Gambling behavior through the lens of big data

A dissertation submitted towards the degree

Doctor of Philosophy

of the Business Informatics Program

of Doctoral School of Economics,
Business and Informatics

of Corvinus University of Budapest

by

Sándor, Máté Csaba

supervised by

Bakó, Barna Ph.D.

Budapest, 2025

Acknowledgements

First of all I want to thank my wife, Klára for pushing me through the hardest parts and providing support over the long weekends, nights and conference travels. During the last few years she also gifted me with two beautiful and smart boys, who also were supporting in their own ways during these years.

I was lucky enough to find a supervisor like Barna Bakó, who besides giving me very high agency over my research topics, helped me to explore my own research interest and style while supporting me with just the right amount of scientific debate where it was needed. He also helped a ton in honing my skills in writing and provided enormous amounts of support in the publication process. For all this I am deeply indebted to him and will not let him get away without scaling some Transilvanian rock faces together first.

As many know, this was not my first attempt to obtain a PhD and while I was not successful in my previous attempt, the support and teachings of Gábor Papp and István Csabai in terms of technological training and overall research ethics were foundational building up to my current attempt. I also want to mention the mentorship and motivation provided by Imre Kondor during this time. He gave me the first push to open my interest towards finance and economics, which led me to amazing intellectual and professional places.

I am thankful to my friends and extended family for providing support and motivation in various forms over the years and continuing.

I would like to dedicate this work to all the individuals and families who have been failed by society by not providing sufficient support or even worsening their fights with inner demons through gambling and thus we lost. I hope humanity over time will triumph over the greedy malpractices of today's gambling industry and through historical clarity we can see it's horrors for what they were.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

My first experiences from childhood about gambling were from around being 8 years old: family gatherings with cheerful afternoon sessions watching the lottery draws on Sundays and matching the numbers on the screen with our tickets hoping for a big win. However, as I grew older, I have learned about the dark side of gambling.

The first big rural wedding that I have participated in and have a good recollection of was the wedding of my mother's childhood friend's. I remember the great feast and dancing well into the night, the warmth of which is still lingers in my memories. About a year later I remember my mother crying while driving us to school in the morning. When I have asked her to explain to me what is the problem, she hesitantly told me that this little brother, who had a great career ahead of him as a fireman and just started his adult life as a husband, have committed suicide over gambling debt, that he recently and rapidly amassed. It was my first experience of an act of suicide in my social surroundings, that got me deeply puzzled, since until that day, gambling meant fun.

As I grew older, but still as a young boy, I have observed more people engage in gambling around me in pubs and other public places.[1] While sometimes it seemed fun, most of the time I have observed either numb or even catatonic people watching the flickering lights, spinning fruits and upbeat sounds, or outright angry and/or sorrowful men arguing with the

[1] Slot machine gambling in pubs was not sanctioned until 2012 in Hungary.

humanized machine. I have seen teens drop their savings into the "one armed thief", and hear stories about resolving to petty crimes to pay up for their debts.[2] These observations and stories, however less severely, have popped up again and again later on in my early adulthood in the context of sports betting.

During my masters, my friend Dani showed me his research looking into the properties of the BitCoin network. We were discussing the weird concept of this digital toy money, looking into the transaction network, and also how people started to use this new tool. We have seen, that the biggest generator of transactions were gambling sites, offering mostly very basic casino games. Estimating the revenue they might had amazed us, since it was an astronomical sum in real value even with BitCoin's then exchange rate of $\sim$ 100$. Around this time the second most popular use of this new currency was that of darker illicit activities, centering around the trafficking of drugs. It was a time when we did think about buying into this new craze, but eventually decided that it's mostly funding illegal and amoral markets.

Besides making the biggest financial mistake of my life I also made notes, that the gambling data stored in BitCoin's blockchain could be an interesting base material to understand why and how all these people get hooked on something so simple and - to the simplest financial terms - so utterly irrational. By this time I got educated in statistics and started to open my interest towards finance and economics. This was also the time when I met my now supervisor, Barna, and discussing the topic with him he encouraged me to start collecting data for later research. I have also armed myself with the toolbox of data professionals: coding, algorithms, visualization tools, thinking in tables, stories and graphs. Fighting technological complexity with even more technology to reduce it. Then the pandemic brought the final push for me to get back to my academic interests and try to get closer to understand one of the great puzzles lingering around in my psyche through both memories of joy and trauma: why and how do people gamble?

[2] "One armed thief" is a Hungarian synonym for the classic mechanical slot machine.

1.2 General overview and problem statement

Problem gambling has been and will continue to be a lasting and extended issue affecting tens of millions worldwide, with significant socio-economic implications. The U.S. National Council on Problem Gambling (2021) concludes that 11% of gamblers have shown at least one sign of problem gambling "many times" in the preceding 12 months. In Europe, problem gambling rates vary between .3% in Ireland to 6.4% in Latvia, according to Carran (2022). The picture globally does not look much better: based on the report of the World Health Organization (2024) estimates suggest that 1.2% of the world's adult population has a gambling disorder. The rise of online gambling (combined with the aftermath of the COVID-19 epidemic), including sports-related and casino games have largely inflated this problem, particularly among young people, with an estimated 21% of online gambling in Europe occurring outside regulated environments.

Although problem and at-risk gamblers cover a fraction of total gambling population, they generate the larger part of gambling revenue, as e.g. Team et al. (2024) shown that 10% of players generated 90% of total casino expenditures in Massachusets in 2024. These losses can lead to an increased incidence of mental illness and suicide, instead of providing leisure and being a source of fun. While disproportionately affecting lower income households, gambling harms also include family violence, financial distress, connection to crimes, neglect of children and erosion of civil institutions via corruption and political activity. While treatments do exist, only .14% of the population is actively seeking formal or informal help for their problems (Bijker et al., 2022).

The global gambling industry have generated $536bn in revenue in 2023 with an expected growth of 7% in 2024 largely driven by digital platforms, and is estimated to reach $700bn by 2028 (World Health Organization, 2024). The normalization of gambling through heavy commercialization and digitization makes the market more accessible, while the industry is also working heavily toward the deregulation of online gambling. An example is how Hungary's national monopoly has been broken up through EU trade regulation principles (see European Gaming Industry News (2023)), which, while in principle meant to push online gamblers toward licensed operators, managed to increase the growth potential in online gambling markets, while left it dominated by unlicensed offshore operators (Vali

(2024)). Besides this, the gambling industry's main response to treatment is the so-called responsible gambling, which - following the model of the energy and garbage (plastic) industry - is effectively blaming the end consumers, ergo the group who mainly suffers the externalities of it.

Addressing the issue of problem gambling requires both economic and psychological understanding as well as political and societal interventions. From an economic perspective, it is crucial to understand the demand for gambling services and the what and how of regulatory impact. Models can help predict the effectiveness and extent of interventions and increase the effectiveness in allocation of resources towards prevention and support. Integrating insights from economics, psychology and technology can enable policy makers to tackle the delicate issue of market regulation. Then again, the question in the end remains if the political will in representing the welfare of the general populace is stronger than the influence of an (almost) trillion dollar industry. To end this section on a brighter tone, there are signs of strengthening general regulations (e.g. Solon and Zuidijk (2024)), that is targeting to reduce the complexity of tackling online gambling with levys, personal limits and timing restrictions.

Modern machine learning technologies are being applied in marketing and widely used in other industries as well, like targeted advertisement and player retention. The basis of these, player tracking gambling datasets, have been mostly left unused by policy making (outside of the observation of generic macro-trends), besides research articles founded on proprietary datasets provided by the market or through regulators. Aligning with the zeitgeist, the attention of industry participants is shifting towards AI solutions around responsible gaming based on tracking data (see Behe (2024)), for better or worse. Thus it is a pivotal moment for research that meant to support the regulatory sides to concentrate on these data intensive issues, so it can keep up with the market competition for the gambler's attention, money and health.

1.3 Aims, objective and research questions

The main goal of my PhD research was to delve into the aspects of gambling research centered around data-intensive, multidisciplinary behavioral tracking approaches, and to

do so in a manner that will only involve datasets and methods that can be made public at the end. To be able to deliver on these aims my foremost task was to find an eligible source and create such dataset that can be made public ethically, while still providing the critical features available in proprietary datasets.

After this, beside familiarizing myself with the research landscape of the area, I wanted to review the methodological toolbox used in these studies and provide support or pose challenge towards them through replication. I also wanted to see if I could provide industry-level tools to the effort of policy making, using or combining the contemporary modeling and intelligence approaches in an open and reproducible manner. Finally I was looking for ways to leverage the special nature of my dataset - being sourced from the crypto-space - that can extend the research focus of the current gambling studies around tracking data. The above objectives have crystallized over time around the three projects that are presented as papers in this work.

1.3.1 Approaching the Hot Hand with a Cool Head

In this project my goal was to leverage the large-scale gambling data mined from the Bitcoin-base online gamble SatoshiDice to challenge the findings of Xu and Harvey (2014) on streak-dependent gambling behavior, particularly focusing on their methodologies in investigation of the hot hand fallacy and the gambler's fallacy. My core research questions can be formalized as follows:

1. Can the observed aggregate-level trends in betting behavior - as reported in Xu and Harvey (2014) - be explained without assuming behavioral biases?
    - Null Hypothesis (H_0): The observed changes in betting behavior after streaks arise due to underlying behavioral biases (e.g., hot hand or gambler's fallacy).
    - Alternative Hypothesis (H_A): The observed changes in betting behavior can emerge without the assumption of underlying behavioral biases, such as the hot hand or gambler's fallacy.
2. Do individual gamblers exhibit consistent behavioral biases (hot hand effect or gambler's fallacy), or do aggregate-level trends emerge from a heterogeneous population

of persistent risk-takers?

- Null Hypothesis (H_0): Individual gamblers frequently adjust their betting strategy in response to winning or losing streaks, supporting the presence of hot hand and gambler's fallacies.
- Alternative Hypothesis (H_A): Individual gamblers tend to maintain their chosen risk levels, and the observed trends in aggregate data emerge from persistent heterogeneity in risk preferences rather than systematic biases.

By analyzing these questions I directly challenge the decision-making theory of Xu and Harvey (2014) and the arguments of Xu and Harvey (2015) while making an effort to highlight the importance of methodological rigor. During the study I was trying to stay as close as possible to the methodological approach of the original article from the field of cognitive psychology. To show the presence of selection bias in their studies I was to apply empirical replication, statistical reasoning, simulation and some - albeit trivial - analytical derivation.

1.3.2 Unmasking Risky Habits: Identifying and Predicting Problem Gamblers Through Machine Learning Techniques

In this study I wanted to explore if machine learning methods can be used to identify and predict problem gambling behaviors without relying on self-labeling or psychological profiling. I also wanted to utilize a diverse set of machine learning methods using an automated modeling approach. My core research questions can be formalized as:

1. Can unsupervised machine learning techniques be used to identify problem gambling behaviors without relying on self-reported labels?
    - Null Hypothesis (H_0): Problem gambling behaviors cannot be meaningfully identified using unsupervised learning methods.
    - Alternative Hypothesis (H_A): Unsupervised learning (e.g., trimmed k-means clustering) can successfully classify gamblers into meaningful behavioral groups based on observed betting patterns.

