MIS 480/580: KNOWLEDGE MANAGEMENT

# Betting in Super Bowl Matchups

## A Sports Data Mining Approach to Knowledge Management

**Gabriel Hazlewood, Josh Hottenstein, Scottie Wang, James Chen**

**Prepared for: Dr. Hsinchun Chen**

Management Information Systems Department

University of Arizona

**5/16/2008**

# Contents

# Introduction

## Research Question

Can patterns in historical game performance allow the bettor to gain a better understanding of what makes a good bet for the Super Bowl each year?

## Purpose

The purpose behind our project is to provide bettors with an "angle" that can be used to exploit certain inefficiencies in the NFL betting market. An "angle," as it refers to sports betting, is a forecasted performance pattern that does not show up in the raw stats. It can be used to better predict future results. An inefficiency in the NFL betting market can be seen as an opportunity. It occurs when the spread for a bet is not derived well-enough to prevent further knowledge from drastically increasing a bettor's odds of winning the bet. We looked for such an inefficiency in the Super Bowl (the NFL's championship game), and a way to exploit it.

## Objective

The objective of our project is analyzing whether there are any exogenous variables that could aid in better determining the outcome of a Super Bowl bet relative to its final line and/or over/under. We used season-long statistics particular to the favorite and underdog in each Super Bowl since 1990 (when the NFL switched to a 17 week regular-season-schedule) as well as each of their statistics in the Super Bowl as our exogenous variables. This allowed us to determine whether any particular statistic, or combination of them, could be used together with the final line and/or over/under to provide a bettor with additional knowledge as to the outcome of a bet on the Super Bowl. The result provides an angle which can be used by a bettor placing a wager on the Super Bowl.

The final line (also referred to as the closing line) is the odds-makers' final prediction as to the outcome of the game, reached just prior to the start of the game. It is listed as a numerical value representing the spread, or predicted winning margin of the favorite team, and bettors can place a wager in favor of the margin (for the favorite) or against it (for the underdog). This value changes over the course of the week leading up to the game as further knowledge pertaining to the game is revealed and/or when bettors' wagering tendencies are discovered to lean heavily in favor of one side of the bet or the other. This is an attempt on the odds-makers' behalf to balance half of the betting populations' wagers on one side of the line, and the other half on the other side. This ensures a profit will be realized on behalf of the odds-maker no matter the outcome of the game.

The over/under is the odds-makers' prediction as to the total amount of points that will be scored in the game. It is represented as a value as well, and bettors can place a wager in favor of the game's total score being over or under the value. This value can change over the course of the week leading up to the game for the same reasons that lead to shifts in the game's line.

## Usefulness

It must be made clear that any finding is not useful as a betting model by itself. This is because the Super Bowl is a game played just once a year, and thus, few opportunities exist for placing a bet on this game. However, any finding can be very useful to a seasoned-bettor looking to add high-probability angles pertaining to different bets to their repertoire for future use. Over time, a bettor could build a repertoire similar to that of an investor's stock portfolio. The repertoire would contain several angles pertaining to bets on certain games that provide a high-probability as to the outcome. In making use of this repertoire a bettor can effectively use sports betting as a means of investment rather than a gamble.

# Literature Review

## *The Man Who Shook Up Vegas*

Sam Walker's article, *The Man Who Shook Up Vegas,* examines experts in the field of sports betting. It introduces Bob Stoll (known as Dr. Bob) as the most highly regarded expert in the sports betting market. It also introduces the notion of thinking of sports betting as an investment rather than a gamble. The article reveals that when betting against a point spread, a bettor must win 52.4% of their wagers to make a profit. However, experts in this field realize closer to 60% winning percentages, which allows them to sustain a very comfortable living through the sports betting market. The article reveals some strategy Dr. Bob uses to determine his bets. Dr. Bob looks for angles that better predict future results of games. One such angle he disclosed was that when a team is favored by 7 or more in a minor bowl game after losing their last game, they fail to cover the spread 77% of the time (Walker).

