University of Reading

Essays on Information & Beliefs in Sports Economics

Doctor of Philosophy

Department of Economics

Philip Michael Ramirez

December 2023

Essays on Information & Beliefs in Sports Economics

Philip Michael Ramirez

Abstract

This thesis presents three chapters that touch on the influence of information beliefs in sports betting markets. Improving on the odds implied outcome forecasts projected by bookmakers, the first chapter uses Wikipedia to reveal systematic mispricing in women's tennis betting markets. Expanding on the first chapter, the second chapter similarly uses the Mincer and Zarnowitz (1969) economic forecast evaluation framework to identify mispricing in women's tennis betting markets due to 1) bookmaker bias towards the more "beautiful" player relative to their opponent 2) bookmaker bias towards the player with the lighter skin tone relative to their opponent. The final chapter uses Twitter and the English Premier League to analyze the impact of belief dynamics and emotional cues on entertainment utility and consumption.

Betting on a buzz, mispricing and inefficiency in online sportsbooks
Bookmakers sell claims to bettors that depend on the outcomes of professional sports events. Like other financial assets, the wisdom of crowds could help sellers to price these claims more efficiently. We use the Wikipedia profile page views of professional tennis players involved in over ten thousand singles matches to construct a buzz factor. This measures the difference between players in their pre-match page views relative to the usual number of views they received over the previous year. The buzz factor significantly predicts mispricing by bookmakers. Using this fact to forecast match outcomes, we demonstrate that a strategy of betting on players who received more pre-match buzz than their opponents can generate substantial profits. These results imply that sportsbooks could price outcomes more efficiently by listening to the buzz.

Beyond the Baseline: Exploring the Impact of Beauty Bias in Women's Tennis Markets
I ask whether bookmakers set prices on the outcomes they offer efficiently, given the physical characteristics of participants in sporting competitions. Using profile photos from the Women's Tennis Association (WTA) and state-of-the-art deep learning facial recognition methods, I construct a Relative Beauty Differential between tennis match participants. Based on the predicted beauty scores, the Relative Beauty Differential measures the proportional difference in beauty between a player and their opponent.

The constructed measure significantly predicts bookmaker implied-odds forecast error (mispricing, in other words). As a test of market efficiency, I use a beauty informed forecasting model to demonstrate how strategic bets on the less beautiful player would yield sustained profits. Adhering to the standards for market efficiency, this result implies inefficiency in tennis betting markets. Furthermore, I use a completely novel machine learning approach to extract skin tone measures. These measures are used to perform the same set of exercises in reference to relative racial bias, returning similar results.

Exploring entertainment utility from football games

Previous research exploring the role of belief dynamics for consumers in the entertainment industry has largely ignored the fact that emotional reactions are a function of the content *and* a consumer's disposition towards certain participants involved in an event. By analyzing 19m tweets in combination with in-play information for 380 football matches played in the English Premier League we contribute to the literature in three ways. First, we present a setting for testing how belief dynamics drive behavior which is characterized by several desireable features for empirical research. Second, we present an approach for detecting *fans* and *haters* of a club as well as *neutrals* via sentiment revealed in Tweets. Third, by looking at behavioral responses to the temporal resolution of uncertainty during a game, we offer a fine-grained empirical test for the popular uncertainty-of-outcome hypothesis in sports.

Lay summary

This thesis is comprised of three chapters that delve into the impact of information beliefs within sports betting markets. The initial chapter derives a relative "buzz" measure driven by the wisdom of crowds to uncover systematic mispricing in women's tennis betting markets. Expanding on this, the second chapter uses machine learning image processing methods to evaluate the odds-implied forecast error in women's tennis betting markets due to both beauty and racial bias. Each of the first two chapters similarly provide evidence of a divergence from market efficiency within women's tennis betting markets. Lastly, the final chapter takes a novel approach to examine the influence of belief dynamics and emotional cues (surprise, shock, suspense) on entertainment utility and consumption. An additional piece of sentiment analysis on Twitter usage allows for a more granular investigation conditional on fandom within the English Premier League.

Betting on a buzz, mispricing and inefficiency in online sportsbooks

In consequence to the global rise in popularity of sports gambling, both online and in-person, bookmakers facing increasingly small margins have to price efficiently. To test how effective their pricing is, we analyze the Wikipedia profile page views of professional tennis players engaged in over ten thousand singles matches. We construct a Wikipedia Relative Buzz Factor that measures the disparity between players based on their pre-match page views relative to the usual number of views they received over the previous year. We conclude that this buzz factor significantly predicts mispricing by bookmakers. Using this information to forecast match outcomes, we demonstrate that a strategy of betting on players who received more pre-match buzz relative to their opponents can generate substantial profits. Drawing from the efficient market, these results imply that sportsbooks could price outcomes more efficiently by factoring in each player's pre-match buzz. Bookmakers sell claims to bettors that depend on the realised outcomes of professional sports events; much the same as other financial assets, the wisdom of crowds (tipped off by Wikipedia) could help sellers to price these claims more efficiently.

Beyond the Baseline: Exploring the Impact of Beauty Bias in Women's Tennis Markets

The chapter starts with an in-depth literature review on beauty in academic literature. With a broad initial scope, the review starts with general evidence drawing from psychological as well as economic literature on the topic. From the commonly known "beauty premium" phenomenon to more granular sports economics papers, the review narrows as it proceeds. After describing the technical methods previously used to measure beauty, the review concludes with the assertion that the methods used in this chapter are more advanced than typical for the area of research. Using images sourced from the Women's Tennis Association (WTA) and state-of-the-art deep learning facial recognition techniques, I construct a Relative Beauty Differential between participants in tennis matches. This metric, based on the predicted beauty scores, measures the proportional difference in attractiveness between a player and their opponent. A key focus of this chapter is to question whether bookmakers efficiently set prices for the outcomes they offer. Remarkably, the constructed measure is a strong predictor of forecast errors in bookmaker implied odds. In a test of market efficiency, I employ a beauty-informed forecasting model to illustrate how strategic bets placed on the less physically attractive player could lead to sustained profits. In accordance with the Efficient Market Hypothesis, these results indicate mispricing and inefficiency. Additionally, I utilize an innovative machine learning approach to extract skin tone measures, facilitating an analogous set of analyses concerning relative racial bias, yielding similar results.

Exploring entertainment utility from football games

By examining data from 19 million tweets in conjunction with in-play event information and odds for 380 football matches within the English Premier League, this chapter provides valuable contributions to the existing literature. Firstly, we start by employing and augmenting a robust framework designed to assess the driving force of belief dynamics on consumer behavior, distinguished by its incorporation of various emotional cues measured throughout a match. Secondly, we introduce a methodological approach enabling the identification of distinct consumer categories, including "fans," "haters," and "neutrals," through an classification of sentiment expressed in the aforementioned tweets. Through our examination we conclude that emotional cues significantly influence Twitter activity in a given minute. Further, emotional response tends to be smallest for *neutrals* and stronger for *haters* and *fans* depending on emotion. Ultimately, we offer a nuanced empirical evaluation of the widely recognized uncertainty-of-outcome hypothesis in the world of sports by investigating the behavioral responses linked to the temporal resolution of uncertainty during a match.

Declaration

Declaration: I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Philip Michael Ramirez

Acknowledgements

I would like to thank my academic supervisors, James Reade and Carl Singleton for their continued guidance and support. I would also like to thank Sarah Jewell and Giovanni Razzu for their poignant feedback throughout my academic tenure. Lastly, I extend my gratitude to Tim Pawlowski, Dooruj Rambaccussing, and Giambattista Rossi for their active collaboration in contributing to this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

My post-graduate journey can be characterized as a convergence of my academic training, intrinsic interests, and technical skill-set. During graduate school, I developed a keen affinity for quantitative analysis, in large part driven by my passion for sports. While my peers channeled the bulk of their effort into areas like international finance or public policy, I elected to cover topics related to sport and competition whenever an opportunity arose. I even went so far as to develop an automated arbitrage algorithm that would predict, detect and capitalizing on pockets of arbitrage opportunities among various sports betting markets. After receiving my M.A. Economics, I found myself working as a data science analyst at Apple Inc., where I was actively engaged in numerous machine learning and deep learning projects related to Natural Language processing (NLP). Unbeknownst to me, this professional experience would refine my abilities and set the stage for my endeavors later in life. After spending three years honing my technical acumen in a corporate setting, I found myself drawn back to the world of academia, rekindling my curiosity for scholarly pursuits. This renewed drive ultimately inspired me to pursue a PhD., with the aim of delving deeper into the intersection of economics, sports, and data science, and contributing further to the field through advanced research.

My eagerness led me to seek a collaborative environment where I could realize the goal of bridging my academic and professional interests, as well as interact with individuals who shared my enthusiasm for competitive sports. This eagerness steered me to an online listing posted by Dr James Reade expressing a need for students with experience in NLP. A rather niche area, especially at the time, I thought it to be an unlikely coincidence and a compelling opportunity. After exploring his work, and that of other researchers at the University of Reading, including Dr Carl Singleton, I became convinced that this was the setting that would fuel my research and intellectual growth while combining my interests in Economics, sports, NLP, and machine learning. Deliberately, all of my efforts culminated in an amalgamation of machine learning approaches for Economics. While my thesis chapters vary in technical implementation, they all revolve around how people's belief preferences interact with available information. Although regression analysis forms the foundation of economic methods and interpretation, I showcase an array of machine learning techniques, spanning from NLP and sentiment analysis to image processing and facial recognition. I am pleased to assert that this thesis uniquely embodies my academic training, technical proficiency, and intrinsic interests in economics, trading, sports, machine learning, and NLP, culminating in a (hopefully) compelling series of works.

In recent years, we've witnessed an unprecedented increase in the size of sports betting markets at global scale. Paralleled by a worldwide increase in the demand for sports betting, namely influenced by the 2018 U.S. Supreme Court ruling to repeal the Professional and Amateur Sports Protection Act (PASPA) of 1992, individual states in the U.S. were given the autonomy legalise online sports betting as they see fit.[1] Accordingly, 33 states have since elected to legalise online sports betting. A congruent wave in the availability of detailed spectator information has enabled a new generation of online betters to more precisely form their expectations on outcomes. In addition to this information wave, increased bookmaker competition and decreased transaction costs[2] have forced existing odds-setters to price their forecasts on outcomes (implicit in their offered) more accurately. The first chapter addresses whether bookmakers do this successfully, taking women's professional tennis as the area of focus. To assess pricing effectiveness, we examine the Wikipedia profile page views of professional tennis players involved in over ten thousand Women's Tennis Association (WTA) singles matches. We create a measure called the Wikipedia Relative Buzz Factor, gauging the difference in players' pre-match page views compared to their typical views over the past year. Our findings suggest that this buzz factor, influenced by crowd wisdom, significantly forecasts inaccuracies in bookmakers' pricing. By utilizing this insight to predict match results, we illustrate that betting on players with higher pre-match buzz compared to their opponents can yield consistent and significant profits. Keeping in mind efficient market theory, it appears that sportsbooks could refine pricing accuracy by factoring in the belief-driven pre-match buzz surrounding individual players, acknowledging the biases in their projections.

Building upon the first chapter, the second chapter adopts the Mincer and Zarnowitz (1969) economic forecast evaluation framework to investigate economic

[1] See Murphy v. National Collegiate Athletic Association, No. 16-476, 584 U.S. (2018).

[2] see (Forrest, 2008) for details on how competition and lowered transaction costs drive down bookmaker profits

efficiency in largely the same setting. Arguably more controversial than the last, the second chapter shifts its focus to mispricing and bias in women's tennis betting markets as it relates to beauty, or physical attractiveness. While the economic methods remain consistent, the subject matter, implementation, discussion, conclusions, and implications of those conclusions differ greatly. Leveraging images obtained from the Women's Tennis Association (WTA) and cutting-edge deep learning facial recognition techniques, I establish a Relative Beauty Differential between tennis match participants. This metric, reliant on predicted beauty scores, quantifies the proportional attractiveness disparity between players. The primary focus of this research is to challenge the efficacy of bookmakers in setting accurate prices for offered outcomes. Most notably, the developed measure emerges as a reliable predictor of forecast errors in bookmaker implied odds. In line with the Efficient Market Hypothesis, evidence of mispricing would suggest market inefficiency in tennis betting markets. Additionally, I employ a novel machine learning approach to extract skin tone measures, enabling an analogous set of analyses regarding relative racial bias. The chapter poignantly uncovers an undeniable presence of racial bias, remnant in the mispricing of the market odds themselves, and the larger perception of beauty in general.

The final chapter studies the impact of belief dynamics on entertainment utility from football matches. Not isolated to entertainment content alone, the study regards utility as a function of a consumer's sentiment toward the participants of a particular event. In other words, while it is well understood that consumer utility is driven by the match content itself, a major contribution of this chapter is its granular examination of the asymmetric responses to emotions experienced during a match. We combine in-game events with betting odds in order to derive emotional cues - surprise, shock, and sentiment. Further, we employ Twitter data for 380 games played in the English Premier League (EPL) to distinguish between different types of individuals using sentiment analysis (a common technique in Natural Language Processing). To understand how football games engage various individuals, we regress the number of Tweets per minute on each emotional cue. Additionally, we observe behavioral differences by disentangling the impact for fans, haters, and neutrals when their respective teams are either winning or losing. Our results indicate that emotional triggers have a notable impact on overall Twitter engagement during a match. Furthermore, we notice distinct asymmetries in behavior between our classified groups, especially concerning the reaction to suspense.

Chapter 2

Betting on a buzz, mispricing and inefficiency in online sportsbooks

Accepted at: International Journal of Forecasting. (2022); DOI: 10.1016/j.jmacro.2015.12.006. This article was co-authored with James Reade and Carl Singleton, both Professors in Economics at the University of Reading; e-mail: j.j.reade@reading.ac.uk and c.a.singleton@reading.ac.uk. James and Carl have agreed that the essay can appear within this thesis, and that it represents a significant contribution on my part. Re-production of the essay here does not infringe the publisher's copyright policies.

2.1 Introduction

The size and ubiquity of online sports betting markets continue to increase. Most notably in recent years, the world's most successful online sportsbooks entered the U.S. after a 2018 Supreme Court ruling allowed states to legalise gambling at their own discretion.[1] As online sports betting markets have grown and replaced more traditional forms of gambling, lower transaction costs have increased competition and driven down bookmaker profit margins (i.e., the overround or vig) (Forrest, 2008). Over the same period, the amount of online information that bettors can use to form expectations about sports outcomes has increased. This includes detailed historical data about the participants and the setting of an event, the commentary and predictions of sports pundits and tipsters, and the so-called 'wisdom of crowds'. This latter term is

[1] See Murphy v. National Collegiate Athletic Association, No. 16-476, 584 U.S. (2018), which ruled that the Professional and Amateur Sports Protection Act of 1992 was unconstitutional. As of 1 April, 2021, 11 states have legalised online sports betting: California, Delaware, Illinois, Indiana, Michigan, Nevada, New Hampshire, New Jersey, Pennsylvania, Rhode Island and West Virginia.

used widely to describe instances where information aggregated from the decisions of many individuals improves forecasting and decision-making processes, compared with relying on a small number of expert positions (Galton, 1907; Surowiecki, 2004). Given the small profit margins and competition with the crowd-based betting exchanges (prediction markets), odds-setters may need to forecast outcomes and price the claims they sell to bettors more efficiently than ever before. It is natural to ask whether bookmakers are doing this successfully. In this chapter, we use a specific practical example to demonstrate how online sportsbooks are vulnerable to information that could represent the wisdom of crowds.

Wikipedia, the free online encyclopedia, is an example of crowd wisdom. It has become the go-to online place for information about almost anything, including the characteristics and form of sports people.[2] We use this fact to construct what we call the Wiki Relative Buzz Factor, for over ten thousand Women's Tennis Association (WTA) singles matches since the beginning of the 2015 season.[3] These matches were all at the elite level of the sport and include the four annual Grand Slam tournaments. The buzz factor uses the numbers of page views on the Wikipedia profiles of players before their matches began. We call it relative because it compares the players within a match. We call it buzz because it uses the profile page views on the day before a match in proportion to the typical numbers over the past 12 months. We then adapt the Mincer and Zarnowitz (1969) forecast evaluation framework, showing that the Wiki Relative Buzz Factor can significantly predict the systematic mispricing of bookmaker odds, with the higher buzz player being under priced. There is no significant evidence of a favourite or longshot bias in these markets, but bookmakers tended to significantly underprice a player who was substantially lower ranked than their opponent. Taking these results together, we can reject a sufficient condition for weak form market efficiency. To prove that these markets are inefficient, we generate probability forecasts of tennis match results by using the same model that detected the mispricing. Combining these forecasts with the Kelly criterion, which can be motivated from expected utility theory, we demonstrate substantial and sustained profits from exploiting the information contained in the Wiki Relative Buzz Factor. Specifically, we found a potential return on investment of 17-29% from applying the forecasting model at Bet365, the world's highest revenue online sportsbook, over five thousand potential bets on WTA matches between the beginning of the 2019 season

[2] Wikipedia is the 7th most visited website worldwide; see https://www.similarweb.com/top-websites/, retrieved 11 May 2022.

[3] We have no particular rationale for focusing on this sport and the women's game only. However, it is convenient that odds on all these events were offered by a large number of online sportsbooks. Further, we had built a dataset containing information about these events for other research projects before using it to explore the questions in this paper.

and March of 2020. In contrast, using probability forecasts from the widely used Elo (1978) rating systems and the Kelly criterion would have generated substantial losses over the same samples of matches.

These results contribute to the growing literature attempting to elicit the value of crowd wisdom from the field and using this to test the Efficient Markets Hypothesis (Fama, 1965, 1970). Relevant to our study of betting markets, research has demonstrated how information from social media can predict what happens in financial markets, including cross-sectional stock returns (e.g., Avery et al., 2016; Chen et al., 2014; Sprenger et al., 2014) and the price movements of cryptocurrencies (Kraaijeveld and De Smedt, 2020). Specifically using Wikipedia, Moat et al. (2013) found that activity on relevant financial pages could provide some early signs of stock market movements. Behrendt et al. (2020) also found that activity on Wikipedia pages could be used to infer collective investor behaviour and design a trading strategy for individual stocks. In a closely related study to our own, Brown et al. (2018) discovered that the aggregate tone extracted from a large number of Twitter posts contained significant information not present in live betting exchange prices during football matches, especially in the aftermath of major events such as goals or red cards. Using a crowd explicitly making predictions, Brown and Reade (2019) found that the aggregated content from a community of online sports tipsters also contained information not present in betting prices. Betting when the majority of the community predicted a particular outcome generated a small average positive return. Peeters (2018) also found that a crowd of sports fans could improve forecasting accuracy and generate profitable opportunities on betting markets. Specifically, forecasts based on the football player transfer market values on *transfermarkt.de* and the implied strengths of international teams proved more accurate than other standard predictors of match results, such as official team rankings or form-based rating systems.

This chapter contributes more generally to the literature on the efficiency of betting and prediction markets, specifically for sports, much of which has focused on the favourite-longshot bias (for reviews see Vaughan Williams, 1999, Ottaviani and Sørensen, 2008 or Newall and Cortis, 2021). There is a small literature focused on the efficiency of tennis match betting markets (Abinzano et al., 2016, 2019; Forrest and McHale, 2007; Lahvička, 2014; Lyócsa and Výrost, 2018). This literature has tended to find evidence of a longshot bias that is not large enough to overcome the bookmaker profit margin and prove inefficiency. The present chapter also contributes to the use of professional sports to learn about the practice of forecasting, in particular to some studies that have focused on professional tennis (e.g., Angelini et al., 2021a; Barnett and Clarke, 2005; Candila and Scognamillo, 2018; del Corral and Prieto-Rodríguez, 2010; Easton and Uylangco, 2010; Knottenbelt et al., 2012; Kovalchik and Reid, 2019; Kovalchik, 2020; McHale and Morton, 2011; Scheibehenne and Broder, 2007; Spanias and Knottenbelt, 2013). The forecasting models introduced by these studies cannot normally outperform bookmakers without shopping around to find the best available odds (Angelini et al., 2021a; Kovalchik, 2016).

The rest of the chapter proceeds as follows: Section 2.2 describes our dataset, a model to detect mispricing, and a simple betting strategy to test market efficiency using the model; Section 2.3 presents the results; and Section 2.4 concludes.

2.2 Data & Method

We collected information from tennis-data.co.uk for all WTA match results from the main draws of all tournaments, including the Grand Slams, between 1 January 2015 and 16 February 2020.[4] This information includes the identity of players and tournaments, as well as when (local date) and where matches took place.[5] The dataset represents 10,522 matches, 443 players and 271 tournaments. It includes the WTA world rankings of the players immediately before each match, which are based on performances over the preceding year and are updated after a tournament is completed. We used the python packages *geopy* and *timezonefinder* to locate the coordinates of each city in the dataset and the time zones for each match location.

The main draw for a WTA tournament normally takes place a few days before the first round begins, after any qualification matches. All tournaments are in a knock-out format and the draw is seeded, except for the end-of-season WTA Tour finals which have a round-robin stage. The seedings are generally based on world rankings going into a tournament. The average length of a WTA tennis match in 2020 was 97 minutes.[6] A player can normally expect one to three days of rest between matches in a tournament. The lineup for a match is usually known at least the day before it starts, either after the first round draw or the completion of players' previous matches in the tournament, at which point betting odds will become available.

We collected betting odds from oddsportal.com for the winner and loser of a match at the time it began. In what follows, we generally use the average odds from the forty to sixty online bookmakers (sportsbooks) that were posted for any given match on oddsportal.com. We also use the highest (or best) available odds from the bookmaker

[4] These tennis match data are readily available before 2015, but our analysis period is restricted by the availability of historical Wikipedia page views data.

[5] The local date gives the match start, which is important since matches can be played over multiple days due to stoppages, for example, due to the weather.

[6] See http://www.tennisabstract.com/blog/category/match-length/.

sample for each match, as well as the specific odds from Bet365, the largest single online bookmaker (sportsbook) in the world by revenue, numbers of customers and visitors, which offered odds on almost every match in the dataset.[7]

2.2.1 The Wikipedia relative buzz factor

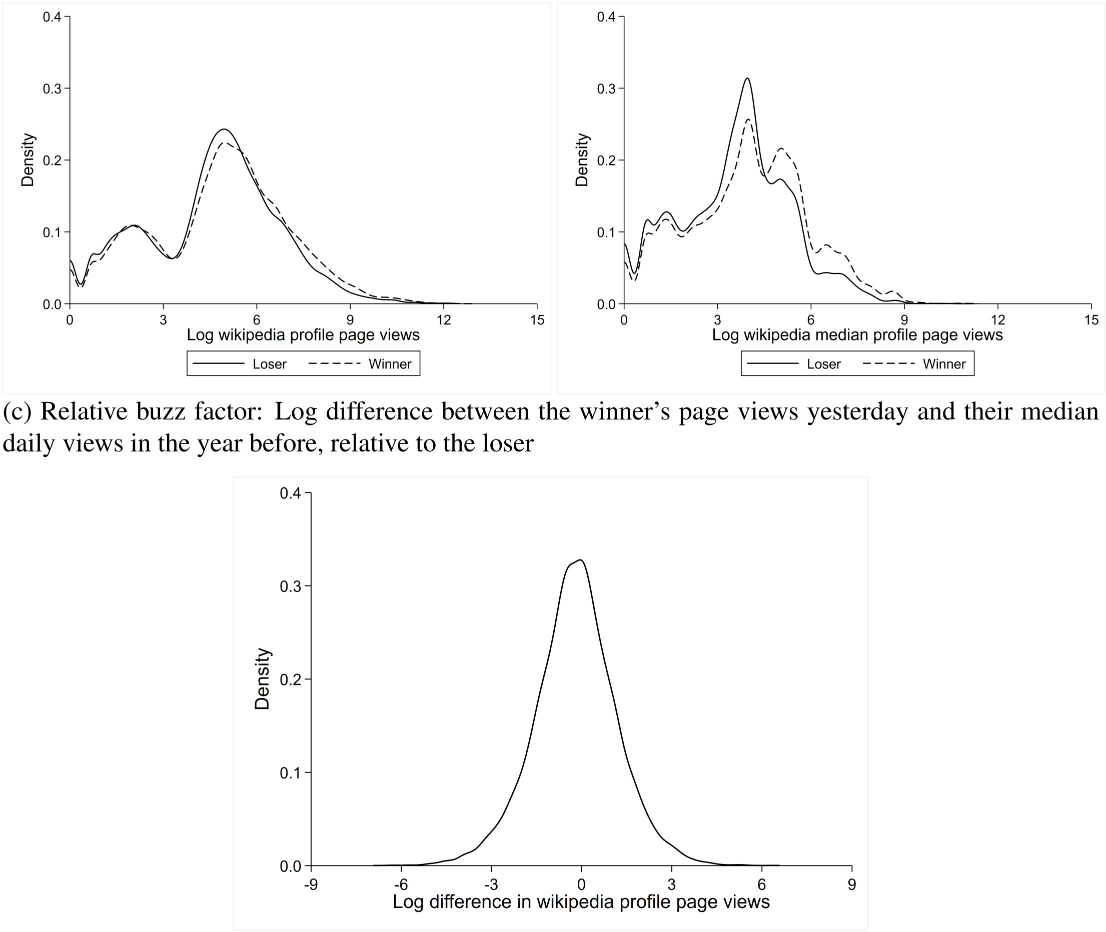
To construct a measure of the pre-match buzz about the players, we collected daily (Coordinated Universal Time, UTC) Wikipedia page views of their English language profiles using the Pageview Application Programming Interface (API), a tool used to query the Wikipedia Foundation pageview data. A small number of observations in the WTA match dataset use maiden names, nicknames, or variations of abbreviations. Therefore, we were careful to ensure every player in the WTA dataset was matched to their Wikipedia profile page views using manual checking. The mean number of page views for players on the day before a match took place was 1,079, with a median of 139, a standard deviation of 6,823 and a maximum of 429,245 (for Naomi Osaka, 7 September 2018, the day before she won the US Open final and her opponent, Serena Williams, accused the umpire of being a “thief”). Panel (a) of Figure 2.1 shows kernel density plots of the log profile page views of players the day before a match took place. The distribution for match winners is generally to the right of that for match losers, suggesting that players with higher levels of interest in their profiles before a match were more likely to win. Panel (b) of Figure 2.1 shows the tighter distributions of the log daily median page views in the past year before a match, though with greater differences between the winner and loser distributions than in panel (a), suggesting that the typical past number of profile page views could be a better predictor of subsequent success in a match.

To generate our ‘Wiki Relative Buzz Factor’ for each player-match observation in the dataset, we combine the information contained in panels (a) and (b) of Figure 2.1. First, we subtract the log median daily page views of a player over the year before a match from the log page views the day before that match for the same player. Second, we subtract from this value the equivalent value for their opponent. As such, our Wiki Relative Buzz Factor measures whether the interest in a player’s Wikipedia profile page was atypical the day before a match, and how much it was atypical relative to their opponent in the match. Precisely, for player i appearing in match j we calculate:

$$\text{WikiBuzz}_{ij} = \ln(w_{ij}/\widetilde{w_{ij}}) - \ln(w_{-ij}/\widetilde{w}_{-ij}) \ , \tag{2.1}$$

[7] see for example https://bestonlinebookmakers.com/largest-bookmakers.html; retrieved 9 June 2022.

FIGURE 2.1: Wikipedia daily page views of tennis players before WTA matches in 2015-2020



(a) The day before the match

(b) Median in the year before the match

(c) Relative buzz factor: Log difference between the winner's page views yesterday and their median daily views in the year before, relative to the loser

Notes: author calculations using Wikipedia Foundation pageview data for the English language profiles of WTA tennis players, collected daily (Coordinated Universal Time (UTC)) using the Pageview Application Programming Interface (API). The densities are estimated with a Gaussian kernel and bandwidth of 0.2.

where w_{ij} is the previous day's page views for the player, $\widetilde{w}_{ij}$ is the median daily page views over the past year before the match, and $-i$ denotes the player's opponent in the match. This measure is plotted in panel (c) of Figure 2.1 only for the winning player observations in the dataset. For the match winners, WikiBuzz is on average negative. Thus, when a player receives a greater log increase in daily pre-match page views relative to the typical number received over the previous year, than their opponent, it tends on average to predict their own defeat in the match (p-value < 0.001). By construction, the 'Wiki Relative Buzz Factor' has zero mean over all winners and losers in the dataset, but we can reject normality with standard tests, due to excess kurtosis of 0.9.

We use the Wikipedia profile page views from the day before the match to construct the buzz factor, instead of the day of the match, because the daily views are in UTC. If we instead used page views from the day of the match, then we could not be confident that the buzz factor was not caused by the outcome of the match (given our data only records when each match began in local time), and we could then not use it to form a realistic betting strategy to test market efficiency. Therefore, by converting all times to UTC format and isolating Wikipedia article views from the day prior to the match, we ensure separation between whenever matches started on a particular day and the Wikipedia data. This rules out the potential for leakage of information about the progress or outcome of a match into the period where we observe and use the Wikipedia profile page views of the players involved.

2.2.2 Detecting mispricing

Let y_{ij} equal one if player $i = 1,2$ won match $j = 1,\dots,J$ and zero otherwise, where i distinguishes between the two players in a match and J gives the total number of matches in a sample, such that the overall sample contains $2J$ player-match observations. Let p_j be the unobserved beliefs of the bookmaker about the probability of $y_{1j} = 1$ happening beforehand, i.e., player 1 winning match j. The bookmaker offers decimal odds o_{ij} on the two potential outcomes, meaning that, on taking a £1 bet, they return o_{ij} to the bettor if the outcome happens and they gain £1 if it does not. Let $z_{ij} = 1/o_{ij}$ be the inverse odds or implied odds-based probability forecast of the bookmaker. For any match, $z_{1j} + z_{2j} = 1 + \kappa_j > 1$, where κ_j has often in the literature been termed as the bookmaker's expected rate of commission or profit margin on a match, also known among sports bettors as the 'overround' or 'vig'. This implies $z_{1j} = p_j + \alpha\kappa_j$ and $z_{2j} = (1 - p_j) + (1 - \alpha)\kappa_j$. If we denote $e_{ij} = y_{ij} - z_{ij}$, then an efficient bookmaker market requires that forecast errors on average are equal to the negative value of some sample average 'overround', $E_{ij}\left[e_{ij}\right] = -\bar{\kappa}$. In other words, the bookmaker is efficient if it makes some average level of commission across matches and outcomes, and no other information can predict e_{ij}, since it would already be priced into the odds.

We consider three potential sources of mispricing and departures from the Efficient Markets Hypothesis in WTA betting markets.

