Profitability in Sports Betting: A Case Study of Men's Tennis

By

Robert Ciobanu

An Undergraduate Thesis submitted in partial fulfillment of the requirements for the

JOSEPH WHARTON SCHOLARS

Faculty Advisor:

Eric T. Bradlow

K.P. Chao Professor, Professor of Marketing, Statistics and Data Science, Education and Economics, Chairperson of Wharton's Marketing Department, and Vice-Dean of Analytics at Wharton

THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA

MARCH 2023

ABSTRACT

This paper evaluates the feasibility of profit generation through sports betting. While sports gambling represents a large and rapidly growing economic sector, few bettors are actually profitable and there is limited evidence of successful publicly available strategies. We investigate how such a strategy can be built for the game of men's tennis. Our methodology for creating a strategy consists of two components. First, it includes a predictive analytics component, in which we combine a large number of observable player, match, and tournament characteristics in order to estimate the probability of either player winning the match. We study both linear and non-linear multivariate combination approaches. Second, our methodology contains a financial strategy component, in which we focus on using money allocation techniques to achieve optimal returns.

Through statistical simulations and back-testing, we find that it is possible to generate positive expected profits at sustainable levels of risk, with both formal and informal strategies. Interestingly, we also establish that building a successful strategy does not necessarily require the bettor's model to have higher predictive accuracy than the betting markets. Instead, bettors can focus on a narrow segment of matches (for example "upsets" – i.e., matches in which the lower-ranked player wins) and outperform the market in that segment alone. We conclude that sports betting can be used as a profitable investment vehicle. Beyond tennis, these techniques can be applied to most other sports, especially those for which large volumes of historical data are publicly available.

INTRODUCTION

This paper focuses on exploring the profitability of sports gambling, particularly as it pertains to the game of men’s singles tennis. The sports betting market was valued at approximately $84 billion in 2022 and is projected to reach $91 billion by the end of 2023, with an estimated compounded annual growth rate of 10.4% for the next eight years (Grand View Research 2023). A study from the Pew Research Center (Gramlich 2022) further found that 19% of U.S. adults admit to having bet money on sports in the past year. Despite this, research estimates that only 3-5% of sports bettors are profitable in the long-run (Bruce 2021). The question begs itself: is there a way to generate consistent returns through sports betting, or is there truly no financial basis for this type of gambling? In order to answer this question, this research evaluates the feasibility of profitably making money by betting on the game of men’s tennis. The specific research approach will be to use a variety of statistical models to predict upsets (matches in which the lower-ranked player wins), and then use financial theory and risk management techniques to determine how to invest. The reasons for solely focusing on men’s tennis are trifold. First, the difficulty of the paper’s objective requires choosing a very narrow segment within sports betting, given that outcome prediction and odds creation require highly specific variables and tailored models. Models in sports betting are not “one-size-fits-all”, and finding even the slightest edge requires in-depth knowledge and analysis of a particular sport. Second, tennis was selected as it is one of the most popular sports in the world, with approximately 1.17% of the world population (87 million people) playing or having played it (Czermak 2021). Lastly, it is the third most popular sport that people bet on, after football and horse racing (Toogood 2022). The following section will provide a brief overview of the generalities of this sport, in order to set the framework for the model creation.

Generalities of the Game of Tennis

Believed to have originated from 12th century France and the "jeu de paume", tennis has since become one of the world's most famous sports with over a billion fans worldwide (Veroutsos 2022). Men's professional tennis is currently divided into three main tiers, ranked by prestige in the following order: ATP[1] Tour, ATP Challenger Tour, and ITF Men's World Tennis Tour. This paper will focus on tournaments within the ATP Tour, as these have the largest datasets available for analytical purposes. Within the ATP Tour, there are eight different types of tournaments: Grand Slams, ATP Finals, ATP Masters 1000s, ATP 500s, ATP 250s, the Davis Cup, the United Cup, and the Olympics (which takes place once every four years). These tournaments take place in over 30 different countries and are played either indoors or outdoors on three different types of surfaces: hard court, clay, and grass.

Player rankings are calculated via a system of points ("ATP points") which are granted after each match won. The specific number of points won depends on the tournament played and the stage of the tournament reached by the player; for example, a player reaching the second round of an ATP 250 will not receive the same number of points as an individual winning a Grand Slam. The numbers "250" or "500" refer to the number of points a player receives if he wins such a tournament. For a Grand Slam, this number is 2,000. Points are accumulated over a 52-week basis, with a given player's 19 best tournaments being counted towards his ranking. These player rankings will then be used to create tournament brackets intended to separate high-ranked players as much as possible; for example, if the world number 1 and the world number 2

[1] ATP stands for Association of Tennis Professionals

decide to compete in the same tournament, they will only be able to meet in the final. In theory, this should lead to the first rounds of a tournament being relatively easy for high-ranked players.

As for scoring, most matches consist of winning 2 sets of 6 games, with Grand Slams being the primary exception to this rule and requiring players to win 3 sets. A game is won by the first player to reach 4 points, with a win-by-two rule applied in the event of 3 points each. Points are counted as 15, 30, 40, Advantage (for whoever won the point at 40-All), and Game. In a given game, a player can either be serving or receiving. The roles switch every game, as the server generally has an advantage over the receiver. If the players reach 6 games each, a tiebreak (usually up to 7 points, win by 2) is played to decide who wins the set.

Mechanics of Sports Betting

Prior to discussing the methods and results of this research, it is also useful to describe how sports betting and odds work in practice. Sports betting refers to placing money on the occurrence of a specific outcome or multiple outcomes in the context of a sports event. In the case of tennis, the outcome of this event can be very varied, from a certain player winning a match to the number of sets won by both players, or to the exact score of the encounter. Bets can be placed either online via an app such as bet365 or in-person in a casino, and can be placed either prior to an event or during an actual event (called "live betting"). The research in this paper focuses on pre-match bets, as there is more data readily available; the payouts from live betting change on a near real-time basis, making it much harder to aggregate and analyze all relevant information. In this paper, the marketplace (online or in-person) that offers the bets will be referred to as the *bookkeeper*, and the individual placing the bets will be called the *bettor*.

After betting on a certain outcome, if that outcome is realized, a certain payout is paid by the bookkeeper to the bettor. This payout is determined by the bookkeeper's published odds. There exist three types of published odds: American, decimal, and fractional. All three of these odds types are equivalent and can be transformed into one another using simple calculations. This paper only looks at decimal odds, as they are generally the easiest to interpret. Decimal odds function by multiplying the bet amount to determine total revenue. For example, betting $100 on odds of 2 means that the bettor will receive $200 back in the case the outcome bet on materializes. If the outcome does not materialize, then the $100 is lost by the bettor. The profit on a given bet is given by the following formula:

$$\text{Profit} = \begin{cases} \text{Bet Amount} \cdot (\text{Odds} - 1), & \text{if success} \\ -\text{Bet Amount}, & \text{if failure} \end{cases} \qquad (1)$$

For the above example, the profit would be $100 · (2 – 1) = $100. The intuition behind this formula is that the bettor will receive the bet amount times the odds of the match in the case of a correct prediction, and will need to subtract out the bet amount originally given to the bookkeeper in order to calculate overall profit. This implies that the lowest value odds can take is 1: this occurs in a situation where the bookkeepers estimate a certain outcome to be nearly guaranteed, meaning that they are not willing to give any profit to the bettor – and hence the bettor would not be willing to place the bet.

Bookkeepers employ odds compilers to formulate their odds. These odds compilers determine their pricing based on analyses of historical data, market movement (i.e., trends and sizes of bets taking place), and quotes from other betting markets. Odds can be thought of as a measure of the probability of an event occurring. Thus, the lower the outcome probability, the higher the posted odds, meaning that bettors are compensated for taking on additional risk.

This study focuses on one type of betting outcome: the winner of a match. This type of bet is offered for every single tennis match, regardless of the difference in ranking between the players, which implies that there will be numerous statistical data points for strategy creation. The exact conversion for this bet type from odds to implied probabilities is as follows, with P1 referring to the higher-ranked player and P2 referring to the lower-ranked player (for a given match between P1 and P2):

$$\text{Odds (P1)} = \mathrm{o}_1 \,;\, \text{Odds (P2)} = \mathrm{o}_2 \qquad (2)$$

$$\text{Probability (P1)} = \frac{\frac{1}{\mathrm{o}_1}}{\frac{1}{\mathrm{o}_1} + \frac{1}{\mathrm{o}_2}} \qquad (3)$$

$$\text{Probability (P2)} = \frac{\frac{1}{\mathrm{o}_2}}{\frac{1}{\mathrm{o}_1} + \frac{1}{\mathrm{o}_2}} \qquad (4)$$

If odds are accurate predictors of the probability of an event, why is it so difficult to make a profit in sports gambling? Therein lies the issue – odds are biased by the odds compiler in order to favor the betting market. The bookkeepers apply a “vigorish”, or “vig”, on any given bet which renders the expected value negative for the bettor. An easy example that showcases this phenomenon is rolling a die. The real probability of rolling a “1” is 1/6. If we decide to bet $100 on every roll on the “1” outcome, the required money multiple, or odds, to achieve an expected value of 0 is 6. In a fair marketplace, this is what the odds should be, given the known probability of success. However, due to the uncertainty of sports events, bookkeepers apply a small cut to this number in order to reduce their risk exposure. In the case of the previous example, they could offer odds of 5 instead of 6, meaning that the expected value for the bettor is negative. For a given tennis match, the vig can be calculated as:

$$\mathrm{Vig} = \frac{\frac{1}{\mathrm{o}_1} + \frac{1}{\mathrm{o}_2} - 1}{\frac{1}{\mathrm{o}_1} + \frac{1}{\mathrm{o}_2}} \qquad (5)$$

This is the mechanism through which betting markets make profits, and why it is difficult for sports bettors to generate consistent returns. Beating the market and creating sustainable strategies will therefore need to overcome this obstacle.

State of the Art: Literature Review

In what follows, we will provide a review of past literature on the subject of sports gambling. How have people approached this problem in the past, and were they successful in beating the market? Before diving into the research, we point out that the idea of investing in sports as an asset class was first popularized by Mark Cuban in 2004 on his personal blog (Cuban 2004). He notably discussed the difference in information availability between stock investing and sports gambling; while we rarely know the inner workings of a company outside of press releases and financial statements, we have large amounts of data for sports events. Gamblers can actually watch matches and they have access to recordings, press interviews, in-depth player statistics, and more. Cuban’s enthusiasm for this asset class gave rise to the first sports betting hedge fund in 2009 – Centaur Galileo. The firm ended up crashing within a few years, after recording $2.5 million in losses (Manfred 2012). Information on other sports betting hedge funds is very limited, with little evidence of truly successful ventures in the space. This confirms the idea that creating strategies to make money through sports gambling is extremely difficult in practice. We will now turn to research in the field to explore the viability of this objective.

There are two components to building a sports betting strategy: a predictive analytics component, which consists of calculating the probability of specific events occurring, and a bankroll management component, which relates to deciding the amount of money to allocate to specific matches. These two components shape the structure of the strategy-creation portion of this research, and their combination is essential for establishing viable betting strategies.

Predictive Analytics

In this first part of the literature review, we will go over past research related to the predictive analytics portion. It is important to note that, according to past studies, it is very difficult to predict matches more accurately than betting markets themselves. Stekler, Sendor, and Verlander (2010) studied the forecasting accuracy of three different types of prediction methods: statistical models, experts ("tipsters"), and the betting markets. They conducted their research across five different sports (horse racing, basketball, football, baseball, and soccer), and found that there exists no evidence of statistical models or expert tips having higher predictive accuracy than betting markets, both in terms of determining the winner of a match and the point spread. Their conclusion supports the theory of market efficiency for sports gambling, as betting markets appear to reflect all information relevant to a particular game, implying that it appears impossible to consistently outperform the odds.

This said, numerous researchers have still attempted to build models with higher predictive accuracy than the betting markets. One of the most exhaustive research papers in the field of predictive analytics for tennis was (Wilkens 2021) on the applications of machine learning. Wilkens' dataset consisted of over 39,000 tennis matches from 2010 to 2019, and his research included tests on the predictive accuracy of a variety of techniques including logistic regression, neural networks, random forests, and gradient boosting. Ultimately, Wilkens found that there were no strategies based on these models that consistently outperformed the market. Total accuracy could not be increased to more than 70%, which is not a large figure considering that 65% accuracy could be achieved simply by betting on the higher-ranked player on any given match. Additional variables such as home advantage or tournament round appeared to add no predictive value, as most information was already reflected in the betting odds and player rankings.

