**Football Betting Prediction Analysis**

By Ty Darnell, Jace Gilbert, and Michael Steffan

## Context

In sports betting, a spread is a chosen number that separates possibilities of betting into two categories. A ‘favorite’ is said to cover the spread if they win the football game by more than the spread number. An ‘underdog’ is said to cover the spread if they lose by less than the spread, tie if the spread is not 0, or win the game. This means that an ideal spread for a bookkeeper would be split the theoretical odds of a team covering into a 50-50 chance. However, if the score difference in the game ends up equaling the spread determined before the game, it is determined a push and the better has their money returned.

The vigorish is the amount charged by the bookkeeper for taking a bet from a better. This means that a bet usually returns less than double your original amount (ie, better must risk $11 to get $21 in a correct prediction). On average, this ensures profitability for the bookkeeper. And on average, the better needs to be 52.38% successful to make a profit due in the long run. Our goal in our analysis, and eventual models, was to create a system that was profitable under these conditions.

## Our Data

Our original dataset included 8 variables; 5 categorical and 3 numerical. One variable in the original dataset, “indicator” was separated out into multiple variables considering it including multiple pieces of information per data entry involving the day of the week, if the game was a night game or not, and the overtime status of the game. After cleaning the data and creating a few more specifically interesting variables as combinations of other variables, as well as the division that each teams plays in, the main, final dataset had the form seen below.

Specifically created variables are ‘spreadcover’ which denotes whether the favorite, underdog, or neither covered the spread of any given game. The variable ‘spreadteam’ specifies which team, if any, covered the specific game. Other variables are notably labeled. Given the sparseness of correlating and contextually strong data, it became immediately necessary to start looking deeper into some specifically useful variables and to create some others.

**Summary Table for Football Dataset**

| Column Name | Column Type | Column Length | Column Label |
|---|---|---|---|
| DOW | char | 2 | Day of Week |
| fav | char | 3 | |
| favored_division | char | 11 | |
| ot | char | 5 | |
| spreadcover | char | 11 | Does favorite, underdog, or neither cover the spread? |
| spreadteam | char | 3 | Which team covered the spread? |
| und | char | 3 | |
| underdog_division | char | 11 | |
| DivGame | num | 8 | Are the two teams from the same division? |
| Night | num | 8 | |
| error | num | 8 | Actual Difference of Scores - Predicted Difference(Spread) |
| favpts | num | 8 | Points Scored by Favored Team |
| ha | num | 8 | Is favored team the home team? |
| ptdiff | num | 8 | Favored Team Points - Underdog Team Points |
| pts | num | 8 | Spread of the Game |
| undpts | num | 8 | Points Scored by Underdog Team |
| wk | num | 8 | |
| year | num | 8 | |

Referencing the table on the next page, some interesting statistics can be seen in the fact that the distribution of the number of times being favored for each division by a team represented in the game is fairly equivalent (note that this removes in-division games where two teams that are

playing are from the same division). We can note that over the 3 years, the spreadcover was fairly close to a 50-50 split, as well as the error of the spread from the actual score difference of the game was only 0.79. However, the standard deviation of the error was 13.26. This means that the bookkeepers were fairly accurate, but the error had a wide spread. Note that specific variables such as 'teamcover', 'wk', 'year', etc., are not include because they are either uninformative and/or excessive in length to include.

| | Level | Measure |
|---|---|---|
| **Total Rows of Data** | | 672 |
| **Categorical Variables** | | |
| **ha - n (%)** | 0 | 224 (33.3) |
| | | |
| **ot - n (%)** | 0 | 640 (95.2) |
| | 1 | 32 ( 4.8) |
| | | |
| **DOW - n (%)** | Mo | 49 ( 7.3) |
| | Sa | 16 ( 2.4) |
| | Su | 599 (89.1) |
| | Th | 8 ( 1.2) |
| | 1 | 448 (66.7) |
| | | |
| **favored_division - n (%)** | AFC CENTRAL | 97 (14.4) |
| | AFC EAST | 106 (15.8) |
| | AFC WEST | 119 (17.7) |
| | NFC CENTRAL | 107 (15.9) |
| | NFC EAST | 127 (18.9) |
| | NFC WEST | 116 (17.3) |

