Football Betting Analysis

Jace Gilbert, Tyree Darnell and Michael Steffan

Our Data

Analysis Variable : favpts

N	N Miss	Mean	Std Dev	Minimum	Maximum
672	0	22.95	9.97	0.00	61.00

Analysis Variable : undpts

N	N Miss	Mean	Std Dev	Minimum	Maximum
672	0	16.86	9.27	0.00	47.00

Analysis Variable : ptdiff

N	N Miss	Mean	Std Dev	Minimum	Maximum
672	0	6.10	13.78	-36.00	54.00

Analysis Variable : pts

N	N Miss	Mean	Std Dev	Minimum	Maximum
672	0	5.31	3.31	0.00	19.50

Summary Table for Dataset SPREAD_FB

Column Name	Column Type	Column Length	Column Label	Column Format
DOW	char	2		$2.
day	char	2		$2.
fav	char	3		$3.
ot	char	5		$5.
spreadcover	char	11		$11.
spreadteam	char	3		
und	char	3		$3.
Night	num	8		1.
favpts	num	8		BEST12.
ha	num	8		BEST12.
ot1	num	8		
ptdiff	num	8		
pts	num	8		BEST12.
undpts	num	8		BEST12.
wk	num	8		BEST12.
year	num	8		BEST12.

Betting Spreads

- A 'spread' is a chosen number that separates possibilities of betting into 2 bets:

  - For a <u>favorite to cover the spread</u>, they must beat the underdog by more than the spread.

  - For an <u>underdog to cover the spread</u>, they must lose to the favorite by less than the spread, tie, or win.

Ideal Spread for the Bookmakers (yr 1989 only through week 10)

- An ideal spread for a bookmaker splits the odds of either team covering into a 50-50 chance.

- A spread that is 'not covered' gets your money back. This occurs when the favorite wins by exactly the spread amount.

year=89

spreadcover	Frequency	Percent
favorite	66	47.14
not covered	5	3.57
underdog	69	49.29

year=90

spreadcover	Frequency	Percent
favorite	117	52.23
not covered	5	2.23
underdog	102	45.54

year=91

| spreadcover	Frequency	Percent
favorite	107	47.77
not covered	9	4.02
underdog	108	48.21

Vigorish

- Vigorish is the amount charged by the bookmaker, for taking a bet from a gambler (ie: Bettor must risk $11 to win $10)

- On average, this ensures profitability for the house regardless of the outcome of the bet (unless it's a push)

- A success rate of 52.38% is required for a bettor to break even, due to the 10% vigorish.

- This was a consideration when evaluating our model.

Violin Plots of Spread Error for Weeks 1-10 for 1989

These violin plots demonstrate the variability and density of the spread errors over the first 10 weeks of the 1989 season. (Game Difference - Spread)

This helps demonstrate both the small range of the majority of errors from the spreads, but also the variability of the errors week to week.
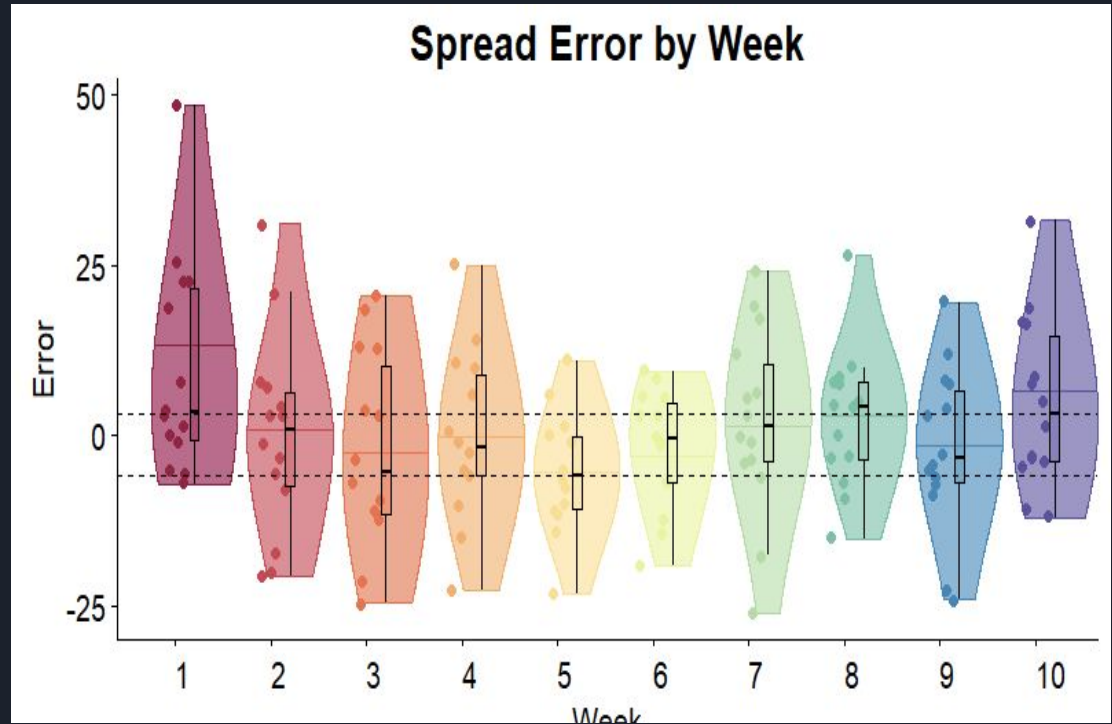
Variables Don't Collectively Explain Covering for Each Team (1989-1991)