Outside of the field of tennis analytics, there exists some evidence of researchers creating successful models to beat the betting markets. One key study was conducted by Egidi, Pauli and Torelli (2018), where the researchers chose to combine historical data with betting odds to predict soccer scores for a number of European leagues. This was the first paper to use betting odds to improve model accuracy. The researchers used a Bayesian Poisson model, which they trained over nine years of soccer data. They interpreted their results from both a probabilistic and a profitability point of view; while both the market and their model had similar levels of aggregate predictive accuracy, they found that by betting on matches with the highest expected returns and by varying the bet amount on the matches' profit variability they could actually generate positive expected profits. The researchers however note that positive expected profits do not guarantee high positive returns. Still, this provides evidence for the hypothesis that outperforming the market in terms of aggregate predictive accuracy may not be a necessary requirement for creating a successful strategy – finding a subset of matches with high expected returns and allocating money astutely may be sufficient. Kaunitz, Zhong, and Kreiner (2017) also found proof of inefficiencies within the soccer betting market. Instead of trying to build complex models that would outperform the bookkeepers – which they mentioned had still not been done convincingly across the literature – they decided to only use the published odds to find mispricing situations. They defined mispricing as an occurrence where the betting odds diverged from their "fair value", which they could estimate using historical data. Why would odds diverge from their fair value? There exist multiple explanations for this situation, including the bookkeepers trying to protect themselves from downside losses or trying to attract clients to their sites, as well as taking advantage of market overreaction and similar psychology-driven effects. The researchers found that by identifying such matches and using fixed amount betting, they could generate positive profits; they notably achieved a 6.2% return from paper trading and real betting in the months following the creation of their strategy.

Bankroll Management

This research signals that certain inefficiencies exist within betting markets. Before diving into the foundational research in bankroll management for sports betting, it is important to review some of the key theoretical underpinnings of the field. Some of the most prevalent theories include Modern Portfolio Theory (MPT), introduced in (Markowitz 1952), and the Kelly Criterion (Kelly 1956). As betting markets became increasingly liquid during the beginning of the 21st century, research was done to understand the similarities between betting markets and financial markets. Researchers such as Fitt (2019) undertook comprehensive studies to illustrate that betting portfolio risk could be minimized by applications of MPT. The general idea behind this theory is that it is possible to find a set of investments that maximizes expected returns for a certain level of risk, which is dependent on the preferences of the investor. Fitt notably found that it is possible to create an optimal betting portfolio if there exist discrepancies between the bettor's model and the betting market. He focused his study on soccer, and assumed that all goals scored follow distinct Poisson distributions with fixed means. Fitt was able to create an "efficient betting frontier" by exploring the variances and correlations between different bets on the same match (i.e., the number of total goals and the number of home goals, for example). It is key to note however that while Fitt's research underlines the possibility of minimizing betting risk, the ability to generate profits in this case still comes down to the bettor's "edge", as determined by the latter's predictive model. Further research was done to generalize the applications of portfolio theory across other sports, and will be included in what follows. As for the Kelly Criterion, researchers found that it could be used in modern betting markets (as well as the stock market) to maximize the expected value of wealth by determining the exact bankroll fraction to invest on a given match, according to the calculated probabilities and actual odds. The exact formula for the Kelly Criterion is shown by Equation 6.

$$f = p - \frac{q}{b} \qquad (6)$$

- $f = \text{fraction of bankroll invested}$ (6.1)
- $p = \text{probability of winning, as estimated by the predictive model}$ (6.2)
- $q = 1 - p$ (6.3)
- $b = \text{odds} - 1$ (6.4)

Several studies, including that by Hung (2010), tested the efficacity of Kelly's Criterion through the use of simulations. Hung's study assumed the probabilities of a certain event to be known, and then modeled total return on capital based on various numbers of bets and amount of initial capital. Hung found that betting according to this strategy is effective and can generate high returns, but is also volatile and requires a high amount of initial capital and/or number of bets. This indicates that a successful betting strategy will require additional risk mitigation strategies to be employed in conjunction with one or both of these foundational theories.

More comprehensive research studies include combinations of both predictive models and risk mitigation strategies based on these overarching theories. One key study employing MPT to maximize returns was conducted by Hubáček, Šourek, and Železný in 2019. In this study, the researchers decided to experiment with three new hypotheses, all of which were designed to contribute heavily to the field of sports gambling. They first chose to suppress the correlation between their own model and the bookkeeper odds, through the use of a number of decorrelation techniques. Though this would reduce the predictive accuracies of their models, they believed that this would allow them to find profitable discrepancies between their estimated probabilities and those implied by the betting odds. Second, they chose to employ convolutional neural networks, which researchers such as Wilkens had mentioned could have potential successful applications. The researchers combined these techniques with elements of MPT to create strategies with optimal risk-return tradeoffs. All three of these hypotheses/strategies were confirmed and validated through back-testing on a dataset including seven years of NBA results and betting odds. The researchers found that the application of MPT to their strategy with a straightforward max-Sharpe selection criterion led to consistent returns with low variability. However, their research simply consisted of measuring cumulative profits and did not assume that the bettor could gradually re-invest his wealth – which would be a more realistic assumption. Uhrin, Šourek, Hubáček, and Železný complemented their previous research on MPT with a new study in 2021. In this study, the researchers chose to apply a variety of relatively straightforward risk minimization techniques (the simplest of which being a maximum bet limit) to formal investment strategies which used the MPT and Kelly Criterion. They argued that the MPT and Kelly Criterion strategies bore too much risk on their own as a result of unrealistic mathematical assumptions regarding the true probability of events. Testing their strategies on large horse racing, basketball, and football datasets, they found that certain modifications were very effective in minimizing the quantity of downside scenarios and helping with wealth progression. The most suitable option of the strategies they tested appeared to be the fractional Kelly (which limits the bet size to a certain fraction of the amount calculated using the Kelly Criterion), as it produced the highest performance across all metrics studied.

All in all, our literature review indicates that while market inefficiencies do exist, there is little evidence of predictive models being able to outperform the bookkeepers, particularly within the field of tennis where no successful models or strategies were found. Most models lack a serious financial underpinning in terms of money allocation (simply focus on the predictive analytics portion), and most studies discussing risk management techniques do not clearly tie back to the output from the initial predictive modeling phase. In what follows, we will present a coherent, comprehensive, start-to-finish strategy that demonstrates the feasibility of making money consistently through sports betting.

DATASET, SELECTION REQUIREMENTS, AND MATCH VARIABLES

Dataset and Match Selection Requirements

The first component essential to our strategy creation was the dataset and match selection. We decided to include eleven years of match data – from 2012 to 2022 – and we formed our dataset by combining three different sources of information, described below:

- Tennis match data from tennis-data.co.uk[2] which contains all Tier 1 matches that occurred within a given year, along with basic game information including the players' respective rankings, the score of the match, the surface of the court, etc. This dataset also contains the published odds for a number of different betting sites.
- Tennis match and player data from (Sackmann 2022) which contains all Tier 1 matches that occurred within a given year, along with player profile information (height, age, country, handedness, etc.) as well as service game statistics.
- Tournament location database (self-produced) which contains the country in which every Tier 1 tournament is played. This information is needed to determine if players have a homecourt advantage.

The first two datasets listed above were merged by requiring matches in either dataset to have the same player ranks, numbers of ATP points, and set-by-set score. In terms of match selection, we decided to retain only the highest-quality tournaments: Grand Slams, ATP Finals, Masters 1000s, and ATP 500s. This was done based on the assumption that lower-stakes tournaments have more variable outcomes and are hence less predictable. We then removed

[2] The tennis-data.co.uk website contains match history and betting information since the year 2000: http://www.tennis-data.co.uk/alldata.php. The variables available in this dataset are listed at: http://www.tennis-data.co.uk/notes.txt

matches with unusual outcomes, which could include players retiring or forfeits, for example. We also removed all matches for which certain betting information is missing, and for which some of the input variables have empty values (see variable list in the next section). To conclude the selection of our final dataset, we also rejected matches in which either player has played no matches in the six months preceding the current match. This is because some of our time-series variables (see next section) aggregate the service game statistics of the previous six months in order to get a measure of recent form and momentum.

Output and Input Variables

Given that we are looking to predict the winner of a given match, we defined our output variable based on whether an upset occurs or not. An upset indicates that the winner had a lower ranking than his opponent going into the match. We therefore created a match variable called "Upset" which contains the match result:

$$\text{Upset} = \begin{cases} 0, & \text{if the winner had the higher ranking going into the match} \\ 1, & \text{if the winner had the lower ranking going into the match} \end{cases} \qquad (7)$$

Both match datasets used were constructed using winner and loser information, which is not suited for predicting match outcome, as the terms "winner" and "loser" only have meaning after a match is completed. This led us to organize our data in terms of "P1" and "P2", where P1 is the player entering the match with the higher ranking. To map (winner, loser) to (P1, P2), we used the Upset variable as shown below:

$$\text{P1} = \begin{cases} \text{Winner}, & \text{if Upset} = 0 \\ \text{Loser}, & \text{if Upset} = 1 \end{cases} \qquad (8)$$

$$\text{P2} = \begin{cases} \text{Winner}, & \text{if Upset} = 1 \\ \text{Loser}, & \text{if Upset} = 0 \end{cases} \qquad (9)$$

In what follows, we note that any player-related variable V (such as player ranking, number of points, etc.) will generate a pair of input variables: V(P1) and V(P2). In our research, rather than using V(P1) and V(P2), we decided to use the equivalent pair V(P1) and ΔV = V(P1) – V(P2). We expected ΔV to be more strongly correlated with the match output than V(P2). This is a simple linear combination of variables which preserves the information contained in the original variables V(P1) and V(P2).

After defining our output variable and deciding how to organize our data, we selected input variables across three categories: player-profile-related variables (Table 1), situational variables (Table 2), and time-series variables (Tables 3 and 4).

Index	Variable	Definition
1	$\text{Points}(\text{P1})$	ATP points of the higher-ranked player
2	$\Delta(\text{Points})$	$\text{Points}(\text{P1}) - \text{Points}(\text{P2})$. This is always positive
3	$\log[\Delta(\text{Points})]$	$\log[\text{Points}(\text{P1}) - \text{Points}(\text{P2})]$
4	$\text{Rank}(\text{P1})$	ATP rank of the higher-ranked player
5	$\Delta(\text{Rank})$	$\text{Rank}(\text{P1}) - \text{Rank}(\text{P2})$. This is always negative
6	$\log[-\Delta(\text{Rank})]$	$\log[\text{Rank}(\text{P2}) - \text{Rank}(\text{P1})]$
7	$\text{Hand}(\text{P1})$	$\begin{cases} 0, & \text{if P1 is right-handed} \\ 1, & \text{if P1 is left-handed} \end{cases}$
8	$\Delta(\text{Hand})$	$\begin{cases} 0, & \text{if both players are right-handed} \\ 1, & \text{if both players are left-handed} \\ 2, & \text{if P1 is right-handed and P2 is left-handed} \\ 3, & \text{if P1 is left-handed and P2 is right-handed} \end{cases}$
9	$\text{Age}(\text{P1})$	Age of the higher-ranked player, in years
10	$\Delta(\text{Age})$	$\text{Age}(\text{P1}) - \text{Age}(\text{P2})$
11	$\text{Height}(\text{P1})$	Height of the higher-ranked player, in centimeters
12	$\Delta(\text{Height})$	$\text{Height}(\text{P1}) - \text{Height}(\text{P2})$

Table 1. Player profile variables and their definitions.

We assigned numbers to categorical variables such as handedness or match surface, for example. The numerical values were chosen so they increase as the probability of an upset increases. We will illustrate the choice process for one variable, match surface. We assumed that upsets are most unlikely on clay courts, as clay is the slowest surface. Supposing that higher-ranked players generally are more skilled than their opponents, clay courts give them more time to apply their skills. Faster surfaces (like hard courts and grass) lead to more unpredictable outcomes, as heavy-hitters and big servers can use speed to derail their opponents. We expect that grass court outcomes are even more unpredictable than hard court outcomes, because: (a) a grass surface is slightly faster, (b) a grass surface is more uneven and makes bounces harder to predict, and (c) grass courts are hard to maintain and thus less widespread than hard courts, so players have less experience on this surface. A similar logic applies to the other categorical variables. Next, we selected time-series variables which are aggregates of service and return game performance over the last 180-day period preceding the match (see Table 3).

