CONSUMER WAGERING BIASES EXPLOITED BY INTERNATIONAL SPORTSBOOKS

A THESIS

Presented to

The Faculty of the Department of Economics and Business

The Colorado College

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Arts

By

Stefan Feiler

February 2024

CONSUMER WAGERING BIASES EXPLOITED BY INTERNATIONAL SPORTSBOOKS

Stefan Feiler

February 2024

Economics

Abstract

Market efficiency, a cornerstone of financial economics, asserts that asset prices fully reflect all available information, thereby facilitating accurate market signals and optimal resource allocation. While the Efficient Market Hypothesis (EMH) underpins numerous economic theories and models, anomalies challenge its assumptions. Behavioral finance posits that psychological biases can lead to market inefficiencies, as evidenced by systematic mispricing biases in various markets, including equities and sports betting. This paper delves into the profitability of sports betting markets, leveraging machine learning and logistic regression to analyze a large sports wagering dataset. The study reveals significant behavioral biases exploited by sportsbooks, shedding light on how advanced analytics can capitalize on market inefficiencies. Insights gleaned from this research not only deepen our understanding of sports wagering dynamics but also offer implications for broader financial markets and the potential transformative role of similar analytic methodologies for asset price discovery.

KEYWORDS: (Efficient Market Hypothesis, Sports Wagering, Gambling, Predictive Analytics)

ON MY HONOR, I HAVE NEITHER GIVEN NOR RECEIVED
UNAUTHORIZED AID ON THIS THESIS

Stefan Feiler

Signature

TABLE OF CONTENTS

Introduction

Market efficiency is one of the foundational concepts in financial economics and is a cornerstone of various interdisciplinary realms that shape critical economic models and the underlying dynamics of competitive markets. The efficient market hypothesis (EMH), introduced by Eugene Fama in the 1960s, states that prices in financial markets correctly reflect all available information, ensuring accurate market signals and guiding optimal resource allocation (Fama, 1970). EMH is also crucial for sustained capital formation, encourages competition and innovation, and provides investor confidence. Indeed, it is applied in nearly every financial decision that takes place in the economy as well as within a multitude of intersecting research domains including economics, finance, business, and even political science. The goal of economic theory is to explain mechanisms by which market forces are driven to produce at equilibrium.

One pertinent mechanism that economic agents utilize to maintain market equilibrium is that of arbitrage. Arbitrage opportunities can help actors find the equilibrium price in the market by buying and selling until price reflects a point at which supply equals demand and price spreads decrease. Actors who respond to prices that do not reflect long run marginal costs and benefits under equilibrium conditions force markets towards equilibrium. Pricing efficiency plays a significant role in determining equilibrium in markets and maintaining economic returns on a risk adjusted basis. Properly structured markets provide a necessary forum through which processes like arbitrage reduces price spreads, and therefore market inefficiencies.

An example of a financial market where it's hypothesized that arbitrage opportunities are frequent is the sports betting market. This market, for all intents and purposes, is a financial market: there are buyers and sellers of fixed-payout contracts whose prices change as information develops. This is not a niche, illiquid financial market either: the industry operators are worth nearly $250 billion (Ibisworld, 2020). In his 2004 paper, economist Steven Levitt argued that the sports betting market creates prices differently from traditional financial markets which lead to inefficiencies. The potential mispricing is due to a key distinction - pricing relies on predictions of human behavior more than on conventionally defined notions of "underlying value" as is the case in equity markets. This connection to complex psychological behavior creates systematic mispricing biases that work to the advantage of sportsbook operators (Levitt, 2004). This paper will analyze and explain potential wagering biases that bookmakers incorporate in their pricing strategies to earn in excess of what should be returned by an efficient market.

This study utilized a combined machine learning and statistical modeling approach to uncover insights from a sports wagering dataset. A proprietary machine learning model was trained on historical NBA wager pricing data to binarily predict winning bets. When applied to previously unseen data from the most recent NBA season (2023-2024), the model has been shown to yield significantly higher return on investment (+7.0%) than expected (-1.0%). Statistical analysis is done in an attempt to elucidate the features that contribute to a higher probability of being positively identified by this machine learning model, providing insights into the pricing models that

sportsbooks use to set irrationally priced odds while still attracting bets, a phenomenon that should not occur in an efficient market.