Spread Cover by Day of Week

Frequency	Table of spreadcover by DOW					
		DOW				
	spreadcover	**Mo**	**Sa**	**Su**	**Th**	**Total**
	favorite	22	9	295	4	330
	not covered	3	0	16	0	19
	underdog	24	7	288	4	323
	Total	49	16	599	8	672

Spread Cover by Home Favorite Status

Frequency	Table of spreadcover by ha		
		ha	
	spreadcover	**0**	**1**	**Total**
	favorite	104	226	330
	not covered	8	11	19
	underdog	112	211	323
	Total	224	448	672

Spread Cover by Night Game Status

Frequency	Table of spreadcover by Night		
		Night	
	spreadcover	**0**	**1**	**Total**
	favorite	287	43	330
	not covered	15	4	19
	underdog	282	41	323
	Total	584	88	672

In-Division Rivalry (years 1989-1991)

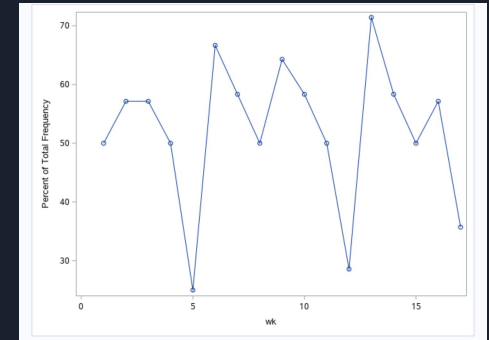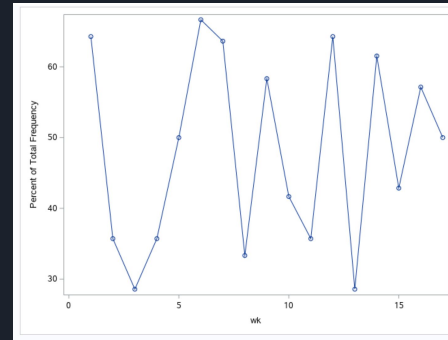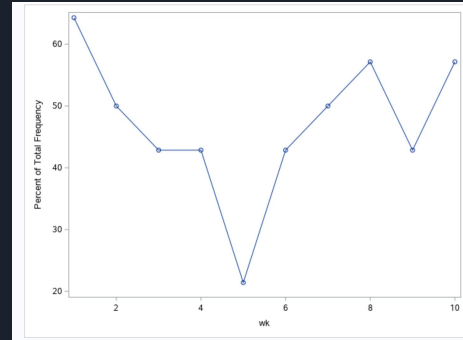In-division games are known to be more-competitive as they are placed twice a year and serve as tie-breakers.