Index	Variable	Definition
13	Surface	$\begin{cases} 0, & \text{if Surface} = \text{Clay} \\ 1, & \text{if Surface} = \text{Hard} \\ 2, & \text{if Surface} = \text{Grass} \end{cases}$
14	Venue	$\begin{cases} 0, & \text{if Court} = \text{Indoor} \\ 1, & \text{if Court} = \text{Outdoor} \end{cases}$
15	Series	$\begin{cases} 0, & \text{if Grand Slam} \\ 0.75, & \text{if ATP Finals} \\ 1, & \text{if Masters 1000} \\ 2, & \text{if ATP 500} \end{cases}$
16	Homecourt(P1)	$\begin{cases} 0, & \text{if tournament country} = \text{P1 country} \\ 1, & \text{if tournament country} \neq \text{P1 country} \end{cases}$
17	Δ(Homecourt)	$\begin{cases} 0, & \text{if P1 plays at home and P2 does not} \\ 1, & \text{if neither player plays at home} \\ 2, & \text{if both players play at home} \\ 3, & \text{if P2 plays at home and P1 does not} \end{cases}$

Table 2. Situational variables and their definitions.

Index	Variable	Definition
18	Ace_p(P1)	Percentage of aces (out of total service points) recorded by P1 in the 180 days prior to the match
19	Δ(Ace_p)	Ace_p(P1) − Ace_p(P2)
20	DoubleFault_p(P1)	Percentage of double faults (out of total service points) recorded by P1 in the 180 days prior to the match
21	Δ(DoubleFault_p)	DoubleFault_p(P1) − DoubleFault_p(P2)
22	FirstIn_p(P1)	Percentage of first serves in (out of total service points) recorded by P1 in the 180 days prior to the match
23	Δ(FirstIn_p)	FirstIn_p(P1) − FirstIn_p(P2)
24	FirstInWon_p(P1)	Percentage of points won (out of total first serve in points) recorded by P1 in the 180 days prior to the match
25	Δ(FirstInWon_p)	FirstInWon_p(P1) − FirstInWon_p(P2)
26	SecondWon_p(P1)	Percentage of second serve points won (out of total second serve points) recorded by P1 in the 180 days prior to the match
27	Δ(SecondWon_p)	SecondWon_p(P1) − SecondWon_p(P2)
28	BreakPointsSaved_p(P1)	Percentage of break points saved (out of total break points faced) recorded by P1 in the 180 days prior to the match
29	Δ(BreakPointsSaved_p)	BreakPointsSaved_p(P1) − BreakPointsSaved_p(P2)
30	ServicePointsWon_p(P1)	Percentage of service points won (out of total service points) recorded by P1 in the 180 days prior to the match
31	Δ(ServicePointsWon_p)	ServicePointsWon_p(P1) − ServicePointsWon_p(P2)
32	ReturnPointsWon_p(P1)	Percentage of return points won (out of total return points) recorded by P1 in the 180 days prior to the match
33	Δ(ReturnPointsWon_p)	ReturnPointsWon_p(P1) − ReturnPointsWon_p(P2)
34	DomRatio(P1)	$\frac{\text{ReturnPointsWon\_p(P1)}}{1 - \text{ServicePointsWon\_p(P1)}}$
35	Δ(DomRatio)	DomRatio(P1) − DomRatio(P2)

Table 3. Service and return variables aggregated over the 180-day period prior to the match.

Finally, we created a number of year-long aggregate variables, which we listed in what follows. We first tried to capture the players' volume of play and win percentage. Second, we tried to obtain measures of the players' mental strength by looking at last-set-played matches and at tiebreak results. Third, we looked at momentum variables such as the winning streak going into the match and number of days elapsed since the last match. These variables are listed in Table 4.

Index	Variable	Definition
36	Nmatch(P1)	Total number of matches played by P1 in the last 365 days
37	Δ(Nmatch)	Nmatch(P1) − Nmatch(P2)
38	Win_p(P1)	Percentage of matches won by P1 out of the Nmatch(P1) played
39	Δ(Win_p)	Win_p(P1) − Win_p(P2)
40	Δ(LSP)	LSP(P1) − LSP(P2) LSP is the number of matches in which the final set was played
41	LSP_Win_p(P1)	Percentage of matches won by P1 out of the LSP(P1) matches played
42	Δ(LSP_Win_p)	LSP_Win_p(P1) − LSP_Win_p(P2)
43	Δ(TB)	TB(P1) − TB(P2) TB is the number of tiebreaks played
44	TB_Win_p(P1)	Percentage of tiebreaks won by P1 out of the TB(P1) played
45	Δ(TB_Win_p)	TB_Win_p(P1) − TB_Win_p(P2)
46	ElapsedDays(P1)	Number of days since P1′s last match
47	Δ(ElapsedDays)	ElapsedDays(P1) − ElapsedDays(P2)
48	Streak(P1)	Number of consecutive wins going into the current match
49	Δ(Streak)	Streak(P1) − Streak(P2)

Table 4. Strength and momentum variables aggregated over the 365-day period prior to the match.

MULTIVARIATE ANALYSIS

Exploratory Analysis

The final dataset selected in the previous section contained 14,079 matches in total, with 4,527 (32.1%) of those matches being upsets. For this dataset, we conducted an exploratory analysis to check the relationships between the input variables and the output.

Figure 1 shows a visual representation of the input-output relationships for several variables. Some of the strongest relationships are captured by the difference in player ranking, Δ(Rank), and the logarithm of the difference in the number of ATP points of the two players. As the difference in ranking/points between the two players approaches 0, the probability of an upset approaches 0.5 (equiprobable result). Other variables such as Series, Surface or Δ(Homecourt) show weak relations to the match result.

We then tested how well the outcome (i.e., the “Upset” variable) can be predicted by each input variable (taken in isolation), by using a univariate ordinary least squares regression. The regression was done using 90% of the dataset, while the remaining data was held out for testing purposes. Figure 2 shows the adjusted R^2 values obtained; the maximum value is 5.7%, which illustrates the difficulty of predicting match outcome using one variable at a time. Next, we investigated the performance of linear and non-linear combinations of multiple input variables.

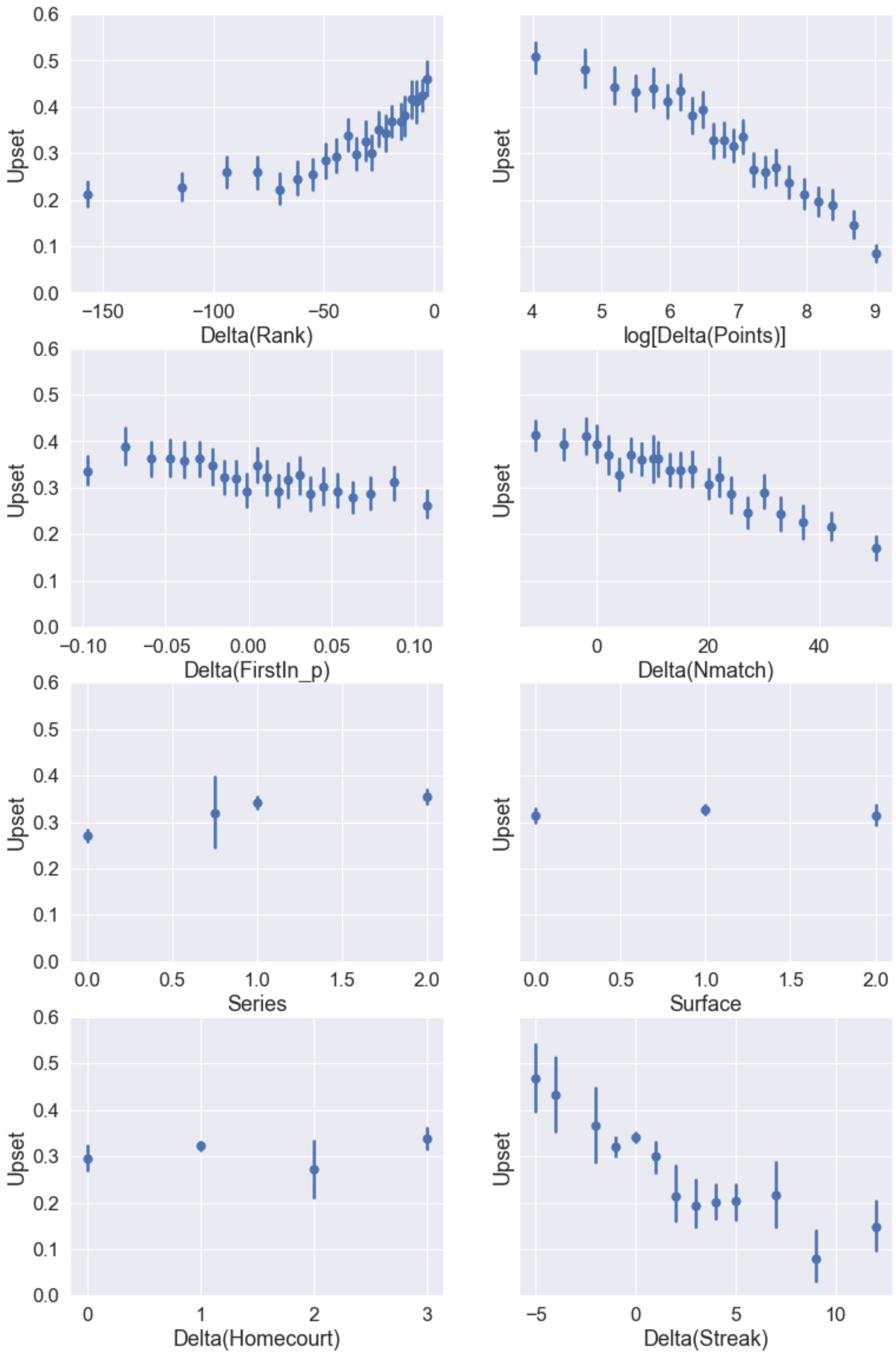

Figure 1. Upset percentage (y-axis) dependence on several input variables. The error bars indicate the 95% confidence interval.