(1) Favourite-longshot bias: There is an empirical irregularity in some prediction and betting markets known as the favourite-longshot bias. When it has been used in the academic literature, this term most typically equates to a longshot bias, whereby the odds offered by bookmakers suggest an underestimation by the market about the chances of the most expected outcomes happening over the least expected outcomes, making bets on favourites generally more profitable than on longshots (see the summaries by Ottaviani and Sørensen, 2008 and Newall and Cortis, 2021). Many studies of professional sports betting markets have identified such a longshot bias, including the seminal study on horse-racing by Ali (1977). Several theoretical contributions, which could be broadly classified as coming from neoclassical economic theory, have demonstrated the sufficient conditions such that this longshot bias can arise in equilibrium, in terms of the preferences, budget constraints and distribution of beliefs among the market participants (e.g., He and Treich, 2017; Manski, 2006; Ottaviani and Sørensen, 2015). The same theoretical frameworks also suggest that high risk aversion among bettors can lead to the bias reversing toward the favourite outcome in the market, which is often termed as a reverse favourite-longshot bias or just favourite bias. Besides the predictions from neoclassical economic theory, a competing set of behavioural explanations has been proposed to explain the favourite-longshot bias, which emphasises the misperception of probabilities by bettors (e.g., Snowberg and Wolfers, 2010; Vaughan Williams et al., 2018). Newall and Cortis (2021) suggest from their review of the empirical literature that sports markets with fewer potential outcomes tend to produce a favourite bias (e.g., team sports or tennis), whereas a longshot bias appears in markets with many outcomes (e.g., horse racing or golf). Nevertheless, previous studies of professional tennis have found a longhsot bias (e.g., Abinzano et al., 2016, 2019; Forrest and McHale, 2007; Lahvička, 2014), though not sufficient to suggest market inefficiency through positive mean returns from consistently betting on match favourites (e.g., Forrest and McHale, 2007; Lyócsa and Výrost, 2018).

(2) Player ranking bias: We consider whether tennis betting markets systematically misprice the outcome of a match according to player rankings. Several studies have demonstrated how the recent performances of tennis players can provide relatively accurate forecasts compared with those implied by bookmaker odds as a benchmark, typically through enhanced Elo (1978) ratings (e.g., Angelini et al., 2021a; Kovalchik and Reid, 2019; Kovalchik, 2020) - we use standard and more advanced Elo ratings later to provide benchmark probability forecasts of match results.[8] There is some suggestive evidence that bookmakers are more risk averse in tennis matches involving lower ranked players and the longshot bias thus increases in these cases (Abinzano et al., 2016; Lahvička, 2014). The WTA world rankings are ordered from one, for the best player cumulatively over the past year, to having no rank, for a player who has not earned enough points at WTA events over the past year to get one. We consider two measures based on these rankings. First, we consider the raw rank

[8] The Elo ratings are computed using all WTA tennis matches between the beginning of the 2007 season and March 2020.

difference between the players in a match, $\text{RankDiff}_{ij} = rank_{ij} - rank_{-ij}$. Second, we assume that the performance difference between two consecutive players in the rankings is decreasing more so as one goes down the ranking list from the top. The difference in ability between the 1st and 2nd ranked players is likely to be more than between the 100th and 101st ranked players, which can be evidenced by how much less often player rankings move at the top compared with the bottom. We construct a ranking distance measure for player i in match j as:

$$\text{RankDist}_{ij} = -\left(\frac{1}{rank_{ij}} - \frac{1}{rank_{-ij}}\right) , \tag{2.2}$$

where we impute $1/rank_{ij} = 0$ if a player was unranked at the time of a match. RankDist_{ij} is bounded by –1, when the player considered is ranked first in the world and is playing somebody unranked, and 1, when it is the other way around, thus having the same sign interpretation as RankDiff_{ij}.

(3) Wikipedia Relative Buzz Factor bias: To the best of our knowledge, this sort of information has not been used to predict the outcome of tennis matches and the efficiency of their betting markets, or at least this has not been documented before. However, there are parallels with studies using information from social media and player evaluations to predict football match outcomes and betting inefficiencies (e.g., Brown et al., 2018; Peeters, 2018).

To detect mispricing and estimate the conditional mean effects on bookmakers' odds implied probability forecast errors, we apply the general Mincer and Zarnowitz (1969) forecast evaluation framework (see Angelini and De Angelis, 2019, Angelini et al., 2021b, and Elaad et al., 2020, who tested for home bias, the favourite-longshot bias and the weak form efficiency of European football betting markets in much the same way). We estimate the following using least squares:

$$e_{ij} = \alpha + \beta_1 z_{ij} + \beta_2 \text{RankDist}_{ij} + \beta_3 \text{WikiBuzz}_{ij} + \psi_{S(j)} + \phi_{T(j)} + \varepsilon_{ij} , \tag{2.3}$$

where $\{\alpha, \beta_1, \beta_2, \beta_3, \psi_{S(j)}, \phi_{T(j)}\}$ are parameters. We expect a significantly negative estimate of α to capture the bookmaker's profit margin (overround). Positive values of β_1, β_2 or β_3 would respectively suggest a longshot bias, a high-rank bias, and a low-buzz bias in the markets, such that betting on a win by the favourite, the lower ranked player, or the one with greater pre-match relative buzz, could be profitable strategies, and vice versa if these parameters are negative. We also consider fixed effects in Equation (2.3) for the season (year), $\psi_{S(j)}$, and tournament of the match, $\phi_{T(j)}$, where $S(j)$ and $T(j)$ are indicator functions, to address the potential heterogeneity over these dimensions in bookmaker overrounds or expected profit margins. The remaining heterogeneity is left in the residual term ε_{ij}. We construct standard errors for the estimates of Equation (2.3) that are robust to clusters at the match and tournament levels. This addresses the heteroskedasticity from including both players in a match in the estimation sample, as well as the possibility that some tournaments may be less predictable than others.[9]

The mean of e_{ij} will be significantly negative for any reasonable sized sample of matches. Therefore, a sufficient condition for the betting market to be weak form efficient, according to Equation (2.3), is given by the null hypothesis: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. If we find that estimates of $\{\beta_1, \beta_2, \beta_3\}$ are significantly positive or negative, then the associated variables provide information that is not fully incorporated in the pre-event prices. In this case, the markets may by inefficient if bettors can use the same information to make sustained positive returns.

2.2.3 Market inefficiency and a simple betting strategy

To test whether the bookmaker markets are inefficient, we use estimation results of the mispricing model in Equation (2.3), an out-of-sample dataset of tennis matches, bookmaker odds and Wikipedia page views data, and the Kelly (1956) criterion. This criterion is the solution to a bettor's maximisation problem on how much of her wealth she should invest in the claim offered by the bookmaker, assuming logarithmic utility and given her beliefs about the outcome of the claim and the odds posted by the bookmaker. Along with simpler strategies, such as "bet one unit when the expected return is positive", the Kelly criterion has been widely used in the literature to evaluate betting market efficiency (e.g., Hvattum and Arntzen, 2010; Peeters, 2018; Ziemba, 2020). We assume that our bettor in this case forms her expectations from estimating Equation (2.3) using ordinary least squares (OLS), though without including the season or tournament fixed effects in the model as these are impractical for forecasting. The other variables in Equation (2.3) are all available to the bettor before a tennis match begins, allowing her to use the estimated model to form probability forecasts of match outcomes. The bettor's out-of-sample expected probability of winning a bet on event i, a specific player to win match j, denoted by $\widetilde{y}_{ij}$, is thus given by:

$$\widetilde{y}_{ij} = \widehat{\alpha} + (1 + \widehat{\beta}_1) z_{ij} + \widehat{\beta}_2 \text{RankDist}_{ij} + \widehat{\beta}_3 \text{WikiBuzz}_{ij} \,, \tag{2.4}$$

[9] As a robustness check, we also considered estimates of Equation (2.3) using weighted least squares, with elements of the diagonal weighting matrix approximated by $z_{1j} \times z_{2j}$, as suggested by Angelini and De Angelis (2019). Although this estimator reduces the influence of more competitive matches, the results that follow are robust to using this instead of ordinary least squares.

where $\{\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3\}$ are in-sample OLS estimators. The Kelly criterion gives the share of a fixed amount of wealth or budget to invest in each bet, remembering that $o_{ij} = 1/z_{ij}$ are the decimal odds offered:

$$x_{ij} = \max\{\widehat{y}_{ij} - \frac{1-\widehat{y}_{ij}}{o_{ij}-1} \text{ , } 0\} \text{ .} \tag{2.5}$$

The bettor's return on investment (ROI) over $N = 2J$ potential bets, expressed as a percentage of the total amount invested over the sample period (some multiple of the per bet budget), is given by:

$$\text{ROI} = \frac{\sum_{ij}^{2J} \left(x_{ij} o_{ij} 1\{y_{ij} = 1\} - x_{ij} 1\{y_{ij} = 0\} \right)}{\sum_{i}^{2J} x_{ij}} \text{ .} \tag{2.6}$$

A substantially positive ROI, over a large out-of-sample number of matches, would provide evidence that tennis match betting markets are weak form inefficient due to some combination of the biases captured by the model. This would suggest that the relatively straightforward model and betting strategy could be applied profitably in real time. To provide benchmark ROIs, we construct alternative estimates of $\widetilde{y}_{ij}$ using the standard player form-based Elo (1978) ratings, with an updating factor (K-factor) of twenty, and using all WTA match results since the beginning of the 2007 season. We also use the more sophisticated W-Elo forecasting model from Angelini et al. (2021a).

2.3 Results

2.3.1 Mispricing

Table 2.1 shows the results of estimating Equation (2.3) for an in-sample period of the 2015-2018 WTA seasons, using as the dependent variable the value of the prediction error according to the mean pre-match odds offered by the K_j (normally 40-60) individual bookmakers ($k = 1, \ldots, K_j$) listed by oddsportal.com for any given match: $\bar{e}_{ij} = y_{ij} - \sum_{k}^{K_j} \left(z_{ijk}/K_j \right)$. Column (I) only tests for a favourite-longshot bias. We find on average a marginal favourite bias, but this is not statistically significant. Column (II) adds the difference in the pre-match WTA rankings of the players, RankDiff_{ij}, as a regressor, which is also not statistically significant. When taken together with the favourite-longshot bias, the null $H_0 : \beta_1 = \beta_2 = 0$ cannot be rejected, and there is no evidence that bookmaker betting markets for WTA tennis matches are mispriced according to the raw difference in ranks and the balance of the odds between players.

In column (III) of Table 2.1, we replace RankDiff_{ij} with our alternative measure of the rank distance between players, RankDist_{ij}. This measure significantly predicts

TABLE 2.1: Model estimates and tests of betting market mispricing for WTA match results, 2015-2018: in-sample period only

	(I)	(II)	(III)	(IV)	(V)
Odds-implied probability	-0.022	-0.061	0.002	0.025	0.025
	(0.024)	(0.040)	(0.027)	(0.029)	(0.029)
WTA rank diff. (player-opponent)		-0.013			
		(0.009)			
WTA rank distance to opponent			0.061^{**}	0.055^{*}	0.055^{**}
			(0.029)	(0.029)	(0.029)
Wiki Relative Buzz Factor				0.009^{**}	0.009^{**}
				(0.004)	(0.004)
Constant	-0.018	0.002	-0.031^{**}	-0.043^{***}	-0.042^{***}
	(0.013)	(0.022)	(0.014)	(0.015)	(0.015)
Year/season fixed effects	Yes	Yes	Yes	Yes	No
Tournament fixed effects	No	No	No	No	Yes
F-test: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$		0.319	0.070	0.022	0.022
N of player-matches	15,854	15,826	15,854	15,854	15,854

Notes.- ***,**,* indicate significance from zero at 1%, 5% and 10% levels, respectively, two-sided tests. Standard errors in parentheses were estimated robust to both match and tournament level clusters.
Column (I): linear regression estimates of Equation (2.3), where the dependent variable is the forecast error implied by average bookmaker odds (oddsportal.com) – test of favourite-longshot bias
Column (II): adds the pre-match raw WTA rank difference to the model in (I)
Column (III): uses the alternative differences in ranks measure described in the text – the coefficient effect should be interpreted as an unranked player against the number one ranked in the world, relative to two hypothetically equally ranked players
Column (IV): adds the Wiki Relative Buzz Factor – preferred results
Column (V): adds tournament fixed effects to the model in (IV)

the average bookmaker odds-implied forecast errors (p-value $= 0.035$) and the null $H_0 : \beta_1 = \beta_2 = 0$ can be rejected at the 10% level. The model estimates suggest that the probability of an unranked player winning against the number one ranked player in the world is 0.061 greater than what bookmaker odds tend to imply. In column (IV), we add the third potential source of mispricing to the model in the form of the Wiki Relative Buzz Factor. This measure positively and significantly predicts the average bookmaker odds-implied forecast errors (p-value $= 0.030$). As mentioned before, on average the player with a relatively larger pre-match increase in Wikipedia profile page views tends to lose a tennis match. However, the model estimates show that bookmaker odds generally imply a further under-prediction of that player's chances, making them more of a longhshot or less of a favourite than they ought to be according to the Wiki Relative Buzz Factor and conditional on the other variables in the model. After including this source of mispricing in the model, the estimated rank distance mispricing remains positive and significant at the 10% level. In this specification, there is a small conditional longshot bias, consistent with the previous literature (Abinzano et al., 2016, 2019; Forrest and McHale, 2007; Lahvička, 2014), though here it is not statistically significant. We can also reject the sufficient condition for weak form market efficiency, $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at the 5% level. In column (V), we add tournament fixed effects to the regression model: the estimates and test results are practically the same.[10] Table 2.2 shows comparable results to Table 2.1 after adding to the estimation samples matches from the 2019 and 2020 (before March) WTA seasons, which we will later use for the out-of-sample forecasting and market efficiency analysis. All of the mispricing test results are robust to extending the sample period in this way.

TABLE 2.2: Model estimates and tests of betting market mispricing for WTA match results, 2015-2020: full sample period

	(I)	(II)	(III)	(IV)	(V)
Odds-implied probability	-0.009	-0.040	0.013	0.036	0.036
	(0.020)	(0.034)	(0.022)	(0.024)	(0.024)
WTA rank diff. (player-opponent)		-0.010			
		(0.008)			
WTA rank distance to opponent			0.054**	0.049**	0.049**
			(0.024)	(0.024)	(0.024)
Wiki Relative Buzz Factor				0.009**	0.009**
				(0.003)	(0.003)
Constant	-0.025**	-0.009	-0.037***	-0.049***	-0.047***
	(0.011)	(0.018)	(0.012)	(0.013)	(0.013)
Year/season fixed effects	Yes	Yes	Yes	Yes	No
Tournament fixed effects	No	No	No	No	Yes
F-test: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$		0.436	0.076	0.016	0.016
N of player-matches	21,044	20,992	21,044	21,044	21,044

Notes.- ***,**,* indicate significance from zero at 1%, 5% and 10% levels, respectively, two-sided tests. Standard errors in parentheses were estimated robust to both match and tournament level clusters. See Table 2.1. Each column model estimates equivalent to the respective columns in Table 2.1 but here matches from the 2019 and 2020 WTA seasons are included in the estimation samples.

Heterogeneity in match location and time differences could perhaps be relevant to the impact of the Wikipedia Relative Buzz Factor. To address this, column (I) of Table 2.3 repeats the model estimates from column (IV) of Table 2.1, and then columns (II)-(IV) show results after cumulatively dropping from the estimation sample matches in time zones from the East, starting with UTC+11&12 (Sydney/Auckland), then UTC+11&12 (Seoul/Tokyo), and finally UTC+7&8 (Singapore/Hong Kong). The influence of the Wiki Relative Buzz Factor and the rejection of the sufficient condition of weak form efficiency are robust to dropping these matches from the estimation sample. After dropping matches from all six of the most eastern time

[10] We checked for misspecification of Equation (2.3) using Ramsey RESET tests and did not reject the null hypothesis; the data generating process was not better approximated by including squared terms for any of the regressors.

zones in the dataset, the mispricing in odds predicted by the buzz factor is greater. This suggests that the Wikipedia profile page views less than 24 hours before the start of a match may be less useful in predicting odds mispricing. This would be consistent with the buzz factor being a proxy for crowd judgements on the relative strengths of players' most recent performances within a tournament. To test whether this could alone explain why the buzz factor can predict bookmaker mispricing, in column (V) of Table 2.3 we re-estimate the model only for matches in the first round of tournaments. The coefficient on the Wiki Relative Buzz Factor remains marginally significant (p-value = 0.069) and is larger than when it is estimated over all matches in tournaments. This suggests that the mispricing is not only driven by whatever happened in the previous round of a tournament, which may have generated interest in a player's Wikipedia profile page. As a further robustness check in this regard, in column (VI) we estimate the model using only first-round matches involving players who had a ranking no greater than 100 and, therefore, were less likely to have come through qualifying rounds in the previous week before entering the main draws of tournaments.[11] In this smaller sample of matches, the coefficient estimate for the Wiki Relative Buzz Factor is even larger than in the previous specifications, but it is also less precisely estimated and thus statistically insignificant at standard levels.

In summary, the results from estimating Equation (2.3), and the tests of mispricing by bookmakers, suggest that there might be inefficiencies in the final result markets of tennis matches. These inefficiencies could be proven by betting on players who are substantially lower ranked then their opponents or who have unusually high interest in their Wikipedia profiles before matches.

2.3.2 Market inefficiency and the betting strategy

Table 2.4 shows the results of applying the simple betting strategy described in Section 2.2.3, by using match outcome probability predictions according to Equation (2.4) and applying the Kelly criterion. We estimated the model up to the end of the 2018 season, used this to forecast match outcomes in the 2019 and 2020 seasons, and then applied the Kelly criterion with these forecasts. Column (I) of Table 2.4 shows the results of the betting strategy for a hypothetical bettor who could place bets at the average pre-match odds offered by the 40-60 bookmakers sampled for each match. The average overround in these markets in 2019 and 2020 (before March) was 5.3%. The out-of-sample probability forecasts and Kelly criterion results suggest betting on 221 of the 5,190 considered odds (2,595 WTA matches in the period), with a total amount

[11] A refinement to this robustness check could less conservatively exclude only matches that exactly included qualifiers, after collecting data on the qualifying events, not least because some 'wildcard' players with a ranking greater than 100 could have entered the tournament directly.

TABLE 2.3: Model estimates and tests of betting market mispricing for WTA match results, 2015-2018: preferred model and dropping time zones, and 1st round matches only

	(I)	(II)	(III)	(IV)	(V)	(VI)
Odds-implied probability	0.025	0.036	0.037	0.043	0.053	0.077
	(0.029)	(0.030)	(0.032)	(0.035)	(0.036)	(0.057)
WTA rank distance to opponent	0.055*	0.061**	0.058*	0.023	0.108*	0.123
	(0.029)	(0.030)	(0.030)	(0.031)	(0.064)	(0.089)
Wiki Relative Buzz Factor	0.009**	0.010**	0.010**	0.014***	0.012*	0.017
	(0.004)	(0.004)	(0.004)	(0.005)	(0.007)	(0.010)
Constant	-0.043***	-0.049***	-0.050***	-0.053***	-0.059***	-0.071**
	(0.015)	(0.016)	(0.018)	(0.018)	(0.019)	(0.030)
Drop UTC+11&12	No	Yes	Yes	Yes	No	No
Drop UTC+9&10	No	No	Yes	Yes	No	No
Drop UTC+7&8	No	No	No	Yes	No	No
Year/season fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
F-test: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$	0.022	0.017	0.019	0.027	0.069	0.165
N of player-matches	15,854	14,448	13,620	11,358	7,208	3,914

Notes.- ***,**,* indicate significance from zero at 1%, 5% and 10% levels, respectively, two-sided tests. Standard errors in parentheses were estimated robust to both match and tournament level clusters.
Column (I): repeats the preferred model estimates from column (IV) of Table 2.1
Column (II)-(IV): each column drops matches from an additional two time zones, starting with UTC+11&12 (Sydney/Auckland) and finally in column (IV) dropping UTC+7&8 (Singapore/Hong Kong)
Column (V): estimates the preferred model from column (I) here but only using matches from the first round of tournaments.
Column (VI): also drops from the estimation sample of column (V) any first round matches which involved a player with a world ranking greater than 100 at the time (i.e., players who were very likely to have come through qualifying rounds in the previous week)

invested equal to 4.3 times the per bet budget and a Return on Investment (ROI) of -6.4%, which is no better than the average bookmaker overround.

For curiosity, column (II) of Table 2.4 presents results whereby the model was estimated and predictions were made using average odds but the best available out-of-sample odds listed on oddsportal.com were used in the Kelly criterion. In this case, a much greater proportion of matches are bet on, the total amount invested over the sample period is 76.6 times the per bet budget, and the ROI is 3.1%. However, despite the existence of ‘oddschecker’ websites being available to the bettor, using the best available odds just before a match begins is not normally realistic due to the transaction costs and time involved with managing a large number of online accounts. Further, there are restrictions that can prevent a bettor from obtaining the best available odds listed on oddsportal.com, such as the location of a bettor affecting which online sportsbooks they can use. This is evidenced by the average overround according to the

TABLE 2.4: Out-of-sample betting strategy results for WTA match results, 2019-20

	Average	Best	Bet365			
				w/out rank	Elo	W-Elo
	(I)	(II)	(III)	(IV)	(V)	(VI)
N odds ($2 \times J$ matches)	5,190	5,188	5,156	5,156	4,796	4,362
Number of bets placed	221	2,350	312	276	1,778	2,058
Mean overround (%)	5.33	-0.23	6.46	6.46	6.48	6.49
Investment (× per bet budget)	4.30	76.63	7.15	4.99	295.36	941.39
Absolute return (× per bet budget)	-0.27	2.34	1.24	1.44	-36.16	-116.50
Return on Investment (%)	-6.37	3.05	17.26	28.82	-12.24	-12.38

Notes.- "Out-of-sample" uses the model from column (IV) of Table 2.1 estimated on matches up to the end of the 2018 season, then uses it to predict match outcomes and apply the Kelly criterion for the 2019 & 2020 seasons. Average odds are always used to estimate the models and generate forecasts, but the odds used in the Kelly criterion are varied.
Column (I): uses the reported average pre-match available odds from oddsportal.com
Column (II): uses the reported best available pre-match odds from oddsportal.com
Column (III): uses pre-match odds from Bet365
Column (IV): uses Bet365 odds but with a version of the preferred model estimated without the rank distance variable
Column (V): uses Bet365 odds but with the standard Elo predicted probability forecast of the match outcome
Column (VI): uses Bet365 odds but with the W-Elo predicted probability forecast of the match outcome as per Angelini et al. (2021a).

best available odds being negative in the 2019 and 2020 WTA seasons, suggesting that theoretical arbitrage opportunities were common if not entirely practical.

As a more realistic test of bookmaker inefficiency, column (III) of Table 2.4 presents results from using the Kelly criterion and the odds from only one online sportsbook. We selected Bet365 because it is the highest revenue sportsbook in the world and had odds listed on oddsportal.com for almost every WTA match since 2015. From using the model's predictions and the Bet365 odds, we find an out-of-sample ROI of 17.3%, which is generated from placing bets according to the criterion on 12% of the main draw WTA matches between the beginning of 2019 and March 2020, equivalent to investing 7.15 times the per bet budget. To check whether these profitable opportunities are driven by the Wiki Relative Buzz Factor, we drop the rank distance measure from the model estimation, with the results shown in column (IV). In this case, fewer matches are bet on and less money is invested according to the Kelly criterion, but the ROI is increased to 28.8% and the absolute return is also greater. To provide a meaningful benchmark ROI using an alternative probability forecasting model of match results, also applied with the Kelly criterion, the same samples of matches and the Bet365 odds, column (V) of Table 2.4 shows results using the standard Elo (1978) ratings model described in Section 2.2.3. The ROI from applying the betting strategy with this alternative set of probability forecasts is -12.2%. This model would also have led to substantial amounts of betting activity and absolute losses over the sample period, because of the frequency and magnitude of differences between the simple Elo predicted probabilities of match outcomes and what bookmaker odds imply, particularly leading to over-betting on longshots.[12] As a further comparison, column (VI) shows betting results using W-Elo, which is a more sophisticated Elo forecasting model of tennis match results that reflects contributions by Kovalchik (2016) and Angelini et al. (2021a). This model gives greater weight to past match wins at prestigious tournaments and takes into account the margins of victory that players achieved.[13] However, the W-Elo model predictions, applied with the Kelly criterion and Bet365 odds, generate a marginally worse ROI and over three times greater absolute losses in our out-of-sample period compared to the standard Elo model in column (V).

Finally, we check whether the betting returns from using the Wiki Relative Buzz Factor are driven by sub-sets of matches expected to be more or less competitive by bookmakers. We estimate the same model over the 2015-2018 seasons and follow the same betting strategy as in column (IV) of Table 2.4, which yielded an out-of-sample ROI of 28.8%, except we consider matches in particular odds ranges. The results in Table 2.5 show that applying the model and betting strategy over matches with intermediate odds, i.e., matches expected to be relatively competitive, generates a marginally higher ROI and a substantially higher absolute return than applying it over all matches, and a substantially higher ROI than applying it over matches expected to be relatively uncompetitive. In this way, the Wiki Relative Buzz factor tends to be a stronger predictor of bookmaker mispricing when matches are expected to be more competitive, and the players involved are by implication more similar in their ability or form.

In summary, a buzz factor about tennis players, constructed from their Wikipedia profile page views data, provides relevant information that is not being fully incorporated into the match result prices offered by bookmakers. This information can

[12] This is not necessarily an indictment of Elo ratings for tennis forecasting and betting. Angelini et al. (2021a) show that standard Elo ratings, a more conservative and sophisticated betting strategy, and the best available odds from a sample of bookmakers, can be used to generate positive betting returns for elite tennis matches.

[13] To generate these ratings, we use an R package associated with Angelini et al. (2021a), *welo* (Candila, 2021). When calculating the W-Elo ratings, we restrict the data to only players who played at least 10 WTA matches since the beginning of 2007, hence the reduced number of odds considered in the betting strategy analysis. The parameters are set to those preferred by Angelini et al. (2021a): player starting points of 1,500, Kovalchik (2016) scale factors, and weights based on the number of games won rather than sets.

TABLE 2.5: Out-of-sample betting strategy results for WTA match result, 2019-20: selecting sample odds based on match competitiveness

	Bet365 odds			
	(I)	(II)	(III)	(IV)
N odds ($2 \times J$ matches)	732	3,459	4,424	1,697
Number of bets placed	4	87	363	263
Mean overround (%)	5.71	6.02	6.58	7.01
Investment (× per bet budget)	0.05	1.03	7.25	9.27
Absolute return (× per bet budget)	-0.002	0.008	1.46	2.72
Return on Investment (%)	-3.02	0.81	20.11	29.38

Notes.- Betting strategy results equivalent to Column (IV) of Table 2.4, varying the sample of match odds used in estimating the model and considered for bets by column. Average odds are always used to estimate the models and generate forecasts, but Bet365 odds are used in the Kelly criterion.
Column (I): uses only odds in the sample which imply a match win probability of $p \in (0,0.2) \cup (0.8,1)$
Column (II): uses only odds in the sample which imply a match win probability of $p \in (0,0.4) \cup (0.6,1)$
Column (III): uses only odds in the sample which imply a match win probability of $p \in [0.2,0.8]$
Column (IV): uses only odds in the sample which imply a match win probability of $p \in [0.4,0.6]$

be used to generate sustained and substantial profits when used in a relatively simple betting strategy.

2.4 Conclusion

In this chapter, we constructed a measure of relative pre-match buzz about tennis players using Wikipedia profile page views data. We found that this Wikipedia Relative Buzz Factor can predict bookmaker odds-implied forecast errors and the significant mispricing of outcomes, suggesting profitable opportunities for bettors who back a player with relatively greater buzz than their opponent going into a match. Using these results to forecast outcome probabilities and the Kelly criterion to select how much to bet on what matches, we found that tennis result betting markets are inefficient. Prices do not fully incorporate the information contained in the buzz factor. The returns on investment from applying the model and betting strategy were sustained and substantial, including when using only the odds of Bet365, the world's highest revenue online sportsbook. Two previous studies also found that online information representing the wisdom of crowds can be used to form profitable betting strategies, though with much smaller rates of return than we have found in tennis markets (Brown and Reade, 2019; Peeters, 2018). However, it is unclear whether correcting these sources of inefficiency would result in greater profits for bookmakers. What we have labelled as mispricing may correlate with unobserved biases and heterogeneity among bettors that bookmakers exploit when setting odds.

There are two natural extensions to this research. The ‘wisdom of crowds’ might explain why a measure constructed from Wikipedia page views data can predict bookmaker mispricing. While this is an appealing and plausible explanation, we have done nothing here to prove it. This would require complementary data sources that capture explicit predictions about tennis match outcomes or evaluations of the players, like the crowd-sourced football transfer market values used by Peeters (2018). The Wikipedia Relative Buzz Factor may only be capturing relative changes in the media interest in tennis players before matches. If that were the case, then our results could perhaps be described more accurately as being driven by the ‘wisdom of the media’, or by a small number of tennis commentators and pundits who selectively draw attention to some players over others. Second, we can think of no good reason why the betting market inefficiencies found here would be constrained to the top level of women’s professional tennis. It would be interesting for others to check whether these results apply to tennis below the WTA level, men’s tennis, or entirely different sports. To this end, we have provided readily adaptable replication code and instructions for all our results on a GitHub page: https://github.com/philiprami/betting_on_a_buzz.

Addendum

Replication study entitled "Not feeling the buzz: Correction study of mispricing and inefficiency in online sportsbooks" submitted to the International Journal of Forecasting. (2023); Manuscript Number: IJF-D-23-00380. The replication highlights a flaw in a dataset used in our published paper that causes an erroneous conclusion: that the Kelly driven, buzz factor informed betting strategy yields positive returns. The authors aptly identify a single "Hercog" (outlier) bet that greatly influences the realistic Bet365 strategy. While the correction is valid, it is only relevant to the Bet365 results since "average" and "best" odds do not have the same vulnerability to outliers. Furthermore, it does not - in any way - impact our main result concerning mispricing (again, since we use average odds to test mispricing). However, the replication emphasises the importance of checking plots; this error would have been identified had we checked the bet-by-bet cumulative returns from the betting strategy.