The information found in this article can aid our project in a couple ways. Since a bettor must win 52.4% of their bets to make a profit, we can only accept findings yielding greater than a 52.4% probability rate. Knowing that experts realize closer to a 60% winning percentage, we can aim for findings yielding a probability closer to that percentage to effectively produce expert-level knowledge. Seeing that Dr. Bob attempts to find angles pertaining to certain betting opportunities, we can have confidence that any finding will be relevant, as it will be knowledge proven useful in the sports betting market on behalf of a renown expert (Walker).

## *Testing Market Efficiency: Evidence From The NFL Sports Betting Market*

Philip and Stephen Gray's writing, *Testing Market Efficiency: Evidence From The NFL Sports Betting Market,* examines the efficiency of the NFL betting market. The reading introduces more sophisticated betting strategies used by experts. Such a strategy is only placing bets only when there is a relatively

high probability of success. Any finding of ours will be useful in such a strategy as it will only be accepted if it provides a relatively high probability of success (Gray).

The reading states that "the model indicates that the market overreacts to a team's recent performance and discounts the overall performance of the team over the season." In our project we use season long statistics, which effectively takes the overall performance of teams into account. The reading also revealed that "exogenous variables such as weather conditions as well as fundamentals such as rushing/passing yards and field-goal kicking success rates could be added to increase the predictive power of the model." Our project uses all available fundamental exogenous variable information, and therefore increases the predictive power of our results. The reading concludes that in the NFL sports betting market "inefficiencies exist, but not all are exploitable." In our project we are looking for an inefficiency in the Super Bowl, and a way to exploit it, so this provides a reason to believe such is attainable (Gray).

## *Testing Rationality In The Point Spread Betting Market*

John Gandar, Richard Zuber, Thomas O'Brien, and Ben Russo combined to create the reading *Testing Rationality In The Point Spread Betting* Market. The reading presents empirical tests of market rationality using data from the point spread betting market on NFL games. It focuses on examining whether, at any point in time, a moving line becomes more significant as to the outcome of a bet. It concludes that "in the NFL, the closing line does not provide a more accurate forecast than does the opening line," and vice-versa. This is very useful to our project as we are using only closing lines pertaining to historical Super Bowl betting opportunities. This information allows us to have confidence in our data set, in that using closing lines will not compromise the validity of any finding (Gandar).

### *Investor Sentiment From Price Paths: The Case of Football Betting*

Christopher Avery and Judith Chevalier created the reading *Investor Sentiment From Price Paths: The Case of Football Betting*. The reading further examines the previous literature's (*Testing Rationality In The Point Spread Betting Market*) findings. It is validated in this reading that the movement of a spread is predictable, and attempting to exploit it yields a very low profit at best. This is deemed true because factors which influence the movement of the spread are usually in the form of knowledge easily and widely available. Therefore the movement only reflects the betting populations' increased level of knowledge as to the outcome of a bet based on common knowledge, and distills the use and effectiveness of that knowledge. This is useful to our project as it further proves the validity of the closing line element included in our data set (Avery).

### What to Take From Literature Review

It has been determined that spreads on NFL games are biased predictors of actual results. This creates inefficiencies in the NFL betting market, some of which are known to be exploitable (Gray). To exploit these inefficiencies most profitably, exogenous variables must be found that together create an angle which allows for increased predictability as to the outcome of a bet (Gray, Walker). The angle must provide a probability of success greater than 52.4%, but should be closer to 60% to be deemed extremely useful and valuable (Walker). Finally, we can also take into account that fluctuations to the line on the game will not matter, and using closing lines will be sufficient in our study (Gandar, Avery).

## Experiment Design

Based upon the literature review we built on the work in an attempt to create a model that accurately predicted team performance in the Super Bowl. The goal of the model as stated is to give better an additional tool they can utilize to find a better, "angle" from which to approach their activities. In

discovering the model we first created a dataset based upon publicly available statistics on team performance.