The analysis indicates that the sportsbooks examined appear to leverage models of consumer behavior to transfer risk and expected losses onto the bettors. Key behavioral factors identified in the data include relative size of game viewership, degree of current forecast uncertainty, and most importantly the time elapsed since game start. These results provide initial empirical evidence that certain characteristics of NBA money line wagers can be used to detect situations where sportsbooks have likely distorted odds in their favor. This paper takes a first step toward identifying properties present within betting data that can reveal when inefficient pricing is probable to exist.

Literature Review

The EMH has great utility in a number of fundamental economic models. Bankers and business people often assume markets are efficient all the time, every day, without properly considering counterfactual evidence. For instance, the influential capital asset pricing model (CAPM), introduced by a Jack Treynor, William Sharpe, John Lintner, and Jan Mossin in the 1960s, is a analytic methodology used across a spectrum of business and financial endeavors to determine the expected return on an investment based on its risk. It is highly dependent on the notion that prices are informed by necessary market information. CAPM is an integral part of the Internal Rate of Return (IRR)[1] and Net Present Value (NPV)[2] calculations, which are the backbone of business investment decisions. In addition, CAPM helps estimate the returns to shareholders for investing capital and as a result, is integral to informed capital allocation decisions throughout the economy. In doing so, it drives proper capital formation and capital allocation decisions in a manner consistent with EMH. Without informational pricing efficiency, the CAPM and asset valuation models would produce misleading results or fail altogether.

---

[1] The internal rate of return is the discount rate that makes the net present value of a project equal to zero. In other words, it is the expected compound annual rate of return that will be earned on a project or investment. IRR allows companies to evaluate and rank alternative investments/projects based on the expected return. It helps companies allocate capital to the projects with the highest returns. A higher IRR generally indicates a more desirable investment. It's an integral component of capital budgeting and investment analysis across corporations and financial institutions.

[2] The net present value is the difference between the present value of the future cash inflows and outflows of a project. It compares the value of money today to the present value of money in the future, taking inflation and returns into account. Investments with positive NPVs are expected to be profitable. NPV factors in the time value of money to help companies make informed comparisons between investment options. It is a key metric used in capital budgeting decisions.

Beyond finance, the EMH and the notion of market efficiency have implications that permeate macroeconomic theory and extend into the realm of political science. Many dynamic stochastic general equilibrium (DSGE) models, crucial in shaping government policy decisions, operate under the assumption that prices and rates convey perfect information about current and future economics trends (Sbordone et. al, 2010). Similarly, rational expectations theory, a cornerstone of macroeconomic models, presumes that expectations of price are unbiased predictors of price (Muth, 1961). The belief in market efficiency underpins macroeconomic frameworks and theories, forming the basis for critical economic models.

When agents react to new data by buying and selling, prices change and assets become fairly valued, allowing markets to direct capital to productive uses. These transactions shift markets towards equilibrium conditions, (as defined by rational actors making rational decisions and bringing returns to their proper risk adjusted rate). Faith in efficient capital markets underpin major economic theories which have great utility for modern industrial-based economies.

The EMH posits that market prices incorporate all available information including, but not limited to, market fundamentals, historical pricing data and future expectations. According to this hypothesis, consistently generating above-market returns by exploiting inefficiencies is impossible in the long run, since transactions force markets back towards equilibrium where supply equals demand. This concept is often illustrated by the analogy that it's not possible for a $100 bill to be lying on the sidewalk

unclaimed, because if it were real, it would have already been picked up by someone. By extension, long-run profitable opportunities do not persist for long in efficient markets.