Chapter 3

Beyond the Baseline: Exploring the Impact of Beauty Bias in Women's Tennis Markets

3.1 Introduction

Beauty is the object of admiration across countless novels, songs, paintings, and works of literature, both academic and non-academic. From artists and aestheticians, to psychologists and mathematicians, the indispensable concept has long captured the attention of those attempting to gain a fuller understanding of the phenomenon. Although vital to the human experience, any formal definition, construction of objective criteria, or quantification of beauty remains an elusive endeavor. Such notable undertakings include the rule of thirds largely used in composition, the geometrical golden ratio apparent in nature and art, and the inherent preference toward the law of symmetry. Beauty permeates through an array of academic disciplines including, but not limited to, Economics (e.g., Langlois et al., 2000; Prusinkiewicz and Lindenmayer, 2012; Grammer et al., 2003). Within the field the majority of research explores the economic implications of beauty on labor markets; however, the intersection of professional sports, emerging sports betting markets, and beauty provides a uniquely underrepresented opportunity for analysis. Concurrent with an unprecedented rise in the popularity of sports gambling, a growing number of betting markets continue to surface. Since the momentous Supreme Court ruling in 2018[1] to end the federal ban on sports betting, U.S. expenditure on sports gambling has increased from 310 million per month to approximately 7 billion per month (O'Brien

[1] See Murphy v. National Collegiate Athletic Association, No. 16-476, 584 U.S. (2018)

and He, 2021). In parallel, the increase in consumer demand for sports entertainment is coinciding with a corresponding increase of relevant sports betting information. This influx of publicly available information allows market participants to more reliably form expectations around market outcomes. These newly informed consumers, along with added competition, subsequently slims down bookmaker profits, forcing price setters to forecast more accurately. As the saying goes, never judge a book by its cover. But do betting markets actually adhere to this parable in practice? While research asserts the existence of pretty privilege in everyday life (e.g., Webster Jr and Driskell Jr, 1983; Berggren et al., 2010), are the price setters of sports betting markets equally biased or do their offered odds accordingly?

To address the question of whether physical attractiveness has any bearing on the expected outcomes of individual matches, I focus my attention on the Woman's Tennis Association (WTA). With over 1650 players in the association, the WTA is unquestionably the most prominent women's tennis league in existence today.[2] Gathering all available WTA profile photos for the competitors featured in every match since 2015, I construct a deep learning beauty mode used to generate unique beauty scores for each available player.[3] With these individual beauty scores, I generate a relative beauty differential comparing the proportional differences in beauty between match opponents. As a test of market mispricing, I then utilize the Mincer and Zarnowitz (1969) economic forecast evaluation framework. The results indicate that the deep learning relative beauty measure can reliably predict market forecast error due to a systematic overvaluation of the player with the higher beauty score. While bookmakers tend to overprice the more relatively beautiful player, they simultaneously undervalue their bookmaker odds on the less beautiful player. The Efficient Markets Hypothesis (Fama, 1965, 1970) defines an efficient market as one in which prices fully reflect all available information. Accounting for this, the results suggest the rejection of a sufficient condition for weak form market efficiency in women's tennis markets. Moreover, to prove this market inefficiency, I use the price forecasting models to employ a simple betting strategy based on the utility maximizing Kelly Criterion. By exploiting the bookmaker omitted information gained through the relative beauty measure, the strategy - informed by the Kelly Criterion - renders a positive return. Assuming the average market odds offered on the out of sample matches, the Kelly motivated strategy demonstrates a potential return on investment of approximately 2%. By placing 309 bets on matches from the 2019 and 2020 seasons, the model is able to

[2] The Women's Tennis Association is the principal organizing body of women's professional tennis which governs the largest worldwide professional tennis tour for women; see https://www.wtatennis.com

[3] "deep learning" refers to neural network algorithms, a specific type of predictive machine learning model

overcome the average overround, or bookmaker commission, of approximately 5%. Assuming the conditions of an efficient market, such a sustained return using publicly available information would not be possible.

This exercise adds to the volume of literature testing the Efficient Market Hypothesis within the context of sports betting markets (see Brown et al., 2018, Brown and Reade, 2019, Peeters, 2018, and Ramirez et al., 2023). In addition to contributing to this collection of research, this chapter opens a larger conversation about beauty and the complex social dynamics surrounding the concept. Originally, the scope of this chapter was limited to the impact of beauty within the confines of women's professional tennis; however, after observing the preliminary deep learning beauty score results, the likelihood of racial bias diluting the measures were hard to ignore, with lighter skin-toned players consistently scoring higher than darker skin-toned players. While convolutional neural networks are the current industry standard for image processing and facial recognition tasks, AI and machine learning models, in general, have historically demonstrated an innate tendency toward racial bias (see Metz, 2021, Verma, 2022, and Raikes, 2023). This observation leads to an open question on whether the apparently racist tendency is a trait unique to the deep learning framework, or that the models are simply adopting the biases introduced during their training. Although this is an unfortunately common trend in the world of machine learning, nevertheless, it speaks to a larger conversation about, as well as our definition of, beauty as humans. Given the imperative to adjust for racial bias (alongside other discriminatory characteristics) in AI and machine learning models, it may be an equal imperative for the public adjust its own bias toward the perception of beauty - whether that be on a cultural, national, or global level.

This chapter proceeds as follows: Section 3.2 presents a literature review identifying formative research regarding 1) beauty in economics and psychology 2) beauty in sports and tennis 3) technological methods in measuring beauty. Additionally, the review offers this chapter as a technological improvement to previous research attempting to quantify beauty. Section 3.3 describes the dataset itself and the deep learning beauty implementation methods used to generate beauty predictions. Section 3.4 illustrates the economic methods used to detect mispricing and offers a Kelly criteria model as a test of market efficiency. Section 3.5 describes results and section 3.6 arrives at conclusions.

3.2 Literature Review

3.2.1 General Evidence

The link between earnings and beauty is a well-documented phenomenon. "Beauty and the Labor Market", the seminal research by Hammermesh and Biddle (1994), examines the relationship between attractiveness and labor market outcomes. The study measures the impact of physical attractiveness on labor market variables such as employment, wages, and job opportunities. Ultimately, the authors identify a "beauty premium" in the labor market i.e. physically attractive individuals enjoy higher wages, higher rates of employment, and more job security compared to their less attractive counterparts when controlling for relevant factors like education and experience. Building atop this pivotal research, Judge et al. (2009) test the influence of physical attractiveness and intelligence on income and financial strain. They assert that general mental ability and attractiveness showed both direct and indirect effects on income, controlling for educational attainment and self-evaluations. Citing a previously limited view of the beauty-to-earnings relationship, (Scholz and Sicinski, 2015) introduce a more comprehensive explanation conditional on natural human ability. In line with the findings of Hammermesh and Biddle, Scholz and Sicinski posit that "facially attractive" individuals tend to earn more over their lifetimes compared to those deemed less attractive (even when controlling for experience, education, and occupation). The authors identify levels of self-confidence and labor market discrimination as likely mechanisms motivating the "beauty premium". Testing the validity of labor market discrimination as a potential driver, Stinebrickner et al. (2019) assert that the assumed connection between physical attractiveness and earnings diminishes significantly when accounting for heterogeneity in job tasks. Their respective analysis of potential drivers uncovers employee self-selection based on job preferences as a probable contributor rather than employer discrimination.

More broadly than the beauty premium alone, physical attractiveness is commonly linked to overall success and productivity. Turning attention to advertising firms, Pfann et al. (2000) find evidence of substantial preferential treatment to more attractive individuals in labor markets. Their study finds that better-looking executives receive higher wages, faster career advancement, more job security, and easier hiring processes. When evaluating factors that contribute to academic achievement such as socioeconomic background and natural intelligence, the authors assert that physical attractiveness has an undoubted influence on personal salary as well as firm revenue. Although confident in their findings, Pfann et al. (2000) acknowledge there are several contributors toward the interrelationship between attractiveness and performance[4]. A holistic accounting would include, but not be limited to, favoritism due to employee attractiveness. Rather than employer discrimination, some literature offers boosted perceptions of competence, self-confidence, and social acumen as driving mechanisms. Naturally, physically attractive workers tend to be more confident; in turn, higher levels of confidence lead to higher wages. At a given level of confidence, the competence of a worker is systematically overestimated by employers (Mobius and Rosenblat, 2006). Pillutla and Murnighan (1996) conduct a comprehensive, meta-analysis of the previous literature assessing the assumed positive relationship between attractiveness and favorable market outcomes. The exhaustive investigation concludes by reinforcing the consistent connection between attractiveness and labor market outcomes; however, the authors emphasize the need for ethical considerations concerning discrimination, special treatment, and equal opportunities. Not isolated to job market outcomes, psychological research suggests benefits related to perceived beauty and physical attractiveness extend to all areas of success (see Hamermesh and Parker, 2005 for evidence of a positive correlation between attractiveness and assumed teaching effectiveness). Challenging the commonplace maxims that diminish the importance of attractiveness in life, Langlois et al. (2000) conduct a comprehensive meta-analysis revealing that attractive individuals are judged more positively, and treated more positively, than those deemed unattractive. Moreover, attractive individuals tend to exhibit more positive traits and behaviors. These findings are demonstrated to be upheld across cultures and age groups. In fact, psychological research on bias and human nature suggest that individuals form their initial impression on such traits within 100 milliseconds of a social interaction. Further, people tend to form opinions surrounding distinct traits such as trustworthiness, friendliness, dominance, competence and attractiveness within a fraction of a second (Willis and Todorov, 2006). In similar fashion, Kramer and Ward (2010) find that physical health, as well as four of the big five personality traits (extroversion, agreeableness, openness, conscientiousness, and neuroticism) can be perceived with some accuracy from facial attributes alone. In effect, three of the five can be accurately understood from the facial features of just one side of the face. Given the far reaching implications of beauty and the natural perception of personality, evidence of bias related to physical attractiveness pervades numerous areas of research.

[4] the authors offer consumer discrimination as one such driver for the relationship between attractiveness and performance

3.2.2 Sport & Tennis

Adding to the growing body of literature on the economics implications of physical attractiveness, previous research investigates the "beauty premium" in the world of sports. In agreement with much of the broader literature, there exists an observable positive relationship between physical attractiveness and earnings in women's professional golf (Ann and Lee, 2014). Measured by tournament earnings, Ann and Lee find that the "more attractive" female golfers tend to earn higher tournament winnings and achieve better rankings relative to their less attractive counterparts. For instance, a change in "beauty" ratings from the 50th percentile to the 70th percentile corresponds to an average golf score decrease of 0.231[5], resulting in a $55,122 increase in tournaments earnings. The potential mechanism the authors emphasize is that more attractive golfers put more effort into competitions due to a higher return to human capital. Turning to the German Bundisliga, professional male footballers seem to experience a beauty premium akin to the effects of attractiveness in women's professional golf. Analyzing 438 players on both bodily and facial attractiveness (using Body Mass Index, and the truth of consensus method respectively), Rosar et al. (2017) reveal economically and statistically significant effects on a player's market value despite controlling for player performance. Explicitly, the authors find that a one point increase of Body Mass Index corresponds to a market value increase of 220,000 euros. Furthermore, each increase in the score for facial attractiveness by one point is consistent with a market value increase of 150,000 euros. Moreover, extensive literature positions sport as a subject of study widely impacted by this attractiveness and success, measured by player performance. Corroborating the findings of more general research depicting physical attractiveness and success, evidence suggests a similar relationship in the Central League Pennant (one of the major Japanese Professional Baseball Leagues). Even when measuring objective performance metrics such as number of runs batted in (RBI) in a season, physical attractiveness, represented by facial width-to-height ratio (FWHR), appears to play a sizable role (Tsujimura and Banissy, 2013). Specifically, the authors highlight a statistically significant correlation between home run percentage and FWHR. Not limited to baseball, previous studies link performance success at the Olympic games to high physical attractiveness (Li et al., 2023). Among female athletes, this "good performance-high attractiveness" effect was observed in the 3-meter diving event and floor exercises that spotlight precise control and aesthetics. Similarly, although not as pronounced as as female athletes, male athletes exhibited a performance-attractiveness effect in the 100-meter sprint (and most notably floor events). In their nuanced approach, the authors reveal

[5] somewhat counter intuitive, lower scores are desired in golf. The aim is to hit the golf ball into the hole in as few strokes as possible

that variations physical fitness requirements for specific skills - and thus specific games - correspond to distinct patterns of attractiveness.

Even more granular than sport in general, there is a volume of research that explores the beauty-success correlation specifically in tennis (and further, female tennis). Historically, demand for the sport has been found to be highly correlated with the physical attractiveness of female tennis players (see Dietl et al., 2019). Using a sample of the top 100 ranking WTA players at a single time, Kiefer and Scharfenkamp (2012) find that physical attractiveness increases social media popularity significantly. The boost in social media activity is also accompanied by increased interest in the player's WTA profile page. In line with research on the beauty and earnings, Bakkenbull and Kiefer (2015) cite evidence of a significant beauty premium amongst professional female tournament play. More specifically, the authors use student ratings of 100 photos of female tennis players on an 8 point scale. They find that a 1 point increase in attractiveness is correlated with prize money winnings by 30% and career winnings by 20% (also see Bakkenbüll, 2017).

3.2.3 Measuring Attractiveness

Previous literature cites FWHR, facial symmetry, and perceived attractiveness acquired through peer surveys as the predominant methodological measures for attractiveness, especially in sport. For instance, Carre and McCormick (2008) use FWHR and penalty box time to measure aggressiveness and dominance. Concretely, their study finds that Canadian Hockey League players with larger FWHR were more aggressive than those with smaller FWHR. Alternatively, Berri et al. (2011) use facial symmetry as their primary measure for attractiveness. Focusing on National Football League (NFL) quarterbacks, the authors determine that more attractive quarterbacks (players with more symmetrical faces) are paid greater salaries on average. Further, the beauty premium is consistent even when controlling for player performance. Building on top of these methods, the prevalence of computer vision in recent years has advanced more sophisticated methods utilizing machine learning methods (Eisenthal et al., 2006 Kagian et al., 2006). In addition to simple linear regression, such methods include K-nearest neighbors (KNN) clustering algorithms, Support Vector Machines (SVM) and decision tree based models. Naturally, the subsequent evolution of artificial intelligence and deep learning methods has been evident in beauty and attractiveness prediction research (Xiao et al., 2021 Anderson et al., 2018). Taking inspiration from psychological research and human anatomy, Convolutional Neural Networks (CNN) have proved most prominent when tasked with delivering a robust, all-encompassing attractiveness prediction (Xu et al., 2017). Most recently, Guo et al. (2023) make use of machine learning, deep learning, and computer vision to isolate facial characteristics within collegiate sports. Pointing their CNN algorithm toward head coaches in American college sports, the authors estimate attractiveness through observable facial characteristics. Their analysis finds a salary discount for attractiveness and a premium on aggressiveness. In method, my chapter largely resembles the latest research conducted by Guo et al. Both studies leverage a CNN algorithm to essentially grade subjects on beauty. In detail, neural networks are used to translate digitized photographs into a continuous attractiveness scores (most previous research uses discrete measures of physical attractiveness). However, my research method contrasts the work of Guo et al. in two major ways. First, facial landmarks and characteristics are treated as implicit features in my predictive modeling exercise. While Guo et al. use computer vision to explicitly provide distinct characteristics to their model, the facial beauty model is expected to learn that these features are distinctive drivers in predicting an all-encompassing attractiveness score. Second, my approach adopts the deep learning framework developed by Parkhi et al., 2015, of the Visual Geometry Group (VGG) at the University of Oxford. Although both studies feature trained algorithms that inherit the core CNN architecture, the model pre-trained by the VGG group, having seen 2.6 million faces, is heavily optimized toward facial recognition tasks. Accordingly, it is reasonable to expect more accurate results from the latter model.

3.2.4 Summary

In summary, this section 1) provides research on the effect of attractiveness on market outcomes 2) outlines more general evidence on attractiveness and overall success 3) delves into research related to the beauty premium, performance, and market outcomes in the world of sports 4) renders a historical reference of the methods used to measure beauty by previous researchers. While much research exists exploring the relationship between beauty and success in more traditional contexts, my findings will reflect uniquely practical evidence of market outcomes due to bias and physical attractiveness in sports. Unlike other areas of research, sports offer an especially contained sandbox favorable to applied exercises since they feature realised results according to rigid success criteria (win or lose). Furthermore, tennis, as an individual sport, removes the noise of team effects and allows for the exploration of relative dynamics between match opponents. Finally, as computational tools have progressed in recent years, cutting edge methods are more readily available than previous research would suggest. A key contribution of this chapter is to employ such methods toward the connection between beauty and market outcomes.

3.3 Dataset

3.3.1 Betting Odds

Betting odds were collected from oddsportal.com for both the winner and loser of every match in the dataset. For analysis, results are typically presented using the average odds across the available forty to sixty bookmakers for each match. Alternatively, best available, and most advantageous, odds are explored. Finally, odds from Bet365 (the world's largest online sportsbook by revenue, customers, and visitors) are used to illustrate potential realistic betting strategy outcomes. Rather than evaluating against unattainable betting strategies, I attempt to exploit market inefficiencies using Bet365 as my primary bookmaker. Its sizable volume and amount of available offerings on a variety of sports make it an ideal sportsbook to test a realistic betting strategy.

3.3.2 Women's Tennis Association

The Women's Tennis Association (WTA) hosts a global audience of over 700 million, making it the premier professional sports organization of its kind. The association consists of more than 1650 players from approximately 85 countries, all jockeying for WTA rankings points and titles for over 50 events, including four Grand Slams[6]. Match results from 20,206 main draws of all WTA tournaments were sourced from tennis-data.co.uk. This primary dataset includes match results, court and surface information, player names, pre-match rankings, match locations, and tournament information. Additionally, the python packages Beautiful Soup - a Python library for pulling data out of HTML and XML files - and Selenium – an open source, automated web testing and browsing tool – were used to collect WTA profile photos, physical attributes such as dominant hand and height, and other demographics from wtatennis.com, the official website of the Women's Tennis Association[7].

Before WTA player images and features were collected, I compiled an exhaustive list of unique player names contained in the tennis-data.co.uk match results data. With the compiled list, I manually combed through the list accounting for edge cases like nicknames, acronyms, and maiden names[8]; creating a many-to-one relationship that ensures uniqueness in player name. I then matched these names with available WTA profile pages using an automated process in which names are digitally inputed into wtatennis.com, matched to the top search result, and recorded (in the form of a link

[6] Typically, tennis players are paid proportional to their performance in a tournament. The prize money allocations are predetermined and compensated according to their final placing i.e. champion, finalist, round lost, etc.

[7] See example, Venus William's WTA profile

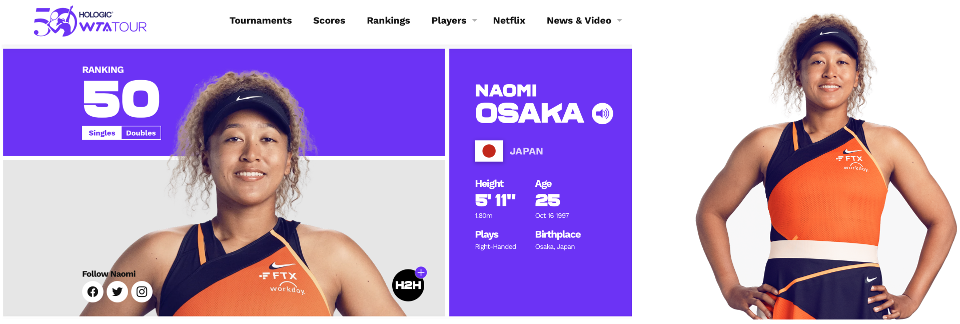
[8] Such name variations include Coco (Cori) Gauff, Kristie (Hyerim) Ahn, and Paula Badosa (Gibert)

to the player profile). Finally, I iterated through the profile links, gathered relevant information, and extracted profile photos, regardless of resolution.

FIGURE 3.1: WTA Profile Photo Extraction

(a) WTA profile

(b) Extracted WTA profile photo



Notes: the heavy lifting for photo extraction was performed with selenium pointed toward wtatennis.com. After ensuring the page is fully loaded, Beautiful Soup is used to isolate key html tags to either scrape appropriate information or locate the image redirect link. The redirect is then pinged in order to download the profile photo, if available.

Figure 3.1 reflects the image extraction process for players via their WTA profiles. Using the WTA profile page belonging to Naomi Osaka, one of the most popular players in the sport, panel (a) shows the selenium rendered profile before extraction. Unlike some WTA profile pages belonging to players of lower popularity and overall profile, Osaka's profile comes equipped with a well-defined, high resolution image. Panel (b) shows the photo post extraction. Fortunately, these photos are extracted void of any background image space, making results more precise. Of the 700 unique players found in the primary dataset, profile photos were sourced for 452 players; of which, 138 were high resolution. Focusing only on matches with participants with high resolution photos, the filtered dataset is comprised of 6,436 matches ranging from 1 January 2011 until 16 February 2020.

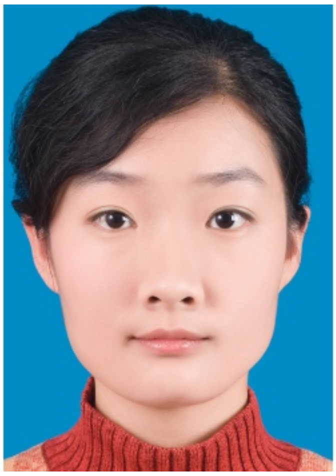
3.3.3 Image Processing

As an exercise in image processing, the training data used to train and validate the deep learning "beauty" model is comprised of images rather than a more traditional set of features. The specific dataset used in this chapter was compiled and released by Liang et al., 2018 at the Human Computer Intelligent Interaction Lab of South China University of Technology (SCUT)[9]. The dataset consists of 5500 frontal face images in total. Compiled by images from the internet, the set includes both male and female

[9] See data release SCUT-FBP5500-Database-Release.

images - although only female images for model training - differing in race and age. Figure 3.2 shows a typical training example image from the SCUT dataset. Although not as high resolution as the WTA profile photos, the SCUT team chose clear, frontal face images without any obstructions or noisy backgrounds. The SCUT data release additionally includes facial landmark coordinates, suggested training/test data splits, and subjective beauty scores ranging from 1 to 5. In order to derive the beauty scores, the SCUT team assembled an ensemble of 60 volunteers tasked with grading each image on a scale from 1 to 5 in terms of attraction. The average of these scores is included as well in the dataset.

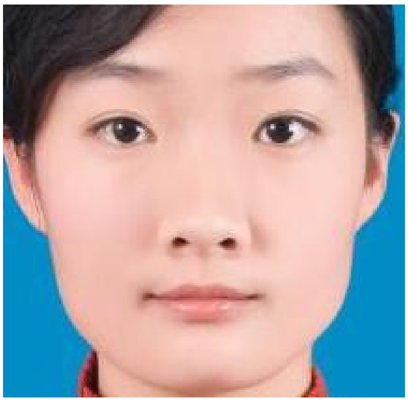
FIGURE 3.2: Training Data Example



Notes: A typical example of the training data included in the SCUT data release. The image is a frontal facing with a clear line of sight to the camera. It features clear resolution void of any obvious obstructions.

Prior to model training a systematic series of image pre processing steps had to take place in order to prepare images as neural net inputs. These steps consist of image cropping, vectorization, resizing, thresholding, and normalization. For more accurate results, I utilize the dlib python package to crop out as much negative space as realistically possible. To do so I locate the 4 corner landmarks on each photo, surround the landmarks with a certain amount of padding, and remove the rest of the background image. The Python Imaging Library (PIL), an open-source python package that supports image processing capabilites across many different file extensions, is used to transform the PNG images into arrays and resize them into the required size - 224 x 224. Once the background is vectorized, resized, and limited to as little pixel space as possible, I then scale the image pixel values using a process called normalization. A common intermediary image processing step, normalization, refers to the practice of pixel value scaling that ensures the best possible comparisons across different methods of data acquisition and texture instances. Figure 3.3 shows the same image depicted in figure 3.2 after cropping, resizing and normalization. As compared to the original image in figure 3.2, the pre processed image in figure 3.3 may seem softer in appearance due to the scaling of extreme pixel values. Finally, a process called thresholding is used to focus all attention on the faces of each photo. Thresholding is a popular image processing practice used to segment images into foreground and background based on each pixels value in comparison to a predetermined threshold. The result of all of these pre processing steps is a 2D array suitable as inputs for machine learning algorithms.

FIGURE 3.3: Training Data Example After Image Processing



Notes: the same image depicted in Figure 3.2 after preprocessing (namely cropping and resizing) methods are applied.

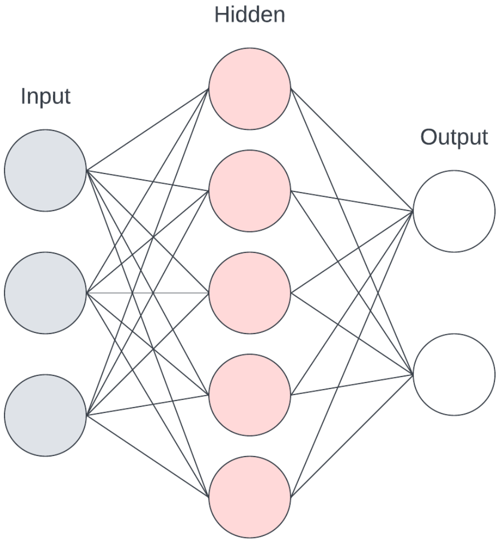
3.3.4 The Deep Learning "Beauty" Model

In order to measure the attractiveness or "beauty" of match participants, I employ and build upon the VGG face model framework[10] developed by the VGG at the University of Oxford. A testament to deep learning artificial neural networks (ANN), the VGG face model is a state-of-the-art, pre-trained convolutional neural network (CNN) popularly used for image processing and facial recognition in particular. In general, the deep learning moniker is derived from the multi-layer architecture associated with ANNs.

Figure 3.4 shows a simple 3 layer artificial neural network consisting of an input layer, a hidden layer, and an output layer. Each layer, comprised of artificial neurons, exchange information with each other through connections holding training weights that are optimized through the training process. Robust to handling machine learning tasks such as speech recognition, machine translation, and image recognition, the

[10] See https://www.robots.ox.ac.uk/ vgg/publications/2015/Parkhi15/poster.pdf

FIGURE 3.4: Artificial Neural Network Architecture

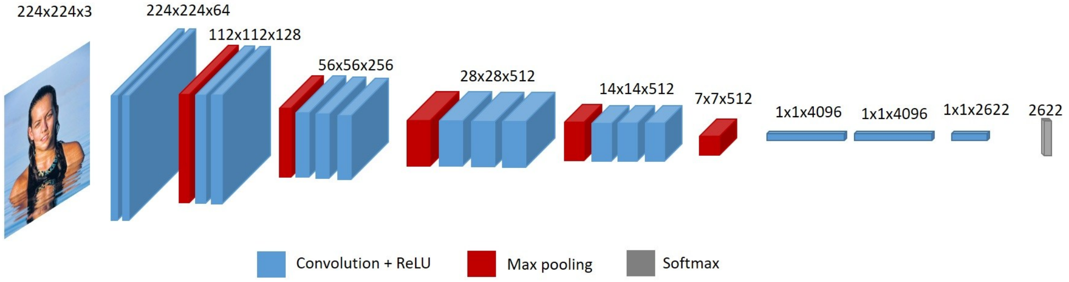


Notes: a simple 3 layer artificial neural network. The first layer (grey) is following by the hidden layer (pink), then the output layer (white).

applications for deep learning ANNs are seemingly endless. Inspired conceptually by biological neuron networks in the human brain, these models are no stranger to sport. In their endeavor to classify actions from videos of football matches, Baccouche et al. (2010) employ the use of a Long Short-Term Memory Neural Network (a high performing variation of a recurrent neural network). Tasked with a similar undertaking, Jiang et al. (2016) use a state of the art, combined Convolutional Neural Network and recurrent neural network approach to deliver superior action classification accuracy. Finally, Rahmad et al. (2019) use a "faster region convolutional neural network" to classify player identities, movement, and position during badminton matches. A powerful subset to ANNs, CNNs are named after the convolution process - an essential mathematical linear operation between matrices. Despite their “black box” nature, convolutional neural nets are often used in the image processing space due to their proficiency in discovering non-linear relationships. The CNN architecture consists of four types of layers; this includes convolutional layers, non-linear layers, pooling layers, and fully connected layers. Used for feature extraction from images, convolutional layers contain a set of features, or kernels filters, with trainable parameters. As the neural network is traversed, spatial redundancy is reduced through repeated convolution (the dot product of the kernel filters and the input pixels). The output of a convolutional layer is a compressed feature representation of the input image. Non-linear layers refer to the non linear triggers or activation functions, i.e.

sigmoid, tanh, and rectified linear units (ReLU), that reside in the layer. Also known as downsampling, pooling layers are used to incrementally reduce the dimensionality of the data in order to reduce the computational overhead of the model. Finally, the features are extracted, downsampled, the fully connected layers transform the output of previous layers into a single dimensional vector to be used in the next layer (see O'Shea and Nash, 2015 for more on CNN architecture). Figure 3.5 shows the base architecture of the VGG face model; a "very deep" series of convolutional, ReLU, and max pooling layers, the architecture is ended by a Softmax activation function (a function that takes the output vector of the previous layers and calculates a vector of probabilities.