## Data source, data extraction and database settings

We first needed to find a data source with the super bowl and football season statistics that we need for the model. Having compared some sports statics websites; we choose the one of the most popular football data websites www.databasefootball.com as our data source. We have collected 38 statistics factors of super bowl teams for the recent 10 years. The most recent 10 years were chosen as this is when sports books started offering both line and over bets for the super bowl. There are two teams each year to attend the super bowl, take the favorite team in 2007 for example; our data includes 20 games statistics the favorite team played in the regular season, playoff and super bowl and all its opponents' statistics each game.

To extract the data, we use the open source software Websphinx from CMU. This tool helps us to collect all the data-related html files. After that, we wrote a Java program to parse the data and put it into the oracle database. Here are the tables in the database (partial screen shot).

| Column Name | Data Type |
|---|---|
| SID | NUMBER(10,0) |
| WINNER | VARCHAR2(100 BYTE) |
| SCOREDIFF | NUMBER(10,0) |
| NFC | VARCHAR2(100 BYTE) |
| AFC | VARCHAR2(100 BYTE) |
| UNDERDOG | VARCHAR2(100 BYTE) |
| FAVORITE | VARCHAR2(100 BYTE) |
| LINE | NUMBER(10,0) |
| OVER | NUMBER(10,0) |

SUPERBOWL DATA

| Column Name | Data Type |
|---|---|
| SID | NUMBER |
| GID | NUMBER |
| TEAM | VARCHAR2(4000 BYTE) |
| Q1 | NUMBER |
| Q2 | NUMBER |
| Q3 | NUMBER |
| Q4 | NUMBER |
| T | NUMBER |
| FIRSTDOWNS | NUMBER |
| TOTALYARDS | NUMBER |
| PASSYARDS | NUMBER |
| PASSYARDSPERPASS | NUMBER |

Favorite(Underdog)

| CQ1 | NUMBER |
|---|---|
| CQ2 | NUMBER |
| CQ3 | NUMBER |
| CQ4 | NUMBER |
| CT | NUMBER |
| C1STDOWNS | NUMBER |
| CTOTALYARDS | NUMBER |
| CPASSYARDS | NUMBER |
| CPASSYARDSPERPASS | NUMBER |
| CRUSHYARDS | NUMBER |

Favorite (underdog) opponent

## Simple Statistics

In our attempt to find how and where the sports books set the line we ran a series of simple statistics consisting of mean and median on the averages for a number of different factors that could contribute to game performance. A example of this information is found below.

| | Total score | First Downs | Total Yards | Rush Attempts | Time of Possession |
|---|---|---|---|---|---|
| **Average** | 28.23529 | 21.390374 | 371.973262 | 29.5828877 | 1.317375966 |
| **Median** | 28.00 | 21.00 | 377.00 | 30.00 | 1.32 |

After running the analysis on the favored and underdog teams in both regular season and super bowl games we found that on average the teams going into the bowl game are fairly evenly matched and usually one team does not have a significant statistical advantage over the other. This data also reinforced the analysis we did on how the line and over are set for the games. Given the bets most of the time the line will behave how the sports books prefer with even bets on either side of the line. This further reinforces what we found in the literature review. Given we could not find factors that attempt to model team behavior through basic statistics we then conducted a series of correlation analysis to find significant factors.

## Attempts to Find Correlation

From our literature review and subject matter expert we know that many sports bettors believe regular season performance and how the teams perform during the post season contributes to the outcome of the super bowl. To find factors from the regular season that could potentially used to model super bowl performance we attempted a series of regressions to see if there was any validity in these beliefs.