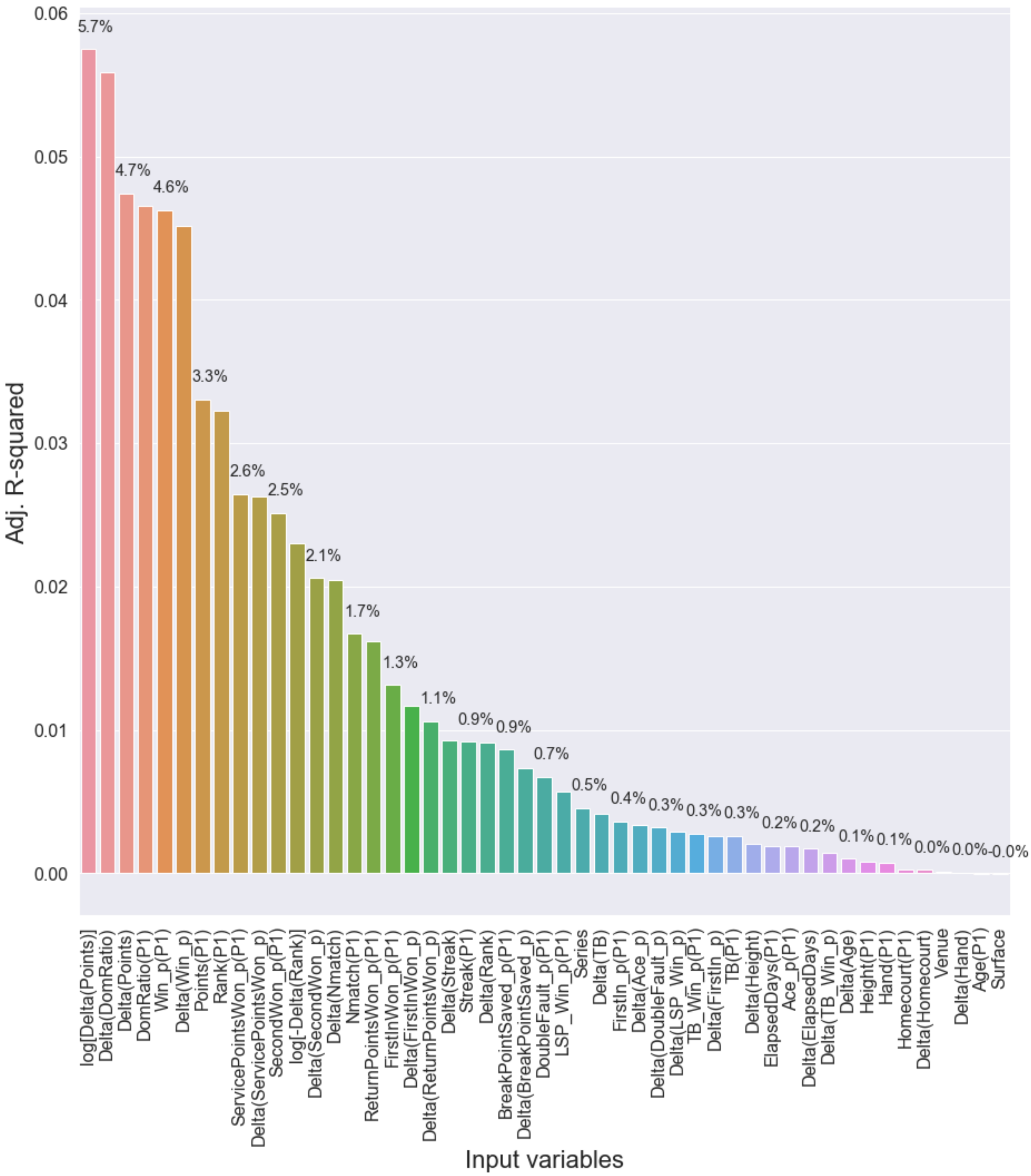

Figure 2. Univariate OLS regression using every input variable in isolation. The best variable log[Δ(Points)] only explains 5.7% of the total variance of the match result.

Multivariate Combinations

Logistic Regression

The binary nature of the output variable (Upset = 0 or Upset = 1) suits well a Logistic Regression (LR) approach. We used Python’s scikit-learn libraries (Pedregosa et al. 2013) to scale the input variables to zero mean and unit standard deviation, and to train an LR model combining the scaled input variables.

Data was split randomly into a training subset (90%) and a testing subset (10%). After the model training was completed, the resulting model was used to predict the class (Upset = 0 or Upset = 1) of each training and testing vector. The prediction results are aggregated in Table 5.

	Training sample	Testing sample	Training + testing
Matches	12,673	1,406	14,079
Upset matches	4,058	469	4,527
Prediction accuracy	69.5%	67.8%	69.3%

Table 5. Linear Regression model accuracy computed as the fraction of correct upset predictions in the training and testing subsamples.

We compared the LR prediction to the simple case of predicting the higher-ranked player to win every time. The latter model would have a very similar accuracy: 67.8%. Thus, it appears that we did not significantly improve upon using a single variable only (Δ(Points) or Δ(Rank)). However, this conclusion is misleading, as can be seen from the confusion matrix in Table 6. The LR model correctly predicts 22.8% of the upsets, and it is this feature we will be exploiting when we will be looking at betting strategies.

	LR model		Simple model (P1 always wins)	
	Pred. non-Upset	Pred. Upset	Pred. non-Upset	Pred. Upset
Actual non-Upset	91.4%	8.6%	100%	0%
Actual Upset	77.2%	22.8%	100%	0%

Table 6. Confusion matrix for the LR model and the simple model of always picking the higher-ranked player to win. The LR model predicts correctly 22.8% of the upsets.

The LR model relies on a linear combination approach. We also tested non-linear methods, among which the Neural Networks show good prediction capabilities, as we will describe in what follows.

Neural Networks

The Neural Networks approach (NN) relies on a similar pre-processing step as that described in the previous section, and on the TensorFlow library (Abadiet et al. 2015). In addition, through experimentation we were able to reduce the input space from 49 variables to 17 variables: {Rank(P1), Homecourt(P1), Δ(Points), Δ(Age), Series, log[-Δ(Rank)], log[Δ(Points)], Win_p(P1), TB(P1), Streak(P1), Δ(Streak), Δ(DoubleFault_p), Δ(FirstInWon_p), Δ(ServicePointsWon_p), Age(P1), ElapsedDays(P1), Δ(Ace_p)]}. Having fewer variables simplifies the neural network convergence, and also prevents over-training which may occur when the ratio between the number of training vectors and the number of trainable parameters drops below 20.

We selected a 3-layer feed forward network with 17 nodes in the input layer, 10 nodes in the middle layer and 1 node in the output layer – for a total of 191 adjustable parameters (weights and thresholds). The output layer uses a sigmoid function and it models upset probability. For the specific choice of the architecture, we note that the number of input nodes is determined by the number of input variables, and the number of output nodes (1) is required by the nature of our 2-class (upset or not) categorization problem. As for the number of intermediate nodes, we tested networks of different values and found that 10 hidden nodes provides a good discrimination power.

To aid with the recognition of upset matches, we decided to train the NN with equal numbers of upset and non-upset matches. On the one hand, this helps with the convergence of the model and improves its performance on upset matches. On the other hand, the model learns less well how to predict non-upset matches. As a result of this choice for the training data composition, the NN output would not estimate well the actual upset probability. To correct this, we binned the NN output and measured the true (actual) upset probability in each bin using the match data. We then fit this dependence (NN output – actual probability) with a quadratic function and used this relationship to scale the NN output for every match[3]. By construction, the new variable, which we referred to as the ScaledNN, would estimate the actual upset probability. The ScaledNN model showed a prediction accuracy of 69.5%, which is close to the value obtained for the LR model. The confusion matrix is shown in Table 7.

	ScaledNN model	
	Predicted non-Upset	Predicted Upset
Actual non-Upset	93.1%	6.9%
Actual Upset	80.3%	19.6%

Table 7. Confusion Matrix for the NN model; the NN model predicts correctly nearly 20% of the upsets. The values correspond to the full match dataset.

Fisher Discriminant

To what extent are the LR and ScaledNN models different? Their outputs (upset probabilities) are shown in Fig. 3 for the entire dataset. While there is a high degree of

[3] We found the following dependence: ScaledNN = $0.2983 \cdot \text{RawNN}^2 + 0.4594 \cdot \text{RawNN} - 0.0027$. More generally, the three coefficients would depend on the model architecture and training sample choices.

correlation between the two models (83.6%), we also see a certain spread in the distribution of matches which indicates a potential gain from combining the two outputs.

We determined that the simplest way to combine the NN and LR output variables was via Fisher's linear discriminant analysis (Fisher discriminant in short), which we implemented using the scikit-learn library (Pedregosa et al. 2013).

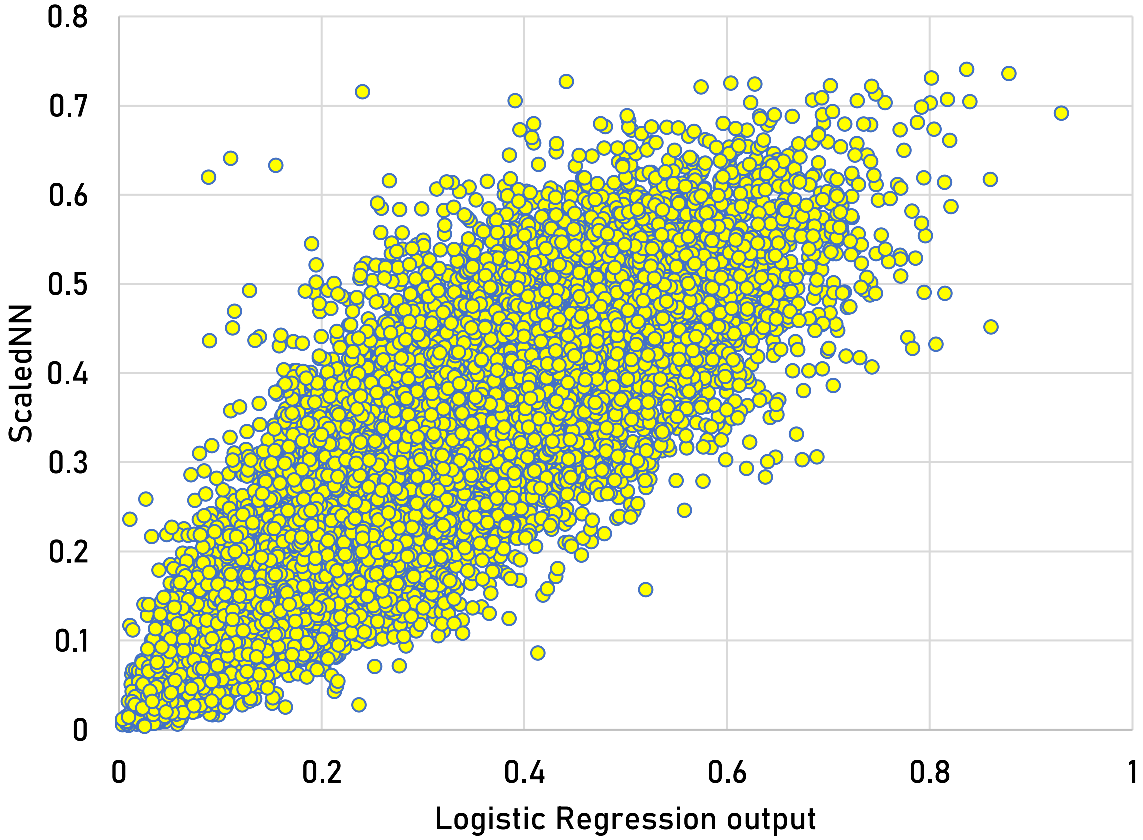

Figure 3. Neural Networks model (y-axis) vs. Logistic Regression model output (x-axis). The correlation coefficient is 83.6%.

The resulting Fisher discriminant is given by Equation 10.

$$\text{Fisher} = 0.54 \cdot \text{ScaledNN} + 0.46 \cdot \text{ LR} \qquad (10)$$

The prediction accuracy over the full match dataset is 69.8%, and the confusion matrix is given in Table 8.

	Fisher discriminant model	
	Predicted non-Upset	Predicted Upset
Actual non-Upset	93.5%	6.5%
Actual Upset	80.3%	19.7%

Table 8. Confusion matrix for the Fisher discriminant; the Fisher discriminant model predicts correctly nearly 20% of the upsets and has the highest non-upset accuracy among all models studied.

We based our match result prediction on this quantity (Fisher), which incorporates the maximum amount of predictive information that we were able to extract out of the 49 input variables.

Comparison with Published Betting Predictions

Before moving on to the strategy creation and bankroll management part of our research, we decided to look at how our match predictions compare with those published by the betting site bet365. We summarized the key figures in Table 9.

Model	Accuracy	Correlation coefficient between model's upset probability and the actual match result
Linear Regression	69.3%	32.1%
ScaledNN	69.5%	32.5%
Fisher Discriminant	69.8%	33.7%
bet365	71.4%	39.0%

Table 9. Summary of accuracy values for the different models and bet365.

The fact that the bookkeepers' accuracies are higher than that of the Fisher discriminant does not imply that it is impossible to make money by betting on the published odds. As discussed previously, our aim was to find a subset of matches where we could predict upsets accurately, and ideally in a different manner than the bookkeepers so as to profit from discrepancies between our model and theirs. Before moving on to the actual betting strategy creation section of our research, we will provide a brief example of how these conditions are sufficient for generating positive expected returns.

First, it is important to point out that the accuracy values recorded in Table 10 are quoted for the entire set of 14,079 matches, meaning that they cover the full range of probabilities. Based on the takeaways from past literature and given how we trained our model, if we were to bet on every single match, then we would certainly lose in expectation. Instead, it is more optimal to focus on a subset of the probability range. For example, we could focus on betting only on predicted upsets, which means all matches that satisfy the requirement Fisher > 0.5.

Overall, there are half as many upsets as non-upsets in our dataset, but the upside of winning is higher given that odds of upsets are higher than those of non-upsets. Predicting upsets accurately is what we sought to achieve by training the Neural Network model with a dataset in which upsets were over-represented.