FIGURE 3.5: VGG Face Model Structure



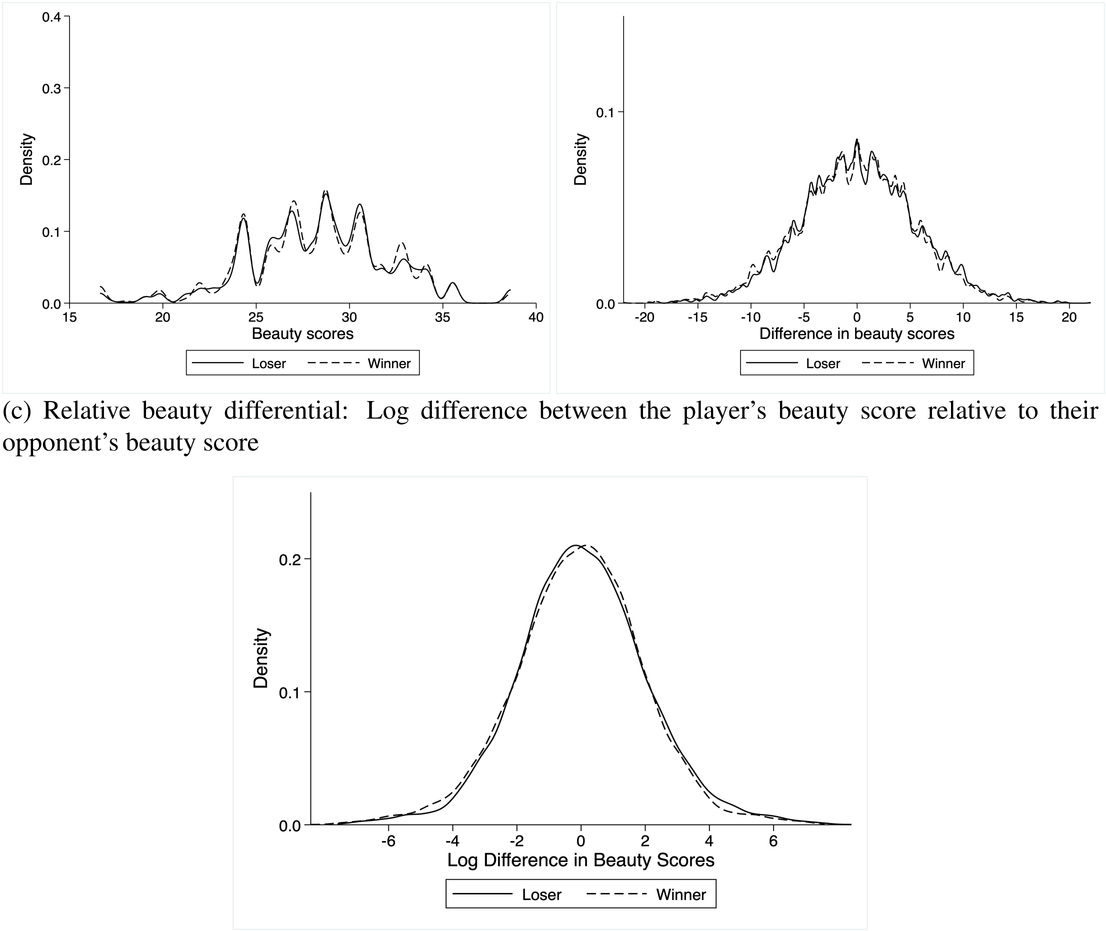
Notes: a visual representation of the VGG Face Model Architecture. From original image to the final activation function, you can see "assembly line" like the medley of convolutional, ReLU, and pooling layers required. Source: sefiks.com/2018/08/06/deep-face-recognition-with-keras/

Using the pre-trained weights provided by the Oxford group, the VGG model is trained on images that were labeled with beauty scores provided by the SCUT survey. Essentially, the highly optimized, out of the box facial recognition model is repurposed toward the task of scoring beauty. This represents the biggest methodological difference between this chapter and those similar on the topic. While prior works have deployed deep learning methods toward measuring beauty, no such research has put forward a cutting edge model developed using a vast amount of resource. In order to ensure consistency across images (both in training and test photos), I isolate participant faces using the dlib python package, a python wrapper for the extensive C++ toolkit equipped with built in image processing and facial recognition functions. To further ensure data integrity, low resolution profile photos are ultimately excluded from the dataset. Evaluating low resolution images in tandem with high resolution images, given their heterogeneity in dimension, could potentially introduce noise that would dilute "beauty" score predictions. The remaining image set includes mostly highly ranked WTA players who are spotlighted with consistent, professionally executed, high resolution head shots. Conveniently, by limiting the image processing modelling to these popular players, I reduce the variance in resolution, image angle, and lighting (all common difficulties with facial recognition). In line with current literature on image processing using convolutional neural networks, although computationally expensive, focusing on high resolution photographs leads to increased performance. After image processing and model training, the most accurate model, identified during cross validation, is chosen to provide the WTA "beauty" scores used in econometric model evaluation.

FIGURE 3.6: Beauty scores of tennis players before WTA matches in 2015-2020

(a) Raw beauty scores

(b) Raw difference in beauty scores



(c) Relative beauty differential: Log difference between the player's beauty score relative to their opponent's beauty score

Notes: beauty scores predicted by the deep learning beauty model are used to generate these distributions. The densities are estimated with a Gaussian kernel and bandwidth of 0.2

After identifying the winning "beauty" model, trained on the previously mentioned 5500 facial images, the model is then pointed toward the accumulated WTA profile photos to provide a "beauty" score for each unique corresponding player. For each match-player observation, these individual scores are used to formulate the difference in "beauty" score for a player relative to their opponent. To do so, I subtract the log beauty score of a player by the log beauty score of their opponent. Concretely, for player i appearing in match j, I calculate:

$$\text{Relative Beauty Differential}_{ij} = \ln(w_{ij}) - \ln(w_{-ij}) \;, \tag{3.1}$$

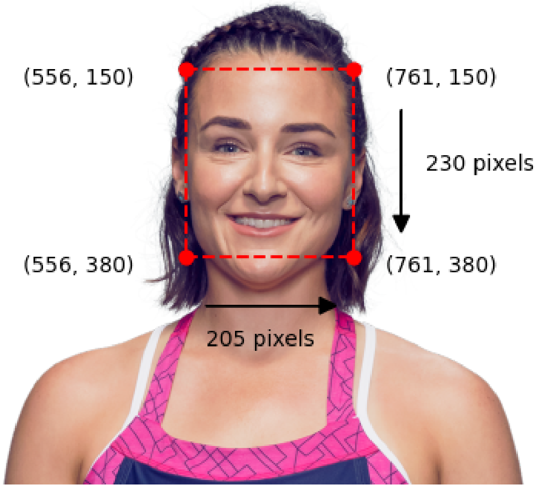
where w_{ij} is the beauty score for the player and *–i* denotes the player's opponent in the match. Consequently, the Relative Beauty Differential shows the proportional difference in beauty between player i and their opponent *–i* in a given match j. The greater the relative beauty differential, the larger the proportional different in beauty between match participants.

3.3.5 Golden Ratio

"The Most Beautiful Number In The Universe", 1.618, is an aesthetics standard related to the rule of thirds. A mathematical expression of beauty, the golden ratio pervades many disciplines in art, including music, architecture, and even poetry. Said to be the ideal proportion to the human body, the ratio inherently coincides with attractiveness in design as well as nature (Thapa and Thapa, 2018). Used with intention in the composition of a photo, the principle effectively draws the eye to the important elements of the image. In practice, the golden ratio pertaining to facial recognition is simply the height to width ratio of the facial landmarks. In theory, the closer the facial height to width ratio is to 1.618, the more aesthetically pleasing. In this chapter, in line with relevant research in facial recognition and aesthetics (see Rossetti et al., 2013 and Ceinos et al., 2017), the facial width to height ratio, being a measure of aesthetics, is used as a "beauty" robustness check. Ideally, the deep learning "beauty" model should encompass any relevant information provided by the measure. To generate the face width to height ratio, I again leverage the dlib python package to acquire the necessary coordinates in pixel space of the facial landmarks. Figure 3.7 is a demonstration of facial width-to-height acquisition process using Abigail Spear's WTA profile photo as an example. The relative four corners, or coordinates, of the landmarks are indicated by the red dots located around the face. The facial height is given by the distance (measured in pixels) between the upper right corner and the lower right corner. The facial width is given by the distance between the lower left corner and the lower right corner of the face. In this case, the width is equal to 205 pixels, while the height is 230 pixels. The formulated facial width-to-height ration is 205/230 or approximately 0.89.

Similarly to the "beauty" score differential, the face width to height ratio differential is a relative measure. First, to calculate a player's facial width to height ratio, I divide the width of a player's face by the height of their face. Second, I subtract

FIGURE 3.7: Facial Width to Height Ratio



Notes: the image depicts the WTA profle photo for Abigail Spears. The four facial landmark corners are indicated by the red dots. The distance between the upper right corner and the lower right corner is taken as the height of the face, while the distance between the lower left corner and the lower right corner is taken as the width of the face. The FWHR formula is facial width/facial height.

the player's facial width to height ratio from the equivalent value for their opponent. Specifically, for player i appearing in match j, I calculate:

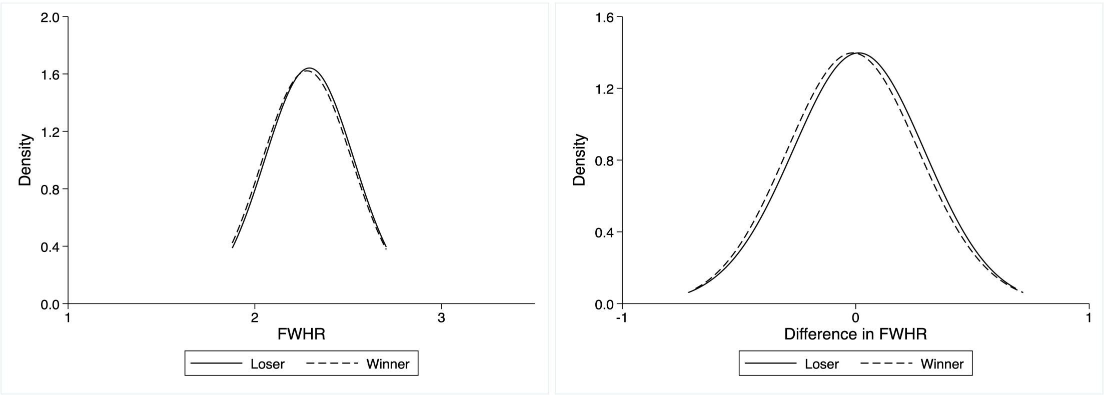
$$\text{FWHR differential}_{ij} = z_{ij}/v_{ij} - z_{-ij}/v_{-ij} \ , \tag{3.2}$$

where z_{ij} is the facial width of player i in match j. v_{ij} similarly denotes the facial height player i in match j. Finally, z_{-ij} represents the facial width of opponent $-i$ and v_{-ij} is the opponent facial height. Accordingly, the facial width to height differential measures the raw difference in FWHR between match opponents. The larger the difference, the larger the relative width of the player's face relative to their opponent.

3.3.6 Skin Tone

After an extensive neural net training process is complete, each WTA player - for which high resolution photos are available - is assigned a predicted "beauty" score. Upon my initial inspection of the results from the deep learning "beauty" model, I noticed an discernable correlation between racial, or ethnic, background and the predicted "beauty" scores. Although the training dataset (SCUT data release) and the WTA profile photos used to predict "beauty" include images of people from a wide array of backgrounds, the deep learning model seemed to systematically favor players with fairer skin tones and penalize players with darker skin tones.

FIGURE 3.8: Facial Width-to-Height Ratios of tennis players before WTA matches in 2015-2020



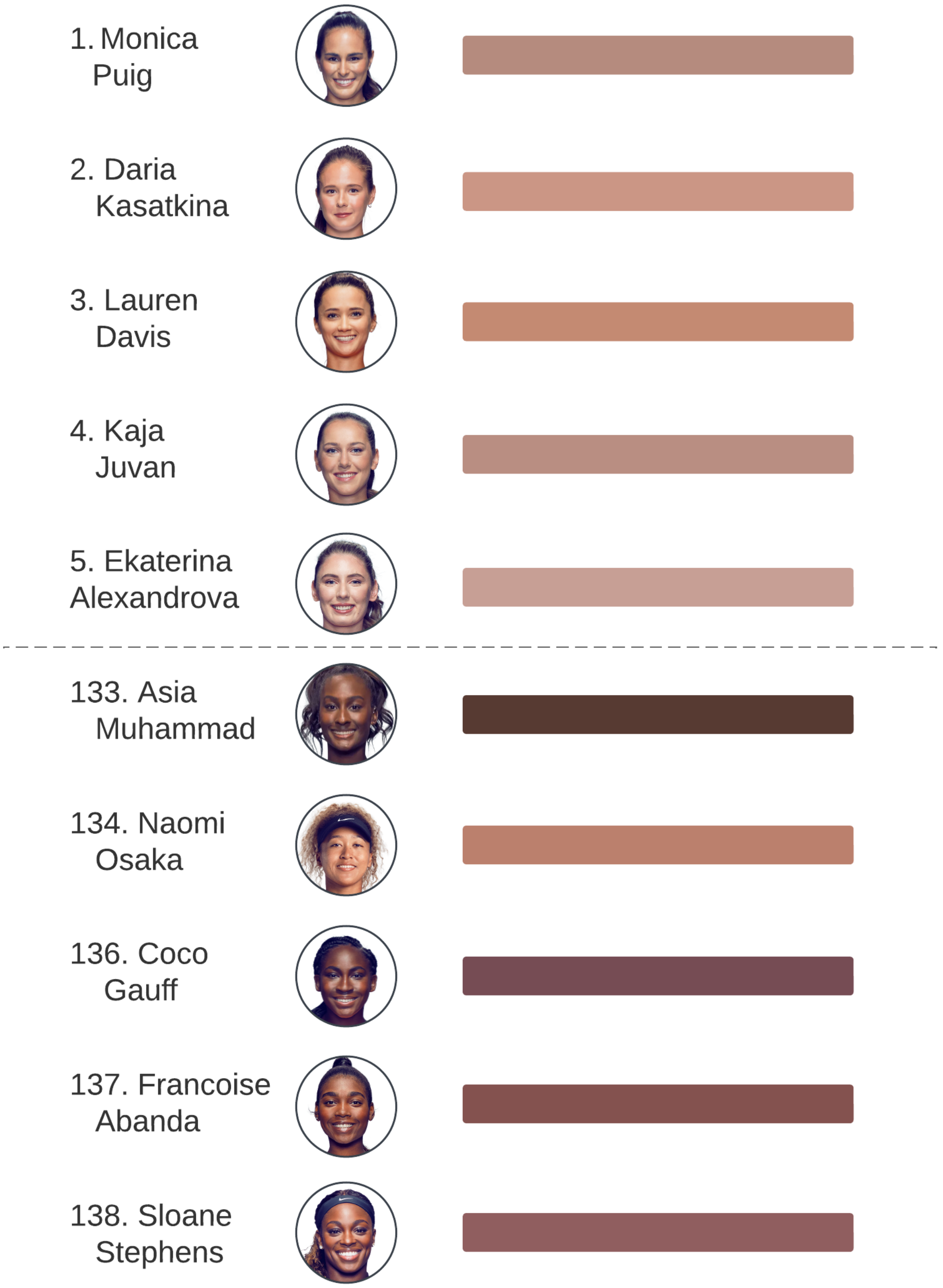
(a) Facial Width-to-Height Ratios

(b) Difference in Facial Width-to-Height Ratios

Notes: formulated facial width-to-height ratios are used to generate these distributions. The densities are estimated with a Gaussian kernel and bandwidth of 0.2

With an apparent "Eurocentric" standard of beauty, the deep learning "beauty" model placed mostly European players with blue or green eyes in the top ten scoring spots. Conversely, players with Asian and African descent tended to occupy the spaces with the lowest "beauty" scores. Used to exemplify this anecdotal observation, figure 3.9 shows the top five highest "beauty" scoring tennis players above the dashed line; conversely five of the ten lowest "beauty" scoring tennis players are highlighted below the dashed line. Of the top five, two players are Russian, one Caucasian American, one Slovenian, and one Puerto Rican. Moreover, the five players with lowest assigned "beauty" scores are all of African descent (Asia Muhammad, Coco Gauff, and Sloane Stephens are all African American. Naomi Osaka is a Japanese national with both Japanese and Haitan ancestry. Franciose Abanda is a Canadian national with Cameroonian parents). From left to right, figure 3.9 shows each player's respective "beauty" score ranking, facial image thumbnail, and dominant pixel color for each player (extraction process detailed below). Directly to the right of each player's thumbnail, the predominant color extracted from the image clearly illustrates the contrast in skin tone between the players with high "beauty" scores and the players with low "beauty" scores. Such observable bias in machine learning is not an uncommon phenomenon. In her analysis of the unsettling implications of algorithmic bias, cybersecurity expert Megan Garcia urges the need for ardent awareness as distorted data introduces unconscious and institutional bias in artificial intelligence models (Garcia, 2016). In their commentary on the state of AI and computer science, Zou and Schiebinger poignantly call out, "Google Translate converts news articles written in Spanish into English, phrases referring to women often become 'he said' or 'he wrote'. Software designed to warn people using Nikon cameras when the

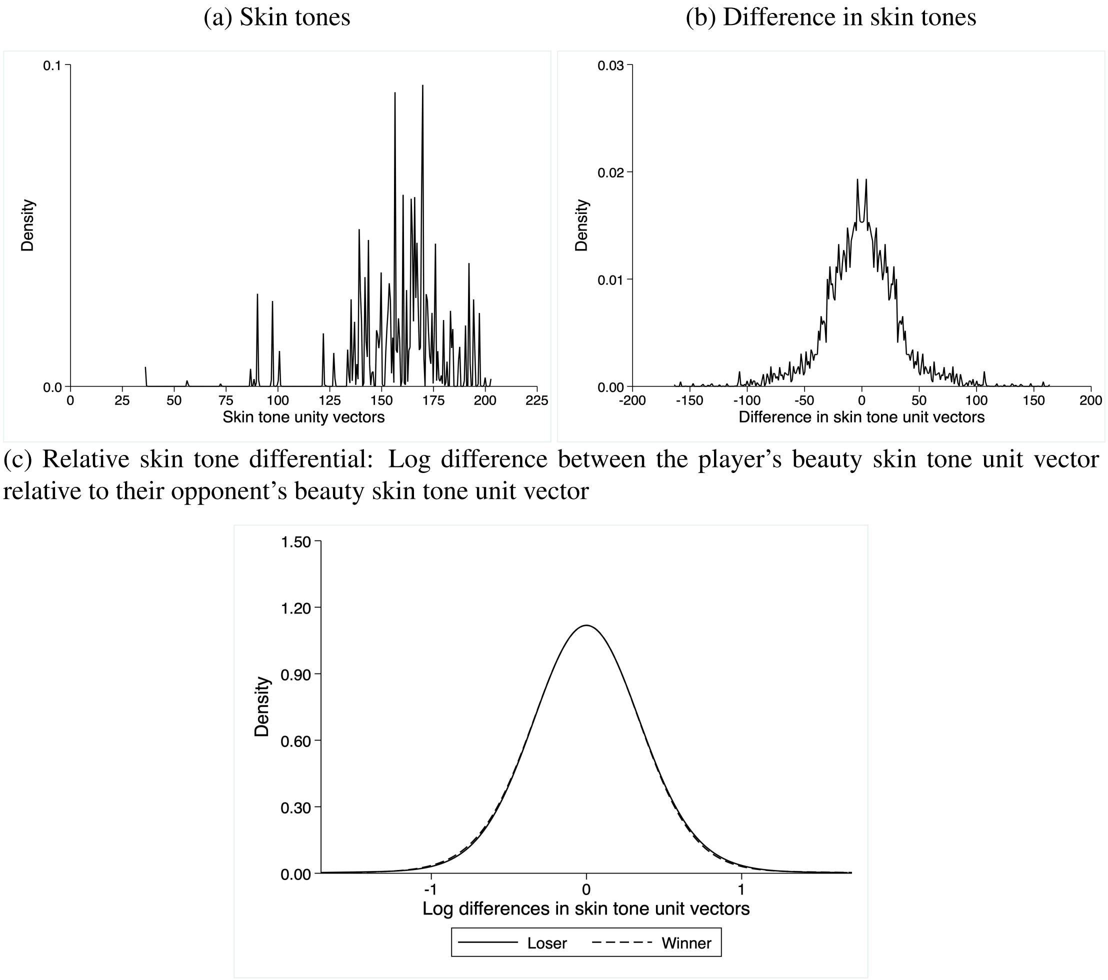
FIGURE 3.9: Observed "Beauty" Results



Notes: a visual representation of the top five and (near) bottom five WTA players by predicted beauty score. The top scores and bottom scores are divided by a dotted line. The 135th ranked beauty score is from European decent and is intentionally omitted from the figure.

person they are photographing seems to be blinking tends to interpret Asians as always blinking. Word embedding, a popular algorithm used to process and analyse large amounts of natural-language data, characterizes European American names as pleasant and African American ones as unpleasant" (Zou and Schiebinger, 2018). Whether it's sexism, racism, or some other source of bias, the systematic introduction of bias is a high profile known issue for machine learning and deep learning tasks.

FIGURE 3.10: Skin tones of tennis players before WTA matches in 2015-2020


(a) Skin tones

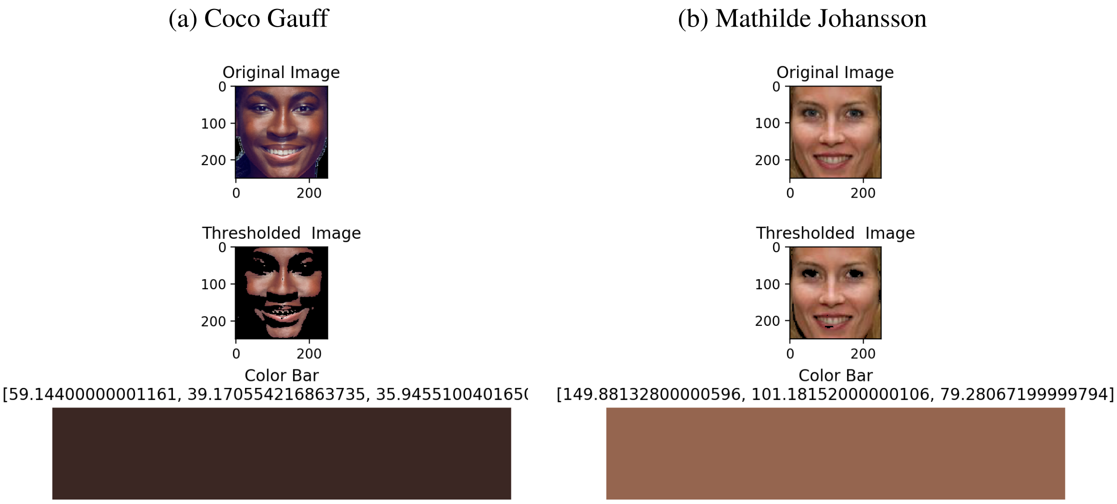
(b) Difference in skin tones

(c) Relative skin tone differential: Log difference between the player's beauty skin tone unit vector relative to their opponent's beauty skin tone unit vector

Notes: derived skin tone unit vectors are used to generate these distributions. The densities are estimated with a Gaussian kernel and bandwidth of 0.2

In an effort to substantiate the discernible relationship between skin tone and predicted "beauty", as well as add a new relevant dimension to the chapter, I transform skin tone into a usable continuous score. As a prerequisite, I first needed to isolate the predominant skin tone for each player's image. Although it is common practice to use a clustered approach, such as a K-means algorithm (see Kamel and Woo-Mora, 2023), to find the predominant color(s) vector(s) in an image; alternatively, I chose to crop, threshold and calculate the average 3-D color vector for each player. More specifically, I first remove all negative space in the WTA profile images using opencv-python, a package developed by Open Source Computer Vision popularly used in deep learning, machine learning, and image processing. The package is used to crop the image to exclude background imagery, hair, and any accessories included in the photo.

FIGURE 3.11: Skin Tone Color Vector Process


(a) Coco Gauff

(b) Mathilde Johansson

Notes: image (a) depicts the WTA profle photo for Coco Gauff, while image (b) is the the WTA profile photo of Mathilde Johansson. Each panel shows 1) the extracted image 2) the same image after thresholding 3) the average color of the resulting image.

Figure 3.11 illustrates the skin tone extraction process starting with the facial detection and cropping step for each player - Coco Gauff (left) and Mathilde Johansson (right). Below each cropped image - labeled as "Original Image" - is each respective player's thresholded image. As previously described, thresholding is used on each image to ensure any uncropped background is ignored. With the negative space removed and the image thresholded, I then calculate the mean RGB (red, green, blue) color vector of the remaining pixel space. In general, each pixel in an image has a corresponding 3-D color vector following the [R, G, B] convention. Following this triplet convention, the values (255, 0, 0) represent red, (0, 255, 0) green, and (0, 0, 255) blue. Each unique color within the color spectrum is represented by some combination of values within an RGB color vector. The mean vector of each of the remaining pixel color vectors is used to represent the predominant skin tone color. Figure 3.11 shows the precise mean color vectors or "Color Bars" for each player. Below each 3-D vector is the corresponding, extracted color. Finally, once the mean skin tone color vector is found, I normalize the vector into a "skin tone unit vector" with lighter (darker) colors at the higher (lower) end of the spectrum. Commonly used in real time image processing applications like video streaming and graphics rendering, unit vectors are leverage as space efficient vector representations. For my purposes, the unit vector is used as a memory footprint that encapsulates the skin tone for each player.

TABLE 3.1: Descriptive Statistics for WTA player attributes

	N	mean	sd	min	max
Age	8,325	27.98	4.097	18	40
Rank	12,878	0.0558	0.109	0	1
Beauty	12,878	28.27	3.862	16.63	38.64
FWHR	11,920	2.289	0.145	1.879	2.704
Skin Tone	12,820	158.6	22.33	36.10	202.7

Notes: the descriptive statistics are taken at the match level. The player attributes include raw ages and ranks, predicted beauty scores, facial width-to-height ratios, and skin tone unit vectors.

Table 3.2 shows the Pearson Correlation for the relevant WTA player attributes. In support of anecdotal evidence, we can see a highly significant, positive relationship between the predicted "beauty" scores and skin tone unit vectors, denoted as "Skin Tone." Again, with lighter tones appearing as higher skin tone unit vectors, table 3.2 indicates that lighter skin tones are significantly correlated with higher "beauty" scores. Age and facial width to height ratio are also notably negatively correlated with "beauty" scores.

TABLE 3.2: Pearson Correlation for WTA player attributes

	Age	Rank	Beauty	FWHR	Skin Tone
Age	1				
Rank	0.0867***	1			
Beauty	-0.183***	-0.0265**	1		
FWHR	-0.221***	-0.0241**	-0.312***	1	
Skin Tone	-0.184***	0.0895***	0.300***	-0.0333***	1

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

Notes: Pearson correlation is generated at the match level. The player attributes include raw ages and ranks, predicted beauty scores, facial width-to-height ratios, and skin tone unit vectors.

To create the comparative measure of skin tone per match-player observation, I subtract the log of the opponent's skin tone unit vector from the log of a player's skin tone unit vector. Precisely, for player i appearing in match j, I calculate:

$$\text{Skin Tone Differential}_{ij} = \ln(q_{ij}) - \ln(q_{-ij}) \;, \tag{3.3}$$

where q_{ij} is the skin tone unit vector for the player and *–i* denotes the player's opponent in the match. As such, the Skin Tone Differential shows the proportional difference in skin tone between player *i* and their opponent *–i* in a given match *j*. The greater the Skin Tone Differential, the higher the percentage difference between the players skin tone and their opponent's. With lighter tones having higher skin tone unit vectors, the Skin Tone Differential favors differences between player's with darker skin tones as they represent larger percentage differences.

3.4 Methodology

3.4.1 Detecting mispricing

Let $y_{ij} = 1$ if player $i = 1,2$ wins match $j = 1,\ldots,J$. Otherwise, $y_{ij} = 0$, where i differentiates between the two players in a given match and J indicates the total number of matches in a sample. Accordingly, the sample contains $2J$ player-match observations in total. p_{ij} depicts the unobserved beliefs of the bookmaker or sportsbook about the probability of $y_{ij} = 1$ happening before the start of a match, i.e., player 1 winning match j. Let o_{ij} represent decimal odds provided by bookmakers, given their beliefs, on the two possible outcomes of match j; o_{1j} are the decimal odds offered for the outcome of $y_{1j} = 1$. For example, a £1 bet on an outcome returns o_{ij} to the better if the outcome is fulfilled. Otherwise, the bookmaker gains £1. From the offered odds I can ascertain the inverse odds, $z_{ij} = 1/o_{ij}$, which represent the bookmaker's forecasted implied odds-based probabilities for each match.

Central in gambling literature, the 'overround' or 'vig' is a bookmaker's expected profit margin, or commission, on a wager. Accounting for the overround ensures, $z_{1j} + z_{2j} = 1 + \kappa_j > 1$, where κ_j is the 'overround'. In other words, the total sum of implied probabilities (inverse odds) on any match is always greater than 1. By result, $z_{1j} = p_j + \alpha\kappa_j$ and $z_{2j} = (1 - p_j) + (1 - \alpha)\kappa_j$. The Efficient Markets Hypothesis states that, in an efficient market, any relevant information toward outcomes would be priced directly into the odds. Accordingly, any errors in forecasting should be absorbed by some average level of bookmaker profit i.e. it should not be possible to reliably predict forecasting errors with publicly available information. As follows, with forecasting error denoted as $e_{ij} = y_{ij} - z_{ij}$, an efficient market requires that e_{ij} is equal to the negative value of the bookmaker sample average 'overround', $E_{ij}\left[e_{ij}\right] = -\bar{\kappa}$.

I evaluate four possible sources of mispricing and departures from the Efficient Markets Hypothesis in WTA betting markets.

(1) Favourite-longshot bias: There is a common systematic irregularity previously identified in academic literature surrounding betting markets. This large body of literature addresses the question: are betters biased toward longshots (outcomes with small changes and big wins) or favorites (outcomes with better changes but smaller winnings). In particular, previous works tend to focus on the longshot bias, whereby betters chronically underestimate the chances of the most expected outcomes (longshots) relative to the changes of least expected outcomes (favorites). This underestimation, reflected in market prices (odds) makes betting on favorites more profitable in general (see Newall and Cortis, 2021). Much research has been carried out on longshot bias within professional sports betting markets. According to several works, longshot bias can arise in equilibrium given heterogeneous preferences, expectations, and budget constraints (see He and Treich, 2017; Ottaviani and Sørensen, 2015). Additionally, within the context of professional sports, researchers have observed that high risk aversion can lead to the existence of reverse favourite-longshot bias, or favourite bias (see Woodland and Woodland, 1994 and Woodland and Woodland, 2011 for observations of reverse favourite-longshot bias within Major League Baseball and the National Hockey League respectively). Aside from the equilibrium conditions suggested by neoclassical theory, Newall and Cortis (2021) assert that sports markets with fewer potential outcomes, such as tennis or football, tend to produce a favourite bias. Conversely, longshot bias tends to appear in markets with many outcomes and match participants (i.e. horse racing or golf). However, while previous research has found evidence of a longhsot bias, and thus mispricing, in professional tennis markets, the magnitude of mispricing is not enough to imply market inefficiency through positive returns from a simple betting strategy of betting on match favourites (see Forrest and McHale, 2007; Lyócsa and Výrost, 2018).

