails to cover in the other half. So we’ll bet on the team that have failed to cover in more than half its games in the season to date, the hungry team (more accurately, the team with hungry fans) and against the teams that have covered in more than half, the fat team.

Notice that we’re exploiting three different kinds of theories. The power indicator is based on the idea that people set the line wrong. Of course there are plenty of smart betting professionals who understand overreaction, and in much more sophisticated forms than we’re using. Nevertheless, the betting market processes this information incorrectly. Back in the 1970s when organized crime controlled the entire national betting system, there was no overreaction. But it is hard to avoid in a dispersed market. This indicator can be exploited with relative impunity, even at a retail betting site, because it’s not immediately recognized as a smart bet.

The turnover indicator is based on bettor errors, which the market understands but adjusts the line anyway to get balance in the action. Most of the large market participants are not gamblers, and are not interested in winning by predicting games better than their customers. They want to have equal bets on both sides of every game so they collect 5% as a risk-free profit, every game, every week. They have no interest in writing checks to people who figure this out, which is why this would be a toxic indicator to use if we wanted to use a retail site and disguise our character.

The hunger indicator is based on economics of the gambling business. It is a doubly toxic indicator to exploit. The bookies are giving up balance in order to direct money preferentially to losers and way from winners. If you pile on to that imbalance you hurt them in two ways, making the imbalance worse and making money. Anyway, we like the diversity of rationales, and it’s important to know whom you are fooling and whom you are helping when you design betting systems.

Feeding the hungry

The visiting team version of the hunger indicator worked in 1,192 games and failed in 1,065, with 685 ties. This is a 52.8% success rate. The home team version worked in 1,155 games and failed in 1,106 for a 51.1% success rate. In the holdout sample, the visiting team version worked 52.7% of the time (1,124 wins versus 1,010 losses with 565 ties) while the home team version was 50.4% (1,062 wins versus 1,045 losses and 592 ties).

The visiting team version of the indicator is strong, but the home team results are marginal. Now there are differences between home and away line setting. Fans bet more on their team when it’s playing at home, and that skews some computations. Since our hunger indicator is based on overall season to date covers by a team, regardless of whether they are home or away games, it is relevant to the subset of fans who bet on the team both home and away. It's more efficient to equalize results for this group when the fans who only bet on home games are not involved. But this is tenuous reasoning. Moreover, I’m suspicious of after-the-fact rationalizations after looking at data, I prefer hypotheses advanced for theoretical reasons that are later confirmed by data. However, in this case, with the visiting version of the hunger indicator strong, and the home team version positive but weak, I prefer to keep both for parsimony.

If our hunger theory is correct, for any team at any point in the season, there should be a negative relation between the net times it has covered the spread in the season to date and the number of times it will cover the spread in the remaining season. So for validation we do a regression, of those two values. In the test data set we find that average future net covers for a team in a season is -0.042 (standard error 0.015, t-stat = 2.9) times the number of past net covers. For example, suppose at the halfway point in the season, a team has covered 2 times and failed 6 times for a net -4 covers. The regression predicts that is will have a net positive 0.168 covers during the remainder of the season. This provides statistical validation, it's unlikely we would see these regression coefficients if our there is no hunger factor, but it's also good to see that the size of the effect is clearly enough to make at least some bets profitable. The validation for the holdout sample was somewhat weaker, but still strong enough to accept the indicator. It gave a coefficent of -0.029 (standard error 0.015, t-stat = 2.0).

Gang of five

At this point, we have five indicators: power, turnover home and away and hunger home and away. Our perfect bet is on a team that the line moved against since the prior week, that had more turnovers in its last game than its last game opponent and has failed to cover more than it has covered in the season to date, playing a team that had fewer turnovers in its last game that its last game opponent and that has covered more often than it has failed in the season to date.

Because we're treating these as factors, we're not going to look for complex patterns among the indicators. If we get at four or five indicators in the right direction, we'll bet; we'll also accept three indicators in the correct direction if the remaining two are neutral. We're also going to use a constant bet size rather than betting more when our signals are stronger or agree better.

At our last step before placing a bet, we're going to check two things. How the system performed in the test and holdout samples, and the marginal contribution of each factor. Our data so far suggest that each factor works (although the evidence was only clear for three of the factors, suggestive but weak for the other two). If the factors work together, we might have a profitable betting rule. But if they overlap or conflict, the combined rule may be no better—or even worse—than the standalone factors. Moreover we don't know how many bets the rule will recommend, it might not be enough to be worth our while.

On the test sample, the rule made bets in 956 of the 2,942 games and had 545 wins versus 411 losses for a 57.0% win rate. Betting one unit on each game, paying full 10% juice, led to a profit of 93 units. In the holdout sample of 2,699 games we had, 459 wins versus 403 losses, a 53.2% win rate and 16 unit profit at full juice. This is below our 55% target, but still profitable.

At the margin

For the marginal contributions, we consider only the bets on which a single factor made a difference. If we made a bet but would have not done so without the factor, we charge the opposite of the bet result the factor; if we did not make a bet but would have without the factor, we charge the result of that bet to the factor. In the test sample, the power factor had a net 41 win contribution—wins minus losses on bets we would not have made without it, plus losses minus wins on bets that we would have made without it. The turnover indicator home team version contributed 62 net wins, and the away team version 63 net wins. The hunger indicator home contributed 44 net wins in each of home and away versions.

We had the turnover indicator on probation, but it had by far the strongest marginal results. So while it is mediocre at predicting outcomes as a standalone indicator, it is superlative at making the close calls.

We were also concerned with the home team version of the hunger indicator, but it also has very strong marginal results.

Note that the sum of the marginal contributions is 254 games, but we only had 134 net wins in the sample. That's because more than one factor could be decisive in a given game. If all five factors agree, no single factor could change the decision, so there is no marginal credit given. But if four factors are positive and one is negative, all four positive factors get marginal credit, because if any one of them had been left out, no bet would have been made. If three factors are positive, one is negative and one is neutral, no bet will be made, but if the negative factor switched to neutral or the neutral one to positive, a bet would have been made, so both of those factors get credit for preventing the bet and earn the opposite of what the bet result would have been.

Despite the double counting, if we were to eliminate an indicator from our system, we would lose the number of wins indicated by its marginal contribution (if we eliminated all five we would not lose the sum of the marginal contributions, obviously, because after we eliminated one, the marginal contributions of the others would decline).

In the holdout sample the marginal contributions are smaller, 30 net wins for the power indicator, 14 for home turnover, 2 for away turnover, 23 for home hunger and 44 for away hunger. These are acceptable numbers, although weaker than we would have liked. The away turnover marginal contribution is near zero, but this was a strong indicator by other measures.

And that's all there is

We have now built a system. All this work took me one afternoon, although to be fair I already had a lot of background knowledge, as well as the necessary data (not just the data that made it into the system, but data for some other factors I tried). This is my point, that it's easy to beat the bookie, or the market, or the house, or City Hall. You use simple quantitative methods, you don't need fancy math or inside information or rare insight. You get a small edge, not the kind of certainty too many people claim, but an edge that can produce reliable profits over long periods of time.

Building a professional quality system is much more work, but there's no secret. You get more and better data, you exploit more and better factors, you bet in more and better ways. The difference is quantity of work, not quality.