Determinants of NFL Spread Pricing:

Incorporation of Google Search Data Over the Course of the Gambling Week

Shiv S. Gidumal and Roland D. Muench

Under the supervision of

Dr. Emma B. Rasiel

Dr. Michelle P. Connolly

Department of Economics, Duke University

Durham, North Carolina

2017

Shiv graduated with High Distinction in Economics and a minor in History in May 2017. Following graduation, he will be working in New York as an Associate Consultant at Bain & Company, a management-consulting group. He can be contacted at shiv.gidumal@gmail.com.

Roland graduated with High Distinction in Economics and a minor in Mathematics in May 2017. Following graduation, he will be working in New York as an Analyst at Jefferies, an investment bank. He can be contacted at roland.d.muench@gmail.com.

Table of Contents

Acknowledgements

We would like to thank our thesis advisors, Dr. Emma Rasiel and Dr. Michelle Connolly, who have guided us from the beginning. We could not have completed this work without their advice, encouragement, and valuable criticism on earlier works. We also want to express our utmost gratitude to our peers from the thesis workshop, who engaged with our ideas even when they may have appeared convoluted and crazy. Without our advisors and peers in the seminar, we would not have been able to shape our ideas.

Abstract

We investigate the factors incorporated by Las Vegas in setting opening spreads for NFL matchups. We include a novel proxy measure for gambler sentiment constructed with Google search data. We then investigate whether changes in this proxy are reflected in the closing spreads for NFL matchups and find that they are. We also reveal bettors' preferences for highly visible teams and teams performing well as of late. We show that the factors that matter in the actual outcome of a game are home field advantage, average points scored for and against, and, most interestingly, our proxy measure for gambler sentiment.

JEL Codes: G14, G17

Keywords**:** market efficiency, NFL Pointspread wagers, Google Trends, Las Vegas

"Writing's like gambling.
Unpredictable and sporadic successes make you more addicted, not less."
-M. John Harrison, 2012

1. Introduction

Las Vegas' sports betting market consists of casinos that each offer a menu of different wagers for bettors to place on outcomes of National Football League (NFL) games. Through these wagers, Las Vegas' sports betting market also presents a convenient venue to test the Efficient Market Hypothesis.[1] To date, many articles have tested whether strategies exist to generate reproducible profits within various sports betting markets (Zuber et al., 1985; Thaler and Ziemba, 1988; Sauer et al., 1988; Camerer, 1989; Woodland and Woodland, 1994; Gray and Gray, 1997; Gandar et al., 1988; Sinha et al., 2013; Fodor et al., 2013). However, the findings have been mixed across different sports gambling markets, with some researchers identifying repeatable, profitable strategies (Zuber et al., 1985; Thaler and Ziemba, 1988; Camerer, 1989; Woodland and Woodland, 1994; Gandar et al., 1988; Sinha et al., 2013; Fodor et al., 2013) and others failing to find any (Sauer et al., 1988; Gray and Gray, 1997). Furthermore, some strategies are found to be profitable when applied in a certain sport's gambling market, yet unprofitable when applied in another (Thaler and Ziemba, 1988; Woodland and Woodland, 1994). Regardless of their position on the efficiency of the sports gambling market, the vast majority of these studies examine if, at a discrete moment, the market incorporates all available information into betting prices, thus eliminating the opportunity to consistently generate above-average, risk-adjusted returns.

This paper first aims to investigate exactly what information Las Vegas incorporates into the initial, or opening, prices it offers on NFL wagers. Las Vegas must attempt to determine the factors that bettors consider when placing wagers in order to construct prices that will attract relatively equal betting from gamblers on both sides of a wager. Therefore, the factors we investigate are ones that we believe bettors consider in their choices of teams and matchups to wager on. These factors include measurements of teams' relative historical and recent performance, home field advantage, national visibility, and preparation time. Lastly, we

[1] We use a definition for Efficient Market Hypothesis taken from Fama (1970) that a "market in which security prices always 'fully reflect' all available information is called 'efficient.'" An implication of an efficient market is that it is impossible to develop a strategy to consistently beat that market and achieve above-average returns.

attempt to incorporate the current gambler sentiment on each team. To capture the current sentiment of the gambling public, we use Google search frequency data leading up to Las Vegas' release of opening prices.

We then expand on past research, which has looked at gambling prices only at discrete points in time, by examining the movement of prices from the opening price, offered at the beginning of the gambling week, to the closing price, offered just before the start of a matchup. We add a measurement of newly revealed gambling sentiment to the factors that determine opening price and test if this change in gambler sentiment is priced into the closing price. We again measure gambler sentiment through Google search frequency data, in this case captured from the time period between the opening and closing prices. In this way, we evaluate the Efficient Market Hypothesis' applicability to the NFL wagering market; we verify whether publicly available Google search data is incorporated into closing prices. Further, we investigate if the factors used by Las Vegas to set the opening price are appropriately priced in at the beginning of the week or if they materially affect the movement of prices over the course of the week. The Efficient Market Hypothesis states that any information available at the time opening prices are released should not affect future price movements.

Next, we test if the factors that determine the closing prices for wagers significantly predict the realized outcomes of NFL matchups (and thus the outcomes of wagers placed). This test allows us to determine the predictive efficacy of factors considered by gamblers when they place bets. Finally, using our findings, we attempt to construct a profitable betting strategy in out of sample data. Should a strategy exist, it would provide evidence contradictory to the Efficient Market Hypothesis, which does not allow for reproducible, profitable strategies.

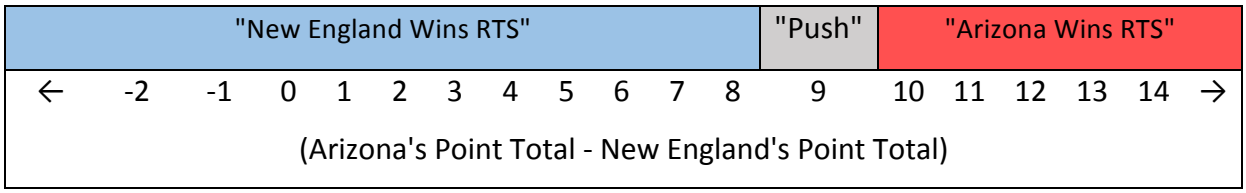
2. NFL Wagering Market Structure

We begin with a discussion of the NFL wagering market and its similarities to other financial markets. The American NFL gambling market exists legally only within the state of Nevada. For this reason, it is referred to as "Las Vegas" for the duration of this paper.

One of the most prominent sports wagers offered by Las Vegas for NFL wagering is the "Pointspread." In a Pointspread wager, one of the teams in a game is assigned a number, named the "spread," which can be positive or negative. At the end of the game, this number will be added to that team's score. Bettors wager on which team will have the higher score after the spread is taken into account. A team listed with a negative spread is the favorite, while a positive spread denotes the underdog. The spread can be defined relative to either team, and the term "team of record" is used to identify which team's spread is being presented (Gray and Gray, 1997).

For example, suppose there is an upcoming game between Arizona and New England, and Arizona is the team of record with a -9 point spread. If Arizona wins by more than 9 points, they are said to have won "relative to the spread" (RTS), while an Arizona win by fewer than 9 points or a New England victory means that New England won RTS. If Arizona wins by exactly 9 points, the game is called a "push," and all money is returned to bettors at no cost to them. This scenario can equivalently be written with New England as the team of record with a +9 point spread. These outcomes relative to the spread are shown graphically in Figure 1.

Figure 1. Outcomes Relative to the Spread of Hypothetical Matchup Between New England and Arizona Where Arizona Has a Spread of -9



The Pointspread bet is an "even-odds bet", where each team is, as assessed by the casinos that comprise Las Vegas, expected to win 50% of the time after adjusting for the spread. Las Vegas achieves a profit by incorporating a fee that forces gamblers to wager slightly

more than they stand to win. The literature refers to this amount by several names, such as "vigorish," "vig," "juice," or "cut," but will be referred to henceforth as "vigorish." Las Vegas quotes the vigorish alongside the spread. The most common vigorish is 10% and, in such case, gamblers must wager 10% more than what they would stand to win (i.e. \$1.10 for every \$1.00 they would win). Las Vegas can set the vigorish higher or lower to entice betting on either side of the bet. For instance, if it were lowered to 5% a gambler would only have to risk \$1.05 to win \$1.00, a more attractive option than wagering \$1.10. A generalized Pointspread bet will be denoted as having a spread of α and a vigorish of β, where α is an integer or an integer $+\frac{1}{2}$ and β is a percentage with $0 \leq \beta \leq 1$.[2] A gambler must wager $\$(1+\beta)$ to win $\$1$. This can equivalently be stated in terms of a $\$1$ wager where a gambler wins $\$\frac{1}{1+\beta}$ for each dollar bet. For a gambler who wagers n_f dollars on the favorite, his/her cash flows are shown in Table 1.

Table 1. Cash Flows for Wager of n_f on Favorite

Game Outcome	Cash Flow at Time of Bet	Cash Flow after Outcome of Game	Net Cash Flow
Favorite wins by amount greater than α i.e. Favorite wins relative to the spread (RTS)	$-n_f$	$n_f + n_f \frac{1}{1+\beta}$	$n_f \frac{1}{1+\beta}$
Favorite wins by α i.e. Push	$-n_f$	n_f	0
Favorite loses or wins by amount less than α	$-n_f$	0	$-n_f$

[2] Using an α that is an integer $+\frac{1}{2}$ allows Las Vegas to, if it wishes, fine tune the spread to be between two consecutive integers. In this case there is no possibility for a "push" to occur, since the difference between the two teams' point totals will always be an integer.

i.e. Favorite loses RTS			

For a gambler who wagers n_u dollars on the underdog, his/her cash flows are shown in Table 2.