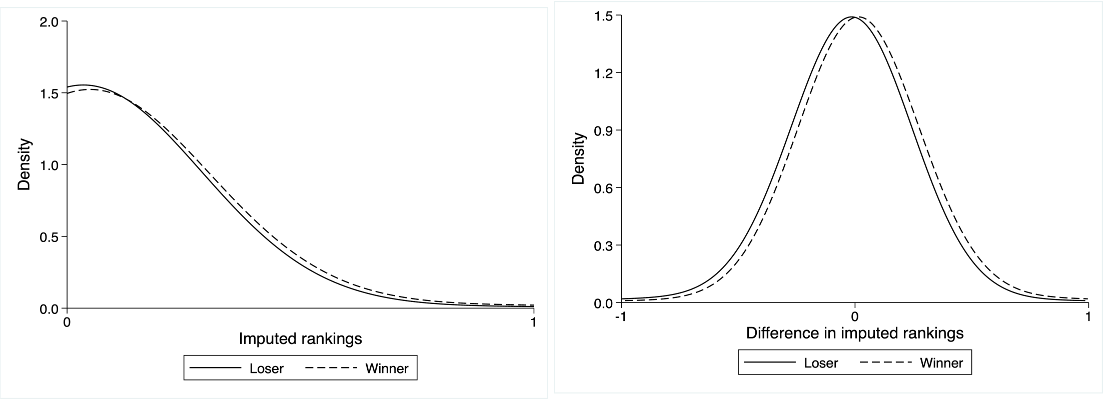
(2) Player ranking bias: In addition to the consideration of mispricing due to favorite-longshot bias, I evaluate any apparent mispricing due to WTA player rankings. Similar studies have been carried out demonstrating how readily available information on recent player performances are not fully implied by bookmaker odds. Angelini et al., 2021a use Elo ratings - a popular method to predict the probability of winning tennis matches - and weighted Elo (WElo) ratings to uncover sustained profits using a simple betting strategy (also see Kovalchik, 2020 extension of Elo ratings on margins of victory). In their analysis of Australian Open tournament matches, Kovalchik and Reid use player performance evaluations to present an in-play forecasting method that "provides a 28% reduction in the error of in-match serve predictions and improves the win prediction accuracy by four percentage points" (Kovalchik and Reid, 2019). Further, some literature suggests that bookmakers' risk aversion to matches with lower ranked players introduce longshot bias, and associated mispricing, in said matches

(Abinzano et al., 2016; Lahvička, 2014). WTA player rankings range from one (corresponding to the best player cumulatively over the past year) to unranked, for players who have not earned enough points to qualify. I consider the raw rank differences between players, formulated by:

$$\text{Rank Differential}_{ij} = 1/rank_{ij} - 1/rank_{-ij} \ , \tag{3.4}$$

where $rank_{ij} = 0$ if player i was unranked at the time of a match. A relative measure, Rank Differential shows the raw difference in rank between player i and their opponent $-i$ in match j.

FIGURE 3.12: Rankings of tennis players before WTA matches in 2015-2020



(a) Rankings day of the match

(b) Rankings differential: difference between the player's imputed ranking and their opponent's

Notes: raw player rankings are used to generate these distributions. The densities are estimated with a Gaussian kernel and bandwidth of 0.2

(3) "Beauty" bias: The link between beauty and the consumption or viewership of sport, and particularly women's tennis, has been well documented in academic literature (see Meier et al., 2016), however aesthetics have not been used to test the efficiency of tennis betting markets. Further, parallel studies exist using social media, player evaluations, and machine learning to predict football match outcomes and betting inefficiencies (e.g., Brown et al., 2018; Peeters, 2018). I adopt the Mincer and Zarnowitz (1969) forecast evaluation framework in order to detect mispricing and estimate the conditional mean effects on bookmakers' odds implied probability forecast errors (see Angelini and De Angelis, 2019, Angelini et al., 2021b, and Elaad et al., 2020, who similarly tested the weak form efficiency of European football betting through home bias and favorite-longshot bias). I estimate the following using least squares:

$$e_{ij} = \alpha + \beta_1 z_{ij} + \beta_2 \text{RankDiff}_{ij} + \beta_3 \text{BeautyDiff}_{ij} + \psi_{S(j)} + \phi_{T(j)} + \varepsilon_{ij} \,, \tag{3.5}$$

where $\{\alpha, \beta_1, \beta_2, \beta_3, \psi_{S(j)}, \phi_{T(j)}\}$ are parameters and $\{\beta_1, \beta_2, \beta_3\}$ are the coefficients denoting bias. e_{ij} indicates odds-implied bookmaker forecast error, which can be reasonably expected to be significantly negative. α represents the bookmaker profit or overround. A positive value of β_1 suggests a longshot bias. A positive value for β_2 represents a high rank bias. In regards to the measure of beauty, a positive value of β_3 indicates a low "beauty" bias in WTA markets. I consider both fixed season effects, $\psi_{S(j)}$, and fixed tournament effects, $\phi_{T(j)}$, on Equation (3.5), where $S(j)$ and $T(j)$ are indicator variables to account for any possible heterogeneity as it relates to these factors. Any unaccounted for heterogeneity is thereby absorbed by the residual, ε_{ij}. Accordingly, a suitable condition for weak form efficiency in the betting market is given by the null hypothesis: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. In the case that any of the estimates $\{\beta_1, \beta_2, \beta_3\}$ are significantly positive or negative, then the parameter's corresponding variable contains specific information that is not, at current, fully incorporated into the price. As previously defined, this is not in line with a weak form efficient market. Systematically betting on favorites (longshots), lower (higher) ranked players, or the player with the higher (lower) relative beauty score could return sustained profits.

(4) Racial bias: Although racial bias is a commonly known issue with machine learning, deep learning, and AI, its implication toward mispricing has not been as well documented. However, similar studies have been conducted related to how betting markets price, or misprice, basic demographics such as gender. Despite their growing representation and performance levels, female jockeys tend to be systematically underestimated by the UK betting market. "Mistake-based discrimination" suggests an adverse impact to the efficiency of horse racing markets in the UK over the last 20 years (Cashmore et al., 2022). Further, research in the professional tennis setting finds evidence of widespread gender bias in the sport and its media covering, leading to asymmetrical pricing (Barrutiabengoa et al., 2022). Of the sparse literature on mispricing due to racial bias, Larsen et al. (2008) document the existence of an own-race bias among NBA officials and its implications toward betting markets. Much the same of the approach of this chapter, the authors exploit the irregularity with a simple betting strategy. Alternatively, research finds no such evidence of racial bias or discrimination of minorities within MLB betting markets (Paul et al., 2018).

To the best of my knowledge, racial bias verified by measured skin tone in betting markets has not been previously explored. Additionally, the methods presented for teasing out skin tones for players are a novel contribution of this research. In order to detect mispricing related to racial bias, I apply the aforementioned forecast evaluation framework. I estimate the following using least squares:

$$e_{ij} = \alpha + \beta_1 z_{ij} + \beta_2 \text{RankDiff}_{ij} + \beta_3 \text{SkinToneDiff}_{ij} + \psi_{S(j)} + \phi_{T(j)} + \varepsilon_{ij} \ , \tag{3.6}$$

where, similarly, $\{\alpha, \beta_1, \beta_2, \beta_3, \psi_{S(j)}, \phi_{T(j)}\}$ are parameters and $\{\beta_1, \beta_2, \beta_3\}$ are the bias coefficients. α represents the bookmaker profit. Positive values for β_1, β_2, and β_3 indicate a longshot bias, high rank bias, and low racial bias respectively. Consistently betting on favorites, lower ranked players, or the player with the lighter skin tone could return sustained profits. Skin tone unit vectors with higher values correspond to lighter colors in comparison to lower values corresponding to darker colors.

3.4.2 Market inefficiency and a simple betting strategy

As a test of market inefficiency in WTA markets, I use the estimation results of Equation (3.5) on an out-of-sample set of tennis matches. I leverage the Kelly (1956) criterion, a commonly used mathematical formula used to optimize the placement and allotment of a wager or investment given current conditions. The Kelly criterion is among the most popular simple betting strategies widely used to exploit price distortions in attempt to evaluate efficiency in betting markets (see Hvattum and Arntzen, 2010; Peeters, 2018; Ziemba, 2020). To conduct a repeatable, realistic test I assume a bettor, with all relevant information prior to the match, forms their beliefs and expectations by estimating Equation (3.5) using least squares, less tournament and fixed effects to ensure a pragmatic approach. With the estimation results used to forecast match outcome probabilities, in coordination with the optimization formula, the bettor places bets appropriately. Let $\widetilde{y}_{ij}$ be the bettor's expected probability of winning a bet on a specific player i to win match j. Thus, $\widetilde{y}_{ij}$ is given by:

$$\widetilde{y}_{ij} = \widehat{\alpha} + (1 + \widehat{\beta}_1) z_{ij} + \widehat{\beta}_2 \text{RankDist}_{ij} + \widehat{\beta}_3 \text{BeautyDiff}_{ij} \ , \tag{3.7}$$

where $\{\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3\}$ are the in-sample OLS estimators. With offered decimal odds $o_{ij} = 1/z_{ij}$, the Kelly Criterion tells the bettor the allotment to allocate in order to maximize profits:

$$x_{ij} = \max\{\widehat{y}_{ij} - \frac{1 - \widehat{y}_{ij}}{o_{ij} - 1} \ , \ 0\} \ . \tag{3.8}$$

The bettor's return on investment (ROI) over $N = 2J$ possible bets is given by:

$$\mathrm{ROI} = \frac{\sum_{ij}^{2J} \left(x_{ij} o_{ij} 1\{y_{ij} = 1\} - x_{ij} 1\{y_{ij} = 0\} \right)}{\sum_{i}^{2J} x_{ij}} . \tag{3.9}$$

In exploiting some combination of bias leading mispricing, sustained profits over time, given a large enough sample, affirms that WTA markets are weak form inefficient. Hence, the outlined model, with the accessible information highlighted and betting strategy could potentially provide real time profits.

3.5 Results

3.5.1 Mispricing

Starting with the estimation results of the mispricing models, Table 3.3 shows the results of estimating Equation (3.5). The dependent variable, given by $\bar{e}_{ij} = y_{ij} - \sum_{k}^{K_j} \left(z_{ijk}/K_j \right)$, uses the average bookmaker mispricing - the observed difference between pre-match odds and actualized results - among the individual bookmakers ($k = 1, \ldots, K_j$) offering odds on a given match. Following conventional literature, I test for a favourite-longshot bias in Column (I). In this case, I find a highly significant favourite bias, however relatively marginal. Column (II) evaluates mispricing due to WTA player rankings- a readily available piece of information accounting for recent player performances. With $\mathrm{RankDiff}_{ij}$, the difference in the pre-match WTA rankings of the players, as an added independent variable, I find a significant, although marginal, rank bias. This specification suggests that the odds of a lower ranked player are actually 0.09 higher with each rank than the pre-match odds imply. In line, the null $H_0 : \beta_1 = \beta_2 = 0$ is rejected at a 5% level since there is evidence that of bookmaker mispricing among WTA tennis matches accounting for difference in rank between opponents.

In column (III) I add in $\mathrm{BeautyDiff}_{ij}$, my featured measure of beauty as predicted by the deep learning attractiveness model. Similar to my estimate in column (II), the specification in column (III) finds a significant ranks bias with an added beauty bias (p-value $= 0.027$). This reflects that average bookmaker odds in fact, overestimate the importance of beauty in their pre-match forecasts. Although much of the literature posits that beauty or attractiveness and success are highly correlated even in sports, my findings suggests that price setters in WTA markets systematically overemphasize the importance of attractiveness relative to opponents when predicting match winners. Accordingly, I can again reject the condition for weak form market efficiency, H_0 :

TABLE 3.3: Model estimates and tests of betting market mispricing for WTA match results

	(I)	(II)	(III)	(IV)
Odds-implied probability	-0.082***	-0.048	-0.047	-0.048
	(0.003)	(0.131)	(0.136)	(0.127)
WTA rank diff. (player-opponent)		-0.09**	-0.093*	-0.093**
		(0.033)	(0.027)	(0.029)
"Beauty" Score diff.			-0.063**	-0.063**
			(0.027)	(0.03)
Constant	0.011	-0.007	-0.007	-0.003
	(0.447)	(0.673)	(0.658)	(0.841)
Year/season fixed effects	Yes	Yes	Yes	No
Tournament fixed effects	No	No	No	Yes
F-test: $H_0 : \beta_1 = \beta_2 = 0$	0.0027	0.0012	0.0004	0.0005
N of player-matches	6,436	6,436	6,436	6,436
Adj R^2	0.0011	0.001	0.0017	0.0026

Notes.- ***,**,* indicate significance from zero at 1%, 5% and 10% levels, respectively, two-sided tests. Standard errors in parentheses were estimated robust to both match and tournament level clusters.
Column (I): linear regression estimates of Equation (3.5), where the dependent variable is the forecast error implied by average bookmaker odds (oddsportal.com) – test of favourite-longshot bias
Column (II): adds the pre-match WTA rank difference to the model in (I)
Column (III): adds the relative difference in "Beauty" Score to model in (II) - preferred results
Column (IV): adds tournament fixed effects to the model in (III)

$\beta_1 = \beta_2 = \beta_3 = 0$ at a 5% level. Finally, in column (IV) I add tournament fixed effects to my specification defined in column (III); results are not markedly different.

In Table 3.4 I explore facial width-to-height ratio (also referred to as the golden ratio), the commonly used measure of attractiveness, as a robustness check for my deep learning attractiveness score. In order for the "Beauty" Score differential to be robust measure of beauty, it should already reflect, if not fully encapsulate, other measures of beauty like facial symmetry, dimensions of prominent facial landmarks, and FWHR. Alternatively, if the model estimate defined in Table 3.3, column (III) is sensitive to the addition FWHR is a regressor, this suggests a probable mispecification. In other words, if BeautyDiff$_{ij}$ is rendered insignificant in response to adding FWHR to the specification, the variable's validity is in question. Following the format outlined in Table 3.3, I first add FWHR to the specification used in column (II) yielding insignificant results. In similar fashion, column (II), table 3.4 uses the relative measure of FWHR. Taking into account the FWHR of a player relative to their opponent's FWHR, there are no significant results. Finally, in column (III) I regress the relative golden ratio measure along with the relative beauty score predicted by the deep learning attractiveness model. While the relative FWHR measure appears to be significant at a 10% level (p-value = 0.074), the "Beauty" Score differential

TABLE 3.4: Model estimates with "Golden Ratio"

	(I)	(II)	(III)
Odds-implied probability	-0.044*	-0.052**	-0.054**
	(0.084)	(0.045)	(0.04)
WTA rank diff. (player-opponent)	-0.07**	-0.062*	-0.066**
	(0.032)	(0.062)	(0.045)
Face Width/Height Ratio	-0.008		
	(0.228)		
Face Width/Height Ratio diff.		-0.008	-0.012**
		(0.09)*	(0.011)
"Beauty" Score diff.			-0.069**
			(0.003)
Constant	-0.008	-0.004	-0.004
	(0.79)	(0.808)	(0.905)
Year/season fixed effects	Yes	Yes	Yes
Tournament fixed effects	No	No	No
F-test: $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$	0.0014	0.0009	0.000
N of player-matches	5,971	5,499	5,499
Adj R^2	0.0006	0.0005	0.0012

Notes.- ***,**,* indicate significance from zero at 1%, 5% and 10% levels, respectively, two-sided tests. Standard errors in parentheses were estimated robust to both match and tournament level clusters.
Column (I): adds WTA player facial width/height (golden) ratio to model (II) in table 3.3
Column (II): replaces WTA player golden ratio with the relative difference in golden ration between player and opponent in model (I)
Column (III): adds the relative difference in "Beauty" Score to the model in (II)

stays approximately consistent with the estimation described by table 3.3, column (III). In summary, the introduction of the commonly used FWHR (also referred to as the golden ratio), has little effect on the efficacy of BeautyDiff$_{ij}$. The results strengthen the validity of the attractiveness score differential as a regressor.

As previously described, BeautyDiff$_{ij}$, or the Beauty Score differential, predicted by the deep learning beauty model seems to be heavily correlated with race or ethnic origin of the respective WTA players. Table 3.5 is an exploration of this observed relationship using SkinToneDiff$_{ij}$, the skin tone differential. Column (I) adds raw WTA player skin tone to the specification in column (II) of table 3.3. Regardless of skin tone relative to opponent, the regression results indicate a nominal, highly significant racial bias. The interpretation of these findings hinges on the direction of the skin tone unit vector. The negative sign of the coefficient corresponding to skin tone indicates a light skin tone bias. In other words, bookmakers' pre-match forecasts tend to underestimate darker skin tones and overestimate lighter skin tones. Column (II) introduces SkinToneDiff$_{ij}$, the relative measure of skin tone. Much the same as the standalone measure, the relative measure of skin tone reflects a light skin tone bias

TABLE 3.5: Mispricing and Racial Bias: Model estimates with Skin Tone Measures

	(I)	(II)	(III)
Odds-implied probability	-0.047	-0.049	-0.049
	(0.134)	(0.125)	(0.125)
WTA rank diff. (player-opponent)	-0.087***	-0.084***	-0.088***
	(0.039)	(0.047)	(0.038)
Skin Tone	-0.016***		
	(0.021)		
Skin Tone diff.		-0.016***	-0.012
		(0.025)	(0.111)
"Beauty" Score diff.			-0.047
			(0.122)
Constant	0.002	-0.007	-0.007
	(0.903)	(0.701)	(0.699)
Year/season fixed effects	Yes	Yes	Yes
Tournament fixed effects	No	No	No
F-test: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$	0.025	0.027	0.022
N of player-matches	6,422	6,378	6,378
Adj R^2	0.0014	0.0018	0.0021

Notes.- ***,**,* indicate significance from zero at 1%, 5% and 10% levels, respectively, two-sided tests. Standard errors in parentheses were estimated robust to both match and tournament level clusters.
Column (I): adds WTA player skin tone to model (II) in table 3.3
Column (II): replaces WTA player skin tone with the relative difference in skin tone between player and opponent in model (I)
Column (III): adds the relative difference in "Beauty" Score to the model in (II)

relative to opponents. Finally, column (III) adds in BeautyDiff$_{ij}$ as a regressor. Not surprisingly, both relative variables drop from statistically significant as independent variables to insignificant. This is most likely due to multicolinearity between the independent variables SkinToneDiff$_{ij}$ and BeautyDiff$_{ij}$ since figure 3.9 reveals that the raw skin tone unit vectors and predicted beauty scores are highly correlated.

Overall, I find consistent match forecasting errors in the WTA market made evident through the estimation of Equation (3.5). These findings reveal there may be significant rank, attractiveness, and racial bias. Further these mispricings could potentially be exploited by systematically betting on the lower rank player, the player with the lower attractiveness scores, and the player with the darker skin tone. Proving reliable returns would provide credible evidence of WTA betting market inefficiency.

3.5.2 Market inefficiency

Turning to the test of market inefficiency, table 3.6 describes the results of a very simple betting strategy in which I wager exactly £1 on the match outcome winner predicted by the baseline model. Alternating between the specification defined in table 3.3, column (III) and the specification in table 3.5 column (II), I consider the betting strategy for BeautyDiff_{ij} and SkinToneDiff_{ij} (denoted as Tone) respectively. First, I evaluate the strategy using average odds, the average of all pre-match odds offered by bookmakers. Second, I test using best odds available. This portion considers all bookmakers that offer odds on a given match and picks the most advantageous of them. This is offered as a conceptual benchmark and not a realistic betting strategy. Third, I demonstrate a realistic betting strategy and test of market inefficiency by adhering strictly to Bet365 [11] odds. Columns (I) and (II), the tests of attractiveness and racial bias using average odds, shows similar negative ROI as well as absolute returns. Naturally, columns (III) and (IV) demonstrate the added advantage of a simple betting strategy when using optimal odds, however returns are marginal. Finally, the realistic betting strategy depicted in columns (V) and (VI) yield negative results not unlike the strategy using average odds. The test of market efficiency using a simple betting strategy with information not fully incorporated in the price provides minimal sustained returns (certainly not enough to overcome the overround); therefore, fails to prove market inefficiency in WTA betting markets.

TABLE 3.6: Very simple betting strategy results for WTA match results

	Average		Best		Bet365	
	Beauty (I)	Tone (II)	Beauty (III)	Tone (IV)	Beauty (V)	Tone (VI)
N odds ($2 \times J$ matches)	6,439	6,439	6,439	6,439	6,439	6,439
Number of bets placed	6,439	6,439	6,439	6,439	6,439	6,439
Mean overround (%)	5.64	5.64	0.08	0.08	6.7	6.7
Investment (× per bet budget)	6,439	6,439	6,439	6,439	6,439	6,439
Absolute return (× per bet budget)	-251.64	-170.72	177.59	268.06	-294.43	-214.36
Return on Investment (%)	-3.91	-2.68	2.76	4.2	-4.57	-3.36

Notes.- "Simple" betting strategy is to wager exactly 1 on the player indicated by the baseline model.
Column (I): Bet on lower "beauty" score - uses the reported average pre-match available odds from oddsportal.com
Column (II): Bet on lower skin tone unit vector - uses the reported average pre-match available odds from oddsportal.com
Column (III): Bet on lower "beauty" score - uses the reported best available pre-match odds from oddsportal.com
Column (IV): Bet on lower skin tone unit vector - uses the reported best available pre-match odds from oddsportal.com
Column (VI): Bet on lower "beauty" score - uses pre-match odds from Bet365
Column (VII): Bet on lower skin tone unit vector - uses pre-match odds from Bet365

Building on the results of table 3.6, table 3.7 shows the results of an alternative betting strategy by applying the Kelly criterion to the probabilities on match outcomes predicted by Equation (3.5). In contrast to my last betting strategy, the strategy

[11] Bet365 - the world's largest online sportsbook in terms of revenue. See https://www.co.bet365.com

defined in table 3.7 serves as a more pragmatic attempt at exploiting bias related to BeautyDiff$_{ij}$, my derived measure of beauty formulated from readily available information. Adhering to realistic standards, I train Equation (3.5) using matches ranging from the 2016-17 WTA season until the 2018-19 season. With three seasons of data dedicated to model training, the test of market efficiency is performed by implementing the Kelly criterion on 2019-20 match probabilities predicted by my trained model. Following the convention exhibited in table 3.6, the more sophisticated strategy will be evaluated using average odds, best odds, and Bet365 odds. The transition from the simple betting strategy to the pickier Kelly criterion driven strategy is made readily apparent by the difference in number of bets placed. The strategy in table 3.6 places a wager of exactly 1 on every match on the likelier winner using the new pieces of information (attractiveness scores and skin tone). Accordingly, every column sees 6,439 wagers placed. In contrast, the Kelly criterion indicates to place 1,521 wagers at most using the best possible odds and only 382 wagers at most when only looking at Bet365 odds. Table 3.7 alternates between the specifications depicting BeautyDiff$_{ij}$ and SkinToneDiff$_{ij}$ and in a similar fashion to table 3.6. The BeautyDiff$_{ij}$ specifications, given by estimating equation (3.5), are denoted as "Beauty". Otherwise, the SkinToneDiff$_{ij}$ specifications, given by estimating equation (3.6), are denoted as "Tone". From left to right, column (I), the Kelly criterion strategy using average odds and the beauty based model, yields some positive returns (approximately 2%), although just enough to outpace the overround. Column (II), given the same overround, more than doubles the ROI by turning focus to exploiting skin tone bias. Despite an average vig of over 5%, the Kelly motivated betting strategy equipped with SkinToneDiff$_{ij}$ information is able to generate a 5% return. Column (III), using best odds informed by the beauty model, places substantially more wagers and demonstrates the highest ROI at 6.34% yet. Column (IV), using best odds while turning focus to skin tone again outpaces the beauty strategy in column (III). Column (V), the most practical strategy using Bet365 odds, conversely reveals negative returns for the beauty model. Finally, column (IV), the reciprocal of (VI) for SkinToneDiff$_{ij}$, demonstrates a similar betting pattern, but returns a marginally positive ROI.

In summary, tables 3.3 and table 3.5 reveal that attractiveness scores and skin tone scalars constructed using publicly available information are not entirely integrated into the prices on WTA matches that bookmakers offer. In this case, bookmakers seem to exhibit a bias toward more attractiveness players and a bias toward players with lighter skin tones. Further, the two reasonably repeatable betting strategies described in table 3.6 and table 3.7 are offered as tests of market inefficiency. The hypothetical existence of sustained, substantial profits generated using BeautyDiff$_{ij}$ and/or SkinToneDiff$_{ij}$ would exploit the evidence that information related to relative

TABLE 3.7: Out-of-sample betting strategy results for WTA match results, 2019-20

	Average		Best		Bet365	
	Beauty (I)	Tone (II)	Beauty (III)	Tone (IV)	Beauty (V)	Tone (VI)
N odds ($2 \times J$ matches)	3,122	3,028	3,122	3,028	3,110	3,028
Number of bets placed	309	408	1,521	1,460	382	441
Mean overround (%)	5.31	5.31	-0.26	-0.26	6.48	6.48
Investment (× per bet budget)	13.09	17.16	54.34	61.61	14.72	18.32
Absolute return (× per bet budget)	0.27	0.85	3.45	4.32	-0.32	0.4
Return on Investment (%)	2.05	5	6.34	7.01	-2.16	2.2

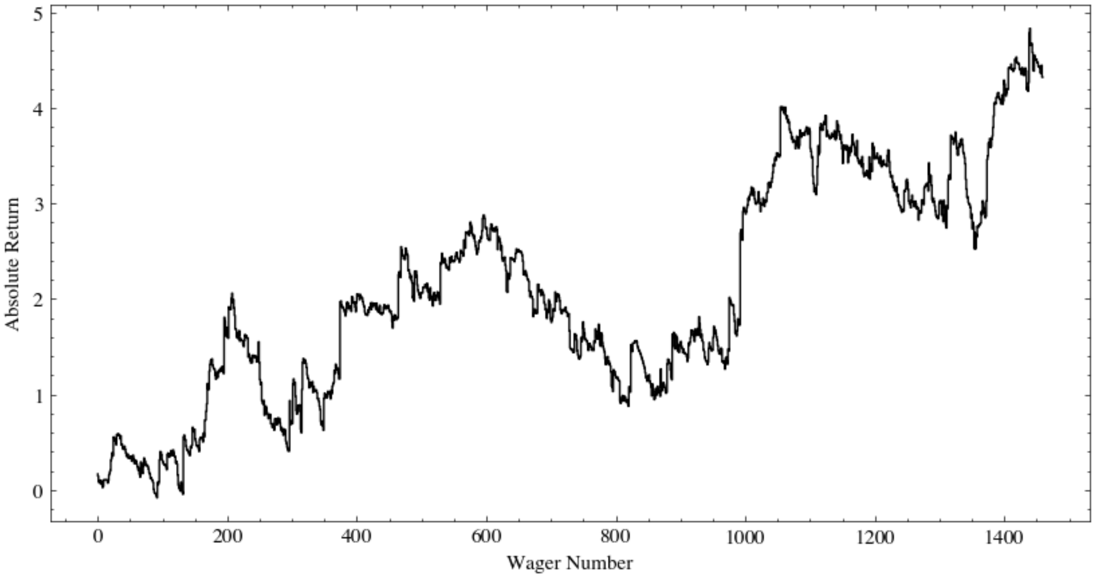
Notes.- "Out-of-sample" uses the model from column (III) of Table 3.3 estimated on matches up to the end of the 2018 season, then uses it to predict match outcomes and apply the Kelly criterion for the 2019 & 2020 seasons. Average odds are always used to estimate the models and generate forecasts, but the odds used in the Kelly criterion are varied.
Column (I): uses the reported average pre-match available odds from oddsportal.com
Column (II): uses the reported best available pre-match odds from oddsportal.com
Column (III): uses pre-match odds from Bet365
Column (IV): uses Bet365 odds but with a version of the preferred model estimated without the rank difference

beauty and skin tone are not fully incorporated into market prices. This identifiable departure from the Efficient Markets Hypothesis are verified by sustained positive returns. Given the existence of systematic mispricing, the strategies leveraging both racial and attractiveness bias consistently overcome the bookmaker overround, with the racial bias strategy generating positive ROI in all three odds settings.

3.6 Conclusion

This chapter develops a deep learning beauty model to predict beauty scores corresponding to many of the major names in the WTA. With these beauty scores, I formulate a relative beauty differential in which I measure the proportional differences in beauty between tennis match opponents. I conclude that the constructed measure of beauty can significantly predict bookmaker forecast error, and therefore bookmaker mispricing, on match outcomes. Bookmakers systematically favor the more beautiful player relative to their opponent; this beauty bias leaves markets vulnerable to betting strategies in which bettors place wagers on the comparatively less beautiful player. Using these results, I conduct a realistic test of market efficiency by leveraging a Kelly criterion motivated betting strategy. When assuming average market pricing, the strategy generates a 2% return on investment. Alternatively, when shopping for best odds, the return on investment climbs to approximately 6%. These results provide evidence of the aforementioned mispricing and inefficiency in tennis markets. In other words, tennis market prices do not fully reflect relevant information in the way of

FIGURE 3.13: Cumulative Absolute Returns Kelly/SkinToneDiff Strategy



Notes: this plot is generated using the strategy described in table 3.5, column (IV). This is the best performing of the Kelly informed strategies depicted in the table. From 0 to 1,460, this plot shows cumulative absolute returns corresponding to each new wager.

beauty. Given the preliminary results of the deep learning beauty model, I subsequently perform the same exercise while shifting focus to racial bias. The latter exercise yields similar results to the former. I find that the relative measure of skin tone can significantly predict bookmaker odds-implied forecast error as well. Moreover, the Kelly criterion driven model demonstrates a substantially larger return on investment (5% assuming average odds and up to 7 % assuming best odds) when exploiting the systematic bookmaker bias toward players with lighter skin tones.

While racial bias was not this chapter's intended area of focus, it arrives at an unquestionably important conversation. In a confined way, this chapter provides more evidence of racial bias in a novel setting. Perhaps more importantly, it offers commentary on the much larger relationship between racial bias and the common understanding of beauty. The questions aroused by this commentary could be addressed in a particularly impactful area of future research. Although the unfortunate connection between beauty and racial bias is largely acknowledged, it is not yet thoroughly understood whether this observed relationship is an inherent defect in machine learning models or a byproduct of humanity's racial bias. Second, the beauty methods outlined in this chapter could be extended to men's leagues as well as other professional sports leagues in general. Even though the WTA offers a uniquely consistent set of high quality photographs for which to derive beauty measures, this barrier could surely be overcome by others in future efforts.

Chapter 4

Exploring entertainment utility from football games

Submitted to: Journal of Economic Behavior and Organization. (2023); Manuscript number: JEBO-D-23-01028. "Revise and Resubmit" received from the journal. This article was co-authored with Tim Pawlowski, Dooruj Rambaccussing, J. James Reade and Giambattista Rossi; e-mail: tim.pawlowski@uni-tuebingen.de d.rambaccussing@dundee.ac.uk j.j.reade@reading.ac.uk and g.rossi@bbk.ac.uk. Tim, Dooruj, James, and Giambattista have agreed that the essay can appear within this thesis, and that it represents a significant contribution on my part.