One of these regressions we ran was an attempt to see if a relationship existed between how the team fared in the regular season and where the line was set for the super bowl. The literature suggested that individuals perceptions of the team based on historical performance would influence their betting behavior. Since the line moves up/down depending on individual's perception of where the team will score we attempted a regression comparing the line to a number of different elements such as score. An example of this is included below. As one can see from the chart there is a very low R Squared value indicating poor fit between the factors. As we continued to run our analysis we found similar poor fit between factors. As a result we decided we would need to do a more complex model of the data based on a multi factor regression to find a combination of factors that could be used to approximate the teems final score.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .171[a] | .029 | -.079 | 10.032 |

a. Predictors: (Constant), Line

## Advanced Data Analysis Model

In the betting market, we can bet that which team the winner is, bet on the score of both teams, on the score difference between the two teams, and on the total score of two teams. And as everyone knows, there are a lot of factors that can influence the football game result. It is very difficult to concisely use one factor to predict the score or the winning team. But it is obvious to find out that the winner team has a 'better statistics' such as more total yards and less turnovers.

As a result, the literature suggests some combinations of different factors have the correlation with the game results such as the score. To test this assumption, we developed a multi linear regression model to find out how different factors correlate and the result. The following is the simple explanation of the model:

$$Y, x_1, \cdots, x_p$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i \qquad i = 1, 2, 3, \ldots, n$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}_{n\times(p+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}_{n\times 1}$$

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e} \qquad X : \text{design matrix}$$

This is a simple application of linear algeber. The column Y is the result to be predicted. In the design matrix X, every comlum is a factor with a set of data. If we put the column 1 in the matrix, it means that we assume that there is a constant,which is the β0 in the model. 'e' is the error of the regression. To apply this model, we wrote a function in the matlab. This function can also calculate the P value and the R square which are two of the most important value to show how precisly the regression result is.

## Process of choosing factors and combinations

As we mention in above section, we have totally 38 factors, they are 19 statistics for the favorite team: scores of quarter 1, quarter 2, quarter 3, quarter 4 and total scores, 1st down score, total yards, passing yards, pass yards per pass, rush yards, rush attempts, rush average, turn interceptions, punts, punting yards, yards per punt, penalties, penalty yards, and time of possessions. The other 19 statistics are for the opponents.

First of all, we need to choose Y. Obviously; we want to find out super bowl result, so we choose the super bowl score for the testing team and the score difference between two teams.

For the choice of factors, it is very difficult to test all the factor combination for the 38 factors. So basd upon discussions with our subject matter expert and work previously done in the literature, we choose the following factors, scores of each quarters, total score, 1st down score, total yards, rush attempts, penalty yards and time of possession as the factor to combine. Each of these factors was shown to stand as a proxy for a team's ability to perform a particular on field function.

Before the super bowl game, all we know is the statistics for the regular season and playoff games of both teams. We don't know anything how teams will perform in the super bowl. So we can only predict the super bowl result using the previous game data. Since we have the data for the last 10 years, to find out the correlation between season games and super bowl, we did some data processing for the data we selected. We calculate the average of game statistics; take total yards for example, there are 18 games exclude the super bowl for the favorite team in one year, and calculate the total yards average of 18 games. So we have 10 average total yards data which is the factor No.1, using the same process, we get other factors.

However, before we run the model, we find out that all the factor are based on the data of the testing team itself, and they are not able to give us an idea of one team's general ability when it faces to different team. As a result, we made another data processing; we also calculated the year-average game statistics for the opponent team. Then we calculate the difference between the correspondent factors, for example, use the year-average total yards for the favorite team to minus year-average total yards of all its opponent teams. Then we get another set of factors.

Later, to better use the quarter scores, we decided to sum the 3$^{rd}$ and 4$^{th}$ quarter score because that's the score of the 2$^{nd}$ half. And then we also calculate the score differences. The reason we did that is because it shows the team ability to adapt the game situation. If the team is able to score more in the 2$^{nd}$ half than its opponent, it means that the coach and the players are able to change the strategies according to actual game condition.