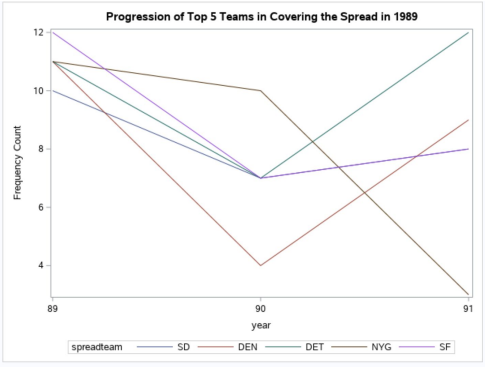
| | | |
|---|---|---|
| **underdog_division - n (%)** | AFC CENTRAL | 95 (14.1) |
| | AFC EAST | 134 (19.9) |
| | AFC WEST | 121 (18.0) |
| | NFC CENTRAL | 133 (19.8) |
| | NFC EAST | 113 (16.8) |
| | NFC WEST | 76 (11.3) |
| | | |
| **DivGame - n (%)** | 0 | 360 (53.6) |
| | 1 | 312 (46.4) |
| | | |
| **Night - n (%)** | 0 | 584 (86.9) |
| | 1 | 88 (13.1) |
| | | |
| **spreadcover - n (%)** | favorite | 330 (49.1) |
| | not covered | 19 ( 2.8) |
| | underdog | 323 (48.1) |
| **Continous Variables** | | |
| **undpts (mean (SD))** | | 16.86 (9.27) |
| **pts (mean (SD))** | | 5.31 (3.31) |
| **ptdiff (mean (SD))** | | 6.10 (13.78) |
| **error (mean (SD))** | | 0.79 (13.26) |

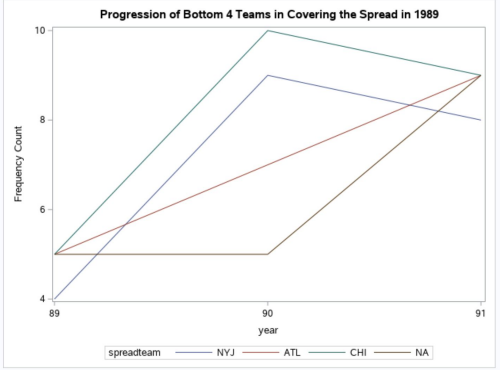
## Variability, Sample Size and Trends

We came to realize that our two major concerns with the data set were variability and the predictability of certain variables for our data. Football outcomes can be volatile and complex in the number of variables that have an impact. There are commonly around 53 players on each team, full coaching staffs, weather conditions, injuries, etc., that have an impact on the outcome. Notably, much of the variability due to all of these measures creates a challenging element to address the balance between minimizing variability but maintaining sample size.

<u>Season to Season</u>

Looking at the data holistically was an immediate concern for us. On the graph below to the left you can see the variability of the top 5 teams in terms of number of times that team covered the spread over the 3 years of our data. Considering teams play 16 games a season, it remarkable to note that each team, only over a course of one season, went from being very reliable to cover a spread, to being a bad bet (less than 50% of the time covering; or 8/16). The opposite trend can be seen on the bottom right graph.



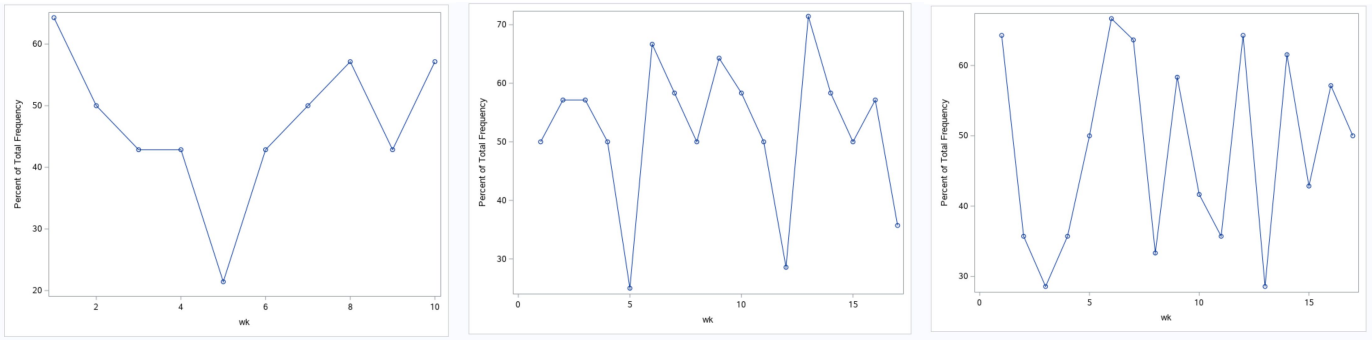



**Football Betting Prediction Analysis**

By Ty Darnell, Jace Gilbert, and Michael Steffan

Within a Season

Our analysis continued by looking through the change of spreadcover over the season, season by season. Below, you can see graphs that demonstrated the percentage of games covered by the favored team each week. The graph on the far left shows the first 10 weeks of the 1989 season, followed by the full seasons of 1990 and 1991. Note that there are many dramatic dips throughout the season demonstrating great variability between the favored team or underdog team covering the spread at a high percentage.