We note that in our dataset there are 1,512 predicted upsets (i.e., matches that have a Fisher value higher than 0.5). Not all these predictions will materialize, and we can group matches into two subsamples: actual upsets, and actual non-upsets. Table 10 contains the accuracy of both our model and the bookkeeper's (in this case bet365) for each subsample. If we were to bet on every one of the 1,512 matches, we would have 894 winning bets and 618 losing bets. As it turns out, this approach yields positive returns. This is demonstrated in Appendix A, where we show that a higher revenue is generated in the 35.9% (= 100% - 64.1%) of actual upsets that the bookkeeper predicts incorrectly, than in the upsets that both parties predict correctly.

Fisher > 0.5	Subsample of actual upsets (894 matches)	Subsample of actual non-upsets (618 matches)	Total (1,512 matches)
Bettor prediction accuracy (%)	100%	0%	59.1%
Bookkeeper prediction accuracy (%)	64.1%	58.6%	61.8%

Table 10. Accuracy for matches having Fisher > 0.5. We use the value of the Fisher discriminant to predict the match outcome ("bettor prediction"), while the bookkeeper's predictions are inferred from the published odds.

We have thus shown why drawing profitability conclusions from a comparison of overall accuracy across different predictive models is misleading. In the example chosen, not only is our accuracy lower in the entire sample of 14,079 matches (Table 9), but it is also lower in our smaller sample of 1,512 matches selected for betting (Table 10). However, the ability to select a sample in which we predict upsets more frequently than the bookkeeper opens a source of revenue which is especially important for the subset of upsets which the bookkeeper mis-predicts. The following section will detail the creation of actual betting strategies, which will exploit this mechanism with the objective of generating consistent profits.

BETTING STRATEGIES IN PRACTICE

In this portion of our research, we investigated two types of strategies:

- Informal strategies, which are based on simple rules of thumb and are often suboptimal. Some key examples of such strategies are betting a fixed amount of money or a fixed fraction of total wealth if certain criteria are met. Though imperfect, these strategies can help understand how effective predictive models are in practice and can aid in visualizing what next steps need to be taken to maximize profits.
- Formal strategies, which are more rigorous in nature. Such strategies have actual theoretical underpinnings, and the betting amount is usually a function of current capital, published odds, and a variety of exposure-to-risk limiting parameters. Examples of formal strategies include variations of MPT and the Kelly Criterion.

In what follows, we will present a number of informal and formal strategies which we have optimized to exploit the output of our Fisher model. All comparisons to betting odds will be made using those published by bet365.

Informal Strategies

The basic decision flow for an informal strategy is the following: first, we decide whether or not to bet on a given match, and second, if we do enter the bet, we determine how much to bet. For all strategies discussed in our research, the decision of whether or not to bet is at least partially determined by whether the output of our Fisher model exceeds a certain threshold. Using this threshold-based approach, we investigated two main informal strategies, which we present below.

Strategy 1: Betting fixed amount if Fisher > Threshold

For our first informal strategy, we decided to look at the simplest money allocation tactic possible. More specifically, we evaluated the results of betting a fixed amount ($100) on every match for which the output of our Fisher model was greater than a certain threshold. In order to find the optimal threshold value, we ran 1,000 simulations for a number of different threshold values between 0 and 1. A single simulation contains 1,280 matches drawn randomly from the dataset. We chose the 1,280 value as this is approximately the number of matches in a year. For the 1,000 simulations made for a given threshold, we computed a number of profitability measures, including for example median profit, profit standard deviation, % of simulations recording a loss, etc. The median profit (± 1 standard deviation) per threshold value is shown in Figure 4.

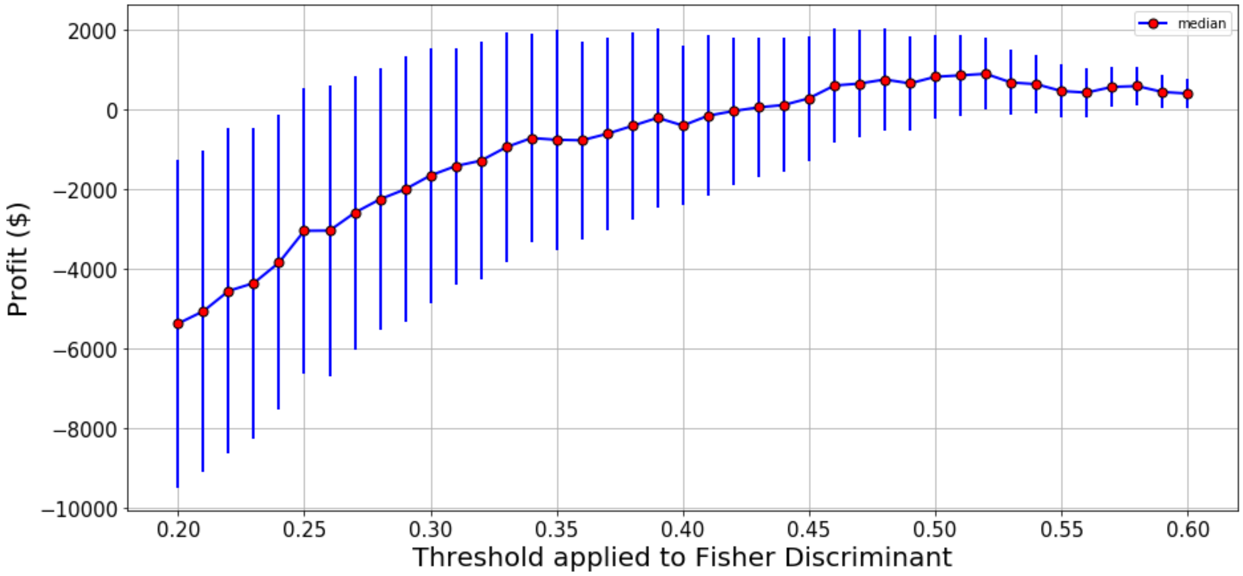

Figure 4. Profit ($) for Strategy 1 per threshold value ranging from 0.2 to 0.6 (with 0.01 increments). The error bars represent ± standard deviations.

As can be seen by the above graph, the optimal threshold value (at least in terms of average profit across all simulations) is 0.52. Further exploration shows the following summary statistics and distribution of simulation profits for a 0.52 threshold strategy.

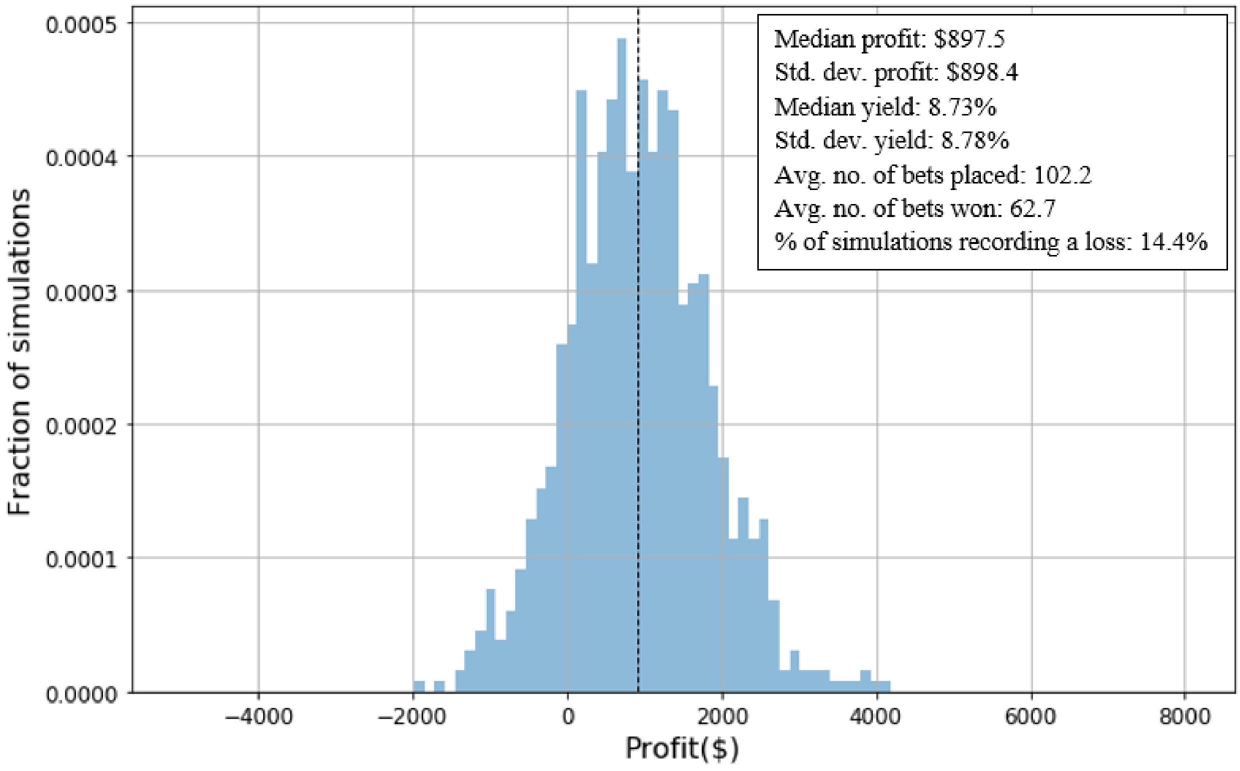

Figure 5. Distribution of profit values for Strategy 1 across 1,000 simulations for a 0.52 threshold value. The dotted line represents the median profit.

The equations below give the formulas for computing profit and yield in Figure 5.

$$\text{Profit} = \text{Total Revenue} - \text{Total Cost} \qquad (11)$$

$$\text{Yield} = \frac{\text{Profit}}{\text{Total Cost}} \qquad (12)$$

Before interpreting our results, it is first important to clarify that "Total Cost" refers to the total amount bet. For Strategy 1, it can simply be calculated as 100 times the number of bets placed.

As we can see from the above summary statistics and distribution, it is possible to generate positive returns using Strategy 1. By design, the strategy works by identifying a small subset of matches where upsets are predicted more accurately than the betting markets. The average number of bets per simulation is approximately 102, which represents less than 8% of matches in a given year. However, the model is accurate more than 60% of the time for this subset of matches, leading to relatively high expected profits and low ruin (only 14.4% of simulations record a loss). Next, we looked at whether we could further increase profits by varying the bet amount.

Strategy 2: Betting a variable amount if Fisher > Threshold

In this informal strategy, we explored two ways to determine the amount invested based on the difference between our model output and the published odds. It is important to note that our decision of whether or not to bet is still solely determined by the output of our model and is not influenced by the betting odds, though these do affect the bet amount.

Strategy 2A: Difference approach to determining bet amount

Strategy 2A consisted of betting an amount proportional to the difference between the output of our Fisher model and the implied probabilities from the betting odds (only when the initial threshold on our Fisher output is met). The intuition behind this approach was that the greater the difference behind our model and the bookkeepers’ model in the high-upset space, the more confident we are in our prediction, and hence the more we bet. The formula for the bet amount is shown in Equation 13.

$$\text{Bet Amount} = 100 \cdot [1 + \Pr(\text{Fisher}) - \Pr(\text{bet365})] \qquad (13)$$

In this equation, Pr(Fisher) refers to the output of our Fisher model and Pr(bet365) refers to the upset probability inferred from the published bet365 betting odds. In order to test this strategy, we followed the same approach as for Strategy 1: we first looked at the optimal threshold value, using the same number of simulations (1,000) and matches per simulation (1,280). The results are given in Figure 6.