2. How accurately can supervised machine learning models predict whether a gambler will develop problematic gambling behavior?

    - Null Hypothesis (H_0): Supervised machine learning models trained on early gambling behaviors do not perform significantly better than chance in predicting problem gambling.
    - Alternative Hypothesis (H_A): Machine learning models can accurately predict whether a gambler will exhibit problem gambling behaviors based on early gambling patterns (e.g., bet frequency, session duration, and bet size variations).

Given the limitations of self-exclusion and self-reports I wanted to explore the a data-driven alternatives in the identification of at-risk gamblers and to highlight the practical implications and power of modern machine learning forecasting methods for operators and policy makers. In this study I have leveraged methods of unsupervised learning and a broader family of classification algorithms combined through ensembles and automation to classify and predict problem gambling tendencies.

1.3.3 How Bitcoin's Ups and Downs Are Changing the Way You Bet

In this research chapter I wanted to explore the relationship between fluctuations in the price of Bitcoin and gambling behavior on the LuckyBit platform. I wanted to assess this relationship from multiple aspects of gambling behavior and to see if the results are consistent across different types of gamblers. My core research questions can be formalized as:

1. How do Bitcoin price changes impact bet sizes and gambling frequency?

    - Null Hypothesis (H_0): Bet sizes and gambling frequency remain unchanged regardless of changes in Bitcoin's price and volatility.
    - Alternative Hypothesis (H_A): Increases or decreases in Bitcoin's price and volatility significantly affect bet sizes and gambling frequency, either encouraging or discouraging betting activity.

2. Do different types of gamblers (casual, committed, extreme) respond differently to Bitcoin price fluctuations?

    - Null Hypothesis (H_0): There are no significant differences in how casual, committed, and extreme gamblers adjust their betting behavior in response to changes in Bitcoin's price and volatility.
    - Alternative Hypothesis (H_A): Different cohorts of gamblers react differently to fluctuations in Bitcoin price and volatility, with extreme gamblers potentially showing a stronger reaction due to risk-seeking tendencies or automated betting strategies.

With the rise of crypto-based gambling, it is essential to understand how external financial factors shape betting decisions. In this study I wanted to link gambling behavior to the broader financial decision-making process, to further our understanding on market-driven risk behavior and perceived wealth effects. In terms of tools I have used econometric linear modeling methods with some advanced validation techniques to control for the effects of larger models and datasets.

1.4 Organization

In Chapter 2. I will give a review of the cross-section of the fields of behavioral economics, decision making under risk and gambling and provide a short overview of tools and methods I have used in the presented papers, but were not described in detail in the respective methodology sections due to the limitations of the article formats.

In Chapter 3. I will present the backbone of my empirical works that is the extraction of transactions of various online gambling sites working in the ecosystem of BitCoin. I will give a detailed analysis of the sites used, the method of data extraction and the structure of the resulting datasets. Some of the discussion may show up again in the articles later to preserve the format they were accepted in.

In Chapter 4, 5 and 6 I present the research papers in their most recently reviewed and submitted (4 and 6) or published (5) state, with minor changes performed to integrate reviewer feedback on the thesis and some extensions provided as appendices.

Chapter 7. provides a collection of the conclusions and designates possible future research directions and developments.

1.5 Division of work

My supervisor, Barna Bakó was a tremendous help and support during the research process. I consider all of the following research projects to be joint work. Barna provided consultation and scientific sparring during the exploratory phase of the research projects, reviewed the first drafts of the research articles and then we reworked the form and wording of all articles together before turning them in for review at targeted journals. He also helped tremendously by managing the review process and driving the discussions on suggested reworks and reviewer responses. The research formulation, data collection, methodological research and design, coding and experimentation, code and data management, visualization and result formulation was carried out by myself for all research articles.

Chapter 2

Research methodology

The technological transformation of gambling made it more easily accessible than ever, thus increasing prevalence of problem gambling (Potenza et al., 2011; Chagas and Gomes, 2017). While the recent pandemic reduced the overall number of gamblers, caused an increase in the number and severity for online and problem gamblers (Wardle et al., 2021; Hodgins and Stevens, 2021) disproportionately affecting the younger generation. The societal costs associated are estimated to have huge impact on the economy (Hofmarcher et al., 2020).

Understanding the behavior of gamblers and the patterns in the way people perceive risk is a crucial task for both psychology, economics, sociology and businesses alike. The study of how gamblers should (economically) rationally behave and how do they actually behave have been a challenging and fruitful topic of sciences for the last few centuries as the development of probability theory itself has roots in games of chance (De Laplace, 1995).

In this chapter, I will outline the general theoretical background of my research interests. I will also present concepts and practices tied to methodological choices. Beside the following, each study chapter will provide it's own introduction to relevant concepts.

2.1 Gambling as a rational choice: prospect theory

The standard model of economic decision making under risk works through the concept of expected utility, where we define a convex utility curve over wealth. This model explains a plethora of economic phenomenon, but fails to work when it comes to gambles with

negative expected value, which these economic agents would never engage in. Whereas we see in real life that people very much do. To resolve this issue the scientists of the field tried introducing concave segments for the utility function (Friedman and Savage, 1948) or add utility to playing the game itself (Conlisk, 1993; Luce et al., 2008) but these approaches led to more inconsistencies with the original theory.

The revolutional idea of (cumulative) prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992; Barberis, 2013) - that came from the field of psychology - have resolved many inconsistencies that the classical economical models had around decision making with the presence of risk. The main idea of prospect theory is to introduce the biases observable in human decision making under risk to the expected utility function in two parts. The value function adjusts the payouts of the various outcomes of the risky situation, introducing a concavity over gains (the wealthier I am, the less joy more money brings me) and convexity over losses (one gets insensitive over additional losses after a big one) as well as loss aversion (we tend to rather avoid losses than to get an equivalent gain in a risky outcome). The probability weighting function introduces an adjustment so that the overestimation of improbable outcomes and the underestimation of highly probable outcomes can be accounted for.

Based on this Barberis (2012) showed that agents with a certain set of parameters describing their value and probability adjustments would actually enter the casino and a subset of them would fail to adhere to their original prospects regarding leaving it. This is highly in line with the observable behavior of casino players. This model although only describes simple gambles and still fails to describe some observed gambler behavior discussed later.

Multiple methodologies have been presented over the years to estimate the model parameters of cumulative prospect theory on both population-wide and individual terms (Murphy and ten Brincke, 2018; Krčál et al., 2016). These studies also suggest that gambler's risk profile may differ from those observed in laboratory experiments, namely loss aversion shows to be much less prevalent. From the researches focusing on betting or gambling data only Krčál et al. (2016) is based on a large cross-sectional dataset, so further studies are much needed. Another question that has not received much focus yet in empirical research is the stability of risk preferences over time or other external factors (Schildberg-Hörisch, 2018).

In the short term the individual risk propensity of agents are found to be stable, although in the life-long term a constant shift is observable with visible impact of external shock events. Regarding gambling, the longitudinal research of the subject is lacking.

2.2 Identifying trends in gambling behavior

2.2.1 Repeated measures ANOVA

A method offered by Demaree et al. (2015) to compare the game choices of gamblers in different streak lengths was using repeated measures analyses of variance (ANOVA) method. Here he would expect a monotonous increase of implied winning pro with a significant difference of observed means as the streak progresses in the gambler groups divided over total streak length. Unfortunately in their response Xu and Harvey (2015) have not performed these tests, but rather provided another test that - as we will show in Chapter 4 - was flawed.

Repeated measures ANOVA is used to analyze data when the same subjects are measured multiple times under different conditions or - as in this case - different time-points or phases of an experiment. However, like many parametric methods, repeated measures ANOVA comes with some key assumptions about the distribution of the underlying data. It expects that the differences between the group means are normally distributed (*normality*), that variances of the differences between all groups - here streak lengths - are equal (*sphericity*) and that the variance within each group is similar (*homogeneity of variances*).

Applying this method to test the main argument of Xu and Harvey (2014) would make us apply it on choices of winning probability, the distribution of which is by it's nature bound on the scale $P_{win} \in [0, 1]$ and thus makes the first assumption of normality very hard to prevail. This can be somewhat mitigated by focusing on relative changes in the choice of P_{win} after each streak and maybe even its logarithmic transformation. Although these steps helped to increase the normality of our target measure, our observational setup provided a discrete choice set of winning probabilities and thus hindered the applicability of this method. A non-parametric methodological alternative could have been the Friedman or Kruskal-Wallis tests.

2.2.2 Bootstrapping

Bootstrapping can be a powerful alternative to traditional significance tests when we want to determine if a measure - here the change in choice of P_{win} after different streak-lengths - is statistically significant. In general to create a bootstrap measure $\Pi(X)$ given a dataset $X = \{x_1, x_2, \ldots, x_i^*, \ldots, x_\beta\}$ we generate β bootstrap samples:

1. We randomly sample n points from X with replacement to create β new datasets $X_\beta = \left\{x_1^*, x_2^*, \ldots, x_i^*, \ldots, x_\beta^*\right\}$
2. We compute the measure $\Pi(X_\beta)$ for each bootstrap sample β
3. We use the empirical distribution of $\Pi(X_\beta)$ to approximate the sampling distribution of $\Pi(X)$

This empirical distribution is then can be compared over different measures and sub-samples without relying on parametric assumptions. This approach can be especially useful when dealing with small sample sizes or complex statistics, where analytical solution may deem impossible to derive. It is also widely used in big data, because in principle it's explanatory power converges to that of the parametric tests with the sample size even if the underlying generator process is indeed parametric, but also keeps its explanatory value should these assumptions fail to hold.

2.2.3 Regularization

In their empirical article Salaghe et al. (2020) presents a much more delicate linear model (compared to Xu and Harvey (2014)) to investigate player response to streaks, including 11 predictors. To evaluate this problem over their huge dataset they utilized simple OLS regression, where the general objective function can be formalized as:

$$\min_{\beta} \|X\beta - y\|_2^2$$

One major drawback of the above linear regression approach is its susceptibility to over-fitting, especially when dealing with such high number of predictors. We call it overfitting when a model learns the noise rather than the true relationship signal from our dataset.

This leads to very high perceived fit (such as inflated R^2 values), that is not robust to cross-validation on sub-samples of our original data, hindering the validity of the relationships derived from the analysis.

Regularization techniques are crucial to mitigate this issue by adding penalty terms to the original OLS objective function, constraining the model's complexity and thus preventing it from fitting noise (Friedrich et al. (2023)). Using these additional terms we can either shrink the coefficient of overly estimated predictors or even turn unimportant or collinear predictors to zero and thus making our models more simple and robust, by introducing a trade-off between model fit and complexity.