In the context of financial markets governed by the Efficient Market Hypothesis (EMH), an intriguing conundrum arises: does the EMH imply that all hedge fund managers, portfolio managers, investment bankers, financial analysts, quants, private equity funds and venture capital investors are modern-day alchemists, aberrations chasing elusive and impossible above-market returns? And are those who boast that they have done it successfully, such as Jim Simons – Founder and CEO of the Renaissance Capital– merely on a very hot, very long, statistically implausible lucky streak?[3] For more than 2 decades, Renaissance's Medallion Fund generated annual returns of 66% and created profits of more than $100 billion. This example challenges the conventional understanding of the EMH: that markets leave little room for sustained long-run above-market returns. Simons' success in employing sophisticated technical analysis to achieve remarkable long-term profitability stands as an intriguing anomaly within the framework of market efficiency.

Examples such as the Medallion Fund prompts an exploration into an opposing view called behavioral finance, which argues that psychological biases and irrational behaviors can lead to mispricing that produces inefficiencies (Barberis & Thaler, 2003). Evidence of such inefficiencies driven by behavior has been found across various

[3] Jim Simons, a brilliant mathematician and code breaker pioneered the use of complex mathematical models and quantitative analysis for investing. He built a powerhouse hedge fund that utilized algorithms and predictive models to trade a variety of securities and consistently generate market-beating returns.

markets, including equities, commodities, and even sports betting. In equity markets, research has found that overreaction to news leads to reversals while underreaction causes momentum - two biases producing mispricing (Bondt & Thaler, 1985; Jegadeesh & Titman, 1993). Numerous studies verify that trading strategies designed around these systematic behavioral errors can generate consistent risk-adjusted excess returns, in stark contrast to the EMH (Basu, 1977; Jegadeesh & Titman, 1993).

Information asymmetry also leads to price imbalances. Eugene Fama, the architect of EMH, coauthored the seminal paper describing the mechanisms by which prices of commodity futures diverge from fundamentals: asymmetric information and inventory constraints (Fama & French, 1987; Kumar & Seppi, 1992). More recently, there has been growth in sophisticated statistical arbitrage and machine learning systems designed to capitalize on informational deficiencies in commodity markets (Ewald et al, 2021; Wang & Yu, 2004). While advanced algorithms trading small statistical discrepancies eliminate some mispricing, Shiller (2003) argues that informational and behavioral inefficiencies persist across a variety of assets.

In regard to the sports betting market, bookmakers have incentives to price contracts in such a way that shifts risk exposure and expected losses to burden the consumer. At the same time, sports bettors often exhibit irrational behavior such as allegiance biases, favorite-underdog biases, home team biases, addiction, and other psychological tendencies that bookmakers can use to price the contracts in their own favor, as opposed to the consumers'. This combination allows sportsbooks to offer betting lines that are intentionally priced to benefit the house rather than reflecting the

"true" probabilistically based line. This exposes a contradiction of the EMH in a financial market: prices are not based perfectly on available information/ probabilistic forecasts, but on gamblers' behavior.

Levitt notes that in a perfectly rational market, the fraction of money wagered on a team to win would equal the "true" probability of that team winning. However, sports bettors have behavioral biases in wagering that create imbalances in the fraction of money wagered on a team to win. For instance, if the Denver Broncos have a 40% chance to win a particular game, in a rational market, 40% of the money wagered should be on the Broncos. Should they lose, that money is transferred from Broncos bettors to those on bet on their opponent. The fair payout for betting on the Broncos should be evenly distributed from the amount wagered on the opponent (in this scenario, the fair payout would be $1.50 for every $1.00 wagered – 60/40 = 1.5). Conversely, if the Broncos win, 60% of the wagers should be transferred from those who bet on the opponent to those who bet on the Broncos.

Betting rarely is symmetric with the true probabilities. For instance, if the Broncos were playing a large market team, such as the New York Giants (whose fan base is multiple times larger than that of the Broncos), an asymmetry may arise where an irrational amount of the wagers are placed on the Giants, say, 75%. In this scenario, the Broncos bettors would receive $3.00 for every $1.00 wagered (75/25 = 3) even though the underlying probability suggests a payout of $1.50 per $1.00 wagered.