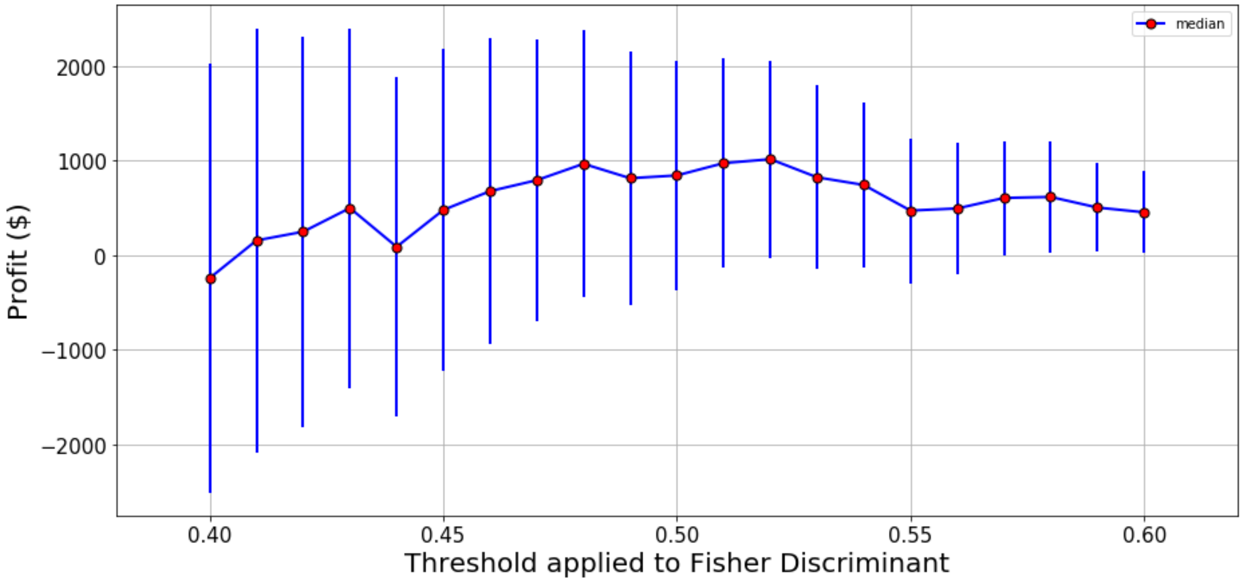

Figure 6. Profit ($) for Strategy 2A per threshold value ranging from 0.4 to 0.6 (with 0.01 increments). The error bars represent ± 1 standard deviation.

Once again, the optimal model threshold value is 0.52. The median profit for this value is very similar to the median profit for 0.48, but the larger standard deviation for this threshold value makes it more risky and hence less attractive. It is also interesting to note that this strategy appears to be overall more risky than Strategy 1, as the error bars are wider. Figure 7 shows the overall profitability of this strategy for a 0.52 threshold value.

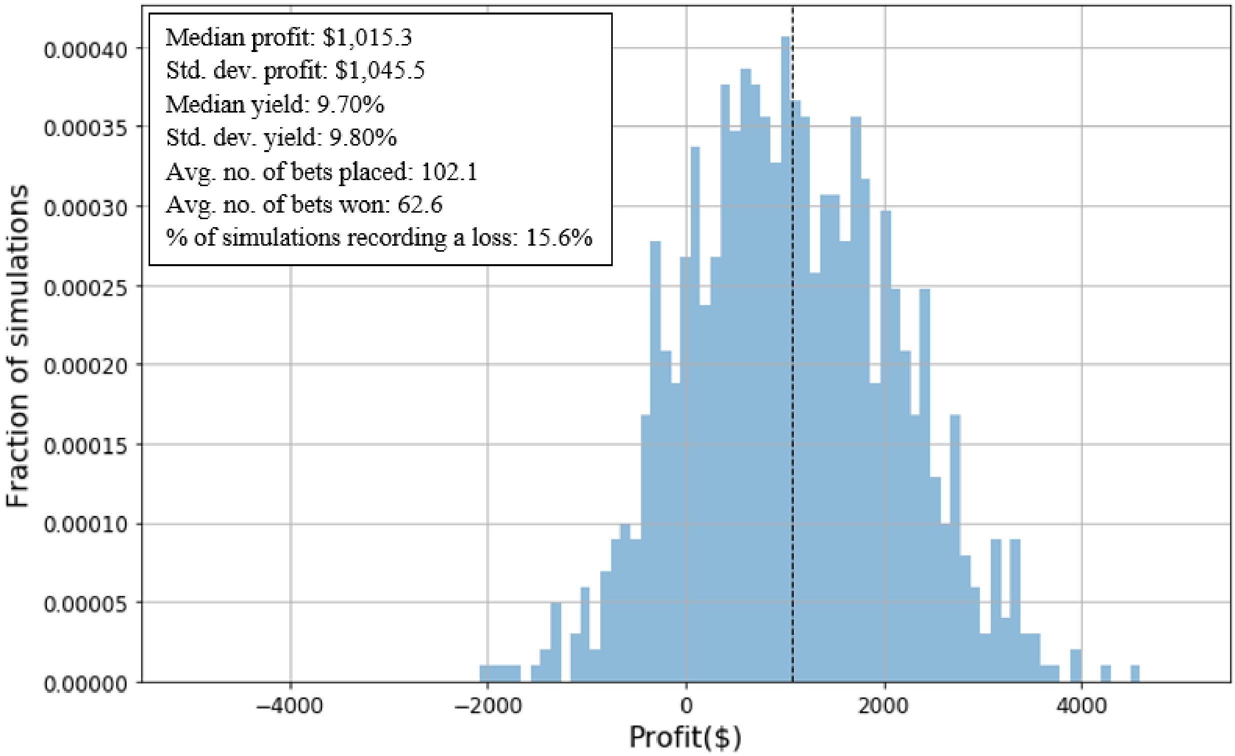

Figure 7. Distribution of profit values for Strategy 2A across 1,000 simulations for a 0.52 threshold value. The dotted line represents the median profit.

The distribution of profits per simulation and the summary statistics in Figure 7 confirm our intuition that this strategy is riskier than Strategy 1. While median profit increases by about $100 and the median yield by roughly 100 basis points, their standard deviations also exhibit slight increases, as does the percentage of simulations recording a loss. Overall, both strategies remain approximately equivalent for a rational bettor. The ratio of median yield to standard deviation of yield (which is a proxy for risk-adjusted returns) is almost identical across both strategies, meaning that the choice between using Strategy 1 or Strategy 2A ultimately comes down to the risk tolerance of the bettor.

Strategy 2B: Ratio approach to determining bet amount

The second way to exploit the dissimilarity between two values is through their ratio; this is the essence of Strategy 2B. Specifically, we decided to use the ratio between the two numbers to determine the amount to bet. Again, the decision to bet is still solely based on the output of our model. The formula we implemented to determine the bet amount in this strategy is shown by Equation 14.

$$\text{Bet Amount} = 100 \cdot \frac{\Pr(\text{Fisher})}{\Pr(\text{bet365})} \qquad (14)$$

In theory, given that odds (and hence probabilities) can exhibit a strong variability for uncertain events like upsets, this approach should allow us to capture a more relative measure of the divergence of our model from market predictions. Implementing the same approach as for the two aforementioned strategies, we found the following distribution of profits per threshold.

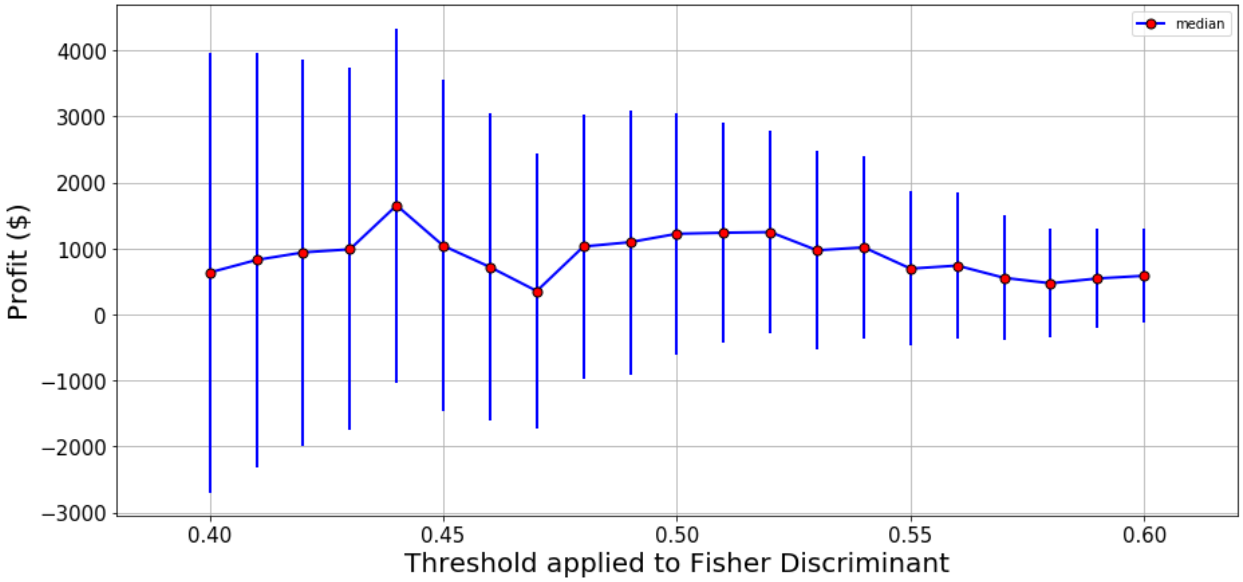

Figure 8. Profit ($) for Strategy 2B per threshold value ranging from 0.4 to 0.6 (with 0.01 increments). The error bars represent ± 1 standard deviation.

While the highest median profit is achieved for a threshold of 0.44, it is clear that this value is not optimal. Implementing such a threshold would lead to a sub-5% median yield and high variability in terms of returns. Looking at this chart more closely, the threshold value of 0.52 proves again to be optimal, as it offers a good balance between median profit value and error bar range. The profitability metrics for this strategy using a 0.52 threshold are highlighted below.

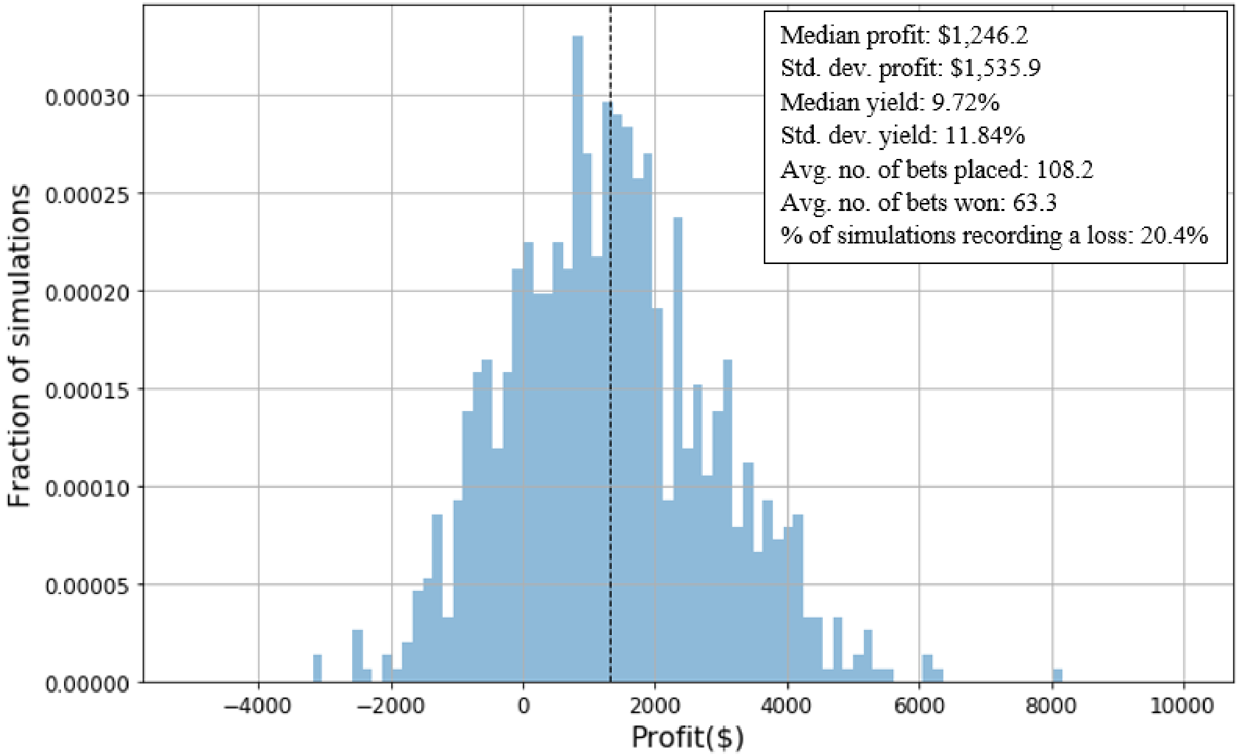

Figure 9. Distribution of profit values for Strategy 2B across 1,000 simulations for a 0.52 threshold value. The dotted line represents the median profit.