The two most popular and useful methods are Ridge and Lasso Regression also known as L2 and L1 regularization, both adding a penalty term proportional to either the square (Ridge, $\|\beta\|^2$) or the absolute value (Lasso, $\|\beta\|$) of the magnitude of the coefficients to the OLS objective function. We can also combine the two:

$$\min_{\beta} \|X\beta - y\|_2^2 + \lambda \left(\alpha\|\beta\| + (1-\alpha)\|\beta\|^2\right)$$

where the regularization parameter λ controls the strength of the penalty (usually optimized through cross-validation) and $\alpha \in [0, 1]$ is adjusting between the two parts. Ridge regression tends to shrink larger coefficients to decrease their influence and thus avoid overfitting on a few large parameters, whereas Lasso can also force some less-important coefficients to be exactly zero, effectively performing feature selection. This makes Lasso ($\alpha = 1$) also particularly useful for identifying the most important predictors and creating a simpler yet more general model that stays robust over sub-samples of our training data and over out of sample data as well. If we consider that co-linearity might be present, we can introduce some Ridge regression into the regularization mix ($\alpha \in [.9, .6]$) to counterbalance the selective tendency of Lasso between.correlated explanatory variables. We have utilized this method in Chapter 6 to select the most important explanatory variables and achieve a simple model that is more easily comparable over different cohort samples.

2.3 Player labeling methods in gambling

A central question of behavioral gambling research is the description and identification of pathological patterns, or "problem gambling". Early psychological approaches focus on identifying characteristics such as the need for excitement, antisocial tendencies or using gambling as an escape, and through focus groups, Rockloff and Dyer (2006) developed methodologies like the four E-s of gambling (Escape, Esteem, Excess, and Excitement), while diagnostic systems like DSM-5 (American Psychiatric Association et al. (2013)) focus on indicators like loss of control and the need to acquire more money to sustain gambling. Many national regulatory frameworks offer through institutions or force market participants in making available to gamblers. These programs enable gamblers to self identify and -exclude from gambling for a certain period of time. While these methods show promising effects in casino gambling (Kotter et al., 2018), for internet gambling the efficacy seems mixed (Caillon et al., 2019; Giroux et al., 2017). A general problem of self-labeling is that it might be biased towards individuals who are already aware of their gambling problem, thus potentially missing a significant chunk from the true size of the affected population. To mitigate these biases associated with self-reporting, unsupervised machine learning techniques offer an alternative to create labels to cluster our player base and potentially separate problem gamblers.

Unsupervised clustering methods, such as k-means have broad applications across disciplines, including market segmentation, bioinformatics, and anomaly detection (Oyewole and Thopil (2023)). In behavioral economics, it is mostly leveraged to group individuals based on spending behaviors or risk preferences. Following the prominent literature of gambling research we were also utilizing this technique to separate groups of our observed gamblers we can label as problem gamblers to target in our prediction study discussed in Chapter 5. We have also used this method to create similar groups of gamblers that we can validate the robustness of our crossectional models introduced in Chapter 6.

2.3.1 Clustering using k-means

k-means clustering is the most popular of the unsupervised machine learning algorithms (Hastie et al. (2017)). It's function is to partition a dataset of n observations into k clusters based on feature similarity over d dimensions of the data. The main optimization measure of the algorithm is within cluster variance, which is measured over the cluster participant's distance from it's centers. The algorithm generally alternates between two steps after the (usually random) initiation of the k centers:

- We create the clusters for each centers, that consists of the points closest to (shortest distance from) them in the feature space
- We calculate the center of mass (the mean of each feature) for each cluster that becomes the new centers of mass

This cycle is then iterated until convergence, ergo when the new cluster centers become identical to the previous iteration's. The algorithm gives us certain liberties in its structure, the largest of which is the number of clusters k that has to be preset before initiation. Another is the definition of how distances are calculated between the participants in the cluster and their centers. The most straightforward choice, Euclidean distance however only provides good performance if the dataset is sufficiently compact in all d dimensions and the scale of within group deviation around the center is on comparable scales.[1] Thus linearization and scaling of the dataset's input features are necessary to balance the influence of the features in clustering importance. Another risk of this measure can be that it loses meaning beside high dimensionality, which is usually treated by its reduction through techniques like Principal Component Analysis before clustering.

Several other methods exists in the space of unsupervised learning that take a fundamentally different approach - like hierarchical clustering, which builds clusters from the ground-up growing them by linking the two closest observations in the feature space each step - or extensions of its basic logic like k-means++ that intelligently initialize centroids, leading to faster and higher quality centroids. Others, like Gaussian Mixture Model can

[1] Euclidean distance: $\delta(x_i, x_i') = \sum_{j=1}^{p}(x_{ij} - xi'j)^2$, where δ stands for distance, i runs over the data points and j over our features.

also be used to generalize the k-means approach to better control for highly imbalanced cluster sizes and/or densities. Despite these limitations it's interpretability and scalability makes it a foundational clustering method, widely used to this day in gambling research as seen in Braverman and Shaffer (2012); Auer and Griffiths (2024). During our analysis we were using the implementation of R Core Team (2013).

2.3.2 Trimmed k-means

One such extension that we have found particularly useful for our use case is the trimmed k-means clustering of Cuesta-Albertos et al. (1997). It's core idea is to make the original algorithm more robust by removing a proportion of potential outliers from the dataset during the clustering process, using "impartial trimming". By randomly trimming a proportion of data (α) as outliers and thus creating new and more robust k-cluster, we exclude data points that would otherwise skew the importance of features in the datasets. The algorithm also needs a predefined trim proportion of $\alpha \in [0, 1]$ to determine the size of trimmed data set A which it will jointly optimize with the best k-set M to minimize the trimmed variation criterion:

$$V_{k,\alpha}(M) = \inf_{A:P(A)=1-\alpha} \inf_{M=\{m_1,m_2,\ldots,m_k\}\subset\mathbb{R}^p} \frac{1}{P(A)} \int_A \min_{1\le i\le k} \|X - m_i\|^2 \, dP(X)$$

This trimmed group can be also treated as a group of outliers, but otherwise it will not adhere for the clustering centrality principle of the k-mean group generated in the end.

I have found this algorithm particularly useful for our special case of clustering our gamblers in Chapter 5 and 6 because in most dimensions of the datasets we have seen a large number of observations outside of the otherwise normal or normalizable distribution of observations. These observations could have resulted from both extreme behaviors as well as the result automated betting bots, both of which we wanted to separate from our sample. During our analysis we were utilizing the algorithms R implementation of Hennig (2020).

2.3.3 Goodness metrics for clustering

The most important parameter to decide when performing k-means clustering is the number of clusters k. Many methods have been created over the last few decades but the most popular and reliable ones we also used are:

- **Elbow Method**: this heuristic approach tries to determine the optimal number of clusters through the calculation of the within cluster scatter $W(k) = \sum_{\kappa} \sum_{i \in \kappa} \delta(x_i, x_\kappa)$ for each κ cluster when the number of cluster is k. We can calculate this measure for increasing numbers of k starting from 1 and visualizing $W(k)$ on a graph and selecting the "elbow" point, where the rate of decrease in the within cluster scatter flattens out, giving us an optimal choice of number of clusters.

- **Silhouette Score**: this measure considers intra-cluster [2] cohesion and inter-cluster [3] separation simultaneously by calculating the coefficient for each point:

  $$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\left(a(x_i), b(x_i)\right)}$$

  then calculating the average over all points. The resulting score will be bound between -1 and 1 with positive scores meaning a clustering better than random classification and values close to 1 signaling well-made clustering.

- **Dunn index**: is an enhanced version of the silhouette score comparing $\delta(C_\kappa, C_{\kappa'})$ minimum cross-cluster distance (between clusters κ and κ') and the Δ_κ maximum intra-cluster distance inside cluster κ:

  $$D = \frac{\min_{\kappa \neq \kappa'} \delta(C_\kappa, C_{\kappa'})}{\max_{\kappa \in k} \Delta_\kappa}$$

By combining the above metrics we can make a more informed decision on choosing k. For both our analyses, where we were clustering using k-means we have ended up with an optimal choice of $k = 2$ which luckily also made the most sense from the behavioral perspective. As a "bonus" we also got the outliers as an extreme group of seemingly truly edge-case gamblers.

[2] Cohesion $a(x_i)$ being the average distance between each point and all the other points in the same cluster

[3] Separation $b(x_i)$ being the average distance between each point and all the center of the nearest neighboring cluster

2.4 Predicting gambling behavior through classification

One central question of behavioral research is predicting human behavior. In gambling research corporations are mostly motivated to predict behavioral changes impacting revenue (e.g. see BetMGM (2023)) like churn that can then be avoided through targeted adverts and other incentives to stay gambling. However, most academic work focuses on forecasting the emergence of pathological and problem gambling behaviors for regulatory and prevention purposes. Nevertheless, once the target is set, and the observational features of our dataset is given, the toolbox of prediction will be the same for both parties. Below I will describe a few modeling methods I have utilized in in Chapter 5 predicting problem gambling status.

2.4.1 Logistic regression using generalized linear model

Binomial Generalized Linear Models (GLMs) provide a flexible probabilistic framework for estimating the likelihood of categorical outcomes based on explanatory variables (Train (2009)). The binomial GLM assumes that the probability p_i of an individual i choosing a particular behavioral category follows:

$$E(Y_i) = p_i, \quad p_i = g^{-1}(X_i\beta),$$

where Y_i is the observed binary outcome, X_i represents explanatory variables, β is a vector of parameters to be estimated, and g^{-1} is the inverse link function. The most common choice of link function in binomial GLMs is the logit function:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = X_i\beta.$$

Estimation of binomial GLMs is performed using maximum likelihood estimation (MLE), optimizing the log-likelihood function:

$$\log L(\beta) = \sum_{i=1}^{N} \left[Y_i \log p_i + (1 - Y_i) \log(1 - p_i)\right].$$

Binomial GLMs are widely applied in behavioral economics for modeling decision-making processes, policy interventions, and consumer behavior (e.g. Percy et al. (2016)).

Unlike standard logit models, they offer a more generalized framework that can accommodate different link functions, such as the probit or complementary log-log link. Despite their flexibility, binomial GLMs assume independent observations, which may not hold in cases of correlated choices. Extensions such as mixed-effects GLMs help address these dependencies. Overall, binomial GLMs provide a robust and - very importantly - easily interpretable approach to modeling categorical outcomes.

In our study we were using the GLM implementation of LeDell et al. (2022), although for explanatory analysis, other standalone implementations like the one of the R stats package of R Core Team (2013) or python's scikit-learn package of Pedregosa et al. (2011) can be utilized just as well.

2.4.2 Random forests

Random Forests (RFs) are an ensemble machine learning method that aggregates multiple decision trees to improve prediction accuracy and robustness (Breiman (2001)). RFs are particularly effective for modeling complex interactions in large datasets.

Each decision tree in an RF is trained on a bootstrapped sample of the data and uses a randomly selected subset of features to split nodes. The final prediction is obtained through majority voting (for classification) or averaging (for regression). A general decision tree model follows:

$$Y_i = f(X_i) + \varepsilon_i,$$

where Y_i is the predicted behavioral outcome, X_i represents the explanatory variables, and ε_i is an error term. By aggregating multiple trees, RFs reduce overfitting and enhance generalizability. While RFs provide high predictive accuracy, they lack the interpretability of parametric models such as GLMs. However, methods like feature importance scores and SHAP values (Lundberg and Lee (2017)) can enable researchers to analyze variable contributions.