If the sportsbooks can model this irrational human behavior, they can artificially set the price of these wagers, i.e., the odds, such that they can still receive an

asymmetric (relative to true probability) amount of bets on one side of the game. Using the previous example, the Giants should be priced at $0.67 dollars profit per $1.00 wagered (40/60 = 0.67), however the sportsbook may price them at $0.43 profit per $1.00 wagered (30/70 = 0.43) knowing that irrational human behavior will drive consumers to buy these contracts at unfair prices. Human irrationality, therefore, is very profitable for the sportsbooks because in this example 60% of the time they will payout $0.43 per $1.00 wagered to the Giants bettors, and 40% of the time they will collect $1.00 from the Broncos bettors. This equates to a profit of $0.14 on each dollar wagered in the long run for the sportsbooks:

(40% * $1.00) - (60% * $0.43) = $0.14

While it has historically been understood that simple betting strategies are unprofitable for the consumer (Moskowitz, 2015), advanced machine learning models have recently produced positive historical returns above market benchmarks (Hubáček, Šourek, & Železný, 2019). Training machine learning models to find patterns in bookmaker pricing data and to identify when this artificial price setting is present seems like it may hold greater promise than conventional betting strategies. This paper will analyze results from a machine learning model that has produced positive historical economic returns in contradiction to the EMH to better understand the behavior that sportsbooks take advantage of when pricing lines in their favor.

This paper will analyze the scope and characteristics of the informational asymmetry that is producing these large sports betting profits. It will examine the forms of information available to sports bettors and sportsbooks, the dynamics of sports

betting timing, team-market information, and other behavioral characteristics of the market that are contributing to these above market returns.

Methods

This study employs a comprehensive approach to analyzing sports-wagering data, leveraging the power of both machine learning and traditional statistical methods. The analysis centers around a dataset comprising 363,068 observations, wherein a proprietary machine learning model has successfully identified 86,083 observations with a higher probability of winning. Notably, this subset of positively identified data exhibits a remarkable +7.0% return on investment (ROI) on wagers, in contrast to the -1.0% ROI observed across the entire dataset.[4] Given the inherent complexity and difficulty of interpreting outputs from neural networks, we utilize logistic regression, a robust statistical technique, to elucidate the characteristics of this profitable subset and identify key features present relative to the broader dataset. Through this methodology, we aim to uncover insights regarding the factors contributing to profitable NBA money-line sports wagering, thus enhancing our understanding of the behavior that sportsbooks often take advantage of to tilt the price of wagers in their favor while still attracting irrational proportions of money wagered.

Logistic regression, a method well-suited for modeling binary decisions, is widely applied in financial and economic research. It helps in understanding the phenomena under study by generating coefficients that distinctly describe the relative impact of each independent variable on the dependent variable while accounting for the influence

[4] ROI was determined using a simulated betting algorithm along with the Kelly Criterion for bet sizing. This sizes bets as a proportion of one's bankroll based on the wager's implied probability and potential payout. Only odds posted only within the most recent 30 seconds before the observation timestamp were used to derive implied probability and potential payout. All observations were run through this algorithm to place simulated bets, then compared against actual results to calculate performance metrics. This methodology was applied to the full 2023-2024 dataset and the profitable ML identified subset.

of other predictive variables. Through regression analysis, the significance of various factors can be ascertained, those of greater importance can be discerned, negligible ones can be identified, and relationships can be delineated.

Data

This research aims to identify significant distinctions between the entirety of the dataset and a specific subset. The full dataset consists of a 40-month times-series of NBA money line contract prices for individual games from 44 international bookmakers as well as game context details. Each datapoint consists of all available market prices for a specific game at a specific time in combination with game context details. Each observation contains 83 features. A proprietary binary prediction neural network was trained on the first 1,500,000 observations consisting of data from the first three seasons of this dataset. This neural network was trained to predict winning bets with a binary output and subsequently applied to 363,068 previously unseen data points from the most recent season (2023-2024). Observations receiving positive predictions from the model, defined as outputs exceeding a probability threshold of 0.5, were then filtered into the subset for analysis. The positively identified subset contains 86,083 observations spanning from October 18, 2023 through Dec 31, 2023. The date range of the positively identified subset is only the most recent season due to data sanitation concerns when training the Neural Net on previous older data.