While we were able to increase median profit and (marginally) median yield with this strategy, the amount of risk taken strongly increases. The standard deviation of profit increases by almost 50% from Strategy 2B, leading to losses in more than 20% of simulations. This additional risk does not compensate for the mere 2 basis points increase in median yield. Thus, for a rational bettor, this strategy is less attractive than the previous two.

In this section, we have shown that it is possible to generate strategies with positive expected profits. Our "edge" mostly comes from our Fisher model, which allows us to predict a subset of matches in the high-upset range more accurately than the betting markets. Even using basic money allocation tools, it is still feasible to achieve consistent returns with relatively low risk. In the next portion of our research, we will explore how using formal tactics can help to further optimize our strategies.

Formal Strategies

As discussed previously, the main formal strategies researched within sports betting literature are variations of MPT and the Kelly Criterion. When considering which formal strategies to include in our research, we decided to eliminate MPT. The premise of this theory is to reduce overall risk by creating a portfolio of investments which are negatively correlated with each other. To our knowledge, this is only possible in tennis for bets on the same match; for example, we can imagine that the number of service games won by one player is negatively correlated with the percentage of break points won by his opponent. However, our strategy relies on betting on one single outcome: the winner of the match. For this reason, using MPT would not fit in the context of our research objective.

By contrast, we expect the Kelly Criterion to work well within the context of our strategy, notably as it is applicable to sequential, independent bets. The inputs of the Kelly Criterion are the odds and the probabilities from our model and the market. Research shows that without modifications this strategy proves to be too risky in practice, as it is founded on the unrealistic assumption that the true probability of an event is known. Uhrin, Šourek, Hubáček, and Železný (2021) studied the performance of eight different formal strategies across three different sports, as measured by a variety of different return and risk metrics, and found that generally the fractional Kelly was the best approach. The fractional Kelly strategy consists of multiplying the Kelly formula given in Equation 6 by a parameter α, in order to reduce the fraction of total wealth invested.

$$\text{Fractional Kelly} = \alpha \cdot \text{f} \qquad (15)$$

Adopting the Kelly strategy requires two changes from the structure of our previous strategies. First, we are now assuming that we start with a certain bankroll. For simplicity purposes, we will set the starting bankroll to $10,000 in the following simulations. This bankroll will be updated after each match that is bet on; either it will go down by the bet amount in the case of a loss, or it will go up by the profit from the bet in the case of a win. The bet amount is shown in Equation 16.

$$\text{Bet Amount} = \text{Bankroll} \cdot \alpha \cdot \text{f} \qquad (16)$$

Second, we only enter a bet if the probability estimated by our Fisher model is higher than the upset probability inferred from the betting odds, as otherwise the Kelly fraction would be negative. Thus, in addition to requiring our model output to exceed a certain threshold, we are also requiring it to exceed the bookkeeper’s upset probability. If these two criteria are simultaneously met, we will bet an amount determined by the fractional Kelly (Equations 6, 15, 16).

Parameter Selection

The first step to creating our formal strategy was to determine the optimal threshold T and the optimal Kelly fraction α. Instead of using median profit to gauge the performance of our strategy for different threshold values, we opted for the median ending position (final bankroll value), as this measure fits well tracking wealth accumulation over time. Using the same approach as for our informal strategies, we ran 1,000 simulations of 1,280 matches each for 360 (T, α) combinations. The results are shown in the contour plot of Figure 10.

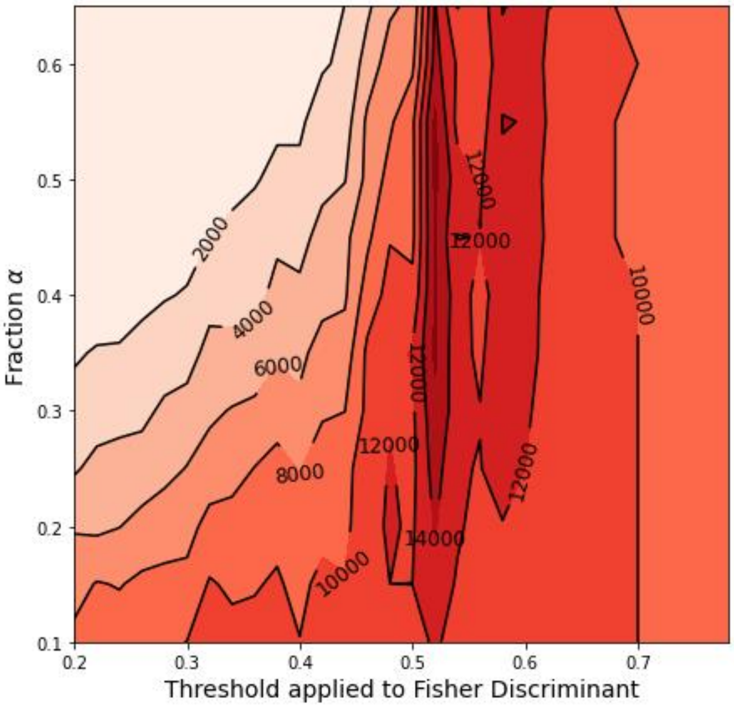

Figure 10. Median bankroll end position in the (T, α) space.

As evidenced by Figure 10, the strategies with the highest median end positions are concentrated within a narrow range of threshold values peaking for 0.52 (and for a wide range of α values).

While choosing the 0.52 threshold is clearly optimal, selecting the right α is a more difficult task. Our first approach to studying the (0.52, α) space was to scan the different profitability metrics for various values of α. The results are presented in Figure 11.

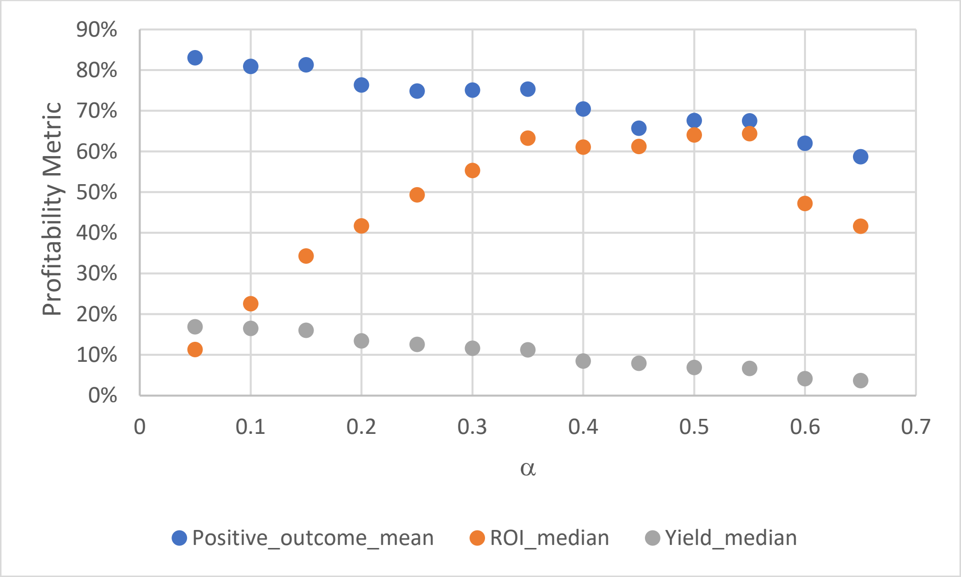

Figure 11. Fractional Kelly performance metrics for different α strategies with a Fisher threshold of 0.52.

In Figure 11, yield is defined in the same way as for the previous informal strategies, while the positive outcome metric is defined as to the percentage of simulations which ended with a bankroll value above the initial bankroll ($10,000). The return on investment metric, or ROI, is calculated as detailed in Equation 18.

$$\text{Positive Outcome} = \begin{cases} 1, & \text{if End Position} \geq \text{Start Position} \\ 0, & \text{if End Position} < \text{Start Position} \end{cases} \qquad (17)$$

$$\text{ROI} = \frac{\text{End Position}}{\text{Start Position}} - 1 \qquad (18)$$

As shown by Figure 11, both the average number of positive outcomes and the median yield generally decrease with larger values of α. Median ROI rises until α = 0.35, at which point it stabilizes then ultimately starts decreasing. One possible approach to selecting a value of α could be to minimize strategy risk whilst maximizing strategy yield. To this end, strategies with values of α ranging 0.05 to 0.15 seem like strong candidates. They all have very similar (high) positive outcome means and median yields, with the strategy for α = 0.15 appearing to have a higher ROI than the other two. Further details on profitability across the (0.52, α) space are included below.

Strategy (0.52, α)	**Positive Outcome Mean**	**Median ROI**	**Median Yield**	**Median End Position**	**Median Cost**
(0.52, 0.05)	83%	11%	17%	$11,127	$6,686
(0.52, 0.10)	81%	23%	17%	$12,253	$13,937
(0.52, 0.15)	81%	34%	16%	$13,426	$21,665
(0.52, 0.20)	76%	42%	13%	$14,165	$29,533
(0.52, 0.25)	75%	49%	13%	$14,934	$37,970
(0.52, 0.30)	75%	55%	12%	$15,534	$47,216
(0.52, 0.35)	75%	63%	11%	$16,328	$57,845
(0.52, 0.40)	70%	61%	8%	$16,104	$65,001
(0.52, 0.45)	66%	61%	8%	$16,120	$71,795
(0.52, 0.50)	68%	64%	7%	$16,406	$82,831
(0.52, 0.55)	68%	64%	7%	$16,433	$93,766
(0.52, 0.60)	62%	47%	4%	$14,716	$97,477
(0.52, 0.65)	59%	42%	4%	$14,160	$102,238

Table 11. Profitability metrics for 0.52 threshold strategies.

Based on the table above, the decision to choose an α value is somewhat subjective and is dependent on the risk tolerance of the bettor. Given that our research objective is to study the feasibility of making money consistently through sports betting and is not necessarily to achieve maximum returns, we chose to adopt a risk-averse approach for our strategy selection. Thus, we decided to select an α value of 0.15, as this strategy possesses an average number of positive outcomes and a median yield quasi-identical to the strategies with smaller α values, while generating a much higher median ROI. Though strategies with higher α values have higher ROIs and median ending positions, we determined that the reduction in average positive outcomes and the extremely elevated median costs make these strategies less appealing, particularly for risk-averse individuals who do not want to gamble large fractions of their portfolio on single matches.

Multi-Year Performance and Back-testing

To this point, we have only studied the profitability of our strategies over simulations of 1,280 matches, a number chosen to represent one single year. After proving that our formal strategy could generate consistent returns over an individual year, our next step was to find out what the profitability metrics would look like over longer periods of time. We therefore chose to calculate the expected returns of our strategy for periods of 5 years and 10 years. In order to do this, we multiplied the number of matches in a given simulation by 5 or 10 respectively. For example, our 5-year horizon simulation would include 6,400 matches (= 5 · 1,280). We then ran 1,000 simulations for each time horizon (5-year and 10-year); the results are presented in Table 12.

Betting Period	Positive Outcome Mean	Median ROI (Annual.)	Median End. Pos.	Avg. Bets Placed	Avg. Bets Won	Median Cost
1 Year	81%	34%	$13,426	59	32	$21,665
5 Years	99%	34%	$43,888	298	163	$224,830
10 Years	100%	34%	$188,703	596	326	$1,188,947

Table 12. Profitability metrics and data for the (0.52, 0.15) strategy over multiple time horizons.

Thus, our strategy performs increasingly well over time, as our returns are compounded by the fractional Kelly approach. Furthermore, our risk decreases as the number of matches increases – the average percentage of positive outcomes reaches 100 for simulations of 10 years. We also asked ourselves how we would have fared had we had this model available 11 years ago, back in 2012. To answer this, we back-tested our strategy starting in January 2012 with a bankroll of $10,000. The following graph shows the bankroll evolution over time.