Table 2. Cash Flows for Wager n_u on Underdog

Game Outcome	Cash Flow at Time of Bet	Cash Flow after Outcome of Game	Net Cash Flow
Favorite wins RTS	$-n_u$	0	$-n_u$
Push	$-n_u$	n_u	0
Favorite loses RTS	$-n_u$	$n_u + n_u \frac{1}{1+\beta}$	$n_u \frac{1}{1+\beta}$

The "house advantage" is the expected profit for the casino as a percentage of the total amount bet, i.e. the expected rate of return to the casino from wagers. First, consider a casino's profit function for a given Pointspread bet

$$\pi = \begin{cases} N_F + N_U - \left(N_F + N_F \frac{1}{1+\beta}\right) = -N_F \frac{1}{1+\beta} + N_U & when\ Favorite\ wins\ RTS \\ N_F + N_U - (N_F + N_U) = 0 & when\ the\ result\ is\ a\ push \\ N_F + N_U - \left(N_U + N_U \frac{1}{1+\beta}\right) = -N_U \frac{1}{1+\beta} + N_F & when\ Underdog\ wins\ RTS \end{cases} \quad (1)$$

where N_F is the total amount wagered on the favorite and N_U is the total amount wagered on the underdog. N_F and N_U are determined by the market's response to the spread offered by Las Vegas. Thus, N_F and N_U are functions of α and β, as well as gambler sentiment, which in turn depends on information available to gamblers. The expected value of the casino's profit function is

$$\begin{aligned} &\mathbb{E}(\pi) = \\ &\left(-N_F \frac{1}{1+\beta} + N_U\right) P(Favorite\ wins\ RTS) + \\ &\left(-N_U \frac{1}{1+\beta} + N_F\right) P(\ Underdog\ wins\ RTS) \end{aligned} \quad (2)$$

For illustration, suppose that Las Vegas can perfectly set the line to attract equal wagering on both sides and give each team a 50% chance of winning RTS. This would result in $N_F = N_U$ and $P(Favorite\ wins\ RTS) = P(Underdog\ wins\ RTS) = \frac{1}{2}$. Expected casino profit becomes

$$\mathbb{E}(\pi) = \left(-N_F \frac{1}{1+\beta} + N_F\right)\frac{1}{2} + \left(-N_F \frac{1}{1+\beta} + N_F\right)\frac{1}{2} = N_F\left(1 - \frac{1}{1+\beta}\right) = \frac{1}{2}N\left(\frac{\beta}{1+\beta}\right) \quad (3)$$

where $N = N_F + N_U$, the total amount bet. The expected house advantage is then

$$\mathbb{E}\left(\frac{\pi}{N}\right) = \frac{\mathbb{E}(\pi)}{N} = \frac{\beta}{2(1+\beta)} \quad (4)$$

Assuming the most common vigorish of $\beta = 10\%$, we get to a house advantage of $\frac{1}{22}$ or 4.55%. Therefore, if a casino can (and chooses to) set the spread so that both teams have a 50% chance of winning RTS and so that equal wagering is placed on either side, it will earn 4.55 cents for each dollar wagered. This is essentially the casino's commission, which is similar to the commission earned by brokers in the financial world.

Under the above two assumptions, it can be shown that $Var\left(\frac{\pi}{N}\right) = 0$, as

$$\left(\mathbb{E}\left(\frac{\pi}{N}\right)\right)^2 = \frac{\beta^2}{4(1+\beta)^2} = \mathbb{E}\left(\left(\frac{\pi}{N}\right)^2\right) \quad (5)$$

Since the variance of the returns is zero, the casino would make riskless profit if these two assumptions hold.

In reality, it is impossible for a casino to perfectly set the spread in this manner. First, it cannot predict the probabilities of the outcomes of the game with perfect accuracy and thus does not know which spread will result in a 50% chance of either team winning RTS. Second, it cannot perfectly forecast the preferences of the wagering market, N_F and N_U, and thus cannot guarantee that there will be equal betting on both sides of the spread set. Further, even if the casino were able to do both perfectly, it is possible that the spread that results in a 50% chance of winning differs from the spread that would attract equal betting on both sides. Once these assumptions are relaxed, we no longer know for certain what $\mathbb{E}\left(\frac{\pi}{N}\right)$ equals and $Var\left(\frac{\pi}{N}\right)$ becomes greater than zero. In practice, skilled professionals with extensive knowledge of the industry set opening spreads and then actively manage them over the course of the week leading up to the game with the goal of achieving the highest $\mathbb{E}\left(\frac{\pi}{N}\right)$ and the lowest $Var\left(\frac{\pi}{N}\right)$.

Due to the black box nature of Las Vegas it is impossible to know whether casinos take active positions in the outcomes of games.[3] Research to date has found contradictory evidence, with Avery and Chevalier (1999) and Paul and Weinbach (2005) pointing towards active positions and Humphreys et al. (2013) finding evidence of neutral positions. Regardless of whether an active position is taken by a casino, spreads will always be adjusted as gambler sentiment, as revealed through newly placed wagers, shifts (Kreiger, 2015). If a higher than anticipated proportion is bet on the favorite, Las Vegas will increase the spread's magnitude to provide an incentive to bet on the underdog. Conversely, if a lower than anticipated proportion is bet on the favorite, bookmakers will decrease the spread's magnitude to provide an incentive to bet on the favorite.

This paper takes advantage of the fact that Las Vegas adjusts spreads over the course of the week as new information is revealed by the betting patterns of gamblers. When determining opening spread, Las Vegas can only use information available at that point in time along with its best estimate of gamblers' sentiment about each team. As the week progresses, gamblers reveal their preferences for a matchup's two teams through the placement of bets. If Las Vegas' estimate of gamblers' sentiment turns out to be inaccurate, Las Vegas adjusts the spread offered accordingly. Using Google search frequency data as a proxy for sentiment, we test whether Las Vegas prices in this gambler sentiment. We examine if Google search frequency data taken from the day before the release of the opening spread is incorporated into that spread. We then see if Google search data taken over the time period from opening spread to closing spread is priced into the closing spread as the Efficient Market Hypothesis would predict.

Further, if Las Vegas' initial forecast of gambler sentiment is incorrect, a bettor who can better predict that sentiment could forecast which direction the spread will move as Las Vegas adjusts. Because a placed bet is locked in and subsequent movements in the spread do not

[3] An active position is defined in comparison to a neutral position. In a neutral position, a casino would stand to make the same profit regardless of which team wins RTS. The casino's return would have no variance and would thus be risk-free. In contrast, a casino with an active position would stand to make more profit if one team wins RTS and less profit (or potentially a loss) if the other team wins RTS. Although maintaining an active position exposes a casino to risk, a casino could choose to take an active position if it believes it can predict the eventual outcome of a matchup better than the gambling public.

affect its standing, this ability to predict spread movement would allow the savvy bettor to wager on a given team at the time when the offered spread is most attractive and, therefore, win a higher percentage of bets. For bettors to overcome the house advantage, they must be correct greater than 52.38% of the time to generate a profit when $\beta = 10\%$.[4] Thus, for a Pointspread strategy to be economically significant it must succeed more than 52.38% of the time.

[4] When a bettor bets \$1.10 to win \$1.00 they must be correct eleven times for every ten times they are wrong to break even. $(11\ wins)\left(\frac{\$100}{win}\right) - (10\ losses)\left(\frac{\$110}{loss}\right) = 0. \frac{11}{21} = 52.38\%$ i.e. they must be correct 52.38% of the time.

3. Literature Review

Sports betting has gained academic interest in recent decades. Just as academics have long sought reproducible strategies to generate profits in securities markets, so have they searched for exploitable patterns within sports wagering markets. Thaler and Ziemba (1988) find that a naïve strategy of wagering on favorites in racetrack betting generates profits in the long term. Conversely, Woodland and Woodland (1994) find the opposite holds in baseball, where betting on underdogs is the profitable strategy. Camerer (1989) finds that if the market perceives that a team is "hot" because of recent wins, the spread can become biased and allow for profitable betting strategies.

The first significant contribution on NFL betting markets comes from Zuber et al. (1985). Zuber tests for market efficiency in the NFL betting market based on the notion that the best unbiased predictor of the actual point spread of an NFL game should be the spread offered by Las Vegas directly before the game. This is equivalent to the financial notion that forward prices are the best unbiased predictors of future spot prices. Using a regression that incorporates the closing spread and team-specific statistics, Zuber is able to predict winning strategies with a 95% confidence level, as well as devise gambling strategies that return profits in excess of the vigorish (Zuber et al. 1985). Later, Sauer et al. (1988) revisit the Zuber strategy and determine that losses would occur to bettors using this strategy if it had been extended past 1985 to 1988.