Unfortunately, there doesn't appear to be any difference between how teams cover between in-division and out of division games.
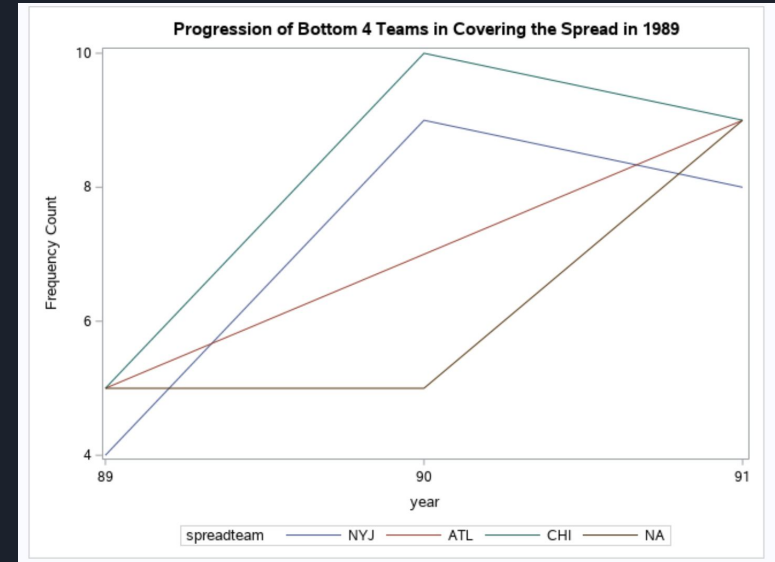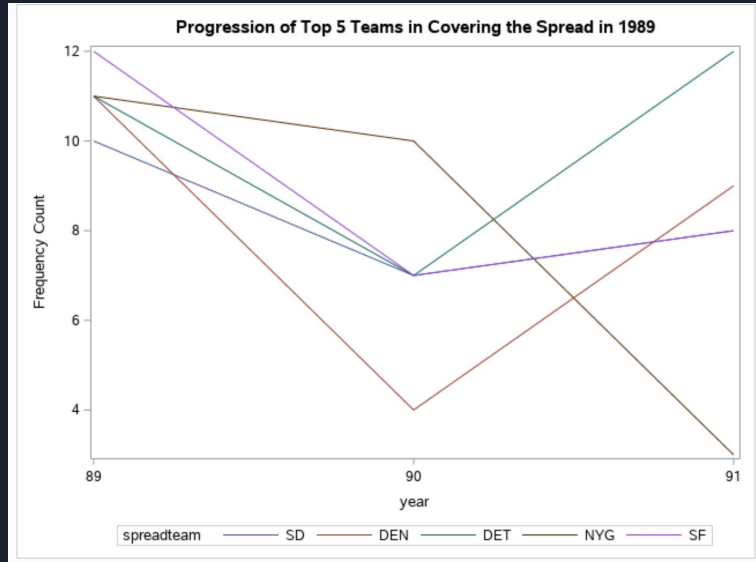
Spread Cover For In-Division Games

Frequency Percent Row Pct Col Pct	Table of DivGame by spreadcover			
	spreadcover			
DivGame	favorite	not covered	underdog	Total
0	181 26.93 50.28 54.85	9 1.34 2.50 47.37	170 25.30 47.22 52.63	360 53.57
1	149 22.17 47.76 45.15	10 1.49 3.21 52.63	153 22.77 49.04 47.37	312 46.43
Total	330 49.11	19 2.83	323 48.07	672 100.00

Percentage of Games Covered by Favorite in 1989 (wk 1-10), 1990 and 1991

Looking through the data, it became immediately obvious that even though the bookmakers were seemingly choosing reasonable spreads on average for the season, that there was significant variability week to week.

Variability Between Seasons for Teams

Progression of Top 5 Teams in Covering the Spread in 1989

Progression of Bottom 4 Teams in Covering the Spread in 1989

Goal

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.

Also, given the inconclusive distributions of spreadcovers across different factors, it made sense to look at <u>individual models</u> 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.

