The Journal of Prediction Markets 2016 Vol 10 No 1 pp 53-67

COLLEGE FOOTBALL AND THE VEGAS LINE: DECONSTRUCTION AND ARBITRAGE

Mark David Witte
College of Charleston
USA
wittem@cofc.edu

McDonald Paul Mirabile
Sports Analytics Consulting, LLC USA
mac@sportsanalyticsconsulting.com

ABSTRACT

We examine the Vegas line in college football games by employing two separate regression models to deconstruct the Vegas line and actual margin of victory for 4,590 unique contests from the 2005 through 2011 seasons. A comparison of these two models suggests which factors represent a true relationship with the margin of victory and which reflect bettor biases. An additional model of the margin of victory illustrates which factors the Vegas line systematically misrepresents. The authors find a number of factors inadequately priced in the Vegas line that help explain variation in the actual margin of victory. Using a holdout dataset comprised of the 2012 and 2013 seasons we identify the magnitude of any mispricing and opportunities for arbitrage. We exploit this mispricing to develop and evaluate profitable betting strategies. A strategy betting on the top 35% mispriced games yields 55% correct picks and a 2.7% APY. A second strategy in which only the top 8% of mispriced games are bet yields 59% correct picks and a 5.9% APY.

JEL Classification: Z2

Keywords: College football; mispricing; Vegas line; bookmakers

1 INTRODUCTION

When placing a bet on a college football game, gamblers must first consider the latest line (aka Vegas line). For example, if Michigan State plays at Nebraska and the latest line is -6.5 then a bet on Michigan State wins only if Michigan State wins by 7 points or more. Likewise, a bet on Nebraska wins if Nebraska wins the game or loses by 6 points or less[1]. Sports bookmakers

[1] This was the Vegas line for the Nov. 16, 2013 matchup between these two teams. The Vegas line is usually quoted as the home team's score minus the away team's score. In this example, the Vegas line is a negative number because the away team (Michigan State) was favored to win the game.

typically charge a 10% fee for placing a bet known as the "juice" or the "vig". If sports bookmakers have equal quantities bet on the two teams, then they are guaranteed profit equal to 10% of all total bets.

However, sports bookmakers do not necessarily balance bets on either side of a game as shown by Levitt (2004), Paul and Weinbach (2007), and Paul and Weinbach (2009). A variety of research has examined whether the Vegas line is "efficient" by pricing in all available information[2]. Fair and Oster (2007), Kuester and Sanders (2011), and Kain and Logan (2014) have found that the market is generally but not always efficient. These findings precede the two key questions this paper seeks to address: 1) if the Vegas line does not operate as a balanced market and if it is not necessarily efficient, then how is the Vegas line set? 2) can any sources of these inefficiencies be readily identified and exploited in a model-based approach?

Using a broad set of statistical indicators for NCAA Football Bowl Subdivision (FBS), we model the latest line. We use the same regression specification to determine if the statistical indicators which formulate the Vegas line also have similar predictive power for the actual margin of victory. Our results suggest that the latest line is a very accurate model for the margin of victory. For the 97 different statistical indicators used in the analysis, we test the hypothesis that each of these parameters is the same for our two dependent variables: the Vegas line and the actual margin of victory. We find a handful of variables that are different at the 95% significance level. First, home teams with larger stadiums tend to perform better than the Vegas line would suggest. Second, visting teams that allow fewer points are underweighted by the Vegas line. Lastly, there are numerous conference based differences between the Vegas line and the actual margin of victory which suggest that the Vegas line does not accurately control for conference level differences in home field advantage. To profit from any inefficiency in the Vegas line, a gambler would need to identify games containing multiple biases that collecitvely move the predicted margin of victory significantly far enough from the current line to capitalize on the relatively small mispricings that are occassionally available. This suggests that any inefficiency in sports bookmaking is small; and gamblers may not have sufficient liquidity as the expected return may be fairly small compared to the variance of the "all-or-nothing" bets being wagered. Sports bookmakers, however, have ample funding to profit from gambler biases by setting the latest line based on forecasts of the margin of victory instead of by balancing the money bet on either side of a game. As such, this data supports the idea that sports bookmakers are very good gamblers espoused in Levitt (2004), Paul and Weinbach (2007) and Paul and Weinbach (2009).

[2] Notable recent research on the efficiency of sports bookmaking can be found in Fair and Oster (2007), Kuester and Sanders (2011), and Kain and Logan (2014).

2 LITERATURE REVIEW

Because this study focuses on the determinants of the Vegas line in football, not necessarily the efficency of the market, this literature review focuses on the most relevant studies. These studies all examine the strategy that bookmakers may be using instead of whether or not the market is efficient. Recent research on the efficiency of college football gambling can be found in Xu (2013), Humphreys, Paul and Weinbach (2013) and Kochman, Gilliam and Goodwin (2013).

Levitt (2004) show that bookmakers achieve substantial profits by estimating the outcome of sporting events better than the average gambler. If bookmakers are more strategic gamblers, then they may not need to balance the money wagered on either side of a bet because they are better at predicting outcomes. The bookmakers themselves are likely very informed gamblers and may be using this to profit from the less informed gambling public. This also suggests that when setting an initial spread that bookmakers are likely attempting to predict the biases of the gambling public.