Examining the period from 1976 to 1994, Gray and Gray (1997) find evidence that Las Vegas has historically overpriced favorites and discounted home teams in NFL wagering markets. Further, Gray and Gray find that teams that beat the spread in a given season typically continue beating spreads in upcoming games more often than teams that had performed poorly relative to the spread. This phenomenon demonstrates that Las Vegas is slow to incorporate a team's historical performance into prices. Gray and Gray also demonstrate that Las Vegas overreacts to teams' recent performance – overshooting when setting gambling prices. This finding suggests that sports bettors possess a short memory and are too quick to discount a team's performance earlier in the season. Using this premise, Gray and Gray construct an economically significant strategy that involves betting on good teams that have not played well as of late.

Other studies have found additional evidence supporting the existence of inefficiencies within the gambling market. Fodor et al. (2013) find evidence of hindsight bias, since too much emphasis is put on teams' performance over the previous season when pricing for the first week of a new season. Davis et al. (2015) expands on this by demonstrating that too much emphasis is put on teams' performance in the first week of a season when pricing spreads for the second week. However, while many of these studies have identified inefficiencies[5] in the gambling market using behavioral biases of bettors, few researchers have explored gambling markets' incorporation of one core informational resource, the Internet.

Recently, academics have started using data from Internet-based resources as proxies for popular sentiment. Sinha et al. (2013) incorporates Twitter data into a predictive model for NFL outcomes and finds that the aggregated "wisdom of crowds" represented by Twitter data has predictive power for NFL outcomes. While Sinha presents a test case of how to utilize social media to directly predict discrete NFL outcomes, neither he nor another academic has tested the implications of changing popular sentiment on gambling price movements over the course of the gambling week.

In addition to Twitter, Google and other search engines are valuable aggregators of the collective pursuit of information. Numerous studies utilize Google Trends and Google Categories, the two publically available forms of Google's search query data, to forecast certain outcomes, including elections (Gayo Avello et al., 2011), unemployment (Askitas and Zimmermann, 2009), the spread of influenza (Ginsberg et al., 2009), and the stock market (Bordino et al., 2010). These studies have proven useful in highlighting the best practices and difficulties associated with using Google Trends and Categories. However, to date no studies have explored the impact of Google's search query data on NFL gambling markets.

Indeed, many studies have identified possible inefficiencies in the gambling market by isolating pieces of information that bettors do not incorporate into their evaluation of a wager

[5] It is important to note that the presence of profitable wagering strategies does not necessarily imply inefficiency in the market. Since Las Vegas reacts to bettors who exhibit certain biases when setting spreads, these biases become incorporated into the price. Thus, even if Las Vegas rationally incorporates the preferences of the majority of the gambling public, a savvy bettor could harness knowledge of these biases to generate profits by going against the herd.

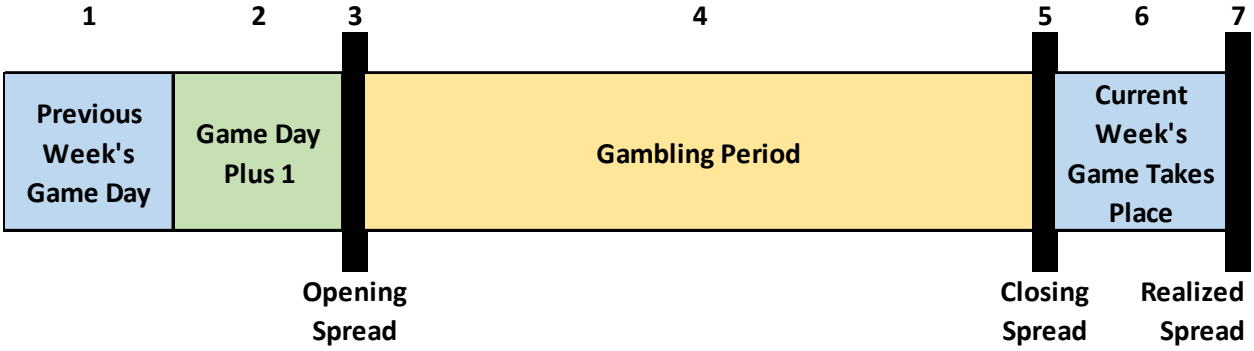
(Zuber et al., 1985; Thaler and Ziemba, 1988; Sauer et al., 1988; Camerer, 1989; Woodland and Woodland, 1994; Gray and Gray, 1997; Gandar et al., 1988; Sinha et al., 2013; Fodor et al., 2013). Other studies have attempted to use social media and Internet search engines to eliminate inefficiencies in predictive models for various outcomes. Still, academics have yet to examine Las Vegas' incorporation of the Internet's information over the course of the gambling week.

This paper expands on the current literature in three ways. First, by using Google search data as a proxy for gambler sentiment, it investigates the factors that Las Vegas uses in setting the opening spread for NFL matchups. Second, again using Google search data, it examines whether newly revealed gambler sentiment becomes incorporated into the closing spread as the Efficient Market Hypothesis predicts. Third, it evaluates the efficacy of gamblers' incorporation of certain factors into wagering choices and the application of the Efficient Market Hypothesis to the NFL wagering market by attempting to construct a profitable betting strategy.

4. Theoretical Framework

This paper tests the semi-strong Efficient Market Hypothesis, which assumes that prices instantaneously reflect all publicly available information (Fama, 1970). In the NFL wagering market, Las Vegas' spreads are equivalent to prices of wagers for bettors to purchase. Therefore, if the semi-strong Efficient Market Hypothesis holds for the NFL wagering market, opening spreads for a given matchup should incorporate all information publically available at the moment they are released by Las Vegas. Additionally, any more information that becomes available over the course of the week should be priced in, becoming evident in the closing spread. To provide clarity on the timing of events that occur during the NFL's gambling week please see Figure 2.

Figure 2. Illustration of Timing of Events During NFL Gambling Week



Period 1	Day that team played the previous week; "Game Day" Google data taken from this day.
Period 2	Day afther the day the team played the previous week; "Game Day Plus 1" Google data taken from this day.
Period 3	**Opening Spread** is set by Las Vegas based on current information available and anticipated gambler sentiment.
Period 4	Gamblers place bets throughout this period revealing gambler sentiment; Las Vegas reacts by adjusting the spread offered; "Weekly" Google data taken from this period.
Period 5	Gambling Period ends right before game begins with final spread offered being the **Closing Spread**; Closing Spread represents the interplay between the Opening Spread and the information revealed by gamblers over the Gambling Period.
Period 6	Game takes place.
Period 7	The actual outcome of the game, "Realized Spread," is known.

Due to the fact that NFL games for a given week occur on Thursday, Saturday, Sunday, and Monday, there is no standard weekly format for each matchup. The specific days of the

week that data are pulled from vary for each matchup, but the rules for defining the NFL gambling week as outlined above are followed for gathering data for each individual matchup.

4.1. Opening Spread Determinants

To begin our analysis, we investigate what factors Las Vegas considers when pricing the opening spread. To do this we run an OLS regression of the opening spread on a variety of factors that bettors would consider in their wagering decisions (Regression 1) and, thus, that Las Vegas would consider in pricing the opening spread. [6] In these independent variables we include dummy variables to control for home field advantage, whether a team played a primetime game the week before (and thus exposure to a national audience), and whether a team had a bye the week before (and thus an extra week to prepare for their opponent).

We also include a variety of winning percentage statistics to control for a team's performance that season. These winning percentages are broken down into two categories. First, we use a team's general winning percentage, i.e. considering whether the team wins or loses a matchup. Second, we include a team's winning percentage relative to the spread, i.e. considering whether the team wins or loses after the spread adjustment is taken into account. We include the win percentage RTS due to the findings of Gray and Gray (1997) that teams who beat the spread in the past tend to beat the spread more often in the future than other teams. Since these findings have been public for so long, it would make sense that they have been incorporated into betting strategies of gamblers and, thus, could be of interest to Las Vegas when setting opening spreads. We further delineate within these winning percentages by splitting them into winning percentages over the last three games played and winning percentages over the entire season excluding the last three games. We do this since it has been shown that gamblers overemphasize recent performance in their decision making (Gray and Gray, 1997).

Next, we include team statistic variables to measure a team's average performance over the course of the current season. In choosing which teams to wager on, bettors take their past performance as indicative of future performance, so it would make sense for Las Vegas to take these statistical measures into consideration when pricing opening spreads. We include a

[6] The specifics of these independent variables and methodology for their calculation will be covered in depth in Section 5.

team's average passing yards per game, average rushing yards per game, average yards allowed on defense per game, average points scored per game, average points allowed per game, and average turnover advantage per game.

Finally, we expect Las Vegas to attempt to measure gambler sentiment about each team when pricing opening spreads. If the gambling market as a whole were to have a more favorable perception of a team, more bets would be placed on that team and, consequently, Las Vegas would attempt to adjust the opening spread to account for that sentiment. To incorporate sentiment, we use a variable constructed from aggregated Google search data, called "Relative Google Search Level", as a proxy for gambler sentiment. For the opening spread we use two such variables, each constructed with data from different time periods in Figure 2. One variable is constructed with Google search frequency data aggregated from the "Game Day" period (Period 1) and the other is constructed with Google search frequency data aggregated from the "Game Day Plus 1" period (Period 2). The coefficient of the "Game Day" variable ultimately is not statistically significant in any regressions, suggesting that it is not a factor considered by Las Vegas in pricing opening spreads. On the other hand, the "Game Day Plus 1" data has statistically significant effects, so we incorporate it into our regression. This makes intuitive sense, since the "Game Day Plus 1" data is taken from a time period closer to the release of opening spreads and Las Vegas would use the most up-to-date information available. Also, interestingly, despite being separated by only 24 hours, "Game Day" and "Game Day Plus 1" data are highly uncorrelated, with a correlation of -0.099 (See Appendix A.1 for cross-correlation matrix).