In our analysis we were utilizing the H2O package's implementation of LeDell et al. (2022) which is mostly based on a distributed implementation of the XRT algorithm of Geurts et al. (2006). In extremely randomized trees (XRT), randomness goes one step

further in the way that splits are computed. As in random forests, a random subset of candidate features is used, but instead of looking for the most discriminative thresholds, thresholds are drawn at random for each candidate feature, and the best of these randomly generated thresholds is picked as the splitting rule. This usually allows to reduce the variance of the model a bit more, at the expense of a slightly greater increase in bias.

2.4.3 Gradient Boosting Machines

Gradient Boosting Machines (GBMs) as introduced by Friedman (2001) are based on forward learning ensembles. The key principle of it is that the trees built sequentially are targeted to correct the errors of the preceding ones instead of targeting the same estimation error, like we seen in random forests. Thus GBMs adaptively built trees can minimize different aspects and sources of estimation error and in the end fully utilize the explanatory power of all layers.

The general gradient boosting framework optimizes a loss function $L\left(y_i, F_m(X_i)\right)$ by iteratively improving the model:

$$F_m\left(X\right) = F_{m-1}(X) + \gamma_m h_m\left(X\right),$$

where $F_m\left(X\right)$ is the boosted model at iteration m, $h_m\left(X\right)$ is the new weak learner (a decision tree), and γ_m is the learning rate controlling the contribution of each new tree.

The most popular version of the algorithm (also used by us) is the XGBoost of Chen and Guestrin (2016). This is an enhanced version of the original, utilizing regularization (see Section 2.2.3), weighted quantile sketching and parallel processing. We were using the earlier mentioned implementation of the H2O R package.

Compared to Random Forests, gradient boosting methods tend to achive high accuracy, but are heavily susceptible to overfit and running times can scale fast with a large dataset. With careful variable selection, tuning of parameters and heavy parallelization we can get around these issues. Other, more lightweight (but less performant) versions also are widely used like AdaBoost from Schapire (2013) or LightGBM of Ke et al. (2017).

2.4.4 Deep learning with Neural Networks

A feedforward artificial neural network (ANN) model, also known as deep neural network (DNN) or multi-layer perceptron (MLP), is the most common type of Deep Neural Network. They are composed of an input layer, hidden layer(s) and an output layer, all of which are composed of neurons applying an activation function to transform the weighted sum of their inputs:

$$Z_j = f\left(\sum_{i=1}^{n} w_{ji} X_i + b_j\right),$$

where Z_j is the activation of neuron j, X_i represents input features of the neuron (output features of the preceding layer if hidden), w_{ji} are the weights of the connection, b_j is a bias term, and f is a (usually) nonlinear activation function (step- , rectified linear unit or ReLU, sigmoid, hyperbolic tangent, etc.).

The key to training these models is backpropagation, where we minimize the loss function $L\left(Y, \hat{Y}\right)$ (where Y are the true categories and $\hat{Y}$ are the outputs from the ANN) through gradient descent optimization:

$$\frac{\partial L}{\partial w_{ji}} = \delta_j X_i,$$

where δ_j represents the error term for neuron j. To prevent overfitting many hyperparameters must be set right like the learning rate, regularization and normalization. ANNs excel in categorization jobs based on tabular data, but other designs exist for special purposes, like Convolutional Neural Networks for image processing and Recurrent Neural Networks for time series forecasting. These models can provide even higher accuracy compared to tree-based modeling but require much larger sample sizes to do so with a significant increase in computational power required for training them. As we will see in 5 our datasets were not adequately large to fully capitalize the power of these models, but as part of an ensemble they still provided significant advances to our bottom line.

As with the earlier algorithms, we will use the implementation by H2O.ai (Candel et al. (2016)) in our calculations, but the most prominent and widely used toolsets to create neural networks are TensorFlow (Pang et al. (2020)) and PyTorch (Paszke et al. (2019)).

2.4.5 Ensemble models

Ensemble methods in machine learning use a combination of multiple learning algorithms to reach higher predictive performance, than a single model can. Actually some of the model families previously discussed are ensembles themselves: RFs are using unweighted averages of trees (bagging) and GBMs are using an adaptive stack of trees (boosting) to obtain high predictive accuracy. Both approaches are great methods to form a single, strong learner using a combination of weaker learners (e.g. single decision trees).

In our analysis we are using the ensemble method called stacking or super learning (Van der Laan et al. (2007)). First we create a set of "weak" learners $\hat{\Psi}_k$ on our input set of covariates W. These can be a variety of the previously proposed methods, with even multiple versions of the same algorithm with different tuning parameters reaching locally efficient estimations. We then create a cross-validation set (randomized training and evaluation subsets of our input data W) and evaluate all candidate models over all the sets. After this step we determine the weights $\hat{\alpha}_k$ that minimizes the combined estimation error over all k candidate models and validation samples by training a metalearner algorithm that can be evaluated over the cross-validation set. The metalearner algorithm can actually be any algorithm that can perform regression tasks (e.g. any of the algorithms we have introduced earlier). After obtaining the optimal model $\mathcal{E}$ we can derive our super learner:

$$\hat{\Psi}_{\mathrm{SL}}(W) = \mathcal{E}\big(\hat{\Psi}_1(W), \hat{\Psi}_2(W), \ldots, \hat{\Psi}_k(W)\big)$$

To use this new predictor we first generate predictions from the base learners, then we feed these to the super learner to generate the ensemble prediction. Stacked ensemble models tend to outperform single models in sample, although on smaller datasets it is not trivial that we will end up with a model that will be more performant out of sample as well. In case of mixing models that perform good and poor on unbalanced categorization tasks, rebalancing prior to training the super learner is suggested. In my research I was utilizing the implementation of LeDell (2015) through the H2O.ai autoML algorithm.

2.4.6 Automated Machine Learning

Automated Machine Learning, or autoML tools are a way to provide a simple interface to non-experts in ML to train a model or even large sets of it, without delving to deep into the ways of parameter optimization, feature engineering, cross validation, writing many lines of code or even spare much experimentation on model selection. This can make a novice machine learning practitioner's learning curve faster or to help an expert by creating excellent baseline models fast.

A powerful implementation of such method we were using is the H2O autoML of LeDell and Poirier (2020), which includes and assists in performing the following ML engineering steps:

- **Data pre-processing**: Some of the repetitive and simple data preparatory steps can be easily automated given the schema of the dataset and the specification of target and explanatory variables, like scale normalization, one-hot encoding, factorization of categorical variables, automatic text encoding, feature extraction and selection and dimensionality reduction.

- **Model selection**: These packages usually bundle up a large set of implementation of various supervised ML algorithms, including the ones we have discussed in this chapter. Beside selecting all the models that are valid candidates for the inputed problem, they facilitate an automated approach to hyperparameter tuning through methods like grid search, adaptive searches. Since the number of possible hyperparameters and the scale of their values can vary highly through the different methods a time constraint is introduced to control for the complexity through which each model family is "searched" through for an optimal model. Beside this externally set runtime constraint the available computational resources are also taken into consideration (RAM, number of CPU cores, GPU parallelization possibility) compared to the size and complexity of the dataset and the target.

- **Ensembles**: Since the algorithm already trains a high number models from a variety of families, with low additional computational cost we can also add ensembles to our model list with low additional tuning complexity.

- **Leaderboard**: After training a swath of models we also have to somehow compare their accuracy, for which a plethora of measurements could be calculated, which should be suitable to compare most model families covering balanced and unbalanced categorizations both well. The most regularly used ranking metrics for classification:

  - *AUC* or the Area Under ROC Curve, where ROC stands for Receiver Operating Characteristic is one of the most common metrics for evaluating binary classification models. It measures the area under the curve which is drawn by piloting the True Positive Rate versus the False Positive Rate for different cutting thresholds of the forecasted category probability by the model. The value of AUC is between 0 and 1 (since both axes are also bound between these values), where the value of 0.5 means random assignment and thus values below signal worse than random accuracy. This characteristic is best to compare models over a balanced dataset where false positives and negatives are of similar importance.
  - *AUCPR* is very similar in design to AUC, but instead of the True and False Positive Rates it compares Precision (True Positives / (True Positives + False Positives)) and Recall (True Positives / (True Positives + False Negatives)). This is necessary, because for unbalanced datasets, overfitting models can easily be perceived as good performers. For example on a dataset of 98% negatives and 2% positives a model predicting only negatives will still have an accuracy of 98% and thus produce a high AUC. Since AUCPR focuses on the positive (or rare) class it both takes into account the predicting power to correctly identify positives (Recall) and also not to mislabel them (Precision). This method also gives a better comparability of explanatory performance of the same model over different unbalanced datasets.
  - *Mean per class error* simply calculates the False Negative Rate or ratio of misidentified labels and the number of observations in the given class. This metric lacks substantial information compared to the earlier methods but in case of a high number of classes (where AUC and AUCPR becomes unfeasible to calculate) it can still provide a suitable ranking.
  - *LogLoss* or logarithmic loss measures the amount of divergence of predicted prob-

ability with the actual label:

$$\text{LogLoss} = -N^{-1} \sum_{n=1...N} \sum_{m=1...M} x_{nm} \log p_{nm}$$

, where N and M stands for the number of data points and classes, x_{ij} stating if the nth observation belongs to class m and p_{nm} indicating the probability output of the model for the same observation and class. As an entropy-like measure, the smaller it's value is the better the model, with a value of zero signing a perfect one. Compared to the earlier categorical counts this measure also quantifies the uncertainty of our estimation.

- *RMSE* or root mean square error measures the average magnitude of the difference between predicted and observed probabilities of class categorization (the latter being either 0 or 1 in this case). It can be calculated as

$$\text{RMSE} = \sqrt{N^{-1} \sum_{n=1...N} \sum_{m=1...M} (p_{nm} - x_{nm})^2}$$

where N and M stands for the number of data points and classes, x_{nm} stating if the nth observation belongs to class m and p_{nm} indicating the probability output of the model for the same observation and class. Lower values indicate better model fit, however for categorization tasks LogLoss generally provides a measure that is moving on a more linear scale when comparing multiple models.

- *F1-score* calculates the harmonic mean of precision and recall and thus combining them into a single score that can even be scaled into multi-class problems. Although easier to implement and explain than the AUCPR, it functions best to find and concentrate on the best performing threshold rather than a threshold-agnostic measure, thus AUCPR is generally preferred to compare multiple models and datasets over various balances.

- **Explain**: All the possible algorithms used can have vastly different approaches in explaining the model contribution and interaction of explanatory variables. autoML solutions usually take the burden off the developer's hands by calculating the most popular methods for each calibrated model family. When it comes to ensemble models however this approach becomes much less trivial or even mathematically unfeasible

due to the difference in fundamental design of these measures for the participating models.

- **Productionalize**: H2O autoML also creates model objects for each trained model and ensemble so the created artifact can be integrated into industrial DevOps and MLOps frameworks to ensure reproducibility and transferability of the resulting predictive product. This can become increasingly important in case of sensitive applications, like when a market regulator enforcing a gambling provider to use such methodologies to forecast and action on preventing the emergence of problem gambling among its player base as proposed in Chapter 5.