Understanding the features within the dataset is essential for analyzing the data and deriving meaningful insights. This section provides a detailed description of the various features included in the dataset.

- Observation timestamp,
- Team on whom the money line wager is placed,
- 2022-2023 final NBA standings rank of team on whom the money line wager is placed,

- 2022 TV market size rank of team on whom the money line wager is placed,
- Opponent of the team on whom the money line wager is placed,
- 2022-2023 final NBA standings rank of opponent of the team on whom the money line wager is placed,
- 2022 TV market size rank of team on whom the money line wager is placed,
- 44 individual odds for this wager collected from international bookmakers (Draftkings, Fanduel, BetMGM, etc.),
- Whether or not this game was played on a weekend (Fri, Sat, Sun.),
- EST hour of the start time of this game,
- Number of minutes since the game has commenced (negative for pre-game observations),
- The average of posted market odds for this game across all bookmakers,
- The standard deviation of posted odds across all bookmakers,
- The best (highest payout) market odds for this game across all bookmakers,
- The difference between the best posted market odds and the average of posted market odds.

Model

The logistic regression model was developed to predict the probability that an observation would belong to the profitable subset based on a set of independent variables. The predictor variables were carefully selected based on both theoretical relevance from subject matter expertise, as well as empirical evidence from previous academic studies and industry research.

The number of minutes since elapsed commencement of the game was included as a key predictor. The rationale is that as a game unfolds in real time, more and more information accumulates regarding the potential outcome. For example, the probable winner may be easier to forecast when there are only 5 minutes left in a game versus trying to predict the outcome days before the game starts.

Another meaningful predictor is the hour when the game started. This accounts for potential differences in the volume of wagers placed on games depending on when bettors are most active. Although no empirical research exists, it's hypothesized that betting activity increases in the evening hours as consumers finish work and have free time. If betting behavior differs across game start times, sportsbooks may adjust pricing models accordingly.

Additionally, a binary indicator variable for whether the game took place on a weekend (Friday-Sunday) helps capture hypothesized viewership and wagering variations. Sports contests on weekends typically attract larger audiences. This increased attention could stimulate different betting behavior compared to weekday games.

Standings of the given team and its opponent serve as a proxy for the relative quality and recent performance of each club. Sports gamblers are known to fall victim to the "hot hand" fallacy, where a team's recent winning or losing streak irrationally impacts betting decisions. The standings variables tie into this tendency.

Finally, one engineered variables describe the dispersion of market opinions. The standard deviation of prices across sportsbooks shows the level of disagreement. Studies demonstrate decision making changes with greater uncertainty, or less consensus in market forecasts. Significant price deviations between sportsbooks from consensus prices may imply genuine informational advantages.

Interactions between team and opponent TV market size groups (high, medium, low) test viewership effects. Games between teams of different market size may exhibit distinct wagering patterns. All interactions between these three groups are included. The medium-medium matchup is the reference level.

In summary, the predictor variables were thoughtfully selected based on theoretical relevance and empirical evidence within sports wagering literature and practice. The goal is to maximize the model's explanatory power for predicting profitable betting scenarios.