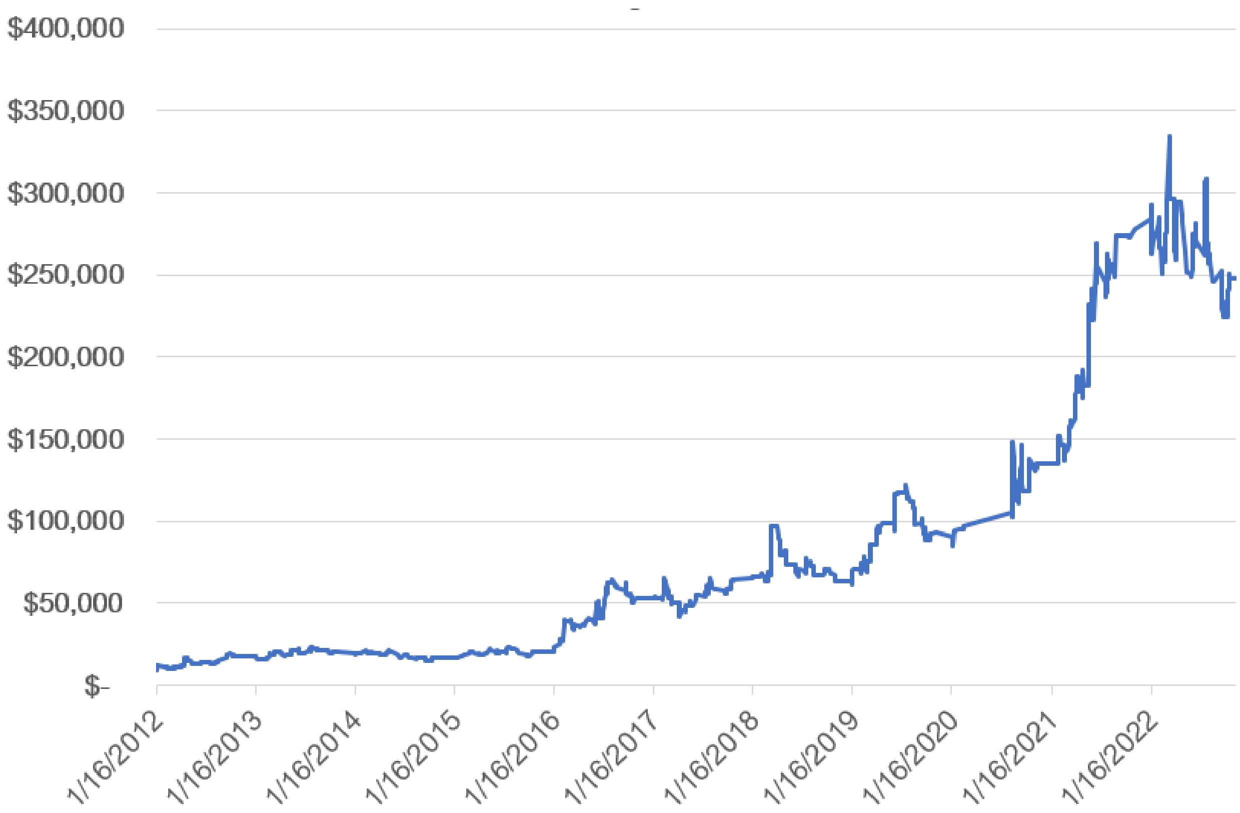

Figure 12. Bankroll evolution from January 16th, 2012, to November 20th, 2021, using the (0.52, 0.15) strategy.

As demonstrated by the graph, the strategy would have been successful over the past 11 years. Our bankroll would have grown slowly for the first 5 years, before exhibiting more rapid increases under the compounding mechanism of the Kelly equation; starting with $10,000, we would have accumulated $248,095 by November 20, 2022.

CONCLUSIONS AND OUTLOOK

In this research we set out to understand if it is possible to generate consistent profits through sports betting. By narrowing our focus to men’s tennis, and by using a combination of predictive analytics and financial strategy, we have shown that this goal is achievable. Our informal and formal strategies show positive median profits and end positions, for relatively sustainable levels of risk. These results are not only significant on a per-year basis, but are also consistent over longer term periods. In the process of obtaining these results, we were able to draw several interesting conclusions.

First, prediction in sports betting is an extremely difficult task. In men’s tennis, the higher-ranked player wins in approximately two-thirds of matches; despite this, the best predictive models are only correct around 70% of the time. In order to build our own predictive model, we included 49 different variables, spanning player profiles, situational information, time-series aggregations, service and return metrics, and others. These variables were combined through a Logistic Regression (LR), as well as a Neural Network (NN) approach. Furthermore, a Fisher discriminant was built from the LR and NN outputs, which improved our discriminative ability.

Second, we found that in order to generate profits, one does not need to outpredict the bookkeepers across all types of matches. A viable strategy can be to focus on a specific subset of matches and tune the predictive models on that subset alone. In our case, we chose to focus on upsets. To this extent, we trained our NN model with equal signal and background proportions to increase our upset recognition capability, despite the fact that this decreased our overall predictive accuracy.

As for the financial portion of the research, we first determined that profits can be made through informal strategies. We showed that varying the bet amount according to either the difference or the ratio between the bettors' and the bookkeepers' probabilities, leads to higher expected profits, at the cost of higher risk. We then formalized our strategies through the application of the fractional Kelly Criterion, which tied the bet amount to the bankroll available at the time of the bet. Our formal strategy research proved that we are able to achieve higher returns, for sustainable levels of risk, when optimizing our Fisher threshold and Kelly fraction α.

We made a number of choices throughout our analysis, and further research is needed to understand how one can improve on this approach. One area of interest could be our choice of sample. For example, more research could be done on lower-stakes tournaments, as well as on other disciplines within tennis, including women's tennis, doubles, mixed doubles, etc. As for the predictive modeling section of our research, one could potentially look into adding more variables as well as aggregating variables over different intervals instead of one year or six months. Additional multivariate methods could be tested. In terms of bankroll management, further research could include studying the effectiveness of the different ways of determining the bet amount, outside of the fractional Kelly Criterion. Furthermore, it could be interesting to build strategies that exploit the availability of multiple betting odds (bet365, Pinnacle Sports, etc.), in order to choose the most favorable ones for every match. Additionally, the spread between different betting odds could give bettors an extra input dimension for their betting decisions.

To conclude, this is an exciting area of research and we look forward to more interesting results in the future.

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to Dr. Eric Bradlow for advising me over the past year. He has provided invaluable advice throughout this research and ensured I progressed steadily between our bi-weekly meetings.

I thank Dr. Catherine Schrand for her guidance throughout the Fall 2022 semester, and especially for helping me narrow down the scope of my research and methodology. Last but not least, I am grateful to Dr. Utsav Schurmans for his continued support during my four years in the Joseph Wharton Scholars program.

REFERENCES

Abadiet Martín et al. 2015. “TensorFlow: Large-scale machine learning on heterogeneous systems”, 2015. Software available from tensorflow.org. We have used the sequential model:

https://www.tensorflow.org/guide/keras/sequential_model?hl=en

Bruce, Aaron, 2021. “What Percentage of Sports Bettors Win?”, Sitpicks.com article archived at: https://web.archive.org/web/20221129211945/https://sitpicks.com/what-percentage-of-sports-bettors-win/

Cuban, Mark. 2004. “My New Hedge Fund”, Blog Maverick article available at: https://blogmaverick.com/2004/11/27/my-new-hedge-fund/

Czermak, Chris. 2021. “Tennis Popularity Statistics 2021”, Tennis Creative article available at: https://tenniscreative.com/tennis-popularity-statistics/

Egidi, Leonardo, Pauli, Franceso, and Torelli, Nicola. 2018. “Combining historical data and bookmakers’ odds in modelling football scores”. Statistical Modelling. 2018;18(5-6):436-459. doi:10.1177/1471082X18798414

Fitt, Alistair D. 2008. “Markowitz portfolio theory for soccer spread betting”, IMA Journal of Management Mathematics, Volume 20, Issue 2, April 2009, Pages 167–184, https://doi.org/10.1093/imaman/dpn028

Gramlich, John. 2022. “As more states legalize the practice, 19% of U.S. adults say they have bet money on sports in the past year”, Pew Research Center article available at: https://www.pewresearch.org/fact-tank/2022/09/14/as-more-states-legalize-the-practice-19-of-u-s-adults-say-they-have-bet-money-on-sports-in-the-past-year/

Grand View Research, 2023. “Sports Betting Market Size, Share & Trends Analysis Report By Platform, By Betting Type (Fixed Odds Wagering, Exchange Betting, Live/In-Play Betting, eSports Betting), By Sports Type, By Region, And Segment Forecasts, 2023 – 2030”, Report ID: GVR-4-68039-539-7, available at: https://www.grandviewresearch.com/industry-analysis/sports-betting-market-report

Hubáček, Ondřej, Šourek, Gustav, and Železný, Filip. 2019. “Exploiting sports-betting market using machine learning”, International Journal of Forecasting, Volume 35, Issue 2, 2019, Pages 783-796, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2019.01.001

Hung, Jane. 2010. “Betting with the Kelly Criterion”, 2010. Article available at:

https://sites.math.washington.edu/~morrow/336_10/papers/jane.pdf

Kelly, John L. 1956. "A new interpretation of information rate," in The Bell System Technical Journal, vol. 35, no. 4, pp. 917-926, July 1956, doi: 10.1002/j.1538-7305.1956.tb03809.x

Manfred, Tony. 2012. “The World's First Sports-Betting Hedge Fund Has Collapsed After Losing $2.5 Million” Business Insider article available at:
https://www.businessinsider.com/sports-betting-hedge-fund-collapses-2012-1

Markowitz, Harry. 1952. “Portfolio Selection.” The Journal of Finance 7, no. 1 (1952): 77–91.
https://doi.org/10.2307/2975974

Pedregosa, Fabian et al. 2011. “Scikit-learn: Machine Learning in Python”, Journal of Machine Learning Research 12, pp. 2825-2830 (2011). We used several libraries including: *sklearn.preprocessing.StandardScaler, sklearn.linear_model.LogisticRegression, sklearn.discriminant_analysis.LinearDiscriminantAnalysis.*

Sackmann, Jeff. 2022. Tennis database by Jeff Sackmann / Tennis Abstract is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Based on a work at: https://github.com/JeffSackmann.

Stekler, H.O., Sendor, David, and Verlander, Richard. 2010. "Issues in sports forecasting", International Journal of Forecasting, Volume 26, Issue 3, 2010, Pages 606-621, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2010.01.003

Uhrín, Matej, Šourek, Gustav, Hubáček, Ondřej, and Železný, Filip. 2021. "Optimal sports betting strategies in practice: an experimental review", IMA Journal of Management Mathematics, Volume 32, Issue 4, October 2021, Pages 465–489, https://doi.org/10.1093/imaman/dpaa029

Toogood, Darren. 2022. "The Most Popular Sports That People Bet On", Islandecho article available at: https://www.islandecho.co.uk/the-most-popular-sports-that-people-bet-on/

Veroutsos, Eleni. 2022. "The Most Popular Sports In The World", WorldAtlas article available at: https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

Wilkens, Sascha. 2021. "Sports Prediction and Betting Models in the Machine Learning Age: The Case of Tennis" (September 1, 2021). Journal of Sports Analytics, vol. 7, no. 2, pp. 99-117, 2021. doi: 10.3233/JSA-200463

APPENDIX A

In this Appendix, we test the profitability of a simple model in which we bet a fixed amount ($100) on all predicted upsets (i.e., matches which have a Fisher discriminant value larger than 0.5). There are 1,512 predicted upsets in our data. Our total cost for entering all these bets is therefore $151,200. The total revenue, computed using the published odds from bet365, is $159,767. The net profit is $8,567, or approximately 5.67% of the total amount bet. Table 13 shows the cost, revenue, and profit for the different match types.

Subsample	Number of matches	Cost	Revenue	Net Profit
Actual non-upset – only bookkeeper's prediction is correct	362	$36,200	-	($36,200)
Actual non-upset – both predictions are incorrect	256	$25,600	-	($25,600)
Actual upset – only bettor's prediction is correct	321	$32,100	$74,344	$42,244
Actual upset – both predictions are correct	573	$57,300	$85,423	$28,123
Total	1,512	$151,200	$159,767	$8,567

Table 13. Betting a fixed $100 on matches which have a Fisher discriminant value larger than 0.5. This strategy, while simple, is profitable. Most of the profit is made on upsets which the bookkeeper predicted incorrectly.

This model demonstrates that positive-return strategies exist. Section 4 presents several strategy frameworks which optimize the financial performance of our approach.