2.5 Reproducible research

As I have mentioned in Chapter 1.2, the replication crisis of modern science extends well into behavioral studies. Psychology, behavioral economics and addiction science (Pearson et al. (2022)) and more specifically gambling research (LaPlante (2019)) has been no exception from this. As it can be persistently shown throughout the research articles cited and showcased in my thesis, when it comes to empirical research articles having the computer codes that they have used to produce their results shared is rather rare (e.g. Auer and Griffiths (2024)). Moreover, none of these papers have been able or willing to share the underlying data to be openly distributed to other researchers. This pattern, unfortunately, holds even for studies where the data is based on otherwise publicly available data like government tax revenues or openly available trading volume data. The replicability of experimental research papers can be even more complex, but we will not address it here, since it is out of scope for this thesis.

To address this issue researchers can choose multiple ways in sharing their work in a reproducible manner with their audience. For the computer codes a good practice is usually to create a version tracked repository, that usually enables to include instructions for installation and execution of the underlying code base. For some coding languages a good practice can be also to create a library for the project which then can properly set particular dependent packages or tools for the underlying scripts. This repository or package is then can be openly shared through public websites like GitHub or BitBucket.

For sharing the input data of one's research, simply hosting from a university webpage or some commercial data sharing services (e.g. DropBox) is deemed to be bad practice, since these solutions do not provide sufficient reliability and longevity for the access. As best practice many research institutions and collaborations have started data repositories that can reliably store and openly share research data for an extended timeframe like Zenodo by CERN, Mendeley Data by Elsevier or Dryad.

Recently consolidated platforms have also been started to become encouraged in social science like the Open Science Framework and Code Ocean where elements of reproducibility like versioned repositories of code and data, pre-publication of study design, and even containerization of the whole process can be facilitated in a single common platform, ensuring end to end reproducibility of the whole research process.

For my studies presented in this thesis I have chosen to share my codes using documented GitHub repositories for the codes used for data extraction and transformation as well as analysis and the production of all graphs and tables featured in the articles.[4] Beside the scripts extracting the gambling datasets from the BitCoin ledger databases I have also shared the tracking datasets for SatoshiDice and LuckyBit to both save the very costly data transformation steps and to provide these cleaned and lean datasets that in themselves are more suitable for further gambling research.

[4]Codebase for Chapter 4, Chapter 5 and Chapter 6

Chapter 3

Research data

3.1 Datasets in empirical gambling research

Empirical studies on gambling behavior can generally be based on different levels of resolution when it comes to their subjects, the gamblers. A general class of studies rely only on national, regional or corporate economic and/or financial indicators. They typically study aggregate signals of gambling like sales data, government tax revenue in comparison with economic or financial indicators like stock price crash risk (Ji et al. (2021)), GDP in general (Baumöhl and Výrostová (2017)) or other welfare indicators to highlight recessions (Horváth and Paap (2012); Eakins (2016)). The underlying data of these analyses are generally made available through governmental data agencies like the Federal Reserve Economic Database for the United States or Eurostat for the European Union. Some governments also put effort into policy making to incentivize or oblige certain market actors to publish their aggregate sales data regularly, like the Commercial Gaming Database of the American Gaming Association. While in theory the general availability of these datasets makes reproduction of these studies feasible, the production codes and the specific data filtering methods were almost never included into these publications.

Studies that focus on the individual gamblers - from the perspective of data collection - are doing so in three ways. The most classic method of psychology, questionnaires and direct analysis is often used in identification and profiling of the problem gambling condition (e.g. Rockloff and Dyer (2006)). The second method is of experimental economics, where

the actions of individuals are observed and studied in an artificially created and controlled environment. This is often used in combination with the questionnaires (e.g. Lindner et al. (2021)). The third method in studying the behaviors of individual gamblers is the realm of natural experiments, where the empirical data is collected in an uncontrolled setting. This group includes tracking the actions of gamblers related to their gambling activities like entering the gambling site itself, to when and how they place bets and initiate gambles (quantity, risk, timing) to when they leave the casino temporarily or when they even report themselves for self-exclusion and/or label themselves as problem gamblers.

As Chagas and Gomes (2017) highlights, most of the research studies on gambling before 2000 were based on experimental settings or self reports from gamblers. An icebreaker in this field was the study of Croson and Sundali (2005), who took on the enormous task of tracking hours of videotape recordings from casinos to track individual roulette sessions creating the first large empirical dataset for gambling. With the birth of online gambling, real life gambling data have started to amass in the databases of facilitators and eventually market regulators. However, very few of these datasets have been made available to the research community to conduct studies on, like Horváth et al. (2010); LaBrie et al. (2007); Braverman and Shaffer (2012); Xu and Harvey (2014) for bets tracking and Kotter et al. (2018) for self-exclusion. Although empirical analysis of behavioral tracking data contributes a lot to our understanding of risky behavior, most research is done on a few instances of sports and online casino datasets that remain largely proprietary. As discussed in Chapter 2.5, this also affects the value of these research papers from the point of reproducibility.

The advent of cryptocurrencies has opened up a new opportunity for this field as well, since user behavior can be tracked through the publicly available ledger of transaction history. Recent articles have only started to be published on the statistical analysis of gambling and gaming datasets collected from the Ethereum blockchain (Scholten et al., 2020, 2019). Some works also suggest that many people are using the trading space of cryptocurrencies as a form of gambling (Mills and Nower, 2019) and Conlon and McGee (2020) even suggested that gambling may have a strong effect influencing the value of BitCoin in fiat currencies.

For me, the lack of publicly available tracking datasets for gambling research provides a gap in the research space to be filled. In the rest of this chapter I will describe in detail how

I have gathered large datasets on gambling based on cryptocurrency data using strategies similar to Conlon and McGee (2020) and (Scholten et al., 2019).

3.2 Bitcoin and blockchain

Satoshi Nakamoto has published his seminal paper on his webpage (Nakamoto, 2008) about a decentralized, virtual payment system in 2008 October 31st. This idea since then have been implemented and facilitated as an open source library by autonomous internet communities. Over the years Bitcoin has become the most successful cryptocurrency, with a market capitalization touching the trillion dollar levels, that spearheaded a new age of finance.

Bitcoin's currency mechanic is based on the decentralized ledger called the blockchain. People initialize transactions by broadcasting a transaction plan. This is then picked up by so called miners who compare the information inside this transaction (multiple ones at a time) with the public ledger of historical transactions and try to add it as a new block to the existing chain. The fact of who can add this new block and receive the attached reward is decided by solving a mathematical puzzle, the complexity of which is adaptive to the pool size of miners, thus creating a competition and ensuring the safety and uniqueness of the transaction history.

From our point of view the most important aspects of Bitcoin are the following:

- **Transparency**: The entirety of the created transaction history is publicly available.
- **Anonymity**: Anybody can create a new Bitcoin account without any supervision of a central entity, thus by nature the account' owner's identity is not logged in the system.
- **Chronology**: The fact that the history of transactions is unalterable after a block has been mined provides an opportunity for the creation of various provability systems.

The nonexistent cost of opening a new Bitcoin account incentivize the users to heavily distribute and circulate their Bitcoin wealth over many accounts, which can make the identification of an individual's activity difficult. Kondor et al. (2014) showed that using

the transaction's information available in the blockchain and the nature and structure of bitcoin transactions, we can identify the Bitcoin addresses that are accessible (thus most probably controlled) by the same entity. Bitcoin transactions are designed to be many-to-many (see Figure 3.1.), meaning that the source of the transacted value can be from multiple addresses and can direct its output to many targets simultaneously. To start such transaction one needs access to all the underlying input accounts. Using these addresses and transactions we can grow a bipartiate network using the ledger. This will result us with anonymous IDs which we can assign to the transaction history. As detailed later, I will rely heavily on the published identification (contraction) datasets of this article. Another possibility this opens up for us is to determine the wealth of any contracted agent in the network at any point in their transaction history.

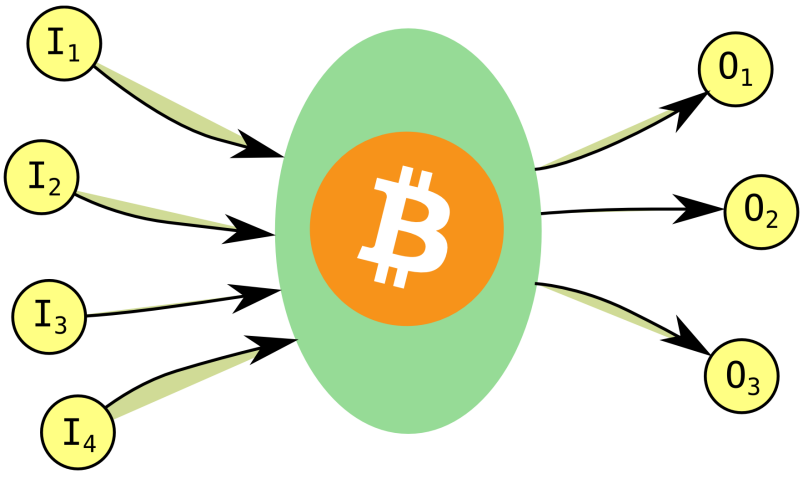

Figure 3.1: Schematic view of a Bitcoin transaction. Here we have four input (I_1–I_4) and three output (O_1–O_3) addresses. In this example we can be sure that all the input addresses are controlled by the same entity. Source: Kondor et al. (2014)

3.3 Bitcoin gambling and data

The low transactional costs, high anonymity and undecided legal stance of Bitcoin in it's early years of existence gave rise to multiple gambling schemes, that started to become popular around 2012. One of these, namely Satoshidice became so popular, that for the

majority of its life, it accounted for more than 40 percent of the overall bitcoin transaction volume (Badev and Chen, 2014). Because of legal problems and increasing market competition, the site's popularity crashed in early 2014 and has been sold and modified many times since. It is still in operation to date (march of 2025), offering a game that is very similar to the original. Based on the success of Satoshidice many gambling websites have started, implementing various forms of online gambling:

- **Dice/slot/roulette**: The game is organized around a single act of risk choice from the user side and then the decision of the payout based on a random outcome, mostly generated using the semi-random numbers generated during the transaction booking process. These type of games exclude human skill since they are purely based on the underlying random generator. They are also can be played in high frequency since the evaluation of the bets scale from nearly instantaneous to the time the bet transaction gets booked which is in the range of 5 to 20 minutes.

- **Lotto**: These games are typically based on a wallet or a chain of wallets, where the players participate in the lotto by placing bets into the pool by transactions. At given times the winner or winners are drawn (using a random number generated from preset secrets and some transaction details) and the pool gets distributed. These games are low intensity ones that also does not include any skill on the player's end.

- **Casino**:These games implement classical forms of casino gambling which can include some from the first category but mostly include elements of user skill as well, like card games. They are usually also include some kind of online visual representation of the real world games and include opening a balance at the site facilitator instead of initializing each bets by different transactions. These games are also high frequency ones.