4.1 Introduction

Modelling dynamic choice problems with an explicit focus on uncertainty attached to a certain point in time goes back to Kreps and Porteus (1978), who explored preferences for the earlier or later resolution of uncertainty. Several scholars have since extended these ideas. For instance, Palacios-Huerta (1999) has focused on the form of the timing of the resolution by explicitly modelling disappointment aversion, as introduced by Gul (1991). This model can explain a preference for the one-shot rather than the sequential resolution of uncertainty (for further extensions, see Kőszegi and Rabin (2009), or Dillenberger (2010)). Caplin and Leahy (2001) more broadly considered both negative and positive anticipatory emotions felt by individuals before uncertainty is resolved. For instance, they define suspense as a positive anticipatory emotion which might explain why fans in sports bet on their favorite team as observed by Babad and Katz (1991), i.e., fans simply want to increase their feelings of suspense.

This literature informed the seminal work by Ely et al. (2015) who modelled the demand for non-instrumental information by focusing on entertainment utility from *suspense* and *surprise*. While suspense is attributed to the variance in the next period's beliefs, thus representing a forward-looking measure, surprise results from an outcome that contradicts anterior beliefs representing a backward-looking measure. The authors close by writing: *"How suspense, surprise, and other aspects of belief dynamics drive demand for noninstrumental information is fundamentally an empirical question, one that we hope will be addressed by future research"* (Ely et al., 2015).

Only a small number of researchers to date have followed their call by empirically exploring this in sports. Bizzozero et al. (2016) used minute-by-minute TV viewing figures from 80 Wimbledon men's singles tennis matches and operationalized suspense and surprise with information coming from betting markets. Buraimo et al. (2020) used minute-by-minute TV viewing figures for 540 Premier League matches and added a further concept, shock. Instead of relying on in-play odds from betting markets, they derived implied probabilities for each outcome in each minute by feeding an in-play model. Richardson et al. (2023) replicated this study using minute-by-minute TV viewing figures for 180 (131) UEFA Champions League games televised in the UK (Spanish) market. Kaplan (2021) used 15-minute interval TV ratings from 477 National Basketball Association (NBA) games during the 2017-18 and 2018-19 seasons and compared the impact of *thrilll* (measured by suspense and surprise) and *skilll* (measured by productivity and popularity). Simonov et al. (2022) used detailed viewership information for a sample of 104 professional eSport tournament games summing up to more than 2,700 rounds played. These data allow modelling the decision-to-join and the decision-to-leave a (Twitch.tv) stream separately. Finally, Liu et al. (2021) used individual-level data about 877 baseball telecasts during the 2018 Japanese Major League season. The granular data which were built, amongst others, upon utilizing a facial recognition algorithm, allow to further disentangle the effects of suspense and surprise for actively versus passively attentive viewers. In fact, many consumers often do not pay full attention to the television programming since, for instance, they might actively search for game-related information and/or just do different things in parallel, such as cooking, or tweeting about the game.

Overall, these studies find that suspense and — at least to some extent — also surprise and shock are important drivers of demand. Detailed findings, however, reveal some interesting and partly contradictory issues. For instance, (i) while Bizzozero et al. (2016) find that surprise has a larger impact than suspense in tennis, Kaplan (2021), Buraimo et al. (2020) and Richardson et al. (2023) as well as Simonov et al. (2022) find the opposite pattern in basketball, football and eSports respectively. (ii) Suspense decreases the probability of leaving a stream while neither surprise nor suspense unfold any effects on the decision to join a stream (Simonov et al., 2022). (iii) Suspense and surprise seem to primarily impact viewership on the intensive margin, i.e., within games. In contrast, skill primarily impacts viewership on the extensive margin, i.e., across games (Kaplan, 2021). (iv) Spectators have a higher probability to turn on games featuring less popular players/teams if they're nearing the end *and* exhibiting sufficiently high suspense (Kaplan, 2021). Finally, (v) postseason games amplify the effects of suspense and surprise while women seem to be less responsive to suspense and surprise than men (Liu et al., 2021).

Despite the contribution of these studies to better understand how entertainment utility translates into the demand for sports, two main shortcomings exist which we intend to address in this study. First, the setting analyzed, i.e., TV/stream viewing behavior, requires a careful distinction between the decision-to-join versus the decision-to leave a program/stream (Simonov et al., 2022) and between active versus passive viewing (Liu et al., 2021). While some studies try to approach these issues with more fine-grained data and complex measures, we propose analyzing a more simple setting: social media behaviour, and in particular behaviour on Twitter, where individuals decide whether to send a Tweet.[1] Second, neither of the studies is able to reveal whether and how fandom is moderating the relation of interest since TV/stream viewing figures do not allow any distinction between fans. However, according to affective disposition theory (Zillmann and Cantor, 1972), emotional reactions by fans are a function of the content *and* a fan's disposition towards athletes/teams in contention (Raney, 2018).

We approach both shortcomings by combining data on in-game events with betting odds and Tweets for 380 games played in the English Premier League (EPL) in season 2013/2014. While the former two data sets are used for operationalizing surprise, suspense, and shock, the latter data allow us to derive temporal sentiment and distinguish between different types of individuals. We start by generating sentiment scores for each Tweet using a random forest estimator trained on data from Stanford's Sentiment Tree Bank. The calculated average post-game sentiment scores for every Twitter user enable us to identify *fans* and *haters* of a club as well as *neutrals* for each game. In order to explore entertainment utility from football games for these different types of individuals, we regress the number of Tweets per minute on surprise, suspense,

[1] Note that exploring the effects of emotional cues on Twitter activity was already proposed by Kaplan (2021) who writes on p. 16: *"Future work can directly assess the relevance of each of these mechanisms using household-level viewership data as well as complementary data from information-providing applications (e.g. Twitter)."* Yet, the only study investigating the effects of emotional cues on complementary activities beyond watching is Fischer et al. (2023). They explored the effects of *suspense* and *surprise* on alcohol consumption during a match.

and shock. Moreover, we explore asymmetries in behavior by disentangling the effects for *fans* and *haters* when 'their' team is losing or winning.

Our findings suggest that emotional cues significantly influence the number of Tweets in a given minute. While both backward-looking measures *increase* the number of Tweets, *suspense* as a forward-looking measure *decreases* the number of Tweets. The latter could be explained by individuals being 'caught in the moment' probably leaving no time to tweet. As could be expected, any response to emotional cues is smallest for *neutrals*. Interestingly, however, *haters* respond more strongly than *fans* to such cues. Further analysis suggests that *goal-induced* effects from *surprise* and *shock* on Twitter activity are the largest, when the favorite (or hated) team either scores or concedes an equaliser. Moreover, we observe asymmetries particularly regarding the response to *suspense*, i.e., very suspenseful moments during a match when 'their' team is losing increase the number of Tweets by *fans* but not by *haters* while the corresponding effects from *suspense* remain negative when 'their' team is winning.

We contribute to the literature in three ways. First, we present a novel setting for testing whether and how belief dynamics drive behavior. This seems highly relevant given the lack of research about immediate emotions and the consequences of a wide range of visceral factors for (immediate) human behavior *in general* (Loewenstein, 2000). Moreover, this seems promising given the identified drawbacks when modelling TV/stream viewing behavior as discussed before. Second, we present an approach for detecting *fans* and *haters* of a club as well as *neutrals* via sentiment revealed in Tweets. From a managerial perspective this approach might help to further develop and implement personalized forms of communication by clubs and sponsors.[2] Third, by looking at behavioral responses to the temporal resolution of uncertainty during the course of a game, we offer a different and fine-grained type of empirical test for the well-known uncertainty-of-outcome hypothesis in sports.[3] This seems relevant from a policy perspective, since the hypothesis still lacks empirical support even though it forms the basic argument for all cross-subsidization measures and labour market interventions in professional sport leagues around the globe (see, for instance, Pawlowski et al. (2018)).[4] Our findings suggest that entertainment utility is influenced

[2] For a recent discussion on the personal, social, and commercial relevance of understanding such behavior, see Jiwa et al. (2021)).

[3] The uncertainty-of-outcome hypothesis (UOH) originates from the seminal works by Rottenberg (1956) and Neale (1964) and suggests a positive relation between the level of uncertainty over the outcome of a sports competition and its attractiveness for spectators and fans.

[4] To the best of our knowledge, only one study exists that has used Twitter data for testing the UOH before. Lucas et al. (2017) use three different types of information about 60 (out of 64) FIFA World Cup games in 2014, i.e., Vegas betting odds in order to measure differences between predicted and actual scores for the two teams in contention, a game's average Tweets per minute as a proxy for attendance by/excitement of the Twitter audience, and the proportion of Tweets which were positive, negative or

by elements which gain in (lagged) *certainty* (such as surprise or shock) as well as elements which gain in *uncertainty* (such as suspense). In particular, we argue that the negative effect of *suspense* on Twitter activity is suggestive of individuals being 'caught in the moment' and as such paying more attention to the match itself. This proposition is fully backed up by studies exploring the demand for sports telecasts which unambiguously reveal a positive effect of *suspense* on viewing figures (see, for instance, Buraimo et al. (2020) or Richardson et al. (2023)). Moreover, it is in line with Fischer et al. (2023) who find that *suspense* reduces alcohol purchases in the stadium during a match.

In Section 4.2 we introduce our data and methodology, in Section 4.3 we present our results, and Section 4.4 concludes.

4.2 Data and Method

4.2.1 Identifying Fans, Haters, and Neutrals

Our data comprise of all worldwide English language Tweets that mention any hashtags associated with a team in the EPL before, during and after all 380 matches played in the 2013/2014 season.[5] This amounts to about 19 million unique Tweets for our analysis.

For identifying *fans*, *haters*, and *neutrals*, one could think of simply using the hashtag used by a particular Twitter user. However, such an approach would be misleading since a neutral consumer may write a Tweet about a match using hashtags for either of the teams, while a fan of one team may tweet and mention a hashtag of another team. We propose a more sophisticated way of identifying *fans*, *haters*, and *neutrals* that uses the sentiment expressed in Tweets. In general, a range of ways of measuring sentiment exist, from simply assigning words a positive or negative number, to classifying particular passages of words as being positive or negative. In this chapter, we generate sentiment scores, ranging from 0 (very negative) to 25 (very positive), for each Tweet using a Random Forest (RF) estimator trained on data from the Stanford's Sentiment Tree Bank. Broadly speaking, the RF estimator produces an ensemble of

neutral during a game. Simple game-level correlations reveal, that games with bigger than expected score differences had higher Tweets per minute and a higher share of negative Tweets. They argue, that the latter finding is in line with the UOH while the former contradicts the UOH. We argue, however, that game-level correlations can hardly reveal any credible and robust evidence on the relation of interest. Moreover, the authors did not make use of the elaborated cue measures as proposed by Ely et al. (2015) and partly even confuse emotional cue and attention measures.

[5] Taking the example of Liverpool, a corresponding Tweet contains one or more of FC Liverpool, @LFC, @lfcbuzztap, @empireofthekop, @liverpool, @Liverpool_FC_, @thisisanfield, #lfc, #liverpool, #liverpoolfc, or #ynwa. For further details about data and methods, please see Appendix A.

decision trees popularly used for Natural Language Processing. In contrast to neural networks – a high performing algorithm though black box approach – the RF estimator shows how important individual features are in determining outcomes. Our model was trained on more than three million features or word tokens with the then most important features being *bad*, *performance*, *best*, *n't*, *funny*, *dull*, *great*, *like*, *good*, and *waste* (see Appendix A.4 for further details on the architecture of the winning model).

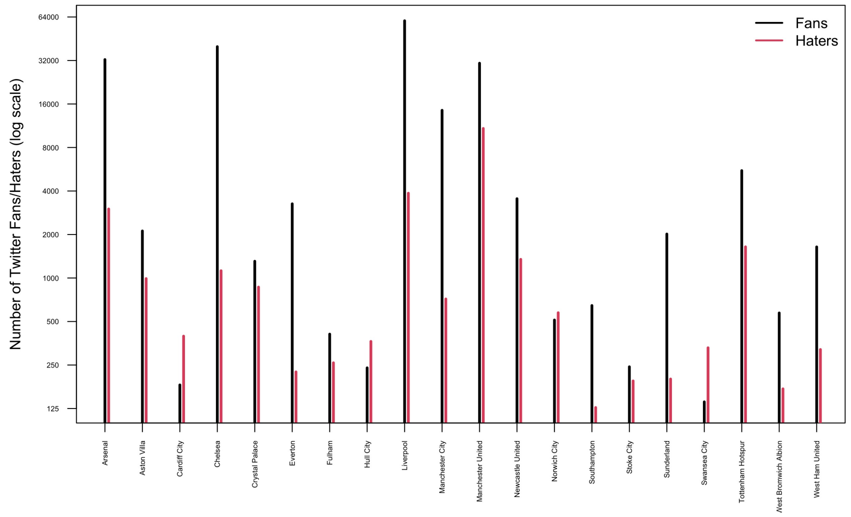
We then isolate post-*win* and post-*loss* sentiment scores for every Twitter user for each game. For identifying fans, we rank the average post-*win* sentiment scores per game and take the most commonly occurring team in a user's top 5. If the user does not comment positively on more than two wins of a particular team, we fail to assign fandom for this user. In other words, a team must appear at least twice in a user's top 5 in order to be considered. Conversely, to determine a hater, we look at Tweets with hashtags associated with the losing team, i.e., post-*loss* sentiment. Again, ranked by average sentiment score of tweets, we take the most commonly occurring team in each user's top 5. The inutition behind this approach is that positive sentiment after a loss probably reflects some kind of *schadenfreude*. Like for fandom, if the Twitter user does not take delight in at least two losses for a team, hateship cannot be found. The remainder of users are neither denoted as a fan nor as a hater and are assumed to be neutral.[6]

In general, most users tweet about a team post win rather than post loss, with eligible users (users with at least 3 tweets) tweeting 1,221,340 times about the winner of a team post-win and only 629,637 times about the loser of a match post-loss. Based on our rule-based approach, the user's top scoring is heavily favored in determining fandom vs hatership. Post win, the average top scoring sentiment is 15.09, well above the average sentiment score overall of 13.43. Post loss, however, the average top scoring sentiment is 13.86 which is just above average. This suggests, more often than not, a user delights on their own team's success more than celebrates another's demise, thus, making it generally easier to assign fandom as opposed to hatership.

Following this approach, we identified 196,270 users as *fans*, 23,747 users as *haters*, 3,792 users as both a *fan* of one team and a *hater* of another team, and 1,096,225 users as *neutrals* amongst the overall 1.3 million Twitter users. Figure 4.1 provides an overview on the number of Twitter users regarded as *fans* and *haters* of particular teams.

[6] If we find overlap between fandom and hateship of the *same* team for a Twitter user (many users regularly comment on just one or two clubs), we assign either fandom or hateship according to the higher absolute value of the post-match sentiment score. If the user has a higher average sentiment score post-*win*, the user is determined to be a fan. If the higher average sentiment score occurs post-*loss*, the user is marked as a hater.

FIGURE 4.1: Number of Twitter users regarded as *fans* and *haters* of a particular team.



In order to see how this classification exercise works, we take an example from the match between Liverpool and Chelsea on matchday 36 of 38. The match was critical for the championship race and ended with a 0-2 home loss leaving Liverpool with considerably reduced chance of winning the title. Out of overall 214,133 'Liverpool' Tweets before, during, and after this match, 28 percent are by Liverpool fans as identified by our approach. As expected, of the 5,577 users retweeting "@LFC LOL!" after the game, only a marginal portion of these users (2.67%) are Liverpool fans as identified by our approach. More generally, we find some strong correlations between the overall number of fans identified by our approach and the average number of spectators attending matches of each team (see Figure 4.2) as well as the number of (actual) followers of the official team accounts (see Figure 4.3) adding some further credibility to our approach.[7]

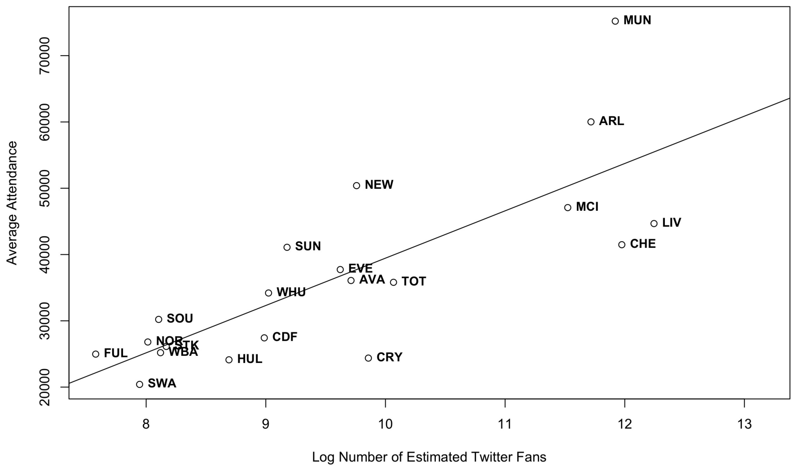
4.2.2 Measuring Emotional Cues

Following Buraimo et al. (2020) we rely on the probability of each of the three outcomes in a football match – i.e., home win (H), draw (D), or away win (A) – at time t, denoted as p_t^H, p_t^D, and p_t^A respectively, for measureing emotional cues.

At first glance, it seems promising to take in-play betting data for deriving these probabilities on a minute-by-minute basis. In this regard, the most comprehensive data come from the *Betfair* betting exchange where offered prices evolve by betting market participants preapred to both buy and sell betting contracts. However,

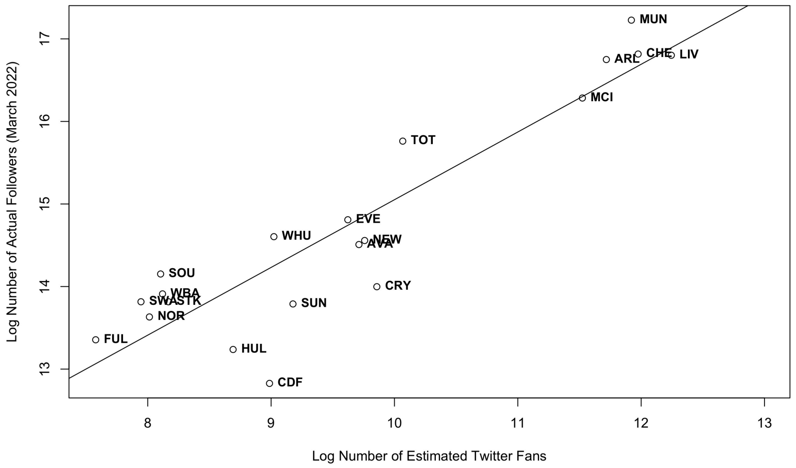
[7] Note that the counts of followers were taken in March 2022, i.e., several years after the Tweets.

FIGURE 4.2: Fans identified from sentiment and average attendance.



Notes: This Figure plots the logarithmized number of identified fans following the method as described in Section 2 and the average attendance at home games in season 2013/14. ARL: Arsenal, AVA: Aston Villa, CDF: Cardiff City, CHE: Chelsea, CRY: Crystal Palace, EVE: Everton, FUL: Fulham, HUL: Hull City, LIV: Liverpool, MCI: Manchester City, MUN: Manchester United, NEW: Newcastle United, NOR: Norwich City, SOU: Southampton, STK: Stoke City, SUN: Sunderland, SWA: Swansea City, TOT: Tottenham Hotspur, WBA: West Bromwich Albion, WHU: West Ham United.

FIGURE 4.3: Fans identified from sentiment and followers of official team accounts.



Notes: This Figure plots the logarithmized number of identified fans following the method as described in Section 2 and the logarithmized number of followers of team accounts as of March 2022. ARL: Arsenal, AVA: Aston Villa, CDF: Cardiff City, CHE: Chelsea, CRY: Crystal Palace, EVE: Everton, FUL: Fulham, HUL: Hull City, LIV: Liverpool, MCI: Manchester City, MUN: Manchester United, NEW: Newcastle United, NOR: Norwich City, SOU: Southampton, STK: Stoke City, SUN: Sunderland, SWA: Swansea City, TOT: Tottenham Hotspur, WBA: West Bromwich Albion, WHU: West Ham United.

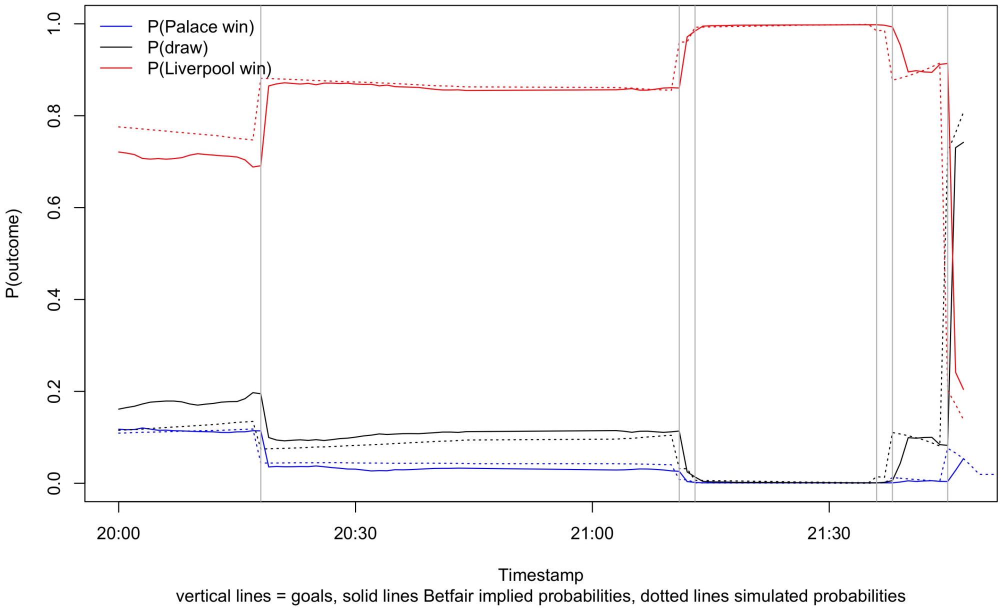
while some studies have shown that *Betfair*, or betting exchange, prices, accurately predict outcomes (Croxson and Reade, 2014), others have rejected the hypothesis of semi-strong market efficiency. For instance, Choi and Hui (2014) found that prices generally underreact to normal news and overreact to surprising news. Such market inefficiencies are also detected by Angelini et al. (2022). In summary, these findings question the overall suitability of using observable (*Betfair*) prices for predicting outcomes in our chapter.

In this chapter, we use in-play odds derived from an in-play model as proposed by Buraimo et al. (2020). The in-play model is built on pre-match closing odds in combination with over-under totals which reflect the strengths of teams in contention as well as other relevant factors such as current form of the teams and their most recent match results. By assuming an independent Poisson distribution for goals scored by both home and away teams and using the empirical goal distribution during EPL games it is possible to generate the probabilities for every scoreline for a given match and calculate the required outcome probabilities p_t^H, p_t^D, and p_t^A (for further details, see Appendix A).

In order to see how both *actual* and *simulated* outcome probabilities develop during the course of a match, we take an example from the match between Crystal Palace and Liverpool on May 5 2014 (matchday 37 of 38). This was the first match after the home loss against Chelsea (mentioned in Section 4.2.1) and was as such also critical for the championship race that season. Liverpool was winning the match 3-0 until the 79th minute when goals by Delaney and Gayle (2) helped Crystal Palace to (unexpectedly) draw. The match ended 3-3 leaving Liverpool with hardly any chance of winning the title. Figure 4.4 shows how actual and simulated probabilities developed during the course of this match.

As could be expected, each goal by Liverpool is decreasing home win and draw probabilities while increasing away win probabilities (at *decreasing* margins). The opposite pattern can be observed for each goal scored by Crystal Palace, i.e., an increase in home win and draw probabilities as well as a fall in away win probabilities (at *increasing* margins). Note that changes in outcome probabilities are not only caused by goals scored (otherwise we would observe just flat lines between any goals scored). Overall, we only observe some minor differences between actual and simulated probabilities by visual inspection. In our analysis we use simulated instead of the actual probabilities for calculating our emotional cue measures for the reasons mentioned earlier.

FIGURE 4.4: Development of outcome probabilities during the course of a match.



Notes: This Figure plots the development of outcome probabilities during the course of the match between Crystal Palace and Liverpool on May 5 2014 (matchday 37 of 38). The outcome probabilities were either derived from *Betfair* exchange data sourced via *Fracsoft* (solid lines) or simulated with our in-play model (dotted lines) as described in Appendix A. Vertical lines indicate goals scored, i.e., 0-1 (Allen, 18'), 0-2 (Delaney own goal, 53'), 0-3 (Suarez, 55'), 1-3 (Delaney, 79'), 2-3 (Gayle, 81'), 3-3 (Gayle, 88').

Recall that surprise is a backward-looking measure which results from an outcome that contradicts anterior beliefs. Considering outcome probabilities as defined before and in line with Buraimo et al. (2020) we define surprise as:

$$Surprise_t = \sqrt{(p_t^H - p_{t-1}^H)^2 + (p_t^D - p_{t-1}^D)^2 + (p_t^A - p_{t-1}^A)^2}. \tag{4.1}$$

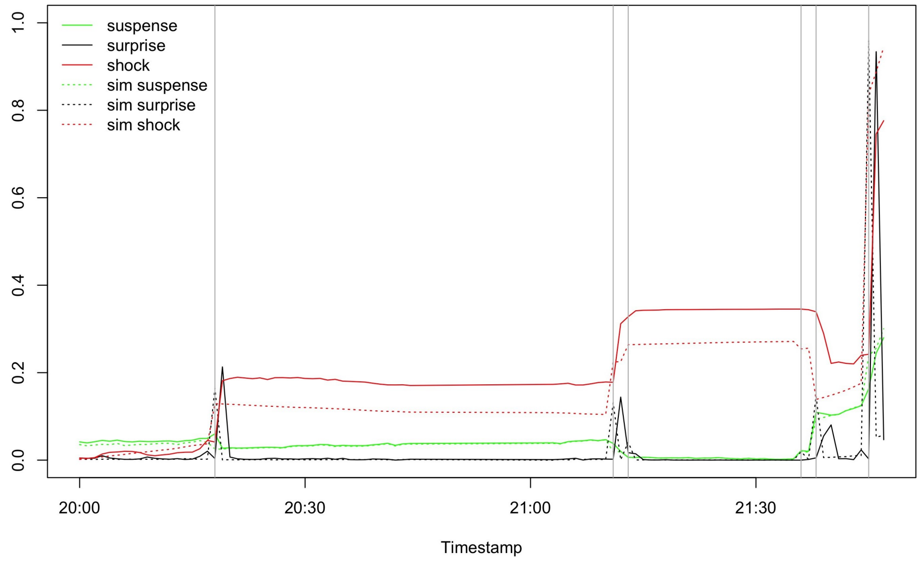
Shock is defined similarly, but with respect to the probabilities at the start of the match:

$$Shock_t = \sqrt{(p_t^H - p_0^H)^2 + (p_t^D - p_0^D)^2 + (p_t^A - p_0^A)^2}. \tag{4.2}$$

In contrast, however, suspense is a forward-looking measure which attempts to capture the impact of a goal scored in the next minute on either of the three outcome probabilities. We thus introduce p_{t+1}^{HS} and p_{t+1}^{AS} to denote the probability of the home and away teams scoring in the next minute. Then suspense is defined as:

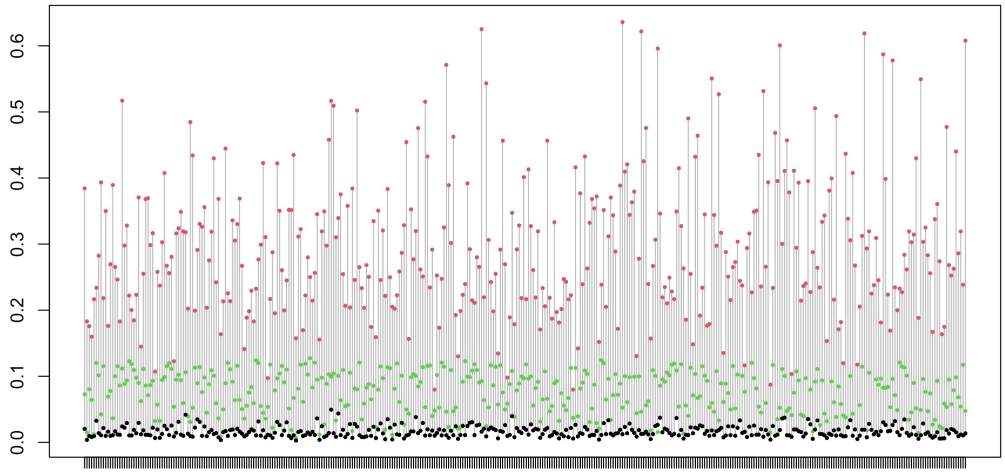
$$Suspense_t = \left(\sum_{i \in H,D,A} p_{t+1}^{HS} \left[(p_{t+1}^i p_{t+1}^{HS}) - p_t^i \right]^2 + \sum_{i \in H,D,A} p_{t+1}^{AS} \left[(p_{t+1}^i p_{t+1}^{AS}) - p_t^i \right]^2 \right)^{1/2} \tag{4.3}$$

FIGURE 4.5: Development of surprise, shock, and suspense during the course of a match.



Notes: This Figure plots the development of surprise (black), shock (red), and suspense (green) during the course of the match between Crystal Palace and Liverpool on May 5 2014 (matchday 37 of 38). Surprise, shock, and suspense were calculated from either *Betfair* exchange data sourced via *Fracsoft* (solid lines) or simulated odds (dotted lines) as described in Section 3 and Appendix A. Vertical lines indicate goals scored, i.e., 0-1 (Allen, 18'), 0-2 (Delaney own goal, 53'), 0-3 (Suarez, 55'), 1-3 (Delaney, 79'), 2-3 (Gayle, 81'), 3-3 (Gayle, 88').

FIGURE 4.6: Mean shock, surprise, and suspense per match.



Notes: This Figure plots the mean surprise (black), shock (red), and suspense (green) per match for all 380 matches played in season 2013/2014 calculated from simulated odds as described in Section 3 and Appendix A.