**Model**

Given that a 'no cover' result simply returns a bet, we chose to define a successful bet as correctly choosing the cover team, or a no cover. Given the inconclusive distributions of spreadcovers across different factors, it made sense to look at individual models (per team) as opposed to collective ones. And given the complexity of factors that change from season to season, it is more appropriate to look at an in-season model. All of these decisions were supported by numerous models in current and past research.

We created individual and updated logistical models. This meant creating a weekly dataset for each team, each week, that would be updated the following week with any new information. For example, week 11 for the Green Bay Packers would include all information of the previous 10 weeks of that season. Week 12 would include the information of the week 11 dataset, with the new information of the week 11 outcomes.

We chose to build our model around weeks 11-16 of the 1989 season with the data from the previous games. This consisted of creating 6 weekly sets of datasets, with 28 team datasets per week. Each game during a week would then have two completing logistic models to determine the individual chance of that team to cover the spread in the game.

In the given model, $P_{ij}$ is the probability that team *i* covers the spread in week *j*. Predictors are the difference of the team and their opponent average victory margins, and the weekly spread for that game.

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \boldsymbol{\beta}_{ij}\boldsymbol{X}_{ij} + \epsilon_{ij}$$

$$\rightarrow P_{ij} = \frac{e^{\boldsymbol{\beta}_{ij}\boldsymbol{X}_{ij}}}{e^{\boldsymbol{\beta}_{ij}\boldsymbol{X}_{ij}+1}} + \epsilon_{ij}^{*}$$

**Football Betting Prediction Analysis**

By Ty Darnell, Jace Gilbert, and Michael Steffan

## Decisions and Outcomes

Decisions

Decisions were made between the two models depending on the highest chance of successfully covering the spread. For each week we chose the top 6 teams with the highest probability to cover the spread according to the model. If two teams face each other in the top 6, we chose the highest probability to bet on.

| team | prob |
|---|---|
| SEA | 0.8620954 |
| NO | 0.8599769 |
| CHI | 0.8421692 |
| DEN | 0.7724969 |
| NYG | 0.7638946 |
| PIT | 0.7235895 |

Outcome

Overall, given the sparseness of informative and significant data, our system of collective logistic models were successful. We ended up winning 55.6% of the games we bet on, and yielded a profit of $320. The obvious note being that the model appeared to have not stabilized, given the lack of data, until week 14.

| Week | Record | Net $ |
|---|---|---|
| 11 | 2-4 | -$220 |
| 12 | 1-5 | -$440 |
| 13 | 2-4 | -$220 |
| 14 | 5-1 | +$400 |
| 15 | 5-1 | +$400 |
| 16 | 5-1 | +$400 |
| Weeks 11-16 | 20-16 | +$320 |

## Assumptions/Concerns of the Model

Sample Size

A rule of thumb to help ensure stability of a logistic model is to have at least 10 events per explanatory variable. So, there is some concern since each model will have between $10*P_{ij}/2$ and $15*P_{ij}/2$ per predictor. We believe that this can be seen by the lack of success of the model in weeks 11-13. But as each model got more data, the model stabilized and was successful in the remaining weeks, providing an overall profit.

Collinearity

Collinearity had to be addressed for 28*6=168 models and so we simply just checked that the variance inflation factor wasn't too high (above 10 by rule of thumb). There no concern of collinearity between intercepts and variables or between the two variables in any logistic model as all the values of the VIF we're below 3.

Independence/Distribution of Consecutive Games

'Momentum' is anecdotally used to describe a team's success in consecutive games. Yet, most of the research has found the distribution of sporting streaks are not significantly different than the distribution of expected streaks. Looking through the distribution of spread covers of the following game, given the previous spread cover for each team, we found results for 1989 were not significant at a p-value of 0.174. So we felt comfortable with the independence assumption game to game.

## Further Considerations

More sophisticated and detailed data would provide more accurate results and most likely better

predictors than we had. Metrics such as DVOA (defense-adjusted value over average) have demonstrated consistent success as a predictor by over 60% per week. In terms of number of predictors, most research has insisted on keeping the model simple with only 1-2. Given the later success of our model, and the concern for sufficient data, a revised system of betting larger on games later in the season would most likely guarantee a more stable model.