Table 1: Regression Results

Variable	Coef	Std. Err	Z	P>|z|	[0.025	0.975]
const	0.6126	0.066000	9.276	0.000	0.483000	0.742
hour_of_start	-0.0056	0.003000	-1.839	0.066	-0.011000	0.000
minutes_since_commence	0.0026	0.000013	203.249	0.000	0.003000	0.003
odds_std	-0.6989	0.020000	-34.453	0.000	-0.739000	-0.659
standings_rank	0.0011	0.001000	1.943	0.052	-0.000009	0.002
standings_rank_opponent	0.0036	0.001000	6.577	0.000	0.003000	0.005
high_tv_rank_v_high_tv_rank	-0.0869	0.019000	-4.533	0.000	-0.125000	-0.049
high_tv_rank_v_mid_tv_rank	0.0626	0.015000	4.066	0.000	0.032000	0.093
high_tv_rank_v_low_tv_rank	0.1635	0.020000	7.995	0.000	0.123000	0.204
low_tv_rank_v_high_tv_rank	-0.0706	0.020000	-3.586	0.000	-0.109000	-0.032
low_tv_rank_v_mid_tv_rank	-0.1769	0.017000	-10.481	0.000	-0.210000	-0.144
low_tv_rank_v_low_tv_rank	0.0069	0.022000	0.319	0.750	-0.036000	0.049
mid_tv_rank_v_high_tv_rank	0.0133	0.015000	0.876	0.381	-0.016000	0.043
mid_tv_rank_v_low_tv_rank	0.2569	0.018000	14.355	0.000	0.222000	0.292
weekend	-0.2869	0.010000	-29.628	0.000	-0.306000	-0.268
Pseudo-R-squared	0.2288					
Log-Likelihood	-144510					

The results of the regression analysis are shown in Table 1. The logistic regression analysis conducted using these variables accounts for 22.88% of the variance in the independent variable, demonstrating statistical significance. With the exception of interactions between teams with low-TV market viewership versus mid-TV market viewership, and mid-TV market viewership versus high-TV market viewership, all variables exhibit statistical significance at a confidence level of $p > 0.01$. These findings suggest the likelihood of the described behaviors. Further implications of these results will be explored in the discussion section.

The regression coefficients are not directly comparable due to differing scales across variables. To better quantify the relative importance of each predictor, additional analysis was conducted which measured the variance explained in the target variable when excluding each feature individually. This same logistic regression was performed 8 times, each time with only 7/8 variable categories included, a different variable category

removed each time. The R-squared value of each new regression was subtracted from the initial R-squared to test the effect of the inclusion of individual variables in the model on the initial R-squared to evaluate the impact of including individual variables on the overall explanatory power of the model. The results are shown in Table 2.

Table 2: R-squared Results

	Initial R-Squared Less R-Squared with Variable Excluded
minutes_since_commence	0.218907
odds_std	0.013067
weekend	0.002362
tv_rank_interaction	0.001577
standings_rank_opponent	0.000115
standings_rank	0.000010
hour_of_start	0.000009
const	0.000000

The minutes since commence variable stands out as having the most substantial impact on explaining the variance of the dependent variable. In fact, it's notably more influential, by an order of magnitude, than the next most significant variable. This analysis suggests that the majority of other variables have limited contribution to the overall explanatory capability of the model.

The logistic regression analysis revealed several significant predictors of profitable NBA money line wagers from the 2023 season. The minutes elapsed since game commence variable exhibited the strongest individual relationship, based on its outsized contribution to model R-squared when excluded from regression. However, this variable selection method has limitations that warrant caution. Specifically, removing highly correlated predictors may underestimate their joint explanatory power if interactions or redundancies exist. For example, the median of the odds_std variable

changes drastically depending on the timing of the observations relative to the minutes_since_commence variable. For pregame observations where minutes_since_commence is < 0, the median of odds_std is 0.03. For all in-game observations where the minutes_since_commence variable is > 0, the median of this variable is 0.15, exemplifying an unaccounted for between these two features specifically. This works to underscore the need for a holistic interpretation synthesizing both selection approaches.

Overall, the results provide some preliminary empirical support for several hypotheses related to observable human gambling behaviors and tendencies contained within the data. As an example, as the standard deviation of prices posted across all bookmakers in the market increased, reflecting greater uncertainty and disagreement surrounding a game outcome, the likelihood of a given wager being filtered into the profitable subset significantly decreased. This aligns logically with a broad literature highlighting deteriorations in judgment, decision quality, and forecasting accuracy amidst high ambiguity and complexity.