4.2. Incorporation of New Information into Spread Prices

After identifying what factors affect the opening spreads offered by Las Vegas, we investigate whether new information revealed over the course of the week is incorporated into the closing spread offered just before the game. To do this we run two OLS regressions, one of the Spread Movement over the course of the gambling period (Regression 2) and one of the Closing Spread (Regression 3). The Spread Movement is simply the difference between the Closing Spread and the Opening Spread. We include the same independent variables used in Regression 1 and add an additional variable to incorporate the new information introduced

over the course of the gambling period. Again, we use a variable constructed from aggregated Google search data, called "Relative Google Search Level," as proxy for gambler sentiment. However, this time, it is constructed from Google data aggregated from the "Weekly" gambling period (period 4 in Figure 2).

If Las Vegas properly prices in the factors that influenced the opening spread, none of these factors should have a statistically significant effect on the spread movement because all of the factors are known at open and do not change. Therefore, in the regression of the spread movement, the Efficient Market Hypothesis predicts that the only variable that could be statistically significant is the "Weekly Relative Google Search Level," since it is information that was not known at the time the opening spread was set. For the regression of the closing spread, we would expect to see similar effects as in the regression of the opening spread. Additionally, we would expect to see the newly captured information represented by "Weekly Relative Google Search Level" incorporated into the Closing Spread price.

4.3. Determinants of Realized Outcomes

Finally, we explore whether factors commonly considered by gamblers actually affect the outcomes of games and bets. To do this we run a regression of the Realized Spread on dummy variables for home field advantage, whether a team played a primetime game the week before, and whether a team had a bye the week before (Regression 4). We also include the team statistic variables of average passing yards per game, average rushing yards per game, average yards allowed on defense per game, average points scored per game, average points allowed per game, and average turnover advantage per game to control for the quality of the teams in a matchup. Due to the fact that we find our relative Google search level variables to be statistically significant in the determination of opening and closing spreads, we also include the relative Google search level variables in this regression to test if they have any predictive power in the outcomes of games and bets. In this regression, the statistical insignificance of a variable's coefficient would suggest that the variable does not impact actual outcomes and, thus, should not be considered by gamblers. We then test Zuber's notion that the Closing Spread serves as the best predictor for actual outcomes of games. To do this we slightly modify Regression 4 by including the Closing Spread as an additional independent variable (Regression

5). For Zuber's claim to hold, the Closing Spread should have a statistically significant coefficient close to 1.00.

5. Data

5.1. Dependent Variables

The dependent variables used in regressions are the opening spread, the closing spread, the spread movement over the course of the week, and the realized spread. For consistency, we define each spread with the initially favored team as the team of record. Because spreads with the favorite as team of record are negative or zero, the opening spread is always less than or equal to zero. However, because the initial favorite determines the team of record, closing spreads can be negative if the favorite remains favored by close, positive if the underdog becomes favorited by close, or zero if betting closes with no favorite or underdog.

In Regression 2, we consider the spread movement, defined as:

$$Spread\ Movement\ = Closing\ Spread - Opening\ Spread \tag{6}$$

Spread movements less than zero suggest that the initially favored team becomes more favored over the course of the week, as the closing spread was more negative than the opening spread. Conversely, positive spread movements suggest that the initially favored team becomes less favored over the course of the week. Intuitively, spread movements equal to zero imply no change.

Our dependent variable in Regressions 4 and 5 is the realized spread, defined as:

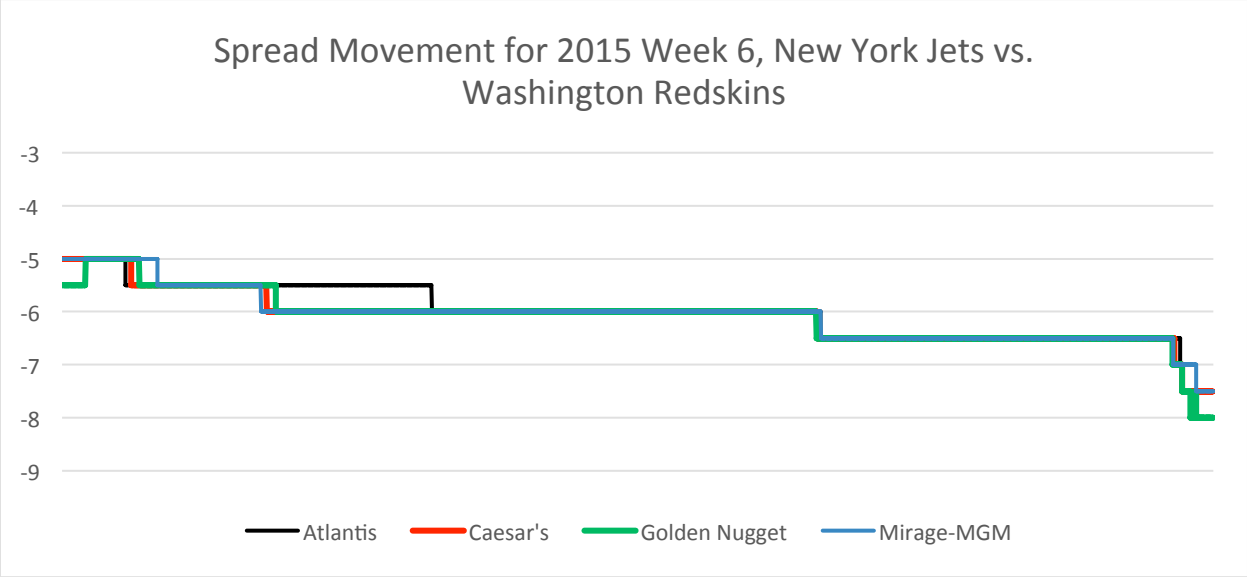
$$Realized\ Spread = Points\ Scored\ by\ Favorite - Points\ Scored\ by\ Underdog \tag{7}$$

While the other dependent variables are determined before the game takes place, the realized spread can be calculated only after the game ends. This variable equals the spread that would have resulted in a push.

A final point is that Las Vegas as a whole can be thought of as offering a standard price even though it is comprised of many different casinos that independently offer their own spreads. If a casino's spread differs too much from another's, it opens up an arbitrage opportunity, which is quickly corrected by market forces. At any given point in time, almost all spreads offered by Las Vegas casinos are the same, at the most differing by 0.5 points. This near equality of prices allows the spread of any given casino to be taken as representative of the spread of Las Vegas as a whole. This fact is of particular relevance since the spreads we use come from only one casino, Caesar's. It allows us to use Caesar's as a representative firm for

the NFL gambling market in Las Vegas. Figure 3 illustrates a sample spread price movement over the course of a gambling period for four casinos; note how closely the four spreads move together.

Figure 3. Sample Spread Movement over Gambling Period



Spread is quoted with New York Jets as team of record

(Source: Vegas Insider; April 2, 2017)

5.2. Google Search Level Construction

Ideally, we could regress a perfect measure of gambler sentiment on Las Vegas' NFL gambling spreads and their movements to understand the extent of the betting market's incorporation of gamblers' preferences. However, due to the impossibility of exactly quantifying the amorphous nature of gambler sentiment, we must create a proxy for this sentiment. Since the bookmakers in Las Vegas cannot perfectly measure gambler sentiment either, they too have to use some proxy.

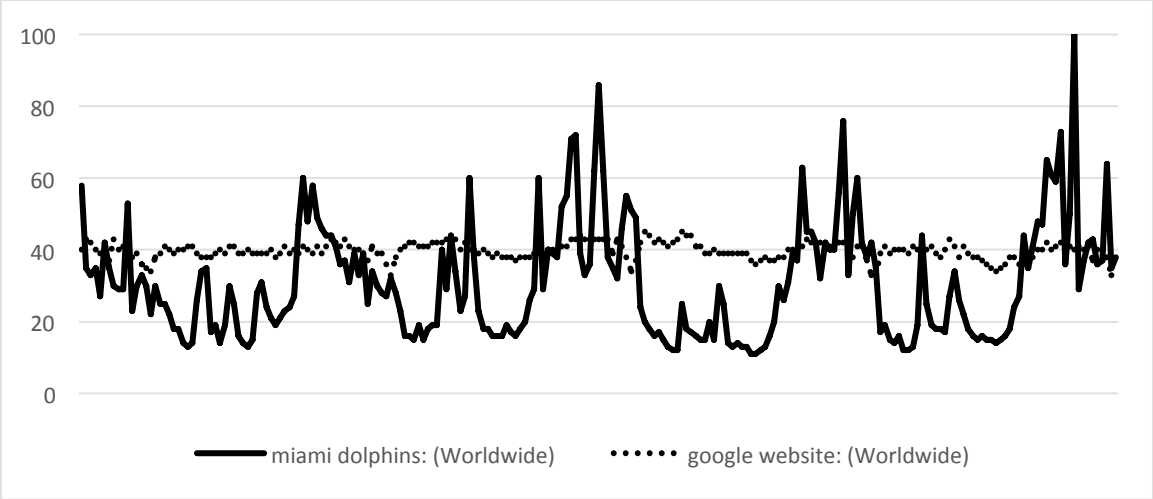
This paper uses a naïve, relative approach to attempt to capture the mood of the gambling market. Specifically, we assume that the frequency of Google search queries can capture the current mood of the overall gambling market. This assumption is based on the idea that Google users will search for a team more frequently if they are interested in the team. Although not every Google-searcher is guaranteed to gamble, most gamblers likely use Google

to research their bets, so increases in search queries should correlate with increased interest in a team and betting on such team.