Paul and Weinbach (2007) test the Levitt model of sports bookmakers. They find evidence that Sportsbook.com does not balance money wagered on either side of a bet confirming the "better gambler" theory of book making. The authors suggest that individual bettors may have insufficient liquidity to maintain a long-term strategy. It may be that by exploiting a contrarian betting strategy that a profitable average return could be achieved. However, in anything as random as a sports outcome, it is possible that a string of losing bets could bankrupt an individual bettor. As such, sports bookmakers could persist in exploiting a specific betting strategy because they would have enough liquidity to maintain that strategy long enough for the law of large numbers to assert itself. The authors also suggest that because bookmakers can limit or refuse any bet then bookmakers may refuse any bet from a gambler who's considered "informed." If bookmakers had to accept every bet, then it is possible that any bias in the price would need to be reduced. In Paul and Weinbach (2009) the authors find similar evidence of the "better gambler" theory of bookmaking by confirming the forecasting abilities of bookmakers and the lack of profitability in strategies that take advantage of potential inefficiencies.

Sinkey and Logan (2013) also test the bookmaker's behavior and find that they consistently price the Vegas line to make it difficult to profit from using unsophisticated strategies like betting on the "hot hand" team (a team that has covered the spread recently). The authors show that the Vegas line is a function of previous betting results even though past performance against the Vegas line is not predictive of future success against the Vegas line. Thus, gamblers betting on a "hot hand" team face a bias in the pricing of the Vegas line. The authors calculate that a strategy of betting against the favored team in a prominent game could yield an average return of 1% after paying the 10%

fee. Arguably, this is a fairly low return when compared to the variance of the win or lose nature of the bet.

Paul et al. (2012) show that the college football gambling market is efficient, in that the Vegas line is an accurate predictor of the final score differential. However, when an automatic qualifying team (from a conference with an automatic bid into the Bowl Championship Series bowl games) plays a non-automatic qualifying team then most gamblers bet on the automatic qualifying team. Sports bookmakers do not attempt to balance their books on either side of the bet. In this case, the gamblers appear to be taking advantage of unsophisticated Vegas lines set by the bookmakers because the automatic qualifying team was the winning bet in 54.45% of games. The authors suggest that automatic qualifying teams may try to impress pollsters by winning these non-conference games by large margin, which would explain why the automatic qualifying teams are more likely to cover the Vegas line. Fodor (2012) also finds automatic qualifying teams cover the line at a statistically significant rate (53.92%) over a ten-year period and the majority of individual seasons from 2002-2011.

Several other studies (Farinella and Moffett, 2013, Paul and Weinbach, 2005, Paul and Weinbach, 2013) have considered the Over-Under bet (the total points scored in the game) and found that profitable strategies for betting the under exist across college football, with more profitable results generally associated with higher profile football programs.

Avery and Chevalier (1999) examine changes in the opening and closing NFL lines and suggest that movement is comprised of two components: 1) a predictable component due to sentiment, and 2) an unpredictable component due to new information arriving during the course of the week. As the sentiment component of movement in the point-spread reflects the predictable bias of the betting public but does not add any value in predicting the actual game outcomes, the authors develop a borderline profitable strategy of betting against teams with high levels of predicted sentiment.

We add to the literature in two ways by examining the strategy of bookmakers in setting the Vegas line. First, we model the Vegas line using a vast array of variables to determine what factors are most important when setting the line. Second, because we use a vast data set of team characteristics that may impact both the Vegas line and/or the margin of victory we can test for a variety of sources of gambler bias. If a variable is a statistically significant predictor of the Vegas line but not the margin of victory, then we have identified a source of gambler bias. If, however, we find a variable that is a statistically significant predictor of the margin of victory but not the Vegas line, then we have identified a source of market inefficiency not exploited by the gambling public; perhaps, because the average return is low compared to the variance of the return. Lastly, if we find that there are very few (or very small) differences between the Vegas line and the margin of victory then we will have evidence of the "better gambler" theory of

bookmaking suggested by Levitt (2004), Paul and Weinbach (2007) and Paul and Weinbach (2009).

3 DATA AND METHODOLOGY

The focus of this paper is two-fold: 1) to develop seminal models for both point spreads and margin of victory in college football, and 2) examine profitable, exploitable opportunities in holdout data, comprised of the 2012 and 2013 seasons.

The data for this analysis are comprised of 4,590 individual college football regular season and post-season matchups from 2005 through 2013 seasons. The data used in this analysis were obtained from two primary sources. All team-specific season-to-date performance statistics were obtained from NCAA.com using a web crawler program invoked before each week's games. The NCAA.com website contains statistics on each team's offense and defense, as well as data on each team's strength of schedule, streaks and game location. The remainder of the data was obtained from Covers.com, a website devoted to sports gambling. The Covers.com website data includes the current Vegas line, game time, injury information for each team, and historical against-the-spread data. Although sports book lines are typically expressed with a negative value to denote the favorite (e.g. a home team that is 7-point favorite would be denoted '-7'), the sign is reversed in this data, so a 7-point home favorite is denoted '+7'. Likewise, the actual of margin of victory is calculated by subtracting the away team's score from the home team's score. Other readily available data such as the current week's Associated Press poll votes and stadium size were obtained from ESPN.com. Table 1 reports the descriptive statistics for the independent and dependent variables used in this analysis.

As can be seen in Table 1, the authors have grouped 98 explanatory variables into nine different variable cohorts: Pre-game Expectations, Expert Opinions, Game Location/Type, In-season Game Metrics, In-season Outcome Metrics, In-season Streaks, Injuries, Matchup Situational, Prior Season Outcome Metrics, Home and Away Conferences.