The model provides empirical evidence that pursuing arbitrage opportunities in inefficient betting markets allows bettors to generate consistent profits over time. These results have important implications for developing wagering strategies that focus on identifying and capitalizing on high levels of market consensus as the end of games draws near. Overall, the findings highlight the potential value of arbitrage betting in markets with pricing inefficiencies and demonstrate a statistically significant way to enhance returns on sports wagers.

Interestingly, some coefficient directions ran counter to ex-ante expectations but support the notion that sportsbooks can and do model irrational gambling behavior. The analysis found an inverse relationship between game estimated viewership size and wager profitability. As estimated viewership increases for games played at night, on weekends, and when teams with high TV market size rank played other teams with high TV market size rank, the probability of profitable wagers decreases. This aligns with efficient market hypothesis predictions that greater participation compresses spreads and arbitrage opportunities. It may also reflects the ability of sportsbooks to leverage larger audiences and attract more recreational bettors through distorted odds. With more market participants, sportsbooks can skew odds to incentivize a higher volume of less informed irrational wagers in their favor. However, we cannot conclusively discern the exact driver of this inverse viewership-profitability relationship from the analysis alone. Further study is needed to disentangle the relative contributions of efficient pricing versus strategic sportsbook odds setting in producing this observed pattern. Regardless of the mechanism, these results suggest that availability of profitable arbitrage opportunities suffer as game visibility and subsequent market participation rise.

Another surprising finding was that matchups with teams of higher TV market size rank against opponents with lower TV market size rank often increased profitability probability. The hypothesized view that games with more viewers would stimulate biased wagering activity and therefore more biased price distortion was not borne out

in the data relating to these specific interaction variables. It may be that books inefficiently price smaller-market team matchups for an unknown reason.

Lastly, as the quality of both the team and its opponent decreased per their standings rank, the likelihood of observations being categorized as profitable increased. This provides some evidence that sportsbooks may capitalize on known bettor tendencies like the “hot hand fallacy” to shade pricing against better teams backed by recency biased perceptions.

Conclusion

This study set out to uncover insights about behaviorally driven market pricing inefficiencies and the presence of arbitrage opportunities in the sports betting market by analyzing a dataset using machine learning and regression modeling. The analysis found several significant predictors of profitable bets suggesting presence of pricing inefficiencies. Key results show a positive relationship between the probability of a wager being profitable and the amount of time left until a game ends. In addition, this research found an inverse relationship between game viewership and profitability, suggesting books exploit larger audiences. The findings provide empirical evidence that certain betting characteristics can reveal situations where wager prices are artificially distorted against bettors. In conclusion, the models explain nuanced market inefficiencies arising from mispricing and irrational behavior, demonstrating ways to systematically identify and capitalize on profitable wagering opportunities.

The findings suggesting bookmakers exploit irrational bettor biases may also extend to smaller equity markets dominated by retail investors. Like sports bettors, individual traders are prone to behavioral biases that sophisticated institutional investors can potentially detect and profit from. The analysis implies similar inefficiencies could arise in equities when pricing is influenced more by retail trader biases versus institutional smart money.

It may be possible that behavioral pricing inefficiencies exist in competitive energy markets as well. Day Ahead-Real Time markets in the competitive wholesale power industry are likely to exhibit some sort of behavioral pricing patterns due to

market making based on weather patterns, transmission constraints and natural gas prices.

While this study provides valuable insights, there are several limitations to consider. First, the data span just four seasons, limiting the observations of market condition trends over time. A longer time horizon may reveal different or more important trends. The proprietary machine learning model used to filter profitable bets is complex and treated as a "black box" without interpretability. This introduces challenges in understanding the specific factors driving model outputs, so logistic regression was used instead of analyzing the source model. The regression modeling also relies on proxy variables like team rank and team TV market size to infer effects like betting volume and game viewership. More direct viewership or wagering data may improve accuracy.