Optimally, we would classify each individual search related to a given team as negative or positive. For example, if a team's quarterback is injured, Google searches for that team may increase because bettors are interested in wagering against that team. Conversely, if a team's quarterback is playing exceptionally well, Google searches for that team may increase because bettors are interested in wagering on that team. Unfortunately, we cannot identify individual searches and evaluate whether they are positive or negative because Google does not provide such granularity in its publically available data. Further, it is impossible to identify every possible search combination about a particular team that can be input into Google. As a result, our measure for search interest for a given NFL team is simply the search frequency of that team's name.

As discussed in existing literature (Gayo Avello et al, 2011; Askitas and Zimmermann, 2009; Ginsberg et al., 2009), Google's publically available historical search query data provides a number for each search term as a percentage of the single largest search volume of all given terms on any given day in the selected period. Consequently, it is not useful to compare the numbers provided by Google Trends across different periods unless they are based on the same scale. Previous authors have therefore found using a benchmark term the best way to compare search volumes for different terms across different time periods. We select the term "Google Website" as our benchmark because its search frequency over our 5-year period is relatively constant and remains close to that of the NFL teams'.

Figure 4. "Miami Dolphins" vs. "Google Website" Search Level



(Source: Google Trends; March 20, 2017)

Figure 4 shows Google Trends' search frequency data for the search terms "Miami Dolphins," represented by the solid line, and "Google Website," represented by the dashed line, from the beginning of the 2011 NFL season, September 8th 2011, until the end of the 2016 NFL season, January 1st, 2016. As the graph illustrates, the search term "Google Website" remains relatively constant and around the average frequency level for the NFL team throughout the relevant period, the 2011 through 2016 NFL seasons.

This consistency allows us to mitigate the possibility of fluctuations in our benchmark term's search frequency arbitrarily affecting the NFL teams' search levels. Because an NFL team's search level must be compared to the benchmark term, fluctuations in the benchmark term alter the relative values for each NFL team's search frequency. Thus, a term with low variance in search volume, such as "Google Website" is an ideal benchmark.

Furthermore, the proximity of the benchmark's average search frequency to the average search frequency level for the NFL teams allows us to capture as much of their variation in search frequency as possible. Because Google Trends does not provide historical search frequency levels for periods smaller than a day, the greatest granularity we can obtain is the daily search frequency level for each term. On each relevant day, we search for two search frequency values, one for "Google website" and one for the NFL team name, the higher of which Google assigns a value of 100 and the lower of which is scaled as a percentage of that

100. Since the frequency of "Google Website" remains similar to the average of our NFL teams, it provides a detailed scale with which to measure the fluctuations in the NFL team's search frequencies. If the benchmark term were vastly higher than the team search term, the observed values would always be 100 for that benchmark term and 0 or 1 for the team search term. Conversely, if the team search term were vastly greater than the benchmark term, the observed values would always be 100 for the team search term and 0 or 1 for that benchmark term. In either scenario, most of the team search term's variation is lost due to the chosen scale. The best benchmark is thus one that is on average at the same search level as the team search term.

We evaluate each team's search frequency level as a multiple of the benchmark "Google Website" on the day after they play a game during the NFL seasons from 2010-2016. Furthermore, we evaluate each team's average search frequency in gambling period between open and close as shown in Figure 2. These are called the Google search levels for team *i* at week *t*, respectively denoted

$$Game\ Day + 1_{i,t} = \frac{Daily\ Team\ Search\ Level_{i,t}}{Daily\ Google\ Website\ Search\ Level_{i,t}}$$

$$Weekly_{i,t} = \frac{Gambling\ Period\ Average\ Team\ Search\ Level_{i,t}}{Gambling\ Period\ Average\ Google\ Website\ Search\ Level_{i,t}} \quad (9)$$

where $Game\ Day + 1_{i,t}$ represents the search frequency level on the day after the previous week's game and $Weekly_{i,t}$ represents the average search frequency during the gambling period leading up to the current game. We have observations for 32 teams spanning 119 weeks. Since Google Trends gives the relative search levels as integers between 1 and 100, $Game\ Day + 1_{i,t}$ and $Weekly_{i,t}$ have a possible range from $\frac{1}{100}$ to 100. If team *i* does not play week *j*, we set $Game\ Day + 1_{i,t}$ and $Weekly_{i,t} = 0$, increasing the range to be from 0 to 100.

Some teams are more popular than others and receive many more daily Google searches on average. As a result, we compare each team's Google search level for a particular week to its search levels for the previous 17 weeks, the length of a full NFL season. For example, we would compare 2013's week 6 to the average from 2012's week 6 through 2013's week 5. This allows us to see how a given week's search level compares to the team's typical

search level over the past season. To do this we divide Google search level by the average Google search level for the past 17 weeks to get a relative Google search level,

$$Relative\ Game\ Day + 1_{i,t} = \frac{Game\ Day + 1_{i,t}}{\frac{1}{X_{i,t}} \sum_{n=t-17}^{t-1} Game\ Day + 1_{i,n}} \tag{10}$$

$$Relative\ Weekly_{i,t} = \frac{Weekly_{i,t}}{\frac{1}{X_{i,t}} \sum_{n=t-17}^{t-1} Weekly_{i,n}} \tag{11}$$

where $X_{i,t}$ is a random variable that counts the number of games played by team *i* in weeks (*t* - 17) to (*t* - 1).[7] We use this constructed relative Google search level to compare Google search activity between different teams. We believe that the difference between the Relative Google Search Levels for the initial favorite and underdog should capture bettors' relative levels of interest in the two teams. As either team's relative Google search level variable increases, we believe that bettors' interest in that team increases. While imperfect, we believe this serves as a reasonable proxy to measure gambler sentiment.

The variable we ultimately use in our regressions is the initial favorite's relative Google search level minus the initial underdog's relative Google search level

$$\begin{aligned} &Relative\ Google\ Search\ Level\ Differential \\ &\quad = Relative\ Google\ Search\ Level_{initial\ favorite} \\ &\quad - Relative\ Google\ Search\ Level_{initial\ underdog} \end{aligned} \tag{12}$$

When this variable takes a positive value it indicates more Google searches for the initially favored team. A negative value indicates more Google searches for the initial underdog.

5.3. Independent Variables

For each matchup, there are two teams, an initial favorite and an initial underdog. In the fifteen cases in our data set where the opening spread is 0, we randomly assign initial favorite/underdog designations to the teams in these matchups. The randomness of the designations should mitigate any noise contributed by these matchups. Since all of our spreads are quoted with the initial favorite as the team of record, in our regressions of opening and closing spread (Regressions 1 and 3), a negative coefficient would mean that increases in the corresponding variable lead to the initial favorite becoming more favored, ceteris paribus**.**

[7] This is to adjust for bye weeks, which result in assigned google search levels of 0.

Similarly, in our regressions of the realized spread (Regressions 4 and 5), a negative coefficient would mean that increases in the corresponding variable lead to the initial favorite scoring more points, ceteris paribus. Our hypothesis for the regression of spread movement (Regression 2) is that all independent variables known at the time of the opening spread will be statistically significant. This leaves us only expecting the weekly relative Google search level variable to be statistically significant, since it represents new information learned from open to close. A statistically significanct coefficient for any of the other independent variables would suggest that Las Vegas does not fully price in the gambling market's preference for that variable in its opening spread. Our hypotheses for the regressions on opening, closing, and realized spread (Regressions 1, 3, and 4) are discussed in relation to each individual variable below.

In Regression 5, we include the Closing spread in the regression of realized spread and expect it to be highly statistically significant as found by Zuber et al. (1985). Since many of our independent variables are incorporated into the formation of the Closing spread, we expect its inclusion to reduce the statistical significance of the variables' coefficients as found in Regression 4.

5.3.1. Dummy Variables

The first dummy variable is equal to 1 if the initial favorite is the home team. We expect this variable to have a negative coefficient in regressions of the opening and closing spreads (Regressions 1 and 3) since Las Vegas most likely prices in home field advantage. We also anticipate a negative coefficient in the first regression of realized spread (Regression 4) since playing at home is advantageous to NFL teams.

The second dummy variable is equal to 1 if the initial favorite played in a primetime game the week before the current matchup.[8] We anticipate that this variable will have a negative coefficient in the regressions of the opening and closing spreads (Regressions 1 and 3). We believe that Las Vegas recognizes that primetime games have more viewers and prices in bettors' preference for wagering on teams with which they are familiar. The third dummy variable equals 1 if the initial underdog played in a primetime game the week before. Similarly, we expect that this variable will have a positive coefficient. We expect a statistically

[8] We define a primetime game to be any game beginning after 8:00PM EST on a Sunday or Monday, as these are the games that draw the highest national audience.

insignificant coefficient for both primetime variables in the regression of the realized spread (Regression 4). Although Las Vegas may attempt to capitalize on bettors' irrational preferences, we find no evidence to suggest, nor is there a logical reason, that playing in a primetime game the week prior should improve or worsen performance in a game.