The supposition of this analysis is that sports books will utilize measurable attributes of each opponent as well as well-known sources of bettor bias to create a line that maximizes the expected profit of the sports book, as suggested by Levitt (2004) and Paul and Weinbach (2007). The measurable attributes of each opponent (the offensive and defensive profile of the team) require some time each season to develop. As such, early season games suffer from a lack of consistent prior gameplay on which to evaluate each team. For the purposes of this analysis, the first three weeks of games from all seasons are excluded from the data.[3]

[3] Regardless of how many weeks of games are excluded, there is an inherent tradeoff between predictive accuracy and the number of useable observations. The adjusted

Table 1: Descriptive Statistics

Variable Cohort	Variable Description	Mean	Std Dev
Game outcome	Margin of Victory (Home score - Away score)	4.19	20.995
Pre-game Expectations	Concensus Line	4.362	13.585
	Concensus Over/Under	53.617	8.189
Expert Opinions	Home Team not ranked=1	0.673	0.469
	Visit Team not ranked=1	0.688	0.464
	Home Team's points in AP Poll	179.644	391.802
	Visit Team's points in AP Poll	170.396	382.803
	Game involves a team ranked in AP Top 25=1	0.361	0.48
Game Location/Type	Post-season bowl game=1	0.024	0.153
	Neutral game site = 1	0.042	0.201
	Night game = 1	0.363	0.481
	Stadium Size (10,000s)	5.364	2.21
In-season Game Metrics	Home Team turnovers gained per game	1.857	0.61
	Home Team turnovers lost per game	1.781	0.605
	Visit Team turnovers gained per game	1.851	0.603
	Visit Team turnovers lost per game	1.787	0.603
	Home Team Points per game	28.369	8.481
	Home Team Rush Yards per game	161.013	57.446
	Home Team Pass Yards per game	228.792	62.702
	Home Team Total Yards per minute	13.057	2.625
	Home Team Points allowed per game	24.909	7.817
	Home Team Rush Yards allowed per game	146.968	46.082
	Home Team Pass Yards allowed per game	219.992	43.105
	Home Team Total Yards allowed per minute	12.246	2.145
	Visit Team Points per game	28.165	8.644
	Visit Team Rush Yards per game	161.183	56.888
	Visit Team Pass Yards per game	226.83	62.889
	Visit Team Total Yards per minute	12.989	2.629
	Visit Team Points allowed per game	25.031	7.814
	Visit Team Rush Yards allowed per game	148.006	46.508
	Visit Team Pass Yards allowed per game	219.776	43.234
	Visit Team Total Yards allowed per minute	12.274	2.142
In-season Outcome Metrics	Home Team winning % season-to-date * week number	4.991	2.963
	Visit Team winning % season-to-date * week number	4.962	2.961
	Home Team schedule strength * week number	4.863	1.83
	Visit Team schedule strength * week number	4.813	1.837
	Home Team covered spread in prior game, Visit Team did not	0.252	0.434
	Visit Team covered spread in prior game, Home Team did not	0.252	0.434
	Home Team and Visit Team both covered spread in prior game	0.25	0.433
	Pct of Games covered by Home Team in last 16 weeks	0.503	0.146
	Pct of Games covered by Visit Team in last 16 weeks	0.498	0.148
	Aggregate MV above line by Home Team in last 16 weeks	0.077	4.525
	Aggregate MV above line by Visit Team in last 16 weeks	-0.024	4.594
In-season Streaks	Home Team Won Last Game = 1	0.498	0.5
	Home Team Winning Streak	1.566	2.696
	Home Team Winning Streak Squared	9.722	40.085
	Home Team Losing Streak	1.297	2.221
	Home Team Losing Streak Squared	6.613	26.861
	Visit Team Won Last Game = 1	0.534	0.499
	Visit Team Winning Streak	1.61	2.66
	Visit Team Winning Streak Squared	9.667	38.304
	Visit Team Losing Streak	1.265	2.327
	Visit Team Losing Streak Squared	7.011	31.017

R-squared of Model 1 ranged from a low of .86 using data from every week to a high of .90, using data from only week 10 on.

Table 1 (continued)

Variable Cohort	Variable Description	Mean	Std Dev
Injuries	Home Team's injured/unavailable players	5.033	3.371
	Visit Team's injured/unavailable players	5	3.388
Matchup Situational	Conference Game = 1	0.809	0.393
	Distance between two campuses	0.445	0.513
	Additional Days of Rest for home team since last game	0.071	3.182
	Additional Days of Rest for home team since last game (for bowl games)	0.008	0.634
	Visit Team travelled East and changed 3+ time zones	0.011	0.105
	Visit Team travelled West and changed 3+ time zones	0.013	0.114
	Home Team's last opponent was ranked in AP Top 10	0.093	0.29
	Home Team's next opponent is ranked in AP Top 10	0.125	0.33
	Home Team's last game was a rivalry game	0.198	0.398
	Home Team's next game is a rivalry game	0.219	0.414
	Visit Team's last opponent was ranked in AP Top 10	0.085	0.278
	Visit Team's next opponent is ranked in AP Top 10	0.121	0.327
	Visit Team's last game was a rivalry game	0.186	0.389
	Visit Team's next game is a rivalry game	0.198	0.399
Prior Season Outcome Metrics	Home Team's Win Pct (1-year lag)	0.523	0.221
	Home Team's Win Pct Squared (1-year lag)	0.323	0.234
	Home Team's Win Pct (2-year lag)	0.52	0.22
	Home Team's Win Pct Squared (2-year lag)	0.318	0.231
	Visit Team's Win Pct (1-year lag)	0.524	0.233
	Visit Team's Win Pct Squared (1-year lag)	0.329	0.254
	Visit Team's Win Pct (2-year lag)	0.514	0.217
	Visit Team's Win Pct Squared (2-year lag)	0.312	0.225
Home and Away Conferences	Home Team is a member of the ACC	0.103	0.304
	Home Team is a member of the Big Ten	0.099	0.298
	Home Team is a member of the Big Twelve	0.1	0.3
	Home Team is a member of the CUSA	0.1	0.3
	Home Team is Independent	0.029	0.168
	Home Team is a member of the MAC	0.096	0.294
	Home Team is a member of the MWC	0.078	0.268
	Home Team is a member of the Pac Ten/Twelve	0.096	0.295
	Home Team is a member of the SEC	0.11	0.313
	Home Team is a member of the Sun Belt	0.056	0.23
	Home Team is a member of the WAC	0.065	0.247
	Visit Team is a member of the ACC	0.093	0.29
	Visit Team is a member of the Big Ten	0.091	0.287
	Visit Team is a member of the Big Twelve	0.096	0.295
	Visit Team is a member of the CUSA	0.107	0.309
	Visit Team is Independent	0.03	0.17
	Visit Team is a member of the MAC	0.111	0.314
	Visit Team is a member of the MWC	0.08	0.271
	Visit Team is a member of the Pac Ten/Twelve	0.096	0.295
	Visit Team is a member of the SEC	0.092	0.29
	Visit Team is a member of the Sun Belt	0.069	0.253
	Visit Team is a member of the WAC	0.07	0.255