From the types of games presented above the first category provides us with the most opportunities for the study of gambling behavior. It should be noted that the method does introduce selection bias that is not present in other online gambling firms, namely the fact that to play the game one has to be legally and technically able to buy cryptocurrencies.

For the sake of simplicity I would now introduce a few key terms that I will use going forward. Every kind of betting activity can be described using a few key variables and

important features. The risk profile of the game is described by three important parts. First of all, the determinant of the outcome, which sometimes is based on real life outcomes, is always going to be some kind of trustworthy random generator, so it can be taken as an independent and identically distributed random variable. Second is the bet amount, which is the maximum limit of our losses in a given game session, but also serves as a basis of our payout in the event of winning. Third and most important is the odds of a given game, which is also referred to as a multiplier, since it implies the multiplicative of the bet amount to be payed out to the bettor in the event of a winning bet.

The odds and the probability of a given winning outcome (P_{win}) are in an inversely proportional relationship to each other for "fair" games.[1] If this relationship is direct, it means that the expected value of the payouts is equal to 1 if the game is repeated an infinite number of times with the unit bet amount. For most casino games, however, this expected value is below one, and the difference is the so-called house cut. This is usually determined in a percentage fashion, which tells us what proportion of the unit bet is kept by the facilitator of the gamble or the "house". For "fair" gambles, it is also expected for the house cut to be uniform over all the risk options offered. If we sum up the bets as monetary outflows and the expected payouts as inflows, if the house cut is larger than zero, we will end up with a negative expected total sum. This pushes players, as economic agents, to the choice dilemma outlined in Section 2.1.

3.3.1 SatoshiDice

SatoshiDice is an online gambling site, launched in 2012, that was the first gambling site to center its business around Bitcoin. In its early days the platform attracted so many players that it generated the majority of the Bitcoin transactions (Badev and Chen, 2014). The game that can be played on the site is rather simple: the webpage lists a set of fair games with a house cut of 1.9%, setting a range of bets between a minimum and a maximum value and providing for every game a Bitcoin address to send the bet to. The bets are evaluated using a random number that is generated from the combination of a secret hash provided by the website and the transaction details. The time between placing a bet and receiving

[1] Most implementation of casino games and all the games discussed here considered as "fair" games.

an answering transaction scaled from seconds to even a block creation long time (6 – 15 minutes) and with a tiny transaction fee attached the process could be fastened.

Figure 3.2: Original layout of SatoshiDice (2012). We can see the list of games and addresses on the bottom.

This game's popularity was wast, but fading. Due to it's high popularity the owner of the webpage had to sell it in early 2014 to another party due to taxation reasons. This caused the game to crash in popularity and the player base moved over to the sprawling number of alternative gambling websites that have been started during the height of SatoshiDice's popularity. The gambling site lives on to this day, but with slight but continuous changes to its game design and fading popularity. Because of this reasons our data analysis will concentrate on the period from the first transaction in the website (2012 February) to the end of 2013.

Since all the wagers and their resulting transactions are kept public in the Bitcoin ledger, a complete history of all the bets placed on the website can be retracted. The detailed code of recreating the dataset is made available in R in the public code repository

github.com/sampaat/hot_hand_cold_head in the file `data_preparation_bitcoin.R`. The process starts with extracting the bet addresses, odds or winning probability and bet range (minimum and maximum accepted) information from the archived version of the website.[2] Then I downloaded the transcribed blockchain datasets of Kondor et al. (2014) and using the SatoshiDice addresses, gathered all transaction ID's that transferred money to any of these addresses (set of bets) and the ones receiving transactions from those addresses (set of answers). [3]

Then I added the information about the other ends of the transactions. This was a nontrivial exercise, given that usual Bitcoin transactions are many to many to keep the participants in relative anonymity. To get around this I used the contraction dataset of Kondor et al. (2014) that used a simple heuristic to identify addresses controlled by the same identity, which in fact turned my transaction list into a one-to-one transactional set.

One problem the Bitcoin protocol creates for us is that transactions do not log time per say. Agents connected to the Bitcoin network log the time when they first receive information about a given transaction and a definite timestamp is added to each transaction that is packed in a block when it is verified. However the true initiation time of the transactions are best modeled by receiving timestamps of agents located centrally in the Bitcoin network. In this case we were lucky enough to be provided with the transaction timestamps of `blockchain.info` thus having a very good estimation available.

After having time-stamped transaction lists for both the bets and the answers of our target period (2012.04.12 to 2013.12.28) the transactions had to be paired up. This I have started with performing a full joint on transactions exchange between the same SatoshiDice and player addresses, conditioning on banning retarded answer transactions (bets cannot be answered before placed, but because our timestamps originate from a third-party, non-central observer, this occasion is theoretically possible). This step creates a many-to-many transaction pair list, which is still non trivial to be made unique, since the player could start a new transaction without receiving an an answer to its earlier ones. For this I have created a subroutine which iterated over the placed bets in time order, selecting it's earliest possible answer. These answers then have been added to a 'used' set not to be used to

[2] see https://web.archive.org/web/20121103121459/http://www.satoshidice.com/

[3] the datasets are continously upgraded and made available at https://datadryad.org/stash/dataset/doi:10.5061/dryad.qz612jmcf

match with future transactions. Using this algorithm I have been able to find a unique pair for above 98% of all transactions.

I have also implemented a sanity check for the algorithm, by checking if the ratio of winning versus losing answers for bets of a certain risk level have matched up with the winning probabilities implied by the odds of them. This test provided reassuring results with very small relative differences between the implied and observed winning probabilities in the sample. This test also shows that the implementation of the game was in fact "fair".

By the end of the exercise I have ended up with 6,145,532 individual bets with answers. During this time period the exchange rate of Bitcoin experienced extreme volatility from exchange rates of $5 to over $1000. For the purposes of this analysis we have chosen five windows with lengths of 3 weeks that are characterized with relative low price volatility to control for the effects of the price surges. Summary statistics of these subsets of the dataset are presented in Table 3.1.

	Start date	Number of bets	Number of users	Total bets placed (BTC)	Median bet size (BTC)	Mean daily price (USD/BTC)
A	2012-05-02	119,399	1002	46,649	0.04	5
B	2012-09-17	129,265	2114	85,280	0.06	12
C	2012-12-17	252,301	3405	407,140	0.04	13
D	2013-05-04	329,155	3432	100,430	0.02	111
E	2013-09-11	86,520	1400	62,920	0.03	123

Table 3.1: Summary statistics of the subsets of the observed gambling history used. The exchange rates have been sourced from the public historical data published by the online cryptocurrency exchange aggregator BitCoinCharts (see http://www.bitcoincharts.com)

From this basic observational set of player ID, risk, bet size, bet time and outcome, we can derive all the variables needed for our tests performed in our papers *Approaching the Hot Hand with a Cool Head* and *Unmasking Risky Habits: Identifying and Predicting Problem Gamblers Through Machine Learning Techniques.* As i have described in detail in 2.5 one of my research goals was to create a public dataset and since all underlying datasets of this product was already collected under a public protocol with users aggreeing into their

data being collected by using it, the resulting dataset is also derived to be public. Thus I have published these prepared datasets in a public data library facilitated by CERN as Sándor (2021).

3.3.2 Luckybit

Luckybit was also a similarly structured, random number based online Bitcoin game, which had a bit more complex payout structure compared to the previous two. The betting system was similar to SatoshiDice, but the games' payout was defined as a set instead of a single value. As described in Conlon and McGee (2020) the game was replicating a Galton box, with 16 outcomes having an occurrence probability defined by the binomial distributions. The payouts added to these outcomes were set in a way to provide four games ranging form safer to riskier providing different odds structure with overall the same house cut percentages.

Bin	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
P_{win}	1.5e-05	2.4e-04	0.0018	0.0085	0.027	0.066	0.12	0.17	0.19	0.17	0.12	0.066	0.027	0.0085	0.0018	2.4e-04	1.5e-05
M_{blue}	3	1.4	1.3	1.2	1.1	0.2	1.1	1.1	1.1	1.1	1.1	0.2	1.1	1.2	1.3	1.4	3
M_{green}	2.2	5	3	2	1.4	1.2	1.1	1	0.4	1	1.1	1.2	1.4	2	3	5	2.2
M_{yellow}	111	38	12	5	3	1.4	1	0.5	0.3	0.5	1	1.4	3	5	12	38	111
M_{red}	999	130	24	9	4	2	0.2	0.2	0.2	0.2	0.2	2	4	9	24	130	999

Table 3.2: Outcome probabilites (P_{bin}) and multipliers (M_{game}) of possible games in LuckyBit. Losing outcomes ($M_{game} < 1$) are higlighted with a slight tint compared to winning ones ($M_{game} > 1$). Note that in this game all outcomes have nonzero payouts. Source: Conlon and McGee (2020)

The advantage of this game compared to SatoshiDice is that while the transactions are just as easily recoverable from the blockchain and the site have been in business from early 2013 to late 2019. The drawback is the more complex risk structure and the much smaller range of available games.

Since the bets are placed through Bitcoin transactions, they are obtainable from the blockchain using the method described in Section 3.3.1. I have performed the extraction using the public dataset of Kondor et al. (2014) namely their 2018 upload. At the time of my data preparation the data extraction method of Conlon and McGee (2020) deemed

Figure 3.3: Interface layout of Luckybit (2015). The colored lines represent the available gambles and the numbers in the bars are the available multipliers. Notice that some are smaller than one - but not zero - thus implementing a partial loss of the bet amount.

to be unavailable due to the underlying public data provider adding a paywall to mass data queries. Nevertheless our original method deemed us a database spanning 1486 days between 2013 October 20th and 2018 February 7th, with a total of 2.060.601 bets placed by 18.220 distinctly identified player of the site.

The timing of the transactions is somewhat limited by blockchain technology, since by the time we performed this extraction, the transaction timestamps of `blockchain.info` were hid behind a paywall. The only universally acceptable time of the transaction thus were tied to the creation of its block, which left us with an accuracy of 1-16 minutes for timing them. This unfortunately also makes the creation of a matched dataset with returning transaction much more error prone: bets placed by the same entity with little time difference may get booked into the same block with both of their answering transactions, making it impossible to resolve their coincidence. Thus for this dataset we have not created this matching.

I have used the same public data library to share this dataset as well: Sándor (2025).

3.3.3 Primedice

Primedice is in principle a very similar game to SatoshiDice, but there is a key difference in the financial framing. While in the predecessor each bet was placed using a Bitcoin transaction, here the users transferred a predetermined amount to an omnibus account of the site and receiving equivalent credit after registration. Then the user could place bets in an almost exactly similar, random number based dice game. Gains could be then requested to be payed out to a custom address provided by the users. The game offered a continuous scale to choose the odds, and an additional checkbox to choose between rolling "under" or "over" the number (between 1 and 99) that was attached to the odds set. Because the gambles' evaluation was independent of the blockchain, it was instantaneous, thus enabling a much faster pace of gaming. Another specialty was that because the user did not need to provide a transaction fee, the size of bets placed in this platform was much lower. It was also possible to roll without any bets placed.