The fourth and fifth dummy variables are equal to 1 if the favorite and underdog, respectively, had a bye week the week prior to a matchup.[9] We expect that that the dummy for the favorite will have a negative coefficient and the dummy for the underdog will have a positive coefficient in the regressions of the opening and closing spreads (Regressions 1 and 3). We hypothesize that Las Vegas anticipates that bettors will disproportionately bet on teams after a bye week because the team has had an extra week of rest and an extra week to prepare for the game. Thus, we expect teams with a bye to be more favored in the opening and closing spreads. We expect similar results in the regression of the realized spread (Regression 4), since teams with an extra week to prepare and rest should perform better.

5.3.2. Winning Percentage Variables

Since some NFL games result in a tie, we calculate Winning Percentage according to equation 13.

$$Winning\ Percentage = \frac{Wins + \frac{Ties}{2}}{Games\ Played} \tag{13}$$

Then, we construct our winning percentage variables as a difference between the initial favorite's winning percentage and the initial underdog's.

$$Winning\ Percentage\ Differential = Winning\ Percentage_{initial\ favorite} - Winning\ Percentage_{initial\ underdog} \tag{14}$$

Our first two variables just use traditional winning percentage, whether a team won or lost. These two variables evaluate two different segments of the season, in order to identify preferences in recent versus past performance. The first measures the winning percentage in all games excluding the most recent three. The second variable measures the team's winning percentage in the last three games.[10]

[9] A "bye" week is a week of rest during which an NFL team does not play a game. Each NFL team is granted one bye week per season.

[10] Due to the construction of these variables we must drop matchups in Weeks 1-4 of each season from our data set.

We expect Las Vegas' spreads to favor a team more as its winning percentage increases. As a result, we expect that the coefficients of both differentials to be negative in the regressions of the opening and closing spread (Regressions 1 and 3). However, we expect the absolute value of the coefficients of the winning percentage over the last three games to be larger than the absolute value of the coefficients of the winning percentage excluding the last three games. This hypothesis stems from the findings of Gray and Gray (1997), as discussed in Section 3, that bettors have a short memory and are too quick to discount early season performance. Because Las Vegas knows bettors focus on recent performance, they most likely incorporate recent winning percentage into their pricing of spreads more than the winning percentage over the rest of the season.

Our third and fourth winning percentage variables are calculated the same way, except with wins and losses calculated relative to the spread. As a result, they determine the percentage of spread bets that a bettor would have won if they had placed bets on this team. We include them since Gray and Gray (1997) find that teams who had performed well against the spread earlier in a season tend to continue to perform well. We do not, however, expect these two winning percentage variables to be statistically significant in the regressions of opening and closing spread (Regressions 1 and 3) since actual winning percentages do a better job of capturing the quality of a given team.

5.3.3. Team Statistic Variables

We include six different team statistic variables. Each variable is calculated as the difference between the initial favorite's average per game statistic and the initial underdog's average per game statistic. Therefore, each variable takes a positive value if the initial favorite's average per game statistic is greater than the initial underdog's average per game statistic, and vice versa. We calculate all of the values using each game played that season up to, but not including, the current week. The six statistics used are average passing yards per game, average rushing yards per game, average yards allowed on defense per game, average points scored, average points given up, and average turnover advantage per game.[11]

[11] Average turnover advantage is the per game amount of turnovers won by a team minus the per game amount of turnovers lost.

We have the same expectation for all of the team statistic variables' coefficients; we expect the team with the superior statistics in each category to be more favored in opening and closing lines. Passing yards, rushing yards, points scored, and turnover advantage are statistics that teams want to maximize. Therefore, we expect that, as the differentials in passing yards, rushing yards, points scored, and turnover advantage increase, Las Vegas favors the favorite more. As a result, these variables should have negative coefficients in the regressions of the opening and closing spreads. Similarly, we expect these variables to have negative coefficients in the regression of realized spreads because these metrics should be predictive of a team's future performance.

Meanwhile, yards allowed and points against are statistics that teams want to minimize, so we expect that Las Vegas favors the favorite less as these variables increase. As a result, we expect positive coefficients in the regressions of the opening, closing, and realized spreads.

6. Results

Table 3. OLS Regressions on Opening Spread, Spread Movement, and Closing Spread

		OLS Regression 1 Opening Spread	OLS Regression 2 Spread Movement	OLS Regression 3 Closing Spread
	Number of Observations	966	774	774
	Prob > F	0.0000	0.0000	0.0000
	R-squared	0.6034	0.1057	0.5743
	Adj R-squared	0.5967	0.0856	0.5647
	Root MSE	2.1702	1.4957	2.4407
Dummy Variables	Favorite is Home Team	-3.4648 ***	0.0686	-3.3952 ***
		(0.1608)	(0.1251)	(0.2041)
	Favorite Played Primetime Last Week	-0.8002 ***	0.2485	-0.5650 *
		(0.2145)	(0.1746)	(0.2849)
	Underdog Played Primetime Last Week	0.6874 **	0.0070	1.0348 **
		(0.2494)	(0.2031)	(0.3315)
	Favorite Had Bye Last Week	-0.1863	-0.3930	-0.6014
		(0.2619)	(0.2062)	(0.3365)
	Underdog Had Bye Last Week	0.3534	0.6513 ***	0.8193 **
		(0.2531)	(0.1936)	(0.3160)
Winning Percentage Variables	Winning Percentage Not Including Last Three Games Differential (Favorite - Underdog)	-0.7503 **	0.5792 **	-0.0518
		(0.2656)	(0.2014)	(0.3287)
	Winning Percentage Last Three Games Differential (Favorite - Underdog)	-1.1032 ***	0.3936 **	-0.4428
		(0.2069)	(0.1632)	(0.2664)
	Winning Percentage RTS Not Including Last Three Games Differential (Favorite - Underdog)	-0.1223	0.1499	-0.0866
		(0.2134)	(0.1653)	(0.2697)
	Winning Percentage RTS Last Three Games Differential (Favorite - Underdog)	0.1037	0.1153	0.1188
		(0.1701)	(0.1341)	(0.2189)
Team Statistic Variables	Average Offensive Passing Yards Per Game Differential (Favorite - Underdog)	-0.0081 ***	-0.0047 **	-0.0083 **
		(0.0022)	(0.0018)	(0.0029)
	Average Offensive Rushing Yards Per Game Differential (Favorite - Underdog)	-0.0075 **	-0.0065 **	-0.0113 ***
		(0.0027)	(0.0021)	(0.0035)
	Average Yards Allowed on Defense Per Game Differential (Favorite - Underdog)	0.0036	0.0042 **	0.0108 ***
		(0.0020)	(0.0016)	(0.0025)
	Average Points Scored Per Game Differential (Favorite - Underdog)	-0.3254 ***	-0.0357	-0.4009 ***
		(0.0231)	(0.0184)	(0.0300)
	Average Points Allowed Per Game Differential (Favorite - Underdog)	0.2428 ***	0.0451 **	0.2736 ***
		(0.0232)	(0.0176)	(0.0287)
	Average Turnover Advantage Per Game Differential (Favorite - Underdog)	0.4005 ***	-0.1175	0.2228
		(0.1023)	(0.0785)	(0.1282)
Google Search Levels	Game Day Plus 1 Relative Google Search Level Differential (Favorite - Underdog)	0.1949 **	-0.0920	0.0333
		(0.0638)	(0.0512)	(0.0835)
	Weekly Relative Google Search Level Differential (Favorite - Underdog)	--	0.1000	0.5224 **
		--	(0.1107)	(0.1807)
	Constant	-0.6826 ***	0.2062	-0.4758 *
		(0.1671)	(0.1304)	(0.2127)

***p<0.001, **p<0.01, * p<0.05

standard errors in parentheses

2011 through 2015 NFL Seasons, not including weeks 1 through 4 for Regression 1

2012 through 2015 NFL Seasons, not including weeks 1 through 4 for Regressions 2 & 3

6.1. Regression 1: Opening Spread Determinants

In Regression 1, we find that Las Vegas incorporates many of the factors we hypothesize to be determinants of the opening spread. Furthermore, for the most part, the signs of each of these factors align with our initial hypotheses.

Two of our variables, however, demonstrate surprising results. First, our bye variables have no statistically significant effects on the opening spread. The second surprising result relates to our Google search variable. Contrary to our expectation of a negative coefficient, which would imply that Las Vegas favors teams with higher Google search frequencies, the coefficient is statistically significant and positive. We propose three possible explanations for this unexpected result.

Explanation 1: It is possible that Las Vegas incorporates a different measure of gambler sentiment that is negatively correlated with our Google search variable. In this way, Las Vegas' setting of opening spreads depends on this alternate measure and its negative correlation with the Google Search variable leads to a positive coefficient.

Explanation 2: Our constructed Google search variable is flawed and actually captures bettors' negative opinions of teams. In this scenario, as a team is searched for more, bettors want to bet against this team, so Las Vegas favors the team less in its spread pricing.

Explanation 3: The Google search variable does not capture gambler sentiment at all, but rather captures some other factor that, when increased, causes Las Vegas to favor favorites less and underdogs more.

In Regression 1, we identify certain factors that Las Vegas incorporates into their setting of opening spreads. As we move to Regression 2, we determine if adding our measurement of revealed gambler sentiment to these factors has predictive power for the spread movement over the course of the week. Regression 3 shows us whether Las Vegas incorporates new information, revealed over the course of the week, into their closing spreads and if any of the factors that influenced the opening spread have a different effect on the closing spread. Using the regressions together, we evaluate whether Las Vegas accurately gauges bettors' preferences for each of these factors.