Each of the models developed in this paper are estimated using a standard ordinary least squares regression. The first model developed, Eq. 1 below, is one that explains the construction of the Vegas line using each opponent's season-to-date offensive and defensive statistics as well as several sources of potential bettor bias such as winning and losing streaks for each team, recent cover performance as well as historical cover performance. Descriptive statistics for each variable in these vectors is given in Table 1.

$$Vegas_i = B_P PreGame_i + B_E Expert_i + B_L Location_i + B_G Game_i + B_O Outcome_i + B_S Streaks_i + B_I Injury_i + B_M Matchup_i + B_R Prior_i + B_C Conference_i + \varepsilon_i \quad (1)$$

The second model developed, Eq. 2, utilizes the same independent variables as Model 1 but explains the actual margin of victory. As has been shown in the literature, the Vegas line is an excellent predictor of the margin of victory, offering only temporary or inconsistent opportunities for arbitrage. By contrasting Model 1 and Model 2, the authors explore how publically available information is used to set the Vegas line and how that same information can be used to predict actual game outcomes.

$$MarginVictory_i = B_P PreGame_i + B_E Expert_i + B_L Location_i + B_G Game_i + B_O Outcome_i + B_S Streaks_i + B_I Injury_i + B_M Matchup_i + B_R Prior_i + B_C Conference_i + \varepsilon_i \quad (2)$$

Lastly, a third model explaining the actual margin of victory is developed with the inclusion of the Vegas line as a predictive variable (Eq. 3). Through the inclusion of this variable, the authors are able to isolate which, if any, other variables retain their statistical significance, suggesting which information relevant to game outcomes is not adequately controlled for with the Vegas Line.

$$MarginVictory_i = B_V Vegas_i + B_P PreGame_i + B_E Expert_i + B_L Location_i + B_G Game_i + B_O Outcome_i + B_S Streaks_i + B_I Injury_i + B_M Matchup_i + B_R Prior_i + B_C Conference_i + \varepsilon_i \quad (3)$$

4 RESULTS AND DISCUSSION

Table 2 reports the model specification and coefficients with their standard errors for each of the three models detailed above. A brief examination of Model 1 (from Eq. 1) shows that the formulation of the Vegas line is quite complex, with seven of the nine variable cohorts contributing multiple statistically significant variables. With 67 statistically significant variables an adjusted R-squared of .879, Model 1 does an exceptional job explaining variation in the Vegas line. Presumably gambler behavior, which is observed only by the bookmaker, represents some of the missing explanatory value of Model 1 (see Paul and Weinbach, 2007). The bookmaker may know that certain teams have a great gambler following and will set the line to profit from those differences.

Model 1 reveals that current season performance is highly influential (In Season Game metrics and In Season Outcome metrics) but insufficient to explain the Vegas line. Expert Opinion and Prior Season Outcome Metrics cohorts also heavily factored into the equation. Additionally, several Game Location/Type cohort variables also show statistical significance, with each of