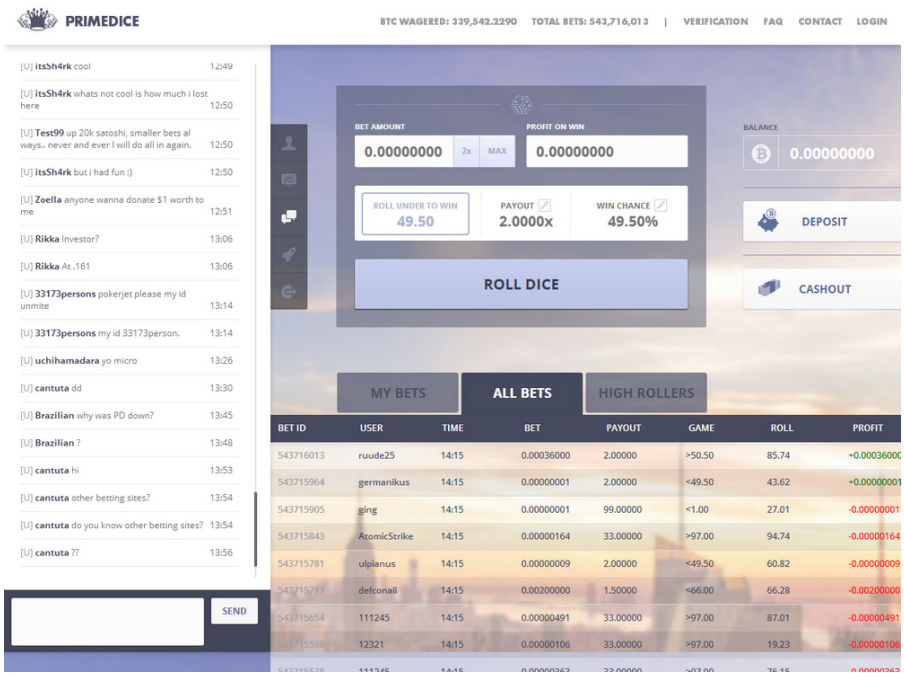

Figure 3.4: Original interface layout of Primedice (2013). We can see the interface to set the risk and bet details.

As I have said before this site implements a game very similar to SatoshiDice in principle,

but instead of operation through blockchain logged transactions for each bet, the site works on a prepay account basis. This denies us the form of data collection we have been working for SatoshiDice. But the interface of the webpage did contain a design weakness I was able to exploit: looking at the bottom of Figure 3.4. we can see a list of recent bets from all users on the site. This list was automatically refreshed using a PHP interface of the server automatically requested every 5 seconds from the client side. I have found that this PHP request can be issued without any authentication or captcha filtering. Thus I have set up a simple bash script that was accessing this interface repeatedly and logged the returned list of the last 5 player bets containing all necessary information about the given bet: user ID, bet ID (sequential), all risk interface settings, time, bet size, roll and payout.

Since the size of the server answer was a fix length of 5, if more than 5 new bets have been placed on the platform between the two requests, we have some lost information in that period. We could get around this problem two ways: the first is to decrease the time between subsequent requests, but by experience we have seen that sending requests more frequent than 1 seconds apart resulted in an IP ban by the firewall of the server. An alternative approach was to run our logger script from multiple machines asynchronously thus getting a more frequent read practically without getting caught by the server firewall. I have managed to run this exploit between 2014.02.04. and 2014.05.13. After consolidating the gathered logs we ended up with 112,251,530 unique bets placed by 53,694 users. Based on the bet IDs we were able to determine how many bets we have missed in a given period. This ended up to be 70% over the whole period, but managed to achieve above 96% success rate for some days.

However giving a due consideration to the ethical and legal aspect of this dataset we have decided not to publicly distribute or base publications on this dataset. One aspect being that the owner of the website explicitly restructured their API so that to avoid scraping and have not made their data publicly accessible ever since. The other is that compared to the explicit publicity consent built into the terms of service of both SatoshiDice and LuckyBit, this website extended their TOS right after our data mining to claim that it collects user data for their services only.

3.4 Comparison with databases from prominent literature

As it was discussed in the beginning of this section, the domain of gambling analysis we have found a lack of availability in public datasets was behavioral tracking. This micro-level observational approach had humble beginnings with classical papers like Croson and Sundali (2005) manually analyzing hours of video footage, but have seen huge advances with the advent of online gambling. Commercial providers have started providing researchers with anonym datasets spanning sometimes years of observations of their player base (e.g. LaBrie et al. (2007)) but the largest databases are mostly provided through market regulators (e.g. Finkenwirth et al. (2021)). If we take a look at Table 3.3 and compare the size of these databases in terms of observed individual gamblers and number of bets between the relevant papers and the datasets I have gathered, we can acknowledge that a comparable size have been assessed. And as it has been highlighted before, in opposition to these commercial datasets, the SatoshiDice and LuckyBit datasets are publicly available and reusable for any researchers.

Dataset	Gamble	Data colleciton	# Players	# Bets	Time span	Year of obs.
Croson and Sundali (2005)	Roulette table	Videotape	139	24131	18 hours	1998
LaBrie et al. (2007)	Mixed online	Online tracking	42647	?	2 years	2005
Xu and Harvey (2014)	Sports bets	Online tracking	776	565915	1 year	2010
Salaghe et al. (2020)	Slot machine	Transaction records	46502	24182076	108 days	2015
Finkenwirth et al. (2021)	Mixed online	Online tracking	30902	575470087	1 year	2014
Auer and Griffiths (2024)	Mixed online	Online tracking	150895	?	3 months	2023
SatoshiDice (total)	Crypto slot	Online tracking	35490	6421645	5.8 years	2012
SatoshiDice (matched)	Crypto slot	Online tracking	11077	803857	105 days	2012
LuckyBit	Crypto slot	Online tracking	18220	2060601	4.3 years	2013
PrimeDice	Crypto slot	Online tracking	53694	12251530	86 days	2014

Table 3.3: Comparison of tracking datasets from prominent articles (top 5) versus my data collection (bottom 4). Year of observation refers to the start of the data gathering period and time span refers to the length of the sample. The *total* version of the SatoshiDice dataset refers to all the bets placed on the gambles while *matched* refers to the subset used in later articles. Where number of bets signed with a question marks were not disclosed in their respective papers.

To be completely objective however we also have to note where our datasets come short

compared to the commercial ones. Although most online tracking data covers a wide range of live betting and various casino games, our collection solely focuses on very simple game designs that shows similarities only with simulated slots, roulette or dice games. Almost all of the datasets in the listed papers have a wide arrange of demographical descriptiors and thus statistical controls can be applied on geographical, age, gender, etc. characteristics, where our datasets lack these descriptors completely. It is important to note that while the outcomes in our analysis are intentionally designed to be independent and identically distributed, this may not hold true for other types of betting such as horse racing, football, or cricket. Moreover, one may argue that subject in our dataset were early adopters of a risky technology who were willing to take significant financial risk or hobbyists who mined the Bitcoin may have viewed it as 'house money'. Moreover, many may have seen the platform as a test of the Bitcoin network's resilience rather than traditional gambling, making their intent challenging to interpret. However taking these shortcomings into considerations, the sheer size and availability of our datasets provide a worthy open access tool to the scientific discourse around behavioral research on gambling.

Chapter 4

Approaching the Hot Hand with a Cool Head

Abstract

In their influential paper Xu and Harvey (2014) claim that gamblers are more likely to win in subsequent games if they are in a winning streak and more likely to lose if they experience a losing streak. They suggest that gamblers create their own hot hand by falling for the gambler's fallacy. In this article we present both theoretical and empirical results that challenge these findings and provide evidence that the presented results can occur without the existence of the hot hand, by replicating their analysis on a large online dataset consisting of 916,640 observations for 10,963 gamblers.

4.1 Introduction

People often see patterns in randomness, which makes them to believe in illusory fallacies. One notable such fallacy is the hot hand, a commonly used explanation for 'being on fire', especially among sport fans and commentators. Many basketball supporters, for example, believe that the player who scored several times in a row (referred to as 'having a hot hand') will score again with a higher probability than the player usually does. In general, the hot hand fallacy refers to the mistaken belief of perceiving streakiness in random processes.

The notion has been used in many areas outside of sport, including gambling and investing. For example, the slot machine player who just had two winning combination of cherries in succession hardly if ever stops playing during the 'lucky streak'. Or, the rookie investor, who examines the stock that had huge returns in the last 3 months, decides to rebuild his portfolio heavily on it. In reality, however, the performance of basketball players does not seem to change substantially during the game (see Gilovich et al. (1985)), the subsequent rolls of a standard slot machine are independent of each other and the momentum factor is not at all a good single predictor of a stock's performance (see Bacidore et al. (1997)). The existence and validity of the hot hand phenomenon has been the focus of many research articles. The majority of the literature provides little if any support for the existence of the phenomenon with the conclusion that the hot hand is more of a cognitive illusion and its existence not something that is supported by the data (see for example Gilovich et al. (1985); Johnson et al. (2005)). However, recent studies cast some shadow on the seminal papers in the topic and argue that the methodology used in those articles might have been misleading. For example, Miller and Sanjurjo (2018) show that a substantial bias is present in the standard measure used to analyze sequential decision making. The authors convincingly argue that the influential paper by Gilovich et al. (1985) and its replications are vulnerable to the so called streak selection bias and if one corrects the bias the results and the conclusions may be reversed. It seems that if there is some sign of the hot hand's existence in sport it may be due to some considerable skill that is not driven by randomness, but has to do with the confidence of the players that unlocks their better performances.[1]

However, there is hardly any study that would support the existence of the hot hand outside of the field of sports. One notable exception is the paper by Xu and Harvey (2014), in which the authors claim that hot hand is indeed a prevalent phenomenon in gambling. By analyzing a large-scale online gambling dataset, the authors suggest that gamblers are more likely to win in subsequent games if they are in a winning streak and more likely to lose their subsequent bets if they are in a losing streak. The claimed mechanism behind these surprising observations is that gamblers with a winning streak gradually choose safer bets,

[1] Another paper that finds statistical traces of the phenomenon in sport is by Yaari and Eisenmann (2011). However, the authors emphasize that it is not clear if the observed patterns are non-random patterns or simply just better and worse periods of the players.

and those with a losing streak opt for more riskier bets in subsequent games, suggesting that gamblers expect their good fortune or bad luck to change if they continue to gamble. To simply put, gamblers create their own hot hand by falling into the trap of the gambler's fallacy. Since the dataset the authors used contains hundreds of thousands of regular sports betting results, one might believe, that there is nothing that suggests that the results are not robust and would not be relevant in a general sense.[2]

These findings, though, were challenged by Demaree et al. (2015), who implied that the results are the consequence of a selection bias present in the statistical analysis used by the authors. In a follow-up paper, however, Xu and Harvey (2015) argue that the results are not attributed to selection effects. They present additional findings that, in their view, further support their initial conclusions.