6.2. Regressions 2 and 3: Closing Spread Determinants and Opening Spread Efficiency

In Regression 2, as we expect, many of the variables' coefficients demonstrate no statistical significance. This finding suggests that Las Vegas sufficiently incorporates these variables into their setting of opening spreads.

However, Las Vegas cannot set the spreads perfectly because they cannot perfectly forecast gamblers' preferences. Consequently, in some cases, Las Vegas misjudges gamblers' preferences for certain statistics. In contrast with the corresponding coefficients in Regression 1, the general winning percentage statistics' coefficients are positive in Regression 2. Although Regression 1 demonstrates that Las Vegas prices winning percentages into the opening spread, this finding suggests that Las Vegas overestimates bettors' preferences for higher winning percentages when they set opening spreads. Furthermore, the variables' statistical insignificance in Regression 3 suggests that bettors' placing of bets over the course of the week eliminates Las Vegas' overestimation by the closing spread.

Similarly, in Regression 2, the statistically significant coefficients of four of the team statistic variables suggest that Las Vegas does not accurately incorporate gamblers' preferences for these statistics into the opening spread. However, these coefficients in the spread movement regression are small in magnitude. Further, when comparing the coefficients on the same variables between the opening spread and closing spread regressions, they are all relatively similar. This suggests that, although not perfect, Las Vegas' estimation of gamblers' preferences for these team statistic variables is relatively accurate at the time of the opening spread. The small change, as evidenced by the statistically significant coefficients in the spread movement regression, can be explained by the impossibility of knowing gamblers' exact preferences ahead of time and does not necessarily imply that the opening spreads were set inefficiently by Las Vegas.

As mentioned in 6.1, the statistical insignificance of the underdog bye dummy's coefficient suggests that Las Vegas does not incorporate this information into their opening spreads. However, the variable has a statistically significant and positive coefficient in Regressions 2 and 3, implying that bettors have a preference for underdogs that had a bye the

previous week. Consequently, we contend that the opening spreads offered by Las Vegas do not efficiently incorporate bettors' preferences for underdogs who had a bye the week before.

Lastly, our Game Day + 1 Google search variable does not have a statistically significant coefficient in Regressions 2 or 3. We interpret this finding to mean that Las Vegas fully incorporates bettors' preferences for Google searches into their opening spreads, whether it is due to *Explanation 1*, *Explanation 2*, or *Explanation 3* from Section 6.1. Furthermore, our Weekly Google search variable has a statistically significant coefficient in Regression 3, meaning Las Vegas incorporates the new information revealed over the course of the week into their closing spreads.

6.3. Regressions 4 and 5: Testing Factors Gamblers Consider to Predict Outcomes

Table 4. OLS Regressions on Realized Spread

		OLS Regression 4 Realized Spread	OLS Regression 5 Realized Spread
	Number of Observations	774	774
	Prob > F	0.0000	0.0000
	R-squared	0.0589	0.0937
	Adj R-squared	0.0428	0.077
	Root MSE	13.757	13.509
Dummy Variables	Favorite is Home Team	-3.0548 **	0.5948
		(1.1462)	(1.3132)
	Favorite Played Primetime Last Week	0.7810	1.3524
		(1.5957)	(1.5705)
	Underdog Played Primetime Last Week	1.9492	0.8692
		(1.8588)	(1.8363)
	Favorite Had Bye Last Week	0.0550	0.6608
		(1.8918)	(1.8612)
	Underdog Had Bye Last Week	1.0301	0.1208
		(1.7788)	(1.7549)
Team Statistic Variables	Average Offensive Passing Yards Per Game	0.0192	0.0282
	Differential (Favorite - Underdog)	(0.0162)	(0.0159)
	Average Offensive Rushing Yards Per Game	0.0137	0.0259
	Differential (Favorite - Underdog)	(0.0196)	(0.0194)
	Average Yards Allowed on Defense Per Game	0.0091	-0.0028
	Differential (Favorite - Underdog)	(0.0142)	(0.0142)
	Average Points Scored Per Game	-0.5274 ***	-0.0802
	Differential (Favorite - Underdog)	(0.1567)	(0.1748)
	Average Points Allowed Per Game	0.2910 *	-0.0164
	Differential (Favorite - Underdog)	(0.1472)	(0.1554)
	Average Turnover Advantage Per Game	-0.3128	-0.5395
	Differential (Favorite - Underdog)	(0.7192)	(0.7076)
Google Search Levels	Game Day Plus 1 Relative Google Search	1.2522 **	1.2329 **
	Level Differential (Favorite - Underdog)	(0.4680)	(0.4596)
	Weekly Relative Google Search	-0.6857	-1.2261
	Level Differential (Favorite - Underdog)	(1.0094)	(0.9963)
	Closing Spread	--	(1.0836) ***
		--	(0.2009)
	Constant	-1.5821	-1.0266
		(1.1882)	(1.1713)

***p<0.001, **p<0.01, * p<0.05

standard errors in parentheses

2012 through 2015 NFL Seasons, not including weeks 1 through 4 for Regressions 4 & 5

In Regression 4, we find most variables' coefficients to be statistically insignificant. This finding suggests that bettors have no reason to consider many of the factors that they incorporate into their gambling decisions. Below, we discuss some of the more interesting findings.

The home team dummy is the only statistically significant dummy variable. Unsurprisingly, we find a basis for home field advantage's existence and estimate that it equates to approximately 3 marginal points scored. When comparing this value to variable's coefficient in the regressions of the opening and closing spreads (Regression 1 and 3) to evaluate whether Las Vegas accurately incorporates home field advantage, we see that Las Vegas gives home teams slightly more of an advantage, closer to 3.5 marginal points scored. Not unexpectedly, we find the coefficients of the primetime dummy variables to be statistically insignificant, and conclude that playing primetime the week before confers no advantage to a team. We also, somewhat unexpectedly, find the bye dummy variables' coefficients to have no statistical significance. This result implies that there is no merit to the claim that teams will perform better when they have two weeks to prepare for games.

One could argue for team coefficients to be statistically significant because they are indicative of a team's historical performance and, therefore, predictive of future performance. However, only average points scored and average points allowed have statistically significant coefficients. We believe that the points scored and allowed statistics already reflect the information conveyed by the remaining offensive and defensive team statistics, respectively. Further, the other statistics do not necessarily convert directly into points put up on the scoreboard by a team and, therefore, the realized spread. Points scored and points allowed, however, do (by definition). Consequently, we are not surprised that only these two variables have predictive power for outcomes.

Lastly, our most interesting finding is that our Game Day + 1 Google search value has a statistically significant, positive coefficient on the realized spread. If we return to our three *Explanations* from Section 6.1, we believe this finding provides more evidence for *Explanation 3*. The unknown factor that Google search values capture could also have an impact on the

outcomes of games. We hypothesize that Las Vegas has identified this factor, recognized its impact on realized spreads, and incorporated it into opening and closing spreads.

In Regression 5, we find that the inclusion of the closing spread in the regression of realized spread pushes all variables' coefficients to statistical insignificance, except the coefficient of the Game Day + 1 Google search value. As expected, these variables, all of which we prove to be incorporated into closing spreads in Regression 3, are overidentified in Regression 5 and, consequently, have no statistically significant effect. However, we show in Regression 3 that the Game Day + 1 Google search variable is not incorporated into closing spreads. Therefore, we are not surprised to find that it remains statistically significant and positive in Regression 5. This final finding provides further evidence that Las Vegas incorporates the Game Day + 1 Google search variable into opening spreads because it captures some unexpected factor that has an impact on outcomes. Indeed, while we recognize the flaws in our hypothesis, we believe this explanation to be the most likely.

For fun, we attempt to use this information to construct a profitable betting strategy in out of sample data collected over the course of the 2016 NFL season. We attempt to harness the predictive power of the Game Day + 1 Google search data to choose which teams to wager on. Although many variations of strategies are devised, we fail to find any strategy that consistently outperforms the hurdle rate of 52.38% in a way that cannot be attributed to chance. The absence of a reproducible profitable strategy adds further evidence that the NFL wagering market is efficient.

7. Summary of Results

Through our five regressions, we answer several of our overarching questions. First, we find that Las Vegas incorporates many factors into its opening spreads, including Google search frequency data. Second, we recognize that Las Vegas incorporates new information revealed over the course of the week into its closing prices. Third, we identify the pertinent factors that affect actual outcomes, but realize that the construction of a profitable betting strategy is not possible using this factors. Overall, we find little evidence to refute that the Efficient Market Hypothesis holds in Las Vegas' NFL wagering market.

8. Suggestions for Further Exploration

We believe that the statistical significance of Game Day + 1 Google search variable's coefficient in the regression of realized spread merits further investigation. Studies could be conducted to identify the mystery "factor" with which the data is correlated in order to understand the implications of our findings. For example, a concurrent analysis of spread movements over smaller time intervals and live Google Trends data could shed light on the potential factors that the data captures.

Furthermore, we believe any subsequent studies should attempt to utilize more granular search data. Due to the limitations of historical Google Trends data, we could not obtain search values for time periods smaller than one day. The replacement of our daily and weekly search values with hourly data or data for smaller time intervals could yield enlightening results. One could also further attempt to break down the Google search data in terms of how positive or negative it is towards a particular team through semantic analysis. This, however, is difficult to do due to the limitations in the data that are released publicly by Google.