Table 2: Results from Eq. 1, Eq. 2 and Eq. 3

		Model 1		Model 2		Model 3	
	Dependent Variable	**Line**		**Margin of Victory**		**Margin of Victory**	
	N	**3,582**		**3,582**		**3,582**	
	Adjusted R-squared	**0.8791**		**0.3963**		**0.4460**	
Cohort	**Independent Variables**	**Coeff**	**SE**	**Coeff**	**SE**	**Coeff**	**SE**
	Intercept	-1.880	1.661	-10.844	5.713*	-8.985	5.474
Pre-game	Concensus Line					0.988	0.056***
	Concensus Over/Under	0.015	0.019	0.118	0.064*	0.103	0.061*
Expert Opinions	Home Team not ranked=1	-1.206	0.279***	0.276	0.960	1.468	0.922
	Visit Team not ranked=1	2.072	0.284***	1.503	0.977	-0.544	0.943
	Home Team's points in AP Poll	0.002	0.000***	0.002	0.001*	0.000	0.001
	Visit Team's points in AP Poll	-0.004	0.000***	-0.003	0.001**	0.001	0.001
	Game involves a team ranked in AP Top 25=1	0.365	0.301	0.027	1.034	-0.334	0.991
Game Location/Type	Post-season bowl game=1	-1.086	0.870	-2.943	2.992	-1.870	2.867
	Neutral game site = 1	-3.509	0.548***	-1.616	1.884	1.852	1.816
	Night game = 1	-0.420	0.173**	-1.047	0.596*	-0.631	0.571
	Stadium Size (10,000s)	0.679	0.061***	1.132	0.209***	0.461	0.204**
In-season Game Metrics	Home Team turnovers gained per game	-0.533	0.142***	-0.610	0.488	-0.083	0.469
	Home Team turnovers lost per game	0.284	0.153*	0.209	0.526	-0.072	0.505
	Visit Team turnovers gained per game	0.968	0.144***	1.506	0.494***	0.549	0.476
	Visit Team turnovers lost per game	-0.220	0.150	-0.262	0.518	-0.045	0.496
	Home Team Points per game	0.134	0.025***	0.096	0.086	-0.036	0.083
	Home Team Rush Yards per game	0.015	0.006***	-0.001	0.020	-0.016	0.020
	Home Team Pass Yards per game	0.013	0.006**	-0.003	0.020	-0.016	0.019
	Home Team Total Yards per minute	0.221	0.164	0.702	0.565	0.484	0.541
	Home Team Points allowed per game	-0.125	0.027***	-0.102	0.091	0.022	0.088
	Home Team Rush Yards allowed per game	-0.005	0.006	-0.003	0.022	0.002	0.021
	Home Team Pass Yards allowed per game	0.004	0.006	0.020	0.022	0.016	0.021
	Home Team Total Yards allowed per minute	-0.363	0.177**	-0.991	0.610	-0.632	0.585
	Visit Team Points per game	-0.163	0.025***	-0.167	0.085**	-0.006	0.082
	Visit Team Rush Yards per game	-0.029	0.006***	-0.042	0.020**	-0.014	0.019
	Visit Team Pass Yards per game	-0.026	0.006***	-0.034	0.020*	-0.008	0.019
	Visit Team Total Yards per minute	0.174	0.163	0.400	0.560	0.228	0.537
	Visit Team Points allowed per game	0.144	0.026***	0.359	0.091***	0.217	0.087**
	Visit Team Rush Yards allowed per game	0.004	0.006	-0.006	0.021	-0.010	0.020
	Visit Team Pass Yards allowed per game	-0.009	0.006	-0.028	0.021	-0.020	0.020
	Visit Team Total Yards allowed per minute	0.504	0.170***	0.647	0.584	0.149	0.560
In-season Outcome Metrics	Home Team winning % season-to-date * week number	0.982	0.067***	0.948	0.230***	-0.023	0.227
	Visit Team winning % season-to-date * week number	-0.898	0.064***	-0.894	0.219***	-0.006	0.216
	Home Team schedule strength * week number	1.038	0.084***	1.205	0.289***	0.179	0.282
	Visit Team schedule strength * week number	-1.095	0.084***	-1.112	0.289***	-0.029	0.284
	Home Team covered spread in prior game, Visit Team did not	0.476	0.251*	-0.507	0.864	-0.978	0.828
	Visit Team covered spread in prior game, Home Team did not	-0.273	0.245	-1.215	0.845	-0.945	0.809
	Home Team and Visit Team both covered spread in prior game	0.157	0.272	-0.345	0.935	-0.501	0.896
	Pct of Games covered by Home Team in last 16 weeks	-1.556	0.933*	-4.752	3.211	-3.214	3.078
	Pct of Games covered by Visit Team in last 16 weeks	1.664	0.914*	4.911	3.146	3.266	3.015
	Aggregate MV above line by Home Team in last 16 weeks	0.367	0.032***	0.496	0.111***	0.133	0.109
	Aggregate MV above line by Visit Team in last 16 weeks	-0.311	0.032***	-0.363	0.111***	-0.056	0.107
	Home Team Won Last Game = 1	-0.557	0.327*	0.404	1.124	0.954	1.077
	Home Team Winning Streak	-0.129	0.092	0.129	0.318	0.256	0.304
	Home Team Winning Streak Squared	0.009	0.004**	0.000	0.015	-0.009	0.014
	Home Team Losing Streak	-0.583	0.119***	-0.556	0.410	0.020	0.394
	Home Team Losing Streak Squared	0.034	0.008***	0.036	0.027	0.003	0.026
	Visit Team Won Last Game = 1	0.533	0.326	1.530	1.121	1.003	1.074
	Visit Team Winning Streak	0.210	0.094**	-0.033	0.325	-0.241	0.312
	Visit Team Winning Streak Squared	-0.014	0.005***	0.001	0.017	0.016	0.016
	Visit Team Losing Streak	0.562	0.111***	0.520	0.382	-0.036	0.368
	Visit Team Losing Streak Squared	-0.026	0.007***	-0.042	0.023*	-0.016	0.022

Note: Robust standard errors are reported. Significance at the 99%, 95% and 90% confidence interval is noted by ***, ** and * respectively.

Post-season bowl game, Neutral game site, and Night game having a negative relationship with the Vegas line. The Vegas line also factors in the size of the stadium as well as many Matchup Situational cohort variables such as the distance between the two teams, days of rest for the home team, and prior and future opponent information. Additionally, the Vegas line incorporates recent performance against the spread of each team, with a better recent rate of covering the spread for the home team increasing the line, and a better recent rate of covering the spread for the away team decreasing the line. Finally, the Home and Away Conferences cohort variables reveals that the conferences of