Base on the set of factor we choose, we further selected the factors that are meaningful in the games. They are all average differences; including 1$^{st}$ down score, total yards, time of possession, 2$^{nd}$ half scores and rush attempts. The reasons are:

1. 1$^{st}$ down: shows the ability of the offence to overwhelm opposing team defense

2. Total yard: shows general ability of the offense to move the ball
3. Time of possession: Shows the ability to control the game
4. 2nd half score: shows the ability to adapt and change
5. Rush attempts: Shows how aggressive the team is

After choosing the factor, we begin the run the model. First we use each factor to test its correlation with the testing team super bowl score and the super bowl score difference. The data shows that the P values are all smaller than 0.01, which means they all have statistical significances. Additionally, the R square is between 0.5and 0.6. But our purpose is to test the multi factors and try to run the factors as many as possible. So we skip the regular process which should follow the rule of testing from 2 factors and then enlarge the quantity to 3 , 4 and 5.

We tested the combination of the 5 factors and the super bowl score, the P value are all between 0.01 and 0.05 which means it still has statistical meaning, and the R square is about 0.65, however, the Coefficient of 1st down is -0.624 which doesn't make any sense. So we deleted this factor and run the rest four with the super bowl score for the favorite team, following is the result:

$$Y=0.129*X1+11.02*X2+1.028*X3+0.792*X4$$

R Square:0.6969 P Values:

| | | | |
|---|---|---|---|
| 0.0026 | 0.00558 | 0.00276 | 0.0124 |

With this model, we can find out the how the factors and the super bowl score for the favorite team correlate, and using the same method, we can also find out the correlation with underdog team score. However, due to the limits of the data, we have only ten years game statistics, the reason we only

choose 10 years is because we need a continued data year by year, and none of the website we found has such a data set. We believe that if we have more than 30 years data, the model will more precise. And that's the future work we are expecting to do.

## Conclusion

Based on our analysis, we found a procedure to model the potential score for a team in the super bowl. This model could be of use to bettors in making their decisions. Furthermore, we found that factors individuals normally "feel" are related to team performance are in fact not significantly related and are of little use in developing a model to proved a better angle on the bet. This angle we provide while basic at this point in time shows potential for future investigation and development

Furthermore, we found both in the literature and from discussion with bettors that individuals are poor decision makers when faced with incomplete information and given that the baseline to turn a profit in betting is ~52% and that expert bettors only average ~60% winning percentage. As a result we have developed a model that has the potential to find a much higher rate of winning and provide additional insight into the outcome of unstructured events. Despite the work we have done much more still needs to be accomplished to give bettors a tool that will be significantly accurate most of the time. The model can be further developed and data validated with actual in game statistics. Additionally, it would be interesting to see the differences between the lines at different sports books and find arbitrage opportunities exist between the sports books.

As a final note we would like to address how the team worked together. The chart below shows the work break down and who accomplished what over the course of the project. While all team members contributed to the project as a whole below are the primary areas of responsibility for each team member.

| | Literature Review | Subject Matter Expert | Data Extraction | Analysis | Statistical Modeling |
|---|---|---|---|---|---|
| Gabe Hazlewood | X | X | | X | |
| Josh Hottenstein | X | | | X | X |
| James Chen | | | X | X | |
| Scottie Wang | | | X | X | X |

## Works Cited


1. Walker, Sam. "The Man Who Shook Up Vegas." The Wall Street Journal. 5 Jan. 2007. 11 March 2008 <http://online.wsj.com/public/article/SB116796079037267731-wjPu4ACcg5J5Qvjh05IYEI_Ooeo_20070112.html>.

2. Gray, Philip K., and Stephen F. Gray. "Testing Market Efficiency: Evidence From The NFL Sports Betting Market." The Journal of Finance, Vol. 52, No. 4, (Sep., 1997), pp. 1725-1737.

3. Gandar, John, Richard Zuber, Thomas O'Brien, and Ben Russo. "Testing Rationality in the Point Spread Betting Market." The Journal of Finance, Vol. 43, No. 4, (Sep., 1988), pp. 995-1008.

4. Avery, Christopher, and Judith Chevalier. "Investor Sentiment From Price Paths: The Case of Football Betting." The Journal of Business, Vol. 72, No. 4, (Oct., 1999), pp. 493-521.