Taking the same example as before, Figure 4.5 indicates how shock, surprise, and suspense develop during the course of the match. Overall, the observed patterns seem reasonable. While suspense gradually *decreases* up to the 79th minute when Crystal Palace scored to make the scoreline 1-3, it substantially *increases* particularly after the third goal scored by Crystal Palace. Likewise, shock and surprise are mainly driven by the goals scored. More broadly speaking, suspense commonly reflects an upward trend over time up to the point when a match is (most likely) decided. In contrast, however, the pattern of surprise is spiky and mainly depends on (un-)expected goals scored. Finally, it is worth noting that we not only observe variation in shock, surprise, and suspense *within* a match but also *between* matches (see Figure 4.6). This must be considered in our empirical model.

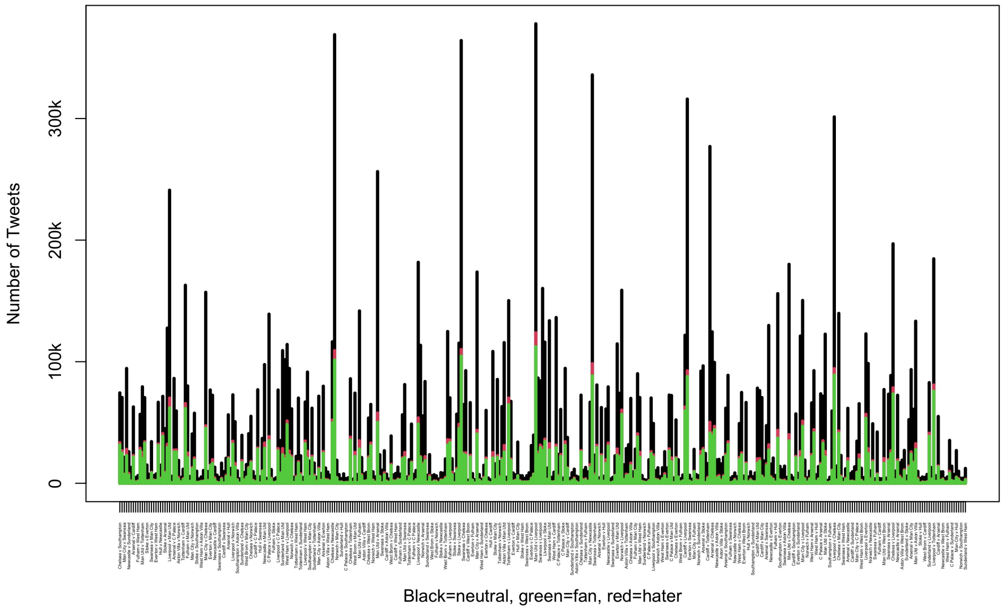
4.2.3 Empirical Model

In this chapter, we intend to model the extent to which emotional cues from football experience, i.e., surprise, shock, and suspense, provoke measurable behavioral responses. As such, the number of Tweets that include *home* team and/or *away* team hashtags in a given minute t of match i serves as the dependent variable y_{it} in our empirical model:

$$\begin{aligned} y_{it} = \beta_0 + \beta_1 y_{i,t-1} + \beta_2 surprise_{it} + \beta_3 shock_{it} \\ + \beta_4 suspense_{it} + \beta_5 X_{it} + \gamma_t + \nu_i + u_{it}. \end{aligned} \tag{4.4}$$

In order to separate the *net* effects of our emotional cue measures, we control for lagged number of Tweets y_{it-1} and a set of in-match events X_{it} like goals scored, shots, corners, cards, or substitutions. Note that as X_{it} includes total goals scored, it could be seen as a kind of basic 'excitement' index. In order to pick up any differences between minutes played and across matches, we control for minute fixed effects γ_t and match fixed effects ν_i. We also run these regressions seperately for Twitter users that we have identified as *fans*, *haters*, and *neutrals*. That is, for a match involving two teams, we count the tweets of fans (haters) of each team separately if they send a tweet using a hashtag for their favoured (hated) team. The count of neutral tweets for a match is made up of both neutral users *and* fans/haters of teams other than the two that are participating in the match, who tweet using any hashtag associated with one of the teams playing. We observe a remarkable variation in number of Tweets by *fans*, *haters*, and *neutrals* across the matches in our sample; in Figure 4.7 we plot these three counts for every match in our dataset.

FIGURE 4.7: Number of tweets per match and type of Twitter user.



Notes: This Figure plots the number of Tweets per match by neutrals (black), haters (red), and fans (green) using home team hashtags for all 380 matches played in season 2013/2014.

4.3 Results

Table 4.1 provides an overview of our regression results separated for *fans*, *haters*, and *neutrals*. Since we control for lagged number of Tweets and excluded extra time, these regressions are based on 89 minutes for 380 games. As we make use of both *home* team and *away* team hastags we end up with 67,590 minute-game observations.[8] While all regressions include minute and match fixed effects as well as lagged number of Tweets, only specifications in columns (2), (4), and (6) also include control variables.

Overall, we find that emotional cues significantly influence the number of Tweets in a given minute. While *surprise* and *shock increase* the number of Tweets, *suspense reduces* the number of Tweets. These findings are robust to the inclusion of the control variables. The only remarkable difference between our specifications with and without control variables is the larger effect size for *suspense* in the *fan* regression with controls. As could be expected, any response to emotional cues is smallest for *neutrals*. Interestingly, however, *haters* respond stronger than *fans* to such cues.

[8] Note, we miss 50 minute-observations. As such, we end up with 67,590 instead of 67,640 minute-game observations (i.e., 89 minutes x 380 games x 2 hashtag types).

TABLE 4.1: Results for tweets by *fans*, *haters*, and *neutrals*

	Dependent variable: log number of Tweets by...					
	(1)	(2)	(3)	(4)	(5)	(6)
	Fans	*Fans*	*Haters*	*Haters*	*Neutrals*	*Neutrals*
Lagged number of Tweets	0.425^{***}	0.423^{***}	0.334^{***}	0.334^{***}	0.471^{***}	0.471^{***}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
Surprise	2.101^{***}	1.776^{***}	2.163^{***}	2.303^{***}	0.703^{***}	0.559^{***}
	(0.381)	(0.665)	(0.501)	(0.876)	(0.114)	(0.199)
Shock	1.572^{***}	1.709^{***}	3.515^{***}	3.471^{***}	0.882^{***}	0.875^{***}
	(0.177)	(0.253)	(0.233)	(0.233)	(0.053)	(0.053)
Suspense	-3.883^{***}	-6.469^{***}	-7.527^{***}	-6.591^{***}	-1.851^{***}	-1.689^{***}
	(0.752)	(0.814)	(0.989)	(1.071)	(0.225)	(0.244)
Minute FEs	x	x	x	x	x	x
Match FEs	x	x	x	x	x	x
Controls		x		x		x
Observations	67,590	67,590	67,590	67,590	67,590	67,590
R^2	0.447	0.448	0.343	0.344	0.784	0.784
Adjusted R^2	0.443	0.444	0.339	0.339	0.782	0.783
Residual Std. Error	5.192	5.187	6.830	6.829	1.555	1.555
	(df = 67,117)	(df = 67,109)	(df = 67,117)	(df = 67,109)	(df = 67,117)	(df = 67,109)

Notes: This Table provides an overview of the effects of emotional cues on the number of Tweets across Twitter users. The logarithmized number of Tweets (as indicated by hashtags associated with the corresponding *home* team or *away* team) by *fans* (Columns 1 and 2), *haters* (Columns 3 and 4), and *neutrals* (Columns 5 and 6) serves as dependent variable in the models. All models include minute and match fixed effects. Specifications in Columns (2), (4), and (6) also include control variabels, i.e., dummy variables indicating goal, shot, shot hit goalframe, corner, yellow card, red card, or substitution, as well as total goals scored. Significance levels are *p<0.1; **p<0.05; ***p<0.01.

These findings are not driven by using simulated cues. As shown in Table 4.2, the results look similar when the cues are based on outcome probabilities derived from *Betfair* exchange data sourced via *Fracsoft* instead of simulated bookmaker probabilities. The only difference is the size of the coefficient for surprise which is about 2–3 times as large compared to our main specification in Table 4.1. A possible reason could be that the effect of surprise takes some time to unfold. As such, it would be better picked up using real odds which commonly reflect a short delay for updating (see Figures 4.4 and 4.5).[9]

Table 4.3 displays the results from our main specification using Poisson regressions instead of OLS. While our main findings seem robust regading the choice of the estimator used, we find that surprise is larger for *haters* than *fans* only when including controls.

Finally, in order to further explore the relevance of a particular course of the match, we add variables measuring whether the favorite (or hated) team is currently winning or losing along with the corresponding interactions between winning/losing and our

[9] Note, that we refrain from further exploring any lagged effects in our setting given econometric concerns caused by the temporal structure of the data with many measurement points.

TABLE 4.2: Results for tweets by *fans*, *haters*, and *neutrals* using bookmaker probabilities

	Dependent variable: log number of Tweets by...					
	(1)	(2)	(3)	(4)	(5)	(6)
	Fans		Haters		Neutrals	
Lagged number of Tweets	0.423^{***}	0.420^{***}	0.331^{***}	0.331^{***}	0.467^{***}	0.467^{***}
	(0.016)	(0.016)	(0.012)	(0.012)	(0.057)	(0.057)
Surprise	5.462^{***}	5.703^{***}	9.179^{***}	8.931^{***}	2.562^{***}	2.511^{***}
	(0.486)	(0.489)	(0.849)	(0.847)	(0.209)	(0.201)
Shock	1.589^{***}	1.744^{***}	3.532^{***}	3.544^{***}	0.940^{***}	0.952^{***}
	(0.193)	(0.211)	(0.290)	(0.288)	(0.136)	(0.135)
Suspense	-3.168^{***}	-6.300^{***}	-6.287^{***}	-6.132^{***}	-1.736^{***}	-1.886^{***}
	(0.707)	(0.976)	(1.014)	(1.159)	(0.380)	(0.414)
Minute FEs	x	x	x	x	x	x
Match FEs	x	x	x	x	x	x
Controls		x		x		x
Observations	67,590	67,590	67,590	67,590	67,590	67,590
R^2	0.449	0.451	0.347	0.347	0.786	0.786
Adjusted R^2	0.445	0.447	0.342	0.343	0.784	0.785
Residual Std. Error	5.184	5.177	6.811	6.810	1.548	1.548
	(df = 67,117)	(df = 67,109)	(df = 67,117)	(df = 67,109)	(df = 67,117)	(df = 67,109)

Notes: This Table provides an overview of the effects of emotional cues on the number of Tweets across Twitter users. In contrast to Table 4.1, all cues are based on outcome probabilities derived from *Betfair* exchange data sourced via *Fracsoft* instead of simulated bookmaker probabilities. The logarithmized number of Tweets (as indicated by hashtags associated with the corresponding *home* team or *away* team) by *fans* (Columns 1 and 2), *haters* (Columns 3 and 4), and *neutrals* (Columns 5 and 6) serves as dependent variable in the models. All models include minute and match fixed effects. Specifications in Columns (2), (4), and (6) also include control variabels, i.e., dummy variables indicating goal, shot, shot hit goalframe, corner, yellow card, red card, or substitution, as well as total goals scored. Significance levels are *p<0.1; **p<0.05; ***p<0.01.

emotional cue measures. Following this approach and given the temporal structure of all measures, the interpretation is as follows: if, for instance, a team scores and goes ahead, that effect on surprise is part of the *surprise*-winning interaction. If a team concedes and goes behind, that effect on surprise is part of the *surprise*-losing interaction. If a team scores (or concedes) an equaliser, that effect is covered in the normal *surprise* coefficient.

From our results in Table 4.4, the main findings remain, i.e., the effects of *surprise* and *shock* are positive while the effects of *suspense* are negative. *Suspense*, however, is not a precise predictor of Twitter activity anymore. Moreover, the interaction effects between winning/losing as well as *surprise* and *shock* are either negative or non-significant suggesting that *surprise* and *shock* unfold their largest effects when the match is currently a tie. Importantly, following the earlier interpretation of our approach, this holds true even in situation when a goal is scored. In other words, our findings suggest that *goal-induced* effects from *surprise* and *shock* on Twitter activity are the largest, when the favorite (or hated) team either scores or concedes an equaliser. Finally, very suspenseful moments during a match when 'their' team is losing seem to

TABLE 4.3: Results for tweets by *fans*, *haters*, and *neutrals* (Poisson regressions)

	Dependent variable: number of Tweets by...					
	(1)	(2)	(3)	(4)	(5)	(6)
	Fans		Haters		Neutrals	
Lagged number of Tweets	0.777*** (0.001)	0.777*** (0.001)	0.161*** (0.001)	0.160*** (0.001)	0.833*** (0.004)	0.834*** (0.004)
Surprise	0.869*** (0.010)	0.212*** (0.016)	0.566*** (0.033)	0.373*** (0.052)	0.650*** (0.007)	0.298*** (0.011)
Shock	0.293*** (0.004)	0.280*** (0.004)	0.871*** (0.011)	0.714*** (0.013)	0.271*** (0.002)	0.263*** (0.003)
Suspense	−1.651*** (0.022)	−1.416*** (0.022)	−3.123*** (0.066)	−2.930*** (0.067)	−0.909*** (0.013)	−0.785*** (0.014)
Minute FEs	x	x	x	x	x	x
Match FEs	x	x	x	x	x	x
Controls		x		x		x
Observations	67,590	67,590	67,590	67,590	67,590	67,590
McFadden's Pseudo-R^2	0.839 (df = 67,117)	0.84 (df = 67,109)	0.521 (df = 67,117)	0.521 (df = 67,109)	0.874 (df = 67,117)	0.875 (df = 67,109)

Notes: This Table provides an overview of the effects of emotional cues on the number of Tweets across Twitter users based on Poisson Regressions. The number of Tweets (as indicated by hashtags associated with the corresponding *home* team or *away* team) by *fans* (Columns 1 and 2), *haters* (Columns 3 and 4), and *neutrals* (Columns 5 and 6) serves as dependent variable in the models. All models include minute and match fixed effects. Specifications in Columns (2), (4), and (6) also include control variabels, i.e., dummy variables indicating goal, shot, shot hit goalframe, corner, yellow card, red card, or substitution, as well as total goals scored. Significance levels are *p<0.1; **p<0.05; ***p<0.01.

increase the number of Tweets by *fans* but not by *haters* while the corresponding effects from *suspense* remain negative for both *fans* and *haters* when 'their' team is winning.

4.4 Discussion and Conclusions

By analyzing 19m tweets in combination with in-play information for overall 380 games played in the English Premier League we provide empirical evidence that emotional cues significantly influence Twitter activity. Our findings suggest that emotional cues significantly influence the number of Tweets in a given minute. While both *surprise* and *shock* increase the number of Tweets, *suspense* - on average - decreases the number of Tweets. As could be expected, any response to emotional cues is smallest for *neutrals*. Interestingly, however, *haters* respond stronger than *fans* to such cues. Further analysis suggests that *goal-induced* effects from *surprise* and *shock* on Twitter activity are the largest, when the favorite (or hated) team either scores or concedes an equaliser. Moreover, we observe some asymmetries regarding the response to *suspense*. Very suspenseful moments during a match when 'their' team is losing increase the number of Tweets by *fans* but not by *haters*. At the same time, however, the corresponding effects from *suspense* remain negative for both *fans* and *haters* when 'their' team is winning.

TABLE 4.4: Results for tweets by *fans* and *haters* considering winning and losing

	Dependent variable: log number of Tweets by...			
	(1)	(2)	(3)	(4)
	Fans	*Fans*	*Haters*	*Haters*
Lagged number of Tweets	0.395^{***}	0.395^{***}	0.321^{***}	0.320^{***}
	(0.015)	(0.015)	(0.011)	(0.011)
Surprise	3.686^{***}	3.170^{***}	3.858^{***}	3.812^{***}
	(0.722)	(0.844)	(1.077)	(1.191)
Shock	2.569^{***}	2.748^{***}	3.939^{**}	4.678^{**}
	(0.966)	(0.977)	(1.920)	(1.978)
Suspense	-6.868^{*}	-6.959^{*}	-6.335	-7.365
	(3.884)	(3.894)	(8.427)	(8.660)
Winning	2.475^{***}	2.454^{***}	2.858^{***}	2.798^{***}
	(0.253)	(0.253)	(0.423)	(0.431)
Losing	-1.433^{***}	-1.536^{***}	-0.528	-0.889^{**}
	(0.246)	(0.253)	(0.413)	(0.427)
Surprise x winning	-2.096^{**}	-2.047^{**}	-1.485	-1.019
	(0.942)	(0.916)	(1.483)	(1.463)
Surprise x losing	-2.786^{***}	-2.727^{***}	-4.031^{***}	-3.532^{***}
	(0.885)	(0.863)	(1.316)	(1.287)
Shock x winning	-4.470^{***}	-4.633^{***}	-5.606^{***}	-6.297^{***}
	(1.134)	(1.137)	(2.083)	(2.125)
Shock x losing	0.786	0.537	1.952	0.956
	(1.136)	(1.142)	(2.081)	(2.134)
Suspense x winning	0.413	1.012	-0.051	2.370
	(4.804)	(4.808)	(8.832)	(9.020)
Suspense x losing	8.014^{*}	9.323^{**}	-0.747	4.246
	(4.163)	(4.187)	(8.841)	(9.008)
Minute FEs	x	x	x	x
Match FEs	x	x	x	x
Controls		x		x
Observations	67,590	67,590	67,590	67,590
R^2	0.459	0.459	0.350	0.350
Adjusted R^2	0.455	0.455	0.345	0.346
Residual Std. Error	5.138 (df = 67,109)	5.137 (df = 67,101)	6.797 (df = 67,109)	6.794 (df = 67,101)

Notes: This Table provides an overview of the effects of emotional cues on the number of Tweets across Twitter users. The logarithmized number of Tweets (as indicated by hashtags associated with the corresponding *home* team or *away* team) by *fans* (Columns 1 and 2) and *haters* (Columns 3 and 4) serves as dependent variable in the models. All models include minute and match fixed effects. Specifications in Columns (2) and (4) also include control variables, i.e., dummy variables indicating goal, shot, shot hit goalframe, corner, yellow card, red card, or substitution, as well as total goals scored. In contrast to results presented in Table 4.1, all specification also include variables which measure whether the favorite (or hated) team is currently winning or losing as well as the corresponding interactions with all emotional cues. Significance levels are $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

A potential criticism of our data is that it is a number of years old, hailing from the 2013/2014 season. We would stress that taking data from such a period allows us to address our research question at a time when chatbots and other potentially manipulating techniques or institutions did not play a major role. Moreover and importantly, even though the way Twitter is used in society has changed over time, we do not see any reason to believe that Twitter activity as a response to emotional cues from sports should have changed systematically. As such, we argue that the data at hand allow for a valid and timely empirical test of the effects of interest.

Overall, these findings could inform the literature in three ways. *First*, we follow the call by Loewenstein (2000) and provide new evidence of how immediate emotions influence immediate human behavior. Our setting seems promising since professional sports is frequently regarded as *the* emotions lab and Tweeting is an easy-to-measure and straight forward activity for millions of people around the world. *Second*, as could be seen in our analysis, *fans*, *haters*, and *neutrals* respond to emotional cues differently. From a managerial perspective this might be relevant to consider when implementing personalized forms of communication by clubs and sponsors during the course of a match. *Third*, by looking at behavioral responses to the temporal resolution of uncertainty during the course of a game, we offer a very fine-grained empirical test for the uncertainty-of-outcome hypothesis in sports. In fact, we find that entertainment utility is driven by both elements which gain in (lagged) *certainty* (such as surprise and shock) as well as elements which gain in *uncertainty* (such as suspense). We argue that the negative effect of *suspense* on Twitter activity is suggestive of individuals being ‘caught in the moment’ and as such paying more attention to the match itself. This proposition is fully backed up by studies exploring the demand for sports telecasts which unambiguously reveal a positive effect of *suspense* on viewing figures (see, for instance, Buraimo et al. (2020) or Richardson et al. (2023)). Moreover, it is in line with a recent study which finds *suspense* to reduce alcohol consumption during a match (Fischer et al. (2023)).

Chapter 5

Conclusion

Collectively, these chapters call attention to the intersection between individuals' belief preferences and their impact on pricing efficacy and decision-making in sports. The first chapter, centered on women's professional tennis, captures this relationship through an analysis of Wikipedia page views ahead of Women's Tennis Association singles matches. With the collected page views, the chapter introduces the concept of "buzz" by deriving the Wikipedia Relative Buzz Factor. This metric, plausibly motivated by wisdom of crowds, demonstrates predictive power for error in bookmaker pricing. By underscoring how belief-driven pre-match buzz significantly predicts forecast error, the chapter's findings suggest odds mispricing (thus, inefficiency) in tennis betting markets. Building upon this, the second chapter employs advanced facial recognition techniques to investigate bookmaker bias related to physical attractiveness and racial markers in tennis betting markets. Here, a Relative Beauty Differential is derived to quantify the proportional difference in beauty between match participants. Guided by the Efficient Market Hypothesis, the chapter posits market inefficiency by validating the predictive power of the Relative Beauty Differential on bookmaker forecast errors in bookmaker implied odds. Drawing attention to racial bias, the chapter executes the same exercise using skin tone unit vectors (extracted with machine learning methods), yielding similar results. Both measures affirm weak form inefficiency in women's tennis betting markets by respectively demonstrating sustained profits through simple betting strategies. Considering efficient market theory, it seems tenable for bookmakers to reduce mispricing by adjusting for bettor anticipation and overall beliefs toward players before matches. Finally, the third chapter expands a line of exploration into football matches, dissecting how individuals' belief dynamics influence entertainment value and consumption. By examining emotional cues derived from in-game events, betting odds, and Twitter data, this chapter reveals how different subsets of individuals respond to emotional influence. This interaction between beliefs, emotional states, and information exposes diver se behavioral responses depending on fan disposition - fans, haters, and neutrals - and team performance, forming engagement levels during matches. Using the sports betting and sports entertainment contexts, each chapter emphasizes how the synthesis of belief-driven preferences and emotions with detailed information reveals mispricing and bias while explaining dynamics and overall entertainment utility.

Regarding future research, the unanticipated conclusion of chapter 3 could serve as a promising launchpad for subsequent chapters aiming to dive into the nuanced interplay between racial bias, beauty standards, machine learning, and Economics. While the existing chapter certainly finds compelling evidence of racial bias within the context of betting, an equally compelling comprehensive investigation between skin tone bias and the perception of beauty seems appropriate. This line of inquiry could extend to other areas of Economics such as labor markets. Where Hammermesh and Biddle (1994) reveal a beauty premium in labor markets, future studies could uncover of a "racial premium" in labor markets. In line with the theme of bias, readily available facial recognition models could be used to answer the questions: Is there a racial premium (penalty) in sports? Do we observe the same mechanism in labor (or consumer) markets? Any conclusive findings toward these questions would undoubtedly generate interest among the respective policy-making regulatory bodies in each of these domains.

Appendix A

Exploring entertainment utility from football games

For our research purposes, we use in-game *Betfair* exchange data sourced via *Fracsoft*. *Betfair*, a peer-to-peer platform, is the largest online betting exchange in the world. Unlike alternative bookmakers, *Betfair* prices (odds) are readily available via the *Betfair* Application Programming Interface (API). Specifically, we work with csv files, compiled by *Fracsoft*, which include match descriptions, scheduled and actual game times, timestamps (UTC) for each price movement, an in-play dummy variable, betting market status (open or closed), and volumes and odds available for each selection (home, away, and draw). Focusing on the 2013/2014 season of the English Premiere League, we collect data for all 380 matches between the 20 EPL teams. With price movements and betting volumes featuring granularity to the millisecond, we use in-match odds to impute the real-time outcome probabilities that are eventually used to measure our emotional cues – shock, surprise, suspense – through the course of a match. Preceding the calculation of these cues, the following procedures are performed using python scripts toward the Fracsoft dataset, aggregated by the minute.

Appendix A.1 Prices and Actual Probabilities

For any given minute within a match there may be more than one unique exercised price match. In this case, a pre-determined aggregation function must be used to determine the value we use. We employ three different methods. Mean, the strategy we default to for our calculations, is simply the mean of every unique price match found in the minute. Similarly, we explore the median of these set of price matches by the minute. Finally, we consider a scheme weighted by volume that we've named "effective odds". Alternative to the other methods, the "effective odds" are proportional to the volumes of each price match. For example, suppose we find that a given minute has price matches for 1.5 and 2. Respectively, the volumes traded are 10 and 90. Using the mean method we would find the aggregated price match to be 1.75 regardless of volume. In contrast, we use volume proportional pricing to calculate 1.95 as our effective odds. Intuitively, these odds tend toward the price match with the higher amount of volume. In practice, however, we found little to no difference in our subsequent results. As a result, we opt to present mean, the more straight-forward aggregation method.

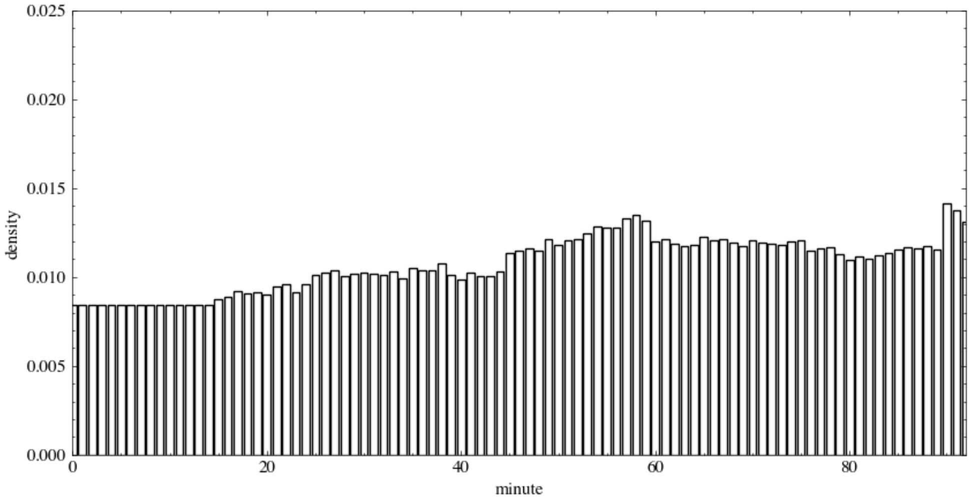
Once the odds are appropriately aggregated, we then derive implied outcome probabilities by taking the inverse odds. In theory, the sum of the inverse odds for every possible outcome should be equal to 1. In reality, however, the sum of these values is slightly above 1. The difference between the sum of implied outcome probabilities and 1 is known as the overround or vig – essentially the bookmakers fee or commission. In order to remove this overround, we proportionally scale the derived outcome probabilities. More specifically, we divide each outcome probability by the sum of all outcome probabilities.

Appendix A.2 Simulated Probabilities

While some studies have shown that *Betfair*, or betting exchange, prices, accurately predict outcomes (Croxson and Reade, 2014), others have rejected the hypothesis of semi-strong market efficiency. For instance, Choi and Hui (2014) found that prices generally underreact to normal news and overreact to surprising news. Such market inefficiencies are also detected by Angelini et al. (2022). In summary, these findings question the overall suitability of using observable (*Betfair*) prices for predicting outcomes in our study. As such, we use in-play odds derived from an in-play model as proposed by Buraimo et al. (2020) in our main specification.

Assuming an independently Poisson distributed number of goals scored by home and away, we estimate team-specific scoring rates by minimizing the squared difference between the bookmaker implied probabilities (imputed using odds on over/under totals) and the outcome probabilities from the in-play model. Furthermore, rather than assuming identical distribution of goals between every minute in a match, we use our goals scored data to distribute historical scoring rates across the minutes. Accounting for the average amount of injury time - and to share out the inflated scoring rates found in the 45th and 90th minute - we presume all matches to be 93 minutes long. Then we calculate a moving average over 15 min to smooth the relative frequency distributions. Using backward interpolation, we fill in the missing values for the first 15 minutes. Finally, we calculate the density function of goals scored per minute.

FIGURE A.1: Goal Scores Density Plot



Notes: This Figure depicts the density plot of the total number of goals scored per minute for all 380 matches played in season 2013/2014.

With goal distributions and team-specific scoring rates as our main ingredients, we execute the match simulations that enable us to calculate hypothetical probabilities. For every match in our dataset, we simulate the number of goals in each minute, sum up the score line, and record the result. We repeat this simulation 100,000 times per minute per match (9 million simulations for every match). In pursuance of runtime reduction, we execute concurrent simulations using python's built-in multithreading packages and distributed computation. For any given minute, the respective outcome probabilities are represented by the number of simulations with each outcome, given the current score, divided by the total number of simulations.

Appendix A.3 Events

Match events depicted throughout our report were sourced from match commentaries supplied by whoscored.com. Other sources considered included BBC and ESPN, however, whoscored.com proved to have the most extensive database. Exactly 12 event types were collected from the public site; events include goal scored, save made, card received, offside, corner, attempt missed, attempt blocked, woodwork hit, substitution off, substitution on, start half, and end half. Most notably, the events "goal scored" and "red card received" were used during analysis and to generate both in-play models (explained in further depth later) respectively. The predominant tools used to acquire this data were python and selenium, a python package primarily used for test automation. Presented in order, data collection included a complete acquisition of urls corresponding to every unique match found in the Odds dataset. Then, using the browser automation enabled by selenium, we systematically scanned each of the gathered urls for in-game commentaries with timestamps, down to the second, included. In addition to having the most exhaustive catalog, whoscored.com commentaries appear to have the most granular timestamps. This added granularity proved to be crucial in our in-game analysis. Finally, the data acquired was packaged into individual xml files corresponding to each match. Using an element tree structure, every commentary entry is presented as a sub element of the larger commentary tree. Root attributes include away team, home team, season, season id, game date and time, league name, sport name, and language. Sub elements include the attributes comment, period, minute, second, expanded minute, and event type. With the complete timestamps included in these xml files, the data is ultimately merged into the master dataset and adjusted by the historical start times found in the Fracsoft dataset. Since we aggregate events by minute as well, we transform single events into a comma delimited string of events in chronological order. Neutral events such as the start and end of a half are regarded as home team events.

Appendix A.4 Sentiment Analysis and Fandom

The random forest - a supervised machine learning model - is an ensemble of decision trees popularly used for sentiment analysis. Although deep learning Neural Nets like the Long Short Term Memory algorithm can hypothetically outperform tree-based models for sentiment analysis, they are “black box” approaches with no discernable feature importances. Alternatively, the random forest shows how important individual features are in determining outcomes. Our machine learning scripts use the python packages nltk and sklearn. Specifically, nltk was used for preprocessing and we employed sklearn for model training and evaluation. Since we’re interested in obtaining sentiment magnitudes, we use a regressor rather than a categorical classifier.