References

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly, 55*(2), 107-120.

Avery, C., & Chevalier, J. (1999). Identifying investor sentiment from price paths: The case of football betting. *Journal of Business, 72*(4), 493-521. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=2467455&site=ehost-live&scope=site

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., & Weber, I. (2012). Web search queries can predict stock market volumes. *PloS One, 7*(7), e40014.

Camerer, C. F. (1989). Does the basketball market believe in the `Hot hand,'? *The American Economic Review, 79*(5), 1257-1261.

Davis, J., Fodor, A., McElfresh, L., & Krieger, K. (2015). Exploiting week 2 bias in the NFL betting markets. *The Journal of Prediction Markets, 9*(1), 53-67.

Fodor, A., DiFilippo, M., Krieger, K., & Davis, J. (2013). Inefficient pricing from holdover bias in NFL point spread markets. *Applied Financial Economics, 23*(17), 1407-1418. doi:10.1080/09603107.2013.829201

Gandar, J., Zuber, R., O'Brien, T., & Russo, B. (1988). Testing rationality in the point spread betting market. *The Journal of Finance, 43*(4), 995-1008. doi:10.1111/j.1540-6261.1988.tb02617.x

Gayo Avello, D., Metaxas, P. T., & Mustafaraj, E. (2011). Limits of electoral predictions using twitter. Paper presented at the *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media,*

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012-1014.

Gray, P. K., & Gray, S. F. (1997). Testing market efficiency: Evidence from the NFL sports betting market. *The Journal of Finance, 52*(4), 1725-1737. doi:10.1111/j.1540-6261.1997.tb01129.x

Humphreys, B., Paul, R., & Weinbach, A. (2013). Bettor biases and the home-underdog bias in the NFL. *International Journal of Sport Finance, 8*(4), 294-311.

Krieger, K., Pace, R. D., Clarke, N., & Girdner, C. (2015). Anchoring, affect, and efficiency of sports gaming markets around playoff positioning. *Financial Services Review, 24*(4), 313.

Paul, R. J., & Weinbach, A. P. (2005). Bettor misperceptions in the NBA: The overbetting of large favorites and the "hot hand". *Journal of Sports Economics, 6*(4), 390-400.

Sauer, R. D. (1988). Hold your bets: Another look at the efficiency of the gambling market for national football league games: Comment. *Journal of Political Economy, 96*(1), 206-213. doi:http://www.jstor.org/action/showPublication?journalCode=jpoliecon

Sinha, S., Dyer, C., Gimpel, K., & Smith, N. A. (2013). Predicting the NFL using twitter.

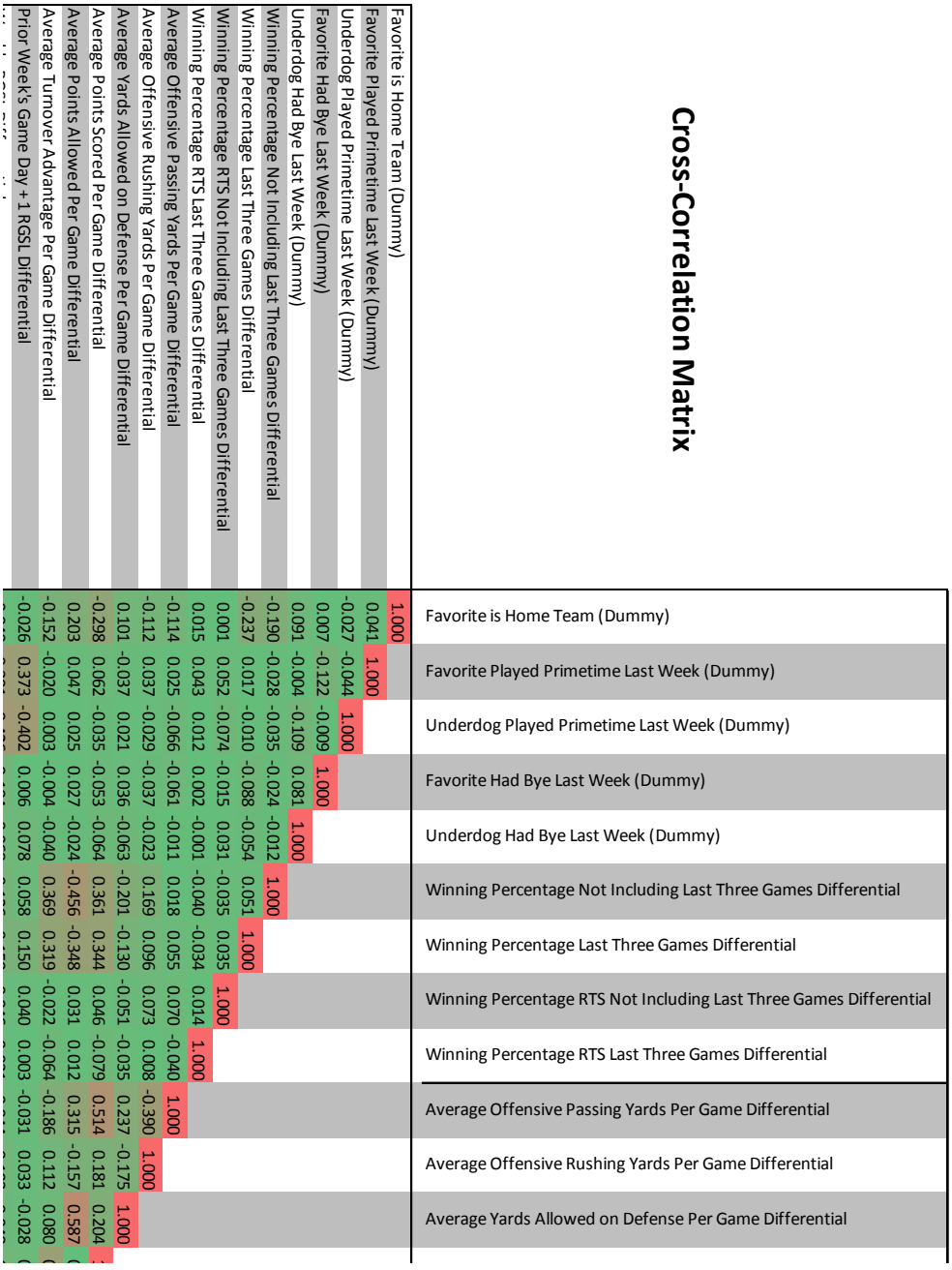
Thaler, R. H., & Ziemba, W. T. (1988). Anomalies parimutuel betting markets: Racetracks and lotteries. *The Journal of Economic Perspectives (1986-1998), 2*(2), 161.

Woodland, L. M., & Woodland, B. M. (1994). Market efficiency and the favorite-longshot bias: The baseball betting market. *Journal of Finance, 49*(1), 269-279. doi:http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291540-6261/issues

Zuber, R. A., Gandar, J. M., & Bowers, B. D. (1985). Beating the spread: Testing the efficiency of the gambling market for national football league games. *Journal of Political Economy, 93*(4), 800-806. Retrieved from http://www.jstor.org/stable/1832139

A. Appendix

A.1. Cross-Correlation Matrix



A.2. Table of Independent Variable Definitions

Dummy Variables

Variable Name	Description
$Home_{f,t}$	1 if the initial favorite is the home team
$Primetime_{f,t-1}$	1 if the initial favorite played a primetime game the previous week
$Primetime_{u,t-1}$	1 if the initial underdog played a primetime game the previous week
$Bye_{f,t-1}$	1 if the initial favorite had a bye the previous week
$Bye_{u,t-1}$	1 if the initial underdog had a bye the previous week

Winning Percentage Variables

Variable Name	Description
$WP\_NotIncluding\_L3_{k,t}$	Initial Favorite's season winning percentage, not including the past three games minus Initial Underdog's season winning percentage, not including the past three games
$WP\_L3_{k,t}$	Initial Favorite's winning percentage only including the last three games minus initial Underdog's winning percentage only including the last three games
$WP\_RTS\_NotIncluding\_L3_{k,t}$	Initial Favorite's season winning percentage relative to the spread, not including the past three games minus initial Underdog's season winning percentage relative to the spread, not including the past three games
$WP\_RTS\_L3_{k,t}$	Initial Favorite's winning percentage relative to the spread only including the last three games minus initial Underdog's winning percentage relative to the

	spread only including the last three games

Team Statistic Variables

Variable Name	Description
$PassingYards_{j,k}$	Average passing yards per game for initial favorite minus that of the initial underdog
$RushingYards_{j,k}$	Average rushing yards per game for initial favorite minus that of the initial underdog
$YardsAllowed_{j,k}$	Average yards allowed on defense per game for initial favorite minus that of the initial underdog
$PointsFor_{j,k}$	Average points scored per game by the initial favorite minus that of the initial underdog
$PointsAgainst_{j,k}$	Average points given up per game by the initial favorite minus that of the initial underdog
$TurnoverDifferential_{j,k}$	Average turnover differential per game for the initial favorite minus that of the initial underdog

Gambler Sentiment Variables

Variable Name	Description
$(Game\ Day\_1_{f,j-1} - Game\ Day\_1_{u,j-1})$	Relative Google Search Level of the initial favorite minus that of the underdog measured on the day after the previous week's game day
$(Weekly_{f,j-1} - Weekly_{u,j-1})$	Relative Google Search Level of the initial favorite minus that of the underdog measured on the day after the previous week's game day