Who is Covering the Spread Over the Quartiles in 1991

Frequency Percent Row Pct Col Pct	Table of spreadcover by qrtl				
			qrtl		
spreadcover	1	2	3	4	Total
favorite	29 12.95 27.10 42.03	19 8.48 17.76 46.34	30 13.39 28.04 54.55	29 12.95 27.10 49.15	107 47.77
not covered	5 2.23 55.56 7.25	1 0.45 11.11 2.44	2 0.89 22.22 3.64	1 0.45 11.11 1.69	9 4.02
underdog	35 15.63 32.41 50.72	21 9.38 19.44 51.22	23 10.27 21.30 41.82	29 12.95 26.85 49.15	108 48.21
Total	69 30.80	41 18.30	55 24.55	59 26.34	224 100.00

Green Bay Covering the Spread Over the Quartiles in 1991

Frequency Percent Row Pct Col Pct	Table of spreadcover by qrtl				
			qrtl		
spreadcover	1	2	3	4	Total
favorite	5 31.25 62.50 83.33	0 0.00 0.00 0.00	2 12.50 25.00 40.00	1 6.25 12.50 33.33	8 50.00
underdog	1 6.25 12.50 16.67	2 12.50 25.00 100.00	3 18.75 37.50 60.00	2 12.50 25.00 66.67	8 50.00
Total	6 37.50	2 12.50	5 31.25	3 18.75	16 100.00

Adjusting In Season Logistic Model

We created individual and updated logistical models.

- Used information provided in previous games
- Started predictions in Week 11 to provide sufficient data
- Following Week's models would be updated (ie, Week 12 model with new Week 11 data)
- Created competing team models for game predictions

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

$$\rightarrow P_{ij} = \frac{e^{\beta_{ij}X_{ij}}}{e^{\beta_{ij}X_{ij}}+1} + \epsilon_{ij}^*$$

- 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.

Who do we bet on each week?

Week 12 1989 season

1)Choose the top 6 teams with highest probability to cover the spread

2) If two teams face each other in top 6, choose highest probability

team	prob
SEA	0.8620954
NO	0.8599769
CHI	0.8421692
DEN	0.7724969
NYG	0.7638946
PIT	0.7235895
BUF	0.2775367
GB	0.2460921
LAA	0.2325032
NYJ	0.2205464
CIN	0.2181877
PHA	0.1118487

How Did the Model Perform in 1989?

Week 14 1989 Season

Team Score	Opponent Score	Spread	Predicted Probability	Covered?	Result
SEA 24	CIN 17	Sea +10.0	0.95	Yes	Win
NO 22	BUF 19	NO + 6.0	0.82	Yes	Win
CLE 17	IND 23	CLE -1.0	0.97	No	Loss (OT)
DET 27	CHI 17	DET + 6.5	0.13	Yes	Did not bet
MIN 43	ATL 17	MIN -12.0	0.41	Yes	Did not bet

How did we do?

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

Concerns of Assumptions?

- 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.

- The concern of collinearity was also addressed as the VIF was less than 10 in our models.

Linearly Independent covariates?

Team	VIF
SF	1.887
TB	1.174
WAS	2.467
SEA	1.519

Independence Continued/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.

	Current Game Non-Cover	Current Game Cover	*Marginal Row Totals*
Previous Game Non-Cover	85 (91.93) [0.52]	112 (105.07) [0.46]	197
Previous Game Cover	111 (104.07) [0.46]	112 (118.93) [0.4]	223
Marginal Column Totals	196	224	420 (Grand Total)

The chi-square statistic is 1.8465. The *p*-value is .174187. This result is *not* significant at *p* < .05.

The chi-square statistic with Yates correction is 1.5898. The *p*-value is .207354. *Not* significant at *p* < .05.

Looking through the distribution of spread covers of the following game, given the previous spread cover for each team, our results for 1989 were not significant.

Continuing Considerations

Moving forward with the model, more sophisticated and detailed data would provide more accurate results.

Metrics such as DVOA (defense-adjusted value over average) have demonstrated consistent success of predictions by over 60% per week.

In terms of adding predictors, most research has insisted on keeping the model simple with only 1-2 predictors.

A State-Space Model for National Football League Scores

Mark E. GLICKMAN and Hal S. STERN

Using Past Performance to Predict NFL Outcomes: A Chartist Approach

March 1997
This Revision: April 1997

The Performance of Betting Lines for Predicting the Outcome of NFL Games

Greg Szalkowski and Michael L. Nelson

Old Dominion University, Department of Computer Science
Norfolk, VA 23529

Abstract. We investigated the performance of the collective intelligence of NFL fans predicting the outcome of games as realized through the Vegas betting lines. Using data from 2560 games (all post-expansion, regular- and post-season games from 2002-2011), we investigated the opening and closing lines, and the margin of victory. We found that the line difference (the difference between the opening and closing line) could