Table 2 (continued) Results from Eq. 1, Eq. 2 and Eq. 3

		Model 1		Model 2		Model 3	
	Dependent Variable	**Line**		**Margin of Victory**		**Margin of Victory**	
	N	**3,582**		**3,582**		**3,582**	
	Adjusted R-squared	**0.8791**		**0.3963**		**0.4460**	
Cohort	**Independent Variables**	**Coeff**	**SE**	**Coeff**	**SE**	**Coeff**	**SE**
Injuries	Home Team's injured/unavailable players	-0.014	0.027	-0.145	0.093	-0.131	0.089
	Visit Team's injured/unavailable players	-0.025	0.027	0.048	0.092	0.074	0.088
Matchup Situational	Conference Game = 1	-0.722	0.270***	0.977	0.929	1.691	0.891*
	Distance between two campuses	1.045	0.238***	0.613	0.819	-0.419	0.787
	Additional Days of Rest for home team since last game	0.043	0.026*	0.025	0.089	-0.018	0.085
	Additional Days of Rest for home team since last game (for bowl games)	-0.080	0.148	0.308	0.510	0.387	0.489
	Visit Team travelled East and changed 3+ time zones	0.412	0.930	-4.486	3.199	-4.893	3.065
	Visit Team travelled West and changed 3+ time zones	-3.332	0.911***	-3.823	3.134	-0.530	3.008
	Home Team's last opponent was ranked in AP Top 10	0.181	0.295	0.992	1.014	0.814	0.972
	Home Team's next opponent is ranked in AP Top 10	-0.273	0.253	-0.539	0.870	-0.270	0.834
	Home Team's last game was a rivalry game	0.441	0.203**	0.522	0.699	0.086	0.670
	Home Team's next game is a rivalry game	0.325	0.194*	-0.006	0.666	-0.328	0.638
	Visit Team's last opponent was ranked in AP Top 10	-0.186	0.306	1.326	1.052	1.510	1.008
	Visit Team's next opponent is ranked in AP Top 10	-0.029	0.262	-0.511	0.903	-0.483	0.865
	Visit Team's last game was a rivalry game	-0.685	0.204***	-0.172	0.701	0.506	0.673
	Visit Team's next game is a rivalry game	-0.552	0.198***	-0.771	0.682	-0.226	0.654
Prior Season Outcome Metrics	Home Team's Win Pct (1-year lag)	11.920	1.625***	8.761	5.591	-3.020	5.397
	Home Team's Win Pct Squared (1-year lag)	-3.680	1.541**	-2.462	5.301	1.176	5.082
	Home Team's Win Pct (2-year lag)	2.372	1.523	10.252	5.240*	7.907	5.021
	Home Team's Win Pct Squared (2-year lag)	1.867	1.467	-6.113	5.048	-7.958	4.837*
	Visit Team's Win Pct (1-year lag)	-15.513	1.489***	-11.275	5.123	4.058	4.983
	Visit Team's Win Pct Squared (1-year lag)	9.352	1.299***	3.869	4.470	-5.375	4.314
	Visit Team's Win Pct (2-year lag)	-0.196	1.517	4.919	5.220	5.113	5.001
	Visit Team's Win Pct Squared (2-year lag)	-4.625	1.467***	-8.771	5.048*	-4.199	4.843
Home and Away Conferences	Home Team is a member of the ACC	2.730	0.602***	-0.775	2.070	-3.473	1.989*
	Home Team is a member of the Big Ten	1.191	0.773	-5.861	2.661**	-7.038	2.550***
	Home Team is a member of the Big Twelve	1.739	0.733**	-2.998	2.521	-4.717	2.417*
	Home Team is a member of the CUSA	-1.162	0.636*	-7.208	2.188***	-6.060	2.097***
	Home Team is Independent	-1.705	0.689**	-4.890	2.372**	-3.205	2.274
	Home Team is a member of the MAC	-1.952	0.670***	-8.737	2.306***	-6.807	2.211***
	Home Team is a member of the MWC	-0.129	0.745	-5.345	2.564**	-5.217	2.456**
	Home Team is a member of the Pac Ten/Twelve	1.646	0.889*	-0.906	3.060	-2.533	2.933
	Home Team is a member of the SEC	3.892	0.652***	-0.341	2.243	-4.188	2.160*
	Home Team is a member of the Sun Belt	-3.334	0.734***	-11.958	2.524***	-8.662	2.425***
	Home Team is a member of the WAC	-1.197	0.779	-5.526	2.679**	-4.343	2.568*
	Visit Team is a member of the ACC	-2.773	0.626***	-0.758	2.154	1.982	2.070
	Visit Team is a member of the Big Ten	-2.068	0.792***	4.007	2.726	6.052	2.614**
	Visit Team is a member of the Big Twelve	-1.512	0.747**	2.686	2.570	4.180	2.464*
	Visit Team is a member of the CUSA	1.577	0.635**	7.233	2.185***	5.674	2.095***
	Visit Team is Independent	0.385	0.732	2.594	2.520	2.213	2.414
	Visit Team is a member of the MAC	3.155	0.654***	8.790	2.251***	5.671	2.164***
	Visit Team is a member of the MWC	0.890	0.747	7.711	2.571***	6.831	2.464***
	Visit Team is a member of the Pac Ten/Twelve	-1.702	0.900*	-0.344	3.096	1.338	2.967
	Visit Team is a member of the SEC	-4.813	0.675***	-1.656	2.323	3.101	2.242
	Visit Team is a member of the Sun Belt	4.379	0.699***	12.543	2.404***	8.215	2.316***
	Visit Team is a member of the WAC	2.965	0.757***	7.466	2.605***	4.535	2.501*

Note: Robust standard errors are reported. Significance at the 99%, 95% and 90% confidence interval is noted by ***, ** and * respectively.

the home and away team are also heavily factored into the formulation of the Vegas line.

Model 2 (from Eq. 2) uses the same variable cohorts to attempt to explain the actual margin of victory. With an adjusted R-squared of 0.3963, Model 2 explains significantly less variation than Model 1, though this is due in part to the large scoring increments of football (most scores occur in three or seven point increments) and the fundamental unpredictable nature of sport (see Stern, 1997).

Although most variable cohorts retain numerous statistically significant variables, the Matchup Situational cohort does not. Indeed, a quick examination of the model reveals that only 31 variables are statically significant. Interestingly, many potential variables that we have hypothesized

as being potential sources of bettor biases, have similar signs and coefficients in both Model 1 and Model 2. While Model 2 is in many ways the analog to Model 1, it is important to recall from the literature that the Vegas line is not explicitly set to predict the Actual Margin of Victory. While it does serve as a *de facto* proxy in many circles for the game's expected outcome, it may also incorporate some information used to exploit bettor biases to the advantage of the bookmaker as suggested in Levitt (2004), Paul and Weinbach (2007) and Paul and Weinbach (2009). By including the actual Vegas line in the original model formulation, we suggest that an improved prediction of the game's outcome can be developed, one that properly weights additional information not fully incorporated into the Vegas line.