Our sentiment analysis model is trained on data from Stanford’s Sentiment Tree Bank. After considering multiple open-source sentiment datasets, we found that the Stanford data consisted of more realistic use cases perhaps more relevant to our own twitter data. Before training, we follow traditional NLP prepossessing protocols. We remove English stop words, replace broken conjugations, and remove noise caused by encoding errors. Again, using nltk, we lemmatize in order to find root words, and use sklearns TfidfVectorizer to extract our feature set. We include our entire twitter dataset into our corpus to ensure all features are extracted. Instead of training on all the observations, each tree of RF is trained on a subset of the observations.

After performing an extensive grid search to tune model hyperparameters, our final specification is used to generate sentiment scores, ranging from 0 to 25, for every tweet associated with every match in our dataset. In summary, in our winning model (accuracy: 70.08%, MAE: 2.49, MSE: 13.88), the maximal depth of a tree, which is defined as the longest path between the root node and the leaf node, was 90. We used 'auto' for the number of features to consider when looking for the best split. The minimum number of samples required to be at a leaf node was two, the minimum number of samples required to split an internal node was nine, and the number of trees in the forest was 550. As described in Section 2, we curate a rule set in order to assign fan association as well as define haters and neutral spectators.

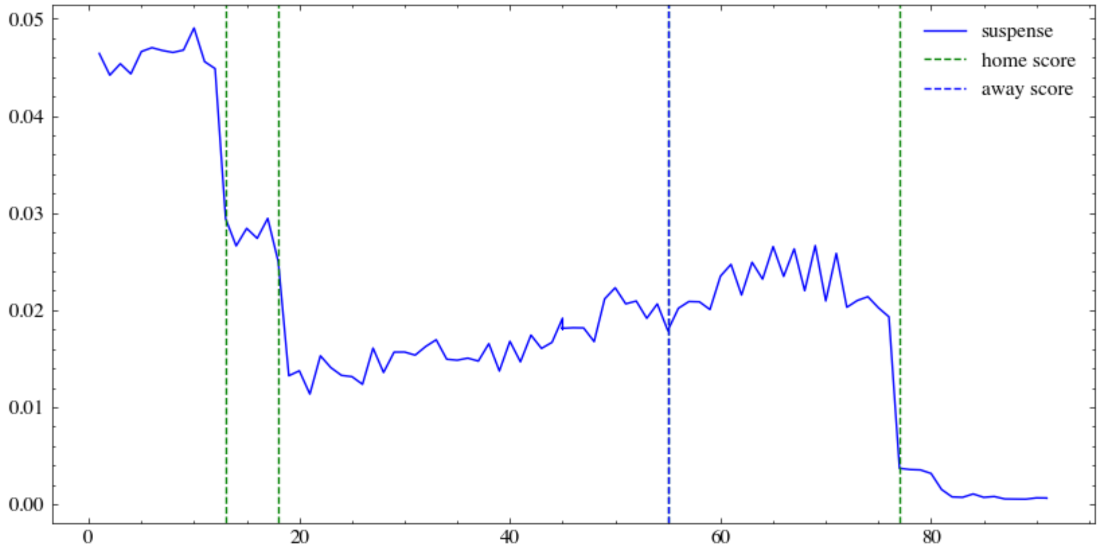
Appendix A.5 Shock and Surprise

The two backward looking emotional cues, shock and surprise, are similar in their calculations. Surprise essentially refers to difference in beliefs relative to preceding events, whereas shock captures the contrast between in-game beliefs and the initial projections. Accordingly, the formula for surprise is the square root of the sum of the squared differences in outcome probabilities and outcome probabilities of the previous time period. Moreover, shock is calculated by taking the square root of the sum of the squared differences in outcome probabilities and outcome probabilities pre match. The key lies in the time dimension of the anterior reference point. For surprise, the outcome probabilities are subtracted by the outcome probabilities of the previous time period. On the other hand, the probabilities are subtracted by a static, pre-match observation. For both measures, we do these operations for all 3 of the aforementioned price aggregations.

Appendix A.6 Suspense

Suspense, the forward-looking measure, requires more computation than its backward-looking counterparts. In this regard, we can't rely on historical observations to formulate an in-play model; instead, we take a simulation-based approach very similar to the procedures described in section A.2. With these procedures in place, we leverage simulated match outcomes to calculate scenario contingent, hypothetical probabilities. Essentially, we find the probabilities for a home win, draw, or away win if either team scores in the next minute. For every match we iterate through each minute and find the likelihood of home (away) win given a home (away) goal. To

FIGURE A.2: Suspense with Events - Liverpool v West Brom



Notes: This Figure uses a match between Liverpool and West Brom in 2013/2014 as an example to display suspense throughout a given match. The plots shows a massive decrease in the forward looking emotional cue following the 2-0 lead for the away team. From there suspense continues to rise until the away team scores again, solidifying their victory.

do so we simply isolate the simulations with home (away) goals appearing within the next minute and find the respective proportion of home wins, draws, and away wins given the current score line. We then square the difference between these and the given in-play model probabilities for home win, draw, and away win for the minute. Next, we multiply this squared difference by the probability of a home (away) goal in the next minute and sum all of values found for each outcome. Finally, we define the minute's suspense measure by the square root of the sum of these sums.

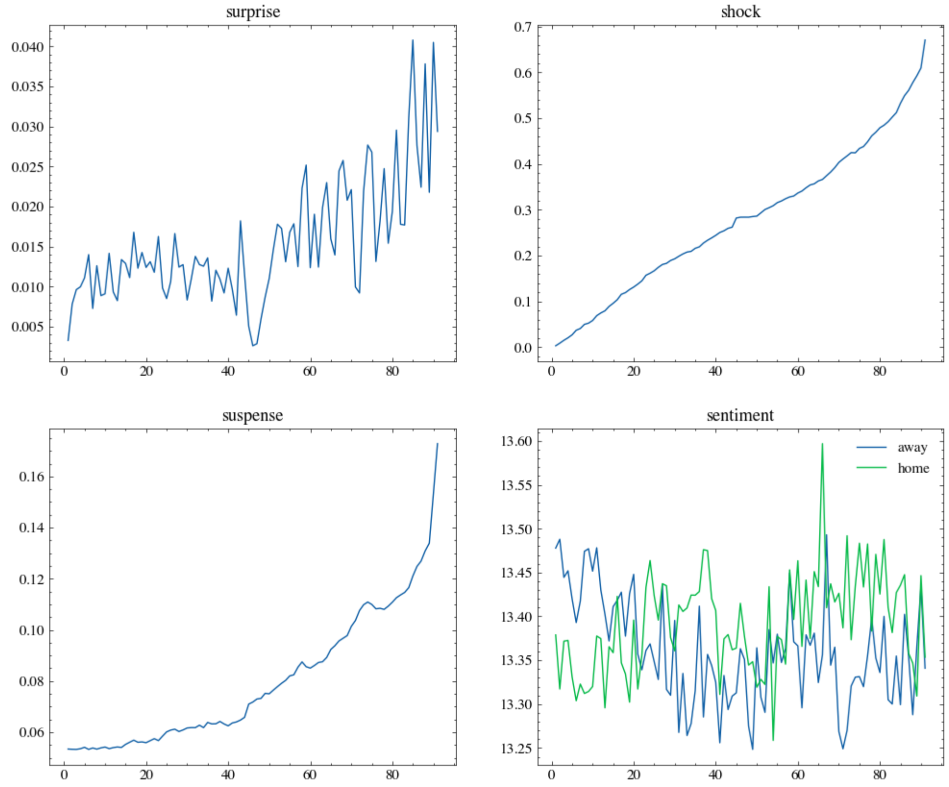
Appendix A.7 All Cues

TABLE A.1: Summary Statistics

	Mean	SD	Min	Med	Max
Suspense	0.080578	0.055328	0	0.069045	0.368522
Surprise	0.014081	0.055565	0	0.004341	1.338340
Shock	0.285513	0.222329	0	0.252127	1.168169

Notes.- generated using the probabilities calculated with the effective price match.

FIGURE A.3: Aggregated Cues By Minute



Notes: This Figure plots the average value for each emotional cue by minute in a match for all 380 matches played in season 2013/2014.

Bibliography

Abinzano, I., L. Muga, and R. Santamaria. 2016. “Game, set and match: the favourite-long shot bias in tennis betting exchanges.” *Applied Economics Letters*, 23(8): 605–608.

Abinzano, I., L. Muga, and R. Santamaria. 2019. “Hidden Power of Trading Activity: The FLB in Tennis Betting Exchanges.” *Journal of Sports Economics*, 20(2): 261–285.

Ali, M. M. 1977. “Probability and Utility Estimates for Racetrack Bettors.” *Journal of Political Economy*, 85(4): 803–815.

Anderson, R., A. P. Gema, Suharjito, and S. M. Isa. 2018. “Facial attractiveness classification using deep learning.” In *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*. 34–38.

Angelini, G., L. De Angelis, and C. Singleton. 2022. “Informational efficiency and behaviour within in-play prediction markets.” *International Journal of Forecasting*, 38(1): 282–299.

Angelini, G., V. Candila, and L. De Angelis. 2021a. “Weighted Elo rating for tennis match predictions.” *European Journal of Operational Research*, Forthcoming.

Angelini, G., and L. De Angelis. 2019. “Efficiency of online football betting markets.” *International Journal of Forecasting*, 35(2): 712–721.

Angelini, G., L. De Angelis, and C. Singleton. 2021b. “Informational efficiency and behaviour within in-play prediction markets.” *International Journal of Forecasting*, Forthcoming.

Ann, B., and J. Lee. 2014. “Beauty and productivity: The case of the ladies professional golf association.” *Economics & Human Biology*, 12 1–12.

Avery, C. N., J. A. Chevalier, and R. J. Zeckhauser. 2016. “The CAPS Prediction System and Stock Market Returns.” *Review of Finance*, 20(4): 1363–1381.

Babad, E., and Y. Katz. 1991. “Wishful thinking—against all odds.” *Journal of Applied Social Psychology*, 21(23): 1921–1938.

Baccouche, M., F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. 2010. “Action classification in soccer videos with long short-term memory recurrent neural networks.” 154–159, 09.

Bakkenbull, J., and E. Kiefer. 2015. “Are attractive female tennis players more successful? an empirical analysis.” *Journal of Sports Science*, 25(4): 567–586.

Bakkenbüll, L.-B. 2017. “The impact of attractiveness on athletic performance of tennis players.” *International Journal of Social Science Studies*, 5, p. 12.

Barnett, T., and S. R. Clarke. 2005. “Combining player statistics to predict outcomes of tennis matches.” *IMA Journal of Management Mathematics*, 16(2): 113–120.

Barrutiabengoa, J. M., P. Corredor, and L. Muga. 2022. “Does the betting industry price gender? evidence from professional tennis.” *Journal of Sports Economics*, 23(7): 881–906.

Behrendt, S., F. J. Peter, and D. J. Zimmermann. 2020. “An encyclopedia for stock markets? Wikipedia searches and stock returns.” *International Review of Financial Analysis*, 72, p. 101563.

Berggren, N., H. Jordahl, and P. Poutvaara. 2010. “The looks of a winner: Beauty and electoral success.” *Journal of public economics*, 94(1-2): 8–15.

Berri, D. J., R. Simmons, J. Van Gilder, and L. O’Neill. 2011. “What does it mean to find the face of the franchise? physical attractiveness and the evaluation of athletic performance.” *International Journal of Sport Finance*, 6(3): 184–202.

Bizzozero, P., R. Flepp, and E. Franck. 2016. “The importance of suspense and surprise in entertainment demand: Evidence from Wimbledon.” *Journal of Economic Behavior & Organization*, 130 47–63.

Brown, A., D. Rambaccussing, J. J. Reade, and G. Rossi. 2018. “Forecasting With Social Media: Evidence From Tweets On Soccer Matches.” *Economic Inquiry*, 56(3): 1748–1763.

Brown, A., and J. J. Reade. 2019. “The wisdom of amateur crowds: Evidence from an online community of sports tipsters.” *European Journal of Operational Research*, 272(3): 1073–1081.

Buraimo, B., D. Forrest, I. McHale, and J. Tena. 2020. “Unscripted drama: Soccer audience response to suspense, surprise, and shock.” *Economic Inquiry*, 58(2): 881–896.

Candila, V. 2021. *welo: Weighted and Standard Elo Rates*. R package version 0.1.0.

Candila, V., and A. Scognamillo. 2018. “Estimating the Implied Probabilities in the Tennis Betting Market: A New Normalization Procedure.” *International Journal of Sport Finance*, 13(3): 225–242.

Caplin, A., and J. Leahy. 2001. “Psychological expected utility theory and anticipatory feelings.” *The Quarterly Journal of Economics*, 116(1): 55–79.

Carre, J. M., and C. M. McCormick. 2008. “In your face: Facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players.” *Proceedings of the Royal Society B: Biological Sciences*, 275(1651): 2651–2656.

Cashmore, V., N. Coster, D. Forrest, I. McHale, and B. Buraimo. 2022. “Female jockeys - what are the odds?” *Journal of Economic Behavior and Organization*, 202 703–713.

Ceinos, R., L. Lupi, A. Tellier, and M. F. Bertrand. 2017. “Three-dimensional stereophotogrammetric analysis of 50 smiles: A study of dento-facial proportions.” *Journal of Esthetic and Restorative Dentistry*, 29(6): 416–423.

Chen, H., P. De, Y. J. Hu, and B.-H. Hwang. 2014. “Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media.” *Review of Financial Studies*, 27(5): 1367–1403.

Choi, D., and S. Hui. 2014. “The role of surprise: Understanding overreaction and underreaction to unanticipated events using in-play soccer betting market.” *Journal of Economic Behavior & Organization*, 107 614–629.

Croxson, K., and J. Reade. 2014. “Information and Efficiency: Goal Arrival in Soccer Betting.” *Economic Journal*, 124 62–91.

del Corral, J., and J. Prieto-Rodríguez. 2010. “Are differences in ranks good predictors for Grand Slam tennis matches?.” *International Journal of Forecasting*, 26(3): 551–563.

Dietl, H., G. Ozdemir, and A. Rendall. 2019. “The role of facial attractiveness in tennis tv-viewership.” *Journal of Sports Economics*, 20(5): 703–720.

Dillenberger, D. 2010. “Preferences for one-shot resolution of uncertainty and Allais-type behavior.” *Econometrica*, 78(6): 1973–2004.

Easton, S., and K. Uylangco. 2010. “Forecasting outcomes in tennis matches using within-match betting markets.” *International Journal of Forecasting*, 26(3): 564–575.

Eisenthal, Y., G. Dror, and E. Ruppin. 2006. “Facial attractiveness: Beauty and the machine.” *Neural computation*, 18 119–42.

Elaad, G., J. J. Reade, and C. Singleton. 2020. “Information, prices and efficiency in an online betting market.” *Finance Research Letters*, 35.

Elo, A. E. 1978. *The rating of chessplayers, past and present*. London Batsford.

Ely, J., A. Frankel, and E. Kamenica. 2015. “Suspense and surprise.” *Journal of Political Economy*, 123(1): 215–260.

Fama, E. F. 1965. “The Behavior of Stock-Market Prices.” *The Journal of Business*, 38(1): 34–105.

Fama, E. F. 1970. “Efficient Capital Markets: A Review of Theory and Empirical Work.” *Journal of Finance*, 25(2): 383–417.

Fischer, L., A. Kelava, M. Nagel, and T. Pawlowski. 2023. “Celebration beats frustration – emotional cues and alcohol use during soccer matches.” *University of Tübingen: mimeo*.

Forrest, D. 2008. “Soccer Betting in Britain.” In *Handbook of Sports and Lottery Markets*. Eds. by D. B. Hausch, and W. T. Ziemba, San Diego Elsevier, 421–446.

Forrest, D., and I. McHale. 2007. “Anyone for Tennis (Betting)?.” *The European Journal of Finance*, 13(8): 751–768.

Galton, F. 1907. “Vox Populi.”

Garcia, M. 2016. “Racist in the machine: The disturbing implications of algorithmic bias.” *World Policy Journal*, 33(4): 111–117.

Grammer, K., B. Fink, A. P. Møller, and R. Thornhill. 2003. “Darwinian aesthetics: sexual selection and the biology of beauty.” *Biological reviews*, 78(3): 385–407.

Gul, F. 1991. “A theory of disappointment aversion.” *Econometrica: Journal of the Econometric Society* 667–686.

Guo, G., B. R. Humphreys, Q. Wang, and Y. Zhou. 2023. “Attractive or aggressive? a face recognition and machine learning approach for estimating returns to visual appearance.” *Journal of Sports Economics*, 0(0): , p. 15270025231160769.

Hamermesh, D. S., and A. M. Parker. 2005. “Beauty in the classroom: instructors’ pulchritude and putative pedagogical productivity.” *Economics of Education Review*, 24(4): 369–376.

Hammermesh, D. S., and J. E. Biddle. 1994. “Beauty and the labor market.” *American Economic Review*, 84(5): 1174–1194.

He, X.-Z., and N. Treich. 2017. “Prediction market prices under risk aversion and heterogeneous beliefs.” *Journal of Mathematical Economics*, 70(C): 105–114.

Hvattum, L. M., and H. Arntzen. 2010. “Using ELO ratings for match result prediction in association football.” *International Journal of Forecasting*, 26(3): 460–470.

Jiang, H., Y. Lu, and J. Xue. 2016. “Automatic soccer video event detection based on a deep neural network combined cnn and rnn.” In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*. 490–494.

Jiwa, M., P. Cooper, T. T.-J. Chong, and S. Bode. 2021. “Choosing increases the value of non-instrumental information.” *Scientific Reports*, 11(8780): .

Judge, T. A., C. Hurst, and L. S. Simon. 2009. “Does it pay to be smart, attractive, or confident (or all three)? relationships among general mental ability, physical attractiveness, core self-evaluations, and income..” *The Journal of applied psychology*, 94 3 742–55.

Kagian, A., G. Dror, T. Leyvand, D. Cohen-or, and E. Ruppin. 2006. “A humanlike predictor of facial attractiveness.” In *Advances in Neural Information Processing Systems*. Eds. by B. Schölkopf, J. Platt, and T. Hoffman, 19 MIT Press.

Kamel, D., and L. G. Woo-Mora. 2023. “Skin tone penalties: Bottom-up discrimination in football.” *Available at SSRN 4537612*.

Kaplan, S. 2021. “Entertainment utility from skill and thrill.” *SSRN Working Paper*, doi.org/10.2139/ssrn.3888785.

Kelly, J. L. 1956. “A new interpretation of information rate.” *The Bell System Technical Journal*, 35(4): 917–926.

Kiefer, S., and K. Scharfenkamp. 2012. “The impact of physical attractiveness on the popularity of female tennis players in online media.” *International Journal of Sport Communication*, 5(1): 1–17.

Knottenbelt, W. J., D. Spanias, and A. M. Madurska. 2012. “A common-opponent stochastic model for predicting the outcome of professional tennis matches.” *Computers & Mathematics with Applications*, 64(12): 3820–3827.

Kőszegi, B., and M. Rabin. 2009. “Reference-dependent consumption plans.” *American Economic Review*, 99(3): 909–36.

Kovalchik, S. 2020. “Extension of the Elo rating system to margin of victory.” *International Journal of Forecasting*, 36(4): 1329–1341.

Kovalchik, S. A. 2016. “Searching for the GOAT of tennis win prediction.” *Journal of Quantitative Analysis in Sports*, 12(3): 127–138.

Kovalchik, S., and M. Reid. 2019. “A calibration method with dynamic updates for within-match forecasting of wins in tennis.” *International Journal of Forecasting*, 35(2): 756–766.

Kraaijeveld, O., and J. De Smedt. 2020. “The predictive power of public Twitter sentiment for forecasting cryptocurrency prices.” *Journal of International Financial Markets, Institutions and Money*, 65(C): .

Kramer, R. S. S., and R. Ward. 2010. “Internal facial features are signals of personality and health.” *Quarterly Journal of Experimental Psychology*, 63(11): 2273–2287, PMID: 20486018.

Kreps, D., and E. Porteus. 1978. “Temporal resolution of uncertainty and dynamic choice theory.” *Econometrica: journal of the Econometric Society* 185–200.

Lahvička, J. 2014. “What causes the favourite-longshot bias? Further evidence from tennis.” *Applied Economics Letters*, 21(2): 90–92.

Langlois, J. H., L. Kalakanis, A. J. Rubenstein, A. Larson, M. Hallam, and M. Smoot. 2000. “Maxims or myths of beauty? a meta-analytic and theoretical review..” *Psychological bulletin*, 126(3): , p. 390.

Langlois, J., L. Kalakanis, A. Rubenstein, A. Larson, M. Hallam, and M. Smoot. 2000. “Maxims or myths of beauty? a meta-analytic and theoretical review.” *Psychological bulletin*, 126 390–423.

Larsen, T., J. Price, and J. Wolfers. 2008. “Racial bias in the nba: Implications in betting markets.” *Journal of Quantitative Analysis in Sports*, 4(2): .

Li, W., H. Zhu, K. Zhao, H. Zhu, X. Wang, and X. He. 2023. “Good performance-high attractiveness effect: an empirical study on the association between athletes’ rankings and their facial attractiveness.” *International Journal of Sport and Exercise Psychology*, 0(0): 1–26.

Liang, L., L. Lin, L. Jin, D. Xie, and M. Li. 2018. “Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction.”

Liu, X., M. Shum, and K. Uetake. 2021. “Passive vs. active attention to baseball telecasts: implications for content (re-)design..” *SSRN Working Paper*, doi.org/10.2139/ssrn.3717894.

Loewenstein, G. 2000. “Emotions in economic theory and economic behavior.” *American Economic Review*, 90(2): 426–432.

Lucas, G. M., J. Gratch, N. Malandrakis, E. Szablowski, E. Fessler, and J. Nichols. 2017. “GOAALLL: Using sentiment in the world cup to explore theories of emotion.” *Image and Vision Computing*, 65 58–65.

Lyócsa, S., and T. Výrost. 2018. “To bet or not to bet: a reality check for tennis betting market efficiency.” *Applied Economics*, 50(20): 2251–2272.

Manski, C. 2006. “Interpreting the predictions of prediction markets.” *Economics Letters*, 91(3): 425–429.

McHale, I., and A. Morton. 2011. “A Bradley-Terry type model for forecasting tennis match results.” *International Journal of Forecasting*, 27(2): 619–630.

Meier, H. E., M. Konjer, and M. Leinwather. 2016. “The demand for women’s league soccer in germany.” *European Sport Management Quarterly*, 16(1): 1–19.

Metz, C. 2021. “Who is making sure the a.i. machines aren’t racist?” *The New York Times*.

Mincer, J., and V. Zarnowitz. 1969. “The evaluation of economic forecasts.” In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*. NBER, 1–46.

Moat, H. S., C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis. 2013. “Quantifying Wikipedia usage patterns before stock market moves.” *Scientific reports*, 3(1): 1–5.

Mobius, M. M., and T. S. Rosenblat. 2006. "Why beauty matters." *American Economic Review*, 96(1): 222–235.

Neale, W. C. 1964. "The Peculiar Economics of Professional Sports." *Quarterly Journal of Economics*, 78(1): 1–14.

Newall, P. W. S., and D. Cortis. 2021. "Are Sports Bettors Biased toward Longshots, Favorites, or Both? A Literature Review." *Risks*, 9(1): , p. 22.

O'Brien, T. L., and E. He. 2021. "The sports gambling gold rush is absolutely off the charts." *Bloomberg*.

O'Shea, K., and R. Nash. 2015. "An introduction to convolutional neural networks." *ArXiv e-prints*.

Ottaviani, M., and P. N. Sørensen. 2008. "The Favorite-Longshot Bias: An Overview of the Main Explanations." In *Handbook of Sports and Lottery Markets*. Eds. by D. B. Hausch, and W. T. Ziemba, San Diego Elsevier, 83–101.

Ottaviani, M., and P. N. Sørensen. 2015. "Price Reaction to Information with Heterogeneous Beliefs and Wealth Effects: Underreaction, Momentum, and Reversal." *American Economic Review*, 105(1): 1–34.

Palacios-Huerta, I. 1999. "The aversion to the sequential resolution of uncertainty." *Journal of Risk and Uncertainty*, 18(3): 249–269.

Parkhi, O. M., A. Vedaldi, and A. Zisserman. 2015. "Deep face recognition." In *British Machine Vision Conference*.

Paul, R., A. Weinbach, and J. Mattingly. 2018. "Tests of racial discrimination in a simple financial market: Managers in major league baseball." *International Journal of Financial Studies*, 6, p. 24.

Pawlowski, T., G. Nalbantis, and D. Coates. 2018. "Perceived game uncertainty, suspense and the demand for sport." *Economic Inquiry*, 56(1): 173–192.

Peeters, T. 2018. "Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results." *International Journal of Forecasting*, 34(1): 17–29.

Pfann, G. A., J. E. Biddle, D. S. Hamermesh, and C. M. Bosman. 2000. "Business success and businesses' beauty capital." *Economics Letters*, 67(2): 201–207.

Pillutla, M. M., and J. K. Murnighan. 1996. "The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies." *Academy of Management Journal*, 39(2): 286–311.

Prusinkiewicz, P., and A. Lindenmayer. 2012. *The algorithmic beauty of plants*. Springer Science & Business Media.

Rahmad, N., N. A. J. Sufri, N. Muzamil, and M. A. As'ari. 2019. "Badminton player detection using faster region convolutional neural network." *Indonesian Journal of Electrical Engineering and Computer Science*, 14 1330–1335.

Raikes, J. 2023. "Ai can be racist: Let's make sure it works for everyone." *Forbes*.

Ramirez, P., J. J. Reade, and C. Singleton. 2023. "Betting on a buzz: Mispricing and inefficiency in online sportsbooks." *International Journal of Forecasting*, 39(3): 1413–1423.

Raney, A. A. 2018. "Why we watch and enjoy mediated sports." In *Handbook of sports and media*. Eds. by A. A. Raney, and J. Bryant, New York, NY Lawrence Erlbaum Associates, 313–329.

Richardson, T., G. Nalbantis, and T. Pawlowski. 2023. "Emotional cues and the demand for televised sports: Evidence from the UEFA Champions League." *Journal of Sports Economics*, forthcoming.

Rosar, U., J. Hagenah, and M. Klein. 2017. "Physical attractiveness and monetary success in german bundesliga." *Soccer & Society*, 18(1): 102–120.

Rossetti, A., M. De Menezes, R. Rosati, V. F. Ferrario, and C. Sforza. 2013. "The role of the golden proportion in the evaluation of facial esthetics." *The Angle Orthodontist*, 83(5): 801–808.

Rottenberg, S. 1956. "The Baseball Players' Labor Market." *The Journal of Political Economy*, 64(3): 242–258.

Scheibehenne, B., and A. Broder. 2007. "Predicting wimbledon 2005 tennis results by mere player name recognition." *International Journal of Forecasting*, 23(3): 415–426.

Scholz, M., and K. Sicinski. 2015. "Facial attractiveness and lifetime earnings: Evidence from a cohort study." *Journal of Business and Psychology*, 30(2): 287–302.

Simonov, A., R. M. Ursu, and C. Zheng. 2022. "EXPRESS: Suspense and surprise in media product design: Evidence from Twitch.tv.." *Journal of Marketing Research*, doi.org/10.1177/00222437221108653.

Snowberg, E., and J. Wolfers. 2010. "Explaining the Favorite-Long Shot Bias: Is it Risk-Love or Misperceptions?." *Journal of Political Economy*, 118(4): 723–746.

Spanias, D., and W. J. Knottenbelt. 2013. "Predicting the outcomes of tennis matches using a low-level point model." *IMA Journal of Management Mathematics*, 24(3): 311–320.

Sprenger, T. O., A. Tumasjan, P. G. Sandner, and I. M. Welpe. 2014. "Tweets and Trades: the Information Content of Stock Microblogs." *European Financial Management*, 20(5): 926–957.

Stinebrickner, T. R., R. Stinebrickner, and W. R. Stinebrickner. 2019. "Beauty, job tasks, and wages: A new conclusion about employer taste-based discrimination." *Journal of Human Resources*, 54(2): 237–260.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Brown Little.

Thapa, G. B., and R. Thapa. 2018. "The relation of golden ratio, mathematics and aesthetics." *Journal of the Institute of Engineering*, 14(1): , p. 188–199.

Tsujimura, H., and M. Banissy. 2013. "Human face structure correlates with professional baseball performance: Insights from professional japanese baseball players." *Biology letters*, 9, p. 20130140.

Vaughan Williams, L. 1999. "Information Efficiency in Betting Markets: A Survey." *Bulletin of Economic Research*, 51(1): 1–30.

Vaughan Williams, L., M. Sung, P. A. F. Fraser-Mackenzie, J. Peirson, and J. E. V. Johnson. 2018. "Towards an Understanding of the Origins of the Favourite–Longshot Bias: Evidence from Online Poker Markets, a Real-money Natural Laboratory." *Economica*, 85(338): 360–382.

Verma, P. 2022. "These robots were trained on ai. they became racist and sexist.." *The Washington Post*.

Webster Jr, M., and J. E. Driskell Jr. 1983. "Beauty as status." *American Journal of sociology*, 89(1): 140–165.

Willis, J., and A. Todorov. 2006. "First impressions: Making up your mind after a 100-ms exposure to a face." *Psychological Science*, 17(7): 592–598.

Woodland, L. M., and B. M. Woodland. 1994. "Market efficiency and the favorite-longshot bias: The baseball betting market." *The Journal of Finance*, 49(1): 269–279.

Woodland, L. M., and B. M. Woodland. 2011. "The reverse favorite-longshot bias in the national hockey league: Do bettors still score on longshots?" *Journal of Sports Economics*, 12(1): 106–117.

Xiao, Q., Y. Wu, D. Wang, Y.-L. Yang, and X. Jin. 2021. "Beauty3dfacenet: Deep geometry and texture fusion for 3d facial attractiveness prediction." *Computers Graphics*, 98 11–18.

Xu, J., L. Jin, L. Liang, Z. Feng, D. Xie, and H. Mao. 2017. "Facial attractiveness prediction using psychologically inspired convolutional neural network (pi-cnn)." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1657–1661.

Ziemba, W. T. 2020. "Parimutuel betting markets: racetracks and lotteries revisited." SRC Discussion Paper 103, Systemic Risk Centre, London School of Economics.

Zillmann, D., and J. R. Cantor. 1972. "Directionality of transitory dominance as a communication variable affecting humor appreciation." *Journal of Personality and Social Psychology*, 24(2): 191–198.

Zou, J., and L. Schiebinger. 2018. “AI can be sexist and racist — it’s time to make it fair.” *Nature*, 559(7714): 324–326.