In this paper, we further challenge the findings of Xu and Harvey (2014) and argue that their conclusion on the existence of hot hand is premature. We present both empirical and theoretical evidence which raises concerns regarding the conclusive nature of the methodology employed in determining the existence of the hot hand phenomenon. By replicating the analyses of Xu and Harvey (2014) on a comparably large dataset we demonstrate that similar results can be obtained even when the phenomenon of hot hand is evidently non-existent. Moreover, we show that their results hold even when gamblers behave exactly in an opposite way to the mechanism suggested by Xu and Harvey (2014).[3] Furthermore, in the appendix we demonstrate theoretically that players with persistent behavior, who do not change their gambling behavior in subsequent gambles, would generate similar outcomes to the ones presented in Xu and Harvey (2014), thus the results may be simply the consequence of the statistical method applied.

[2]We must emphasize that the question analyzed by Xu and Harvey (2014) is only loosely related to the traditional concept of the hot hand, as understood in the existing literature. We find it unfortunate that they labeled their observed phenomenon as hot hand, given that their article does not explore the conventional idea of success leading to future success. Instead, they investigate situations where bettors may opt for lower-risk gambles, leading to an increase in the probability of a payoff. Comparing this to the hot hand in basketball, it would be akin to saying that basketball players have the hot hand simply because they move closer to the rim after a streak of success. Nevertheless, we acknowledge the importance of analyzing how gamblers alter their behavior during winning or losing streaks of varying lengths. This question warrants thorough investigation and could provide valuable insights into the dynamics of gambling behavior.

[3]The scripts used for this study are publicly available at github.com/sampaat/hot_hand_cold_head.

4.2 Streak dependent behaviour on population level

In this section we follow the same methodology applied by Xu and Harvey (2014) and by replicating their analysis on our online gambling dataset from SatoshiDice, consisting of 916,640 observations for 10,963 gamblers, we show that similar observations can be made as in their analysis, however, without the existence of the hot hand phenomenon. [4]

SatoshiDice is an online gambling site, launched in 2012, that was the leading gambling site at the time of interest to center its business around the cryptocurrency Bitcoin. In its early days the platform attracted so many players that it generated the majority of the Bitcoin transactions (Kondor et al., 2014). The game that can be played on the site is rather simple: the webpage lists a set of fair games with a house cut of 1.9%, setting a range of wager size between a minimum and a maximum value and providing for every game a Bitcoin address to send the bet to. The bets are evaluated using a random number that is generated from the combination of a secret hash provided by the website and the transaction details.

Since all the wagers and their resulting transactions are kept public in the Bitcoin ledger, a complete history of all the bets placed on the website can be extracted. Following the method presented in Kondor et al. (2014) we acquired the outcomes of each game played on the platform by matching the return transactions sent from the game addresses to the users. The dataset we use in this analysis spans from 2012.04.12 to 2013.12.28, which covers the most popular period of SatoshiDice.[5] During this period the exchange rate of Bitcoin to US dollar experienced extreme volatility from exchange rates of $5 to over $1000. Therefore, to control for the impact of price surges, we have selected five windows, each spanning three weeks, that exhibit relatively low price volatility. Summary statistics of these subsets of the dataset are presented in Table 4.1.

On SatoshiDice 27 games were available during the observed period, 3 of which were added after the first period. In general, these games encompass a range of winning probabilities spanning from 0% to 100%, with a greater concentration observed in the higher risk range of 0-1% (see Table 4.2). Each game has a minimum and a maximum accepted

[4] see https://web.archive.org/web/20121103121459/http://www.satoshidice.com/

[5] The dataset used for the analysis can be accessed at DOI: 10.5281/zenodo.5600259.

	Start date	Number of bets	Number of users	Total bets placed (BTC)	Median bet size (BTC)	Mean daily price (USD/BTC)	Relative daily price deviation (%)
A	2012-05-02	119,399	1002	46,649	0.04	5	1.16
B	2012-09-17	129,265	2114	85,280	0.06	12	2.29
C	2012-12-17	252,301	3405	407,140	0.04	13	0.6
D	2013-05-04	329,155	3432	100,430	0.02	111	4.07
E	2013-09-11	86,520	1400	62,920	0.03	123	1.33

Table 4.1: Summary statistics of the subsets of the observed gambling history used. The exchange rates have been sourced from the public historical data published by the online cryptocurrency exchange aggregator BitCoinCharts (see http://www.bitcoincharts.com).The average relative price deviation over the 21 day periods between 2012-04-21 and 2014-04-21 were 9.9%, which is much higher than what can be observed in the highlighted periods.

bet amount, which have changed once in the analyzed periods. The observed dispersion in risk preferences among gamblers over time is evident from both Table 4.1 and Table 4.2. Moreover, a significant surge is evident in the exchange rate of the underlying cryptocurrency when compared to fiat currencies. We explicitly highlight these variations across the analyzed periods in our sample. The consistency of our findings across all observed periods provides additional support and strengthens the robustness of our conclusions.[6]

As one can see from Table 4.1 and Table 4.2 there is a clear dispersion in risk preference among gamblers over time. Showing that our findings hold for all of the observed periods gives further reinforcement and robustness to our conclusions.

Moreover, a substantial surge can be observed in the exchange rate of the underlying cryptocurrency versus fiat currencies. We do emphasize these differences between the analyzed periods in our sample. Showing that our findings hold for all of the observed periods gives further reinforcement and robustness to our conclusions.

Following Xu and Harvey (2014), first we have sorted out bets with a winning streak of length $n = 1, \ldots, 6$ and calculated their population-wide observed winning probability.

[6]It is worth noting that while a later version of SatoshiDice offered automated betting strategies on its site, the period we examined lacked this option. Therefore, it is reasonable to assume that the behavior revealed by analyzing the dataset is about human behavior and not an artifact of bot behavior.

Next, we have taken all the bets without a winning streak of length n and determined their population-wide observed winning probability as well, exactly in the same way as in Xu and Harvey (2014).[7] As we can see in Figure 4.1.A the winning probability observed at a population-wide level generally increases with the length of the winning streaks. In contrast, for the 'off streak' population the probability of winning is close to the streak-independent winning probability observed at a population-wide level. The same analysis was carried out for the losing streaks (see Figure 4.1.B). These results are similar to Xu and Harvey (2014): the subsets of the population with different losing streaks show a decreasing winning probability observed at a population-wide level as the length of losing streaks increases, while in the respective disjunct population the probability of winning observed at a population-wide level again converges to streak-independent observed winning probability. These results are robust over the different analyzed periods in our sample.

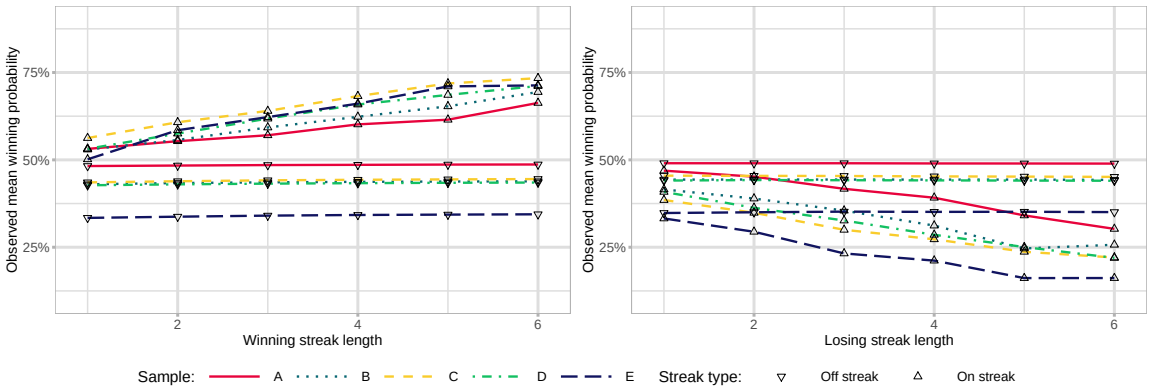

Figure 4.1: A: Population-wide observed winning probabilities for different lengths of winning streaks and after not obtaining winning streaks of those lengths. B: Population-wide observed winning probabilities for different lengths of losing streaks. The different colors represent the observation periods described in Table 4.1.

To examine the choice of risk for players experiencing a winning or a losing streak with a specific length we can take a look at the odds, or more specifically the implied winning probability.[8] Calculating the mean implied probability for the games gamblers

[7] We are using the term 'population-wide observed winning probability' to clearly differentiate it from the ρ values of individual games, i.e. the observed winning probability of the individual players.

[8] Instead of odds, we use the implied winning probability since it has natural bounds (0%-100%) and provides a better comparison for low and high risk games.

were choosing after experiencing a winning or a losing streak of length n, we find that the mean implied winning probability monotonically increases for winners and decreases for losers as the length of the streaks increases. In other words, as the length of the winning streak increases, the risk associated with the game played by the average player decreases, while the opposite holds true for losing streaks (see Figure 4.2).

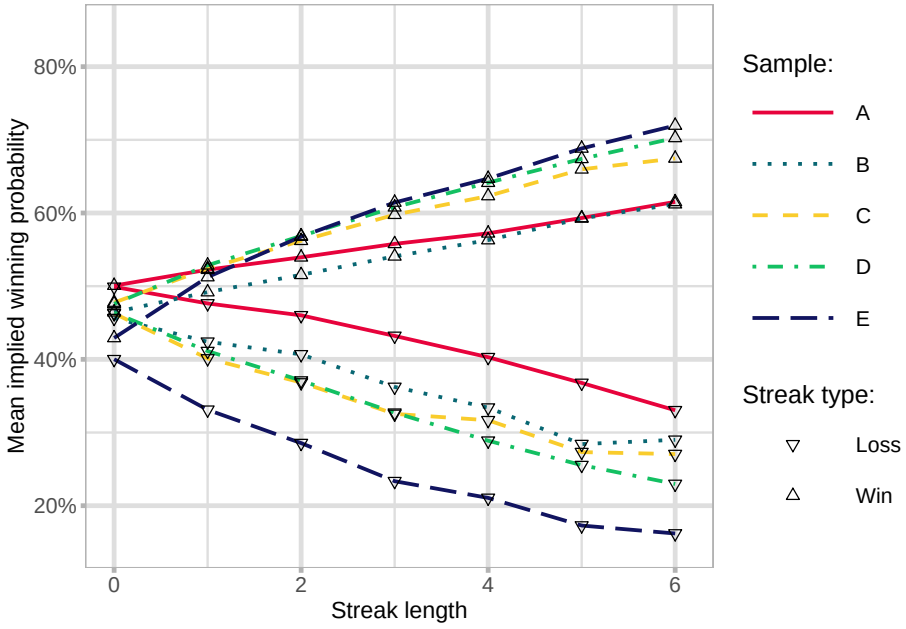

Figure 4.2: Mean winning probability for players after experiencing a winning or losing streak of length n. The different colors represent the results for the different observed periods.

4.3 Streak-dependent behaviour of the individual gamblers

The findings presented in the previous section align with those of Xu and Harvey (2014). Nevertheless, in this section, we aim to provide an analysis centered on individual-level data to demonstrate that Xu and Harvey (2014)'s results could be misleading and their conclusions potentially erroneous. Xu and Harvey (2014) in explaining their results, reason that gamblers select safer odds after winning a bet and they choose riskier bets after losing one, and by doing so, they create circumstances in which they are more likely to win or lose in the subsequent game, i.e. to have a hot hand.