The omitted variable bias may also be present if key drivers are absent from the model. For example, inclusion of variables like injuries or referee assignments may provide new insights. Furthermore, while the model suggests inefficiencies and potential mispricing, it cannot definitively prove intentional manipulation by sportsbooks. Research regarding sportsbook operations and pricing decisions could corroborate findings. Finally, the betting simulation uses simplified assumptions for bet sizing and timing which may differ from real-world conditions.

While the analysis provides valuable evidence of market inefficiencies in sports betting, the constraints warrant caution in interpreting findings as definitive proof of intentional mispricing by bookmakers. As with any modeling exercise, limitations exist,

but results can guide future data collection and analyses to further illuminate sports wagering dynamics. Expanding the dataset to include additional sports, leagues, and a longer time horizon could improve generalizability of the findings. Access to proprietary sportsbook data on pricing decisions and bet volumes would also be invaluable for corroborating the hypothesized intentional manipulation.

Sources Cited

Angelini, G., & De Angelis, L. (2019). Efficiency of online football betting markets. *International Journal of Forecasting*, *35*(2), 712–721. https://doi.org/10.1016/j.ijforecast.2018.07.008

Barberis, N., & Thaler, R. (2002). *A Survey of Behavioral Finance*. https://doi.org/10.3386/w9222

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance*, *32*(3), 663. https://doi.org/10.2307/2326304

Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, *40*(3), 793. https://doi.org/10.2307/2327804

Elaad, G., Reade, J. J., & Singleton, C. (2020). Information, prices and efficiency in an online betting market. *Finance Research Letters*, *35*, 101291. https://doi.org/10.1016/j.frl.2019.09.006

Ewald, C.-O., Haugom, E., Lien, G., Størdal, S., & Wu, Y. (2021). Trading time seasonality in commodity futures: An opportunity for arbitrage in the natural gas and crude oil markets? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3792028

Fama, E. F. (1970). Efficient Capital Markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383. https://doi.org/10.2307/2325486

Fama, E. F., & French, K. R. (1987). Commodity futures prices: Some evidence on forecast power, premiums, and the theory of storage. *The Journal of Business*, *60*(1), 55. https://doi.org/10.1086/296385

Hubáček, O., Šourek, G., & Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, *35*(2), 783–796. https://doi.org/10.1016/j.ijforecast.2019.01.001

Ibisworld (November 28, 2020). Global sports betting & lotteries industry - market research report. https://www.ibisworld.com/global/market-research-reports/global-sports-betting-lotteries-industry/

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for Stock Market Efficiency. *The Journal of Finance*, *48*(1), 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

Kumar, P., & Seppi, D. J. (1992). Futures manipulation with “cash settlement.” *The Journal of Finance*, *47*(4), 1485. https://doi.org/10.2307/2328948

Levitt, S. D. (2004). Why are gambling markets organised so differently from financial markets? *The Economic Journal*, *114*(495), 223–246. https://doi.org/10.1111/j.1468-0297.2004.00207.x

McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education data science: Past, present, future. *AERA Open*, *7*, 233285842110520. https://doi.org/10.1177/23328584211052055

MOSKOWITZ, T. J. (2021). Asset pricing and sports betting. *The Journal of Finance*, *76*(6), 3153–3209. https://doi.org/10.1111/jofi.13082

Muth, J. F. (1961). Rational expectations and the theory of Price Movements. *Econometrica*, *29*(3), 315. https://doi.org/10.2307/1909635

Sbordone, A. M., Tambalotti, A., Rao, K., & Walsh, K. (2010). Policy Analysis using DSGE models: An introduction. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1692896

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, *17*(1), 83–104. https://doi.org/10.1257/089533003321164967

Smith, M. A., Paton, D., & Williams, L. V. (2009). Do bookmakers possess superior skills to bettors in predicting outcomes? *Journal of Economic Behavior & Organization*, *71*(2), 539–549. https://doi.org/10.1016/j.jebo.2009.03.016

Wang, C., & Yu, M. (2004). Trading activity and price reversals in futures markets. *Journal of Banking & Finance*, *28*(6), 1337–1361. https://doi.org/10.1016/s0378-4266(03)00120-1