Model 3 (from Eq. 3) is the result of this enhancement of Model 2 (the inclusion of the Vegas line as an explanatory variable). The improvement in the adjusted R-squared is readily apparent – an increase from 0.396 to 0.446. This Model is comprised of 97 explantory variables and the Vegas line with just 22 of the independent variables being statistically significant. Due to the inclusion of the Vegas line in this model, we are able to determine which variables and to what extent different types of information is predictive of the margin of victory (which could roughly be construed as being "missing" or "misrepresented" by the Vegas line). As example, consider the Stadium Size (10,000s) coefficient of 0.46. An interpretation of this variable is that an additional 10,000 seat capacity is worth an additional ½ point in the predicted margin of victory for the home team. Investigation of additional significant variables suggests that the Vegas line is effectively mispricing different conference home and away matchups by inaccurately estimating the degree to home field advantage for different games. Specifically, teams from certain conferences may travel "better" and negate home field advantage. Likewise, certain conferences may have more (or less) home field adavantage than the Vegas line would estimate.

To determine the usefulness of Model 3 in identifying arbitrage opportunities, the authors next examine the model's performance predicting game outcomes on data that was not used to train the models, two "holdout" seasons.

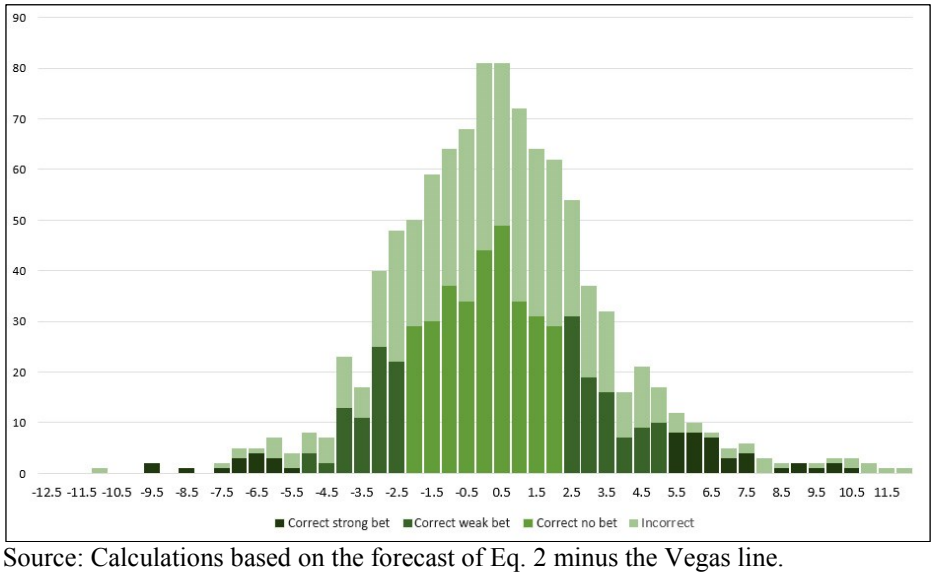
5 HOLDOUT DATA SCORING

To illustrate how applicable the model's predictions are to the real world, we employ a true out-of-sample dataset to effectively grade the model's performance. The authors' elected to use the two later seasons (2012 and 2013) rather than a simple random selection of 25% of the games from the 2005 through 2013 seasons as it represented a truer representation of how a gambler would approach mispricing in the Vegas line. The holdout data is comprised of 1,008 games after the elimination of all game outcomes in which the actual margin of victory was equal to the Vegas line. In such

instances, the gambler's money is fully refunded, so these outcomes do not materially impact the profitability of the proposed strategies.

To establish a profitable betting strategy, the authors create a variable, *PredMoV-Vegas*, equal to the difference between the model's predicted margin of victory and the Vegas line. This variable represents the extent that the game is mispriced, with a negative value suggesting an undervaluation of the home team (relative to the visiting team) and a positive value suggesting an overvaluation of the home team. Each game is then categorized into one of Bet Type three categories: 1) a **strong bet** where the game is mispriced by 5.5 or more points; 2) a **weak bet** where the game is mispriced by between 2.5 and 5.5 points; and 3) **no bet**, where the game is more or less accurately priced (a mispricing of less than 2.5 points). Figure 1 shows the distribution of *PredMoV-Vegas* for all 1,008 games in the holdout data as a stacked bar chart, with light green bars indicate the number of incorrect bets and the other green bars identifying correct bets for the **weak bet** (darker) and **strong bet** (darkest). As shown in Figure 1, the vast majority of games are more or less accurately priced with a few vastly mispriced games, representing the model's suggested opportunities for profit.

Figure 1: Holdout Games: Distribution of *PredMoV-Vegas*



Source: Calculations based on the forecast of Eq. 2 minus the Vegas line.

Table 3 shows the bet categories and associated payoffs for each model-dictated Bet Type category. For each game, a hypothetical $110 was wagered to win $100, with the $10 difference representing the vigorish, or price of placing a bet charged by the sports book.

Table 3Arbitrage opportunities calculated in holdout data

Bet Type	Eligible Games	Bets Placed	Bets Won	Correct Bets %	Bet Cost	Bet Winnings	ROI	Profit
All	35.6%	359	198	55.2%	$ 39,490	$ 41,580	5.3%	$ 2,090
Strong	7.9%	80	47	58.8%	$ 8,800	$ 9,870	12.2%	$ 1,070
Weak	27.7%	279	151	54.1%	$ 30,690	$ 31,710	3.3%	$ 1,020
No Bet	64.4%	649	340	52.4%	$ 71,390	$ 71,400	0.0%	$ 10

Source: Author's calculations based on Eq. 2 to determine the predicted margin of victory relative to the latest line. The coefficients for the forecasted margin of victory are based on games played from 2005-2011 while the holdout data is comprised of games played in 2012 and 2013.

As would be expected, strong bet games have the largest return on investment (ROI), with the model correctly picking the correct team to cover the spread 58.8% of the time. Just 7.9% of the games qualified as strong bet games, however, which equates to 3 per week, on average. Over the course of the two seasons, this 12.2% return for strong bet games is roughly equivalent to a 5.9% annual percentage yield. Additional but smaller arbitrage opportunities exist for weak bet games, which represent an additional 27.7% of the data. The ROI for these games is considerably smaller, consistent with a barely profitable 54.1% correct pick percentage. When combined, 35.6% of the games were classified as either a strong bet or a weak bet, with a 55.2% correct pick percentage yielding a 5.3% ROI roughly equivalent to a 2.7% annual percentage yield over the two seasons[4]. By comparison, in the holdout years of 2012 and 2013 the return on the S&P 500 was 15.9% and 32.2%, respectively, and the return on Barclay's U.S. Aggregate Bond index was 4.2% and -2.0%, respectively.

6 CONCLUSION

Our results suggest that the Vegas line accounts for most, though not all, relevant information to act as an efficient prediction market. However, the Vegas line is imperfect in this respect as we find some indicators (stadium size, a range of conference home/away field advantages) that provide profitable betting opportunities.

[4] Liquidity constraints do apply to these betting strategies due to the all or nothing payouts. In order to make all the strong bets a gambler would need to make as many as 11 bets in a week (roughly 20% of games played) while to make both strong and weak bets a gambler would need to make as many as 26 bets in a week (roughly 50% of games played).

There is also evidence that the Vegas line incorporates many different potential indicators of a college football game in order to predict the outcome of the game. This is consistent with the literature that suggests that bookmakers attempt to profit, not by balancing bets on either side of the Vegas line, but by making better estimates than the average gambler.

One caveat to this research is that there is a profitable reason for bookmakers to revise their methodology. The publication of this article may be sufficient for bookmakers to incorporate the additional indicators that we highlight into their models; the gambling public may find that any inefficiency that is highlighted here may not exist in the future. Conversely, arbitrage opportunities may exist in spite of the Model's identified mispricing due to the ability of the sportsbook to limit the size of the sophisticated bettor's wagers. Models used to create the Vegas line evolve regularly as offenses and defenses evolve and changes to the rule book are implemented. Factors excluded from setting the Vegas line in the past may be incorporated now and in future lines. Future research should examine different weighting of games (exponential smoothing) from different parts of the season, player values for starters and their impact on the line, travel distance, altitude (sea-level vs. high altitude teams), weather (snow and wind) and calculation of statistics based on an entire season-to-date or only more recent games.

7 REFERENCES

Fair, Ray C., and John F. Oster, 2007. College Football Rankings and Market Efficiency, *Journal of Sports Economics*, Vol. 8, No. 1, February, 3-18.

Farinella, Joseph, and Clay M. Moffett, 2013. Market Efficiency and Behavioral Biases in SEC Football: The Over-Under Wager, *Academy of Economics and Finance Journal*, Vol. 4, 15-20.

Fodor, Andy, 2012. The Power of Wagering on Power Conferences, *The Journal of Prediction Markets*, Vol.7, No. 1, 13-25.

Kain, Kyle .J. and Trevon D. Logan, 2014. Are Sports Betting Markets Prediction Markets?: Evidence From a New Test, *Journal of Sports Economics*, Vol. 15, No. 1, February, 45-63.

Kuester, Daniel D., and Shane Sanders, 2011. Regional information and market efficiency: the case
of spread betting in United States college football, *Journal of Economics and Finance*, Vol. 35, No. 1, January, 116-122.

Levitt, Steven, 2004. Why are Gambling Markets Organised so Differently from Financial Markets?, *The Economic Journal*, Vol. 114, No. 1, April, 223-246.

Paul, Rodney J., and Andy P. Weinbach, 2005. Bettor preferences and market efficiency in football totals markets, *Journal of Economics and Finance*, Vol. 29, No. 3, September, 409-415.

Paul, Rodney J., and Andy P. Weinbach, 2007. Does Sportsbook.Com Set Pointspreads to Maximize Profits? Tests of the Levitt Model of Sportsbook Behavior, *The Journal of Prediction Markets*, Vol. 1, No. 3, 209-218.

Paul, Rodney J., and Andy P. Weinbach, 2009. Sportsbook Behavior in the Ncaa Football Betting Market: Tests of the Traditional and Levitt Models of

Sportsbook Behavior, *The Journal of Prediction Markets,* Vol. 3, No. 2, August, 21-37.

Paul, Rodney J., and Andy P. Weinbach, 2009. National television coverage and the behavioral bias of bettors: The American college football totals market, *International Gambling Studies*, Vol. 9, No. 1, March, 55-66.

Paul, Rodney J., Andy P. Weinbach and Eric Higger, 2012. The "large-firm" effect? Bettor preferences and market prices in NCAA football, *The Journal of Prediction Markets*, Vol. 7, No. 2, 29-41.

Sinkey, Michael, and Trevon Logan, 2013. Does the Hot Hand Drive the Market? Evidence from Betting Markets, *Eastern Economic Journal*, Vol. 40, No. 4, September, 1-21.

Stern, Hal S., 1997. How Accurately Can Sports Outcomes Be Predicted?, in the column "A Statistician Reads the Sports Pages, *Chance,* Vol. 10, No. 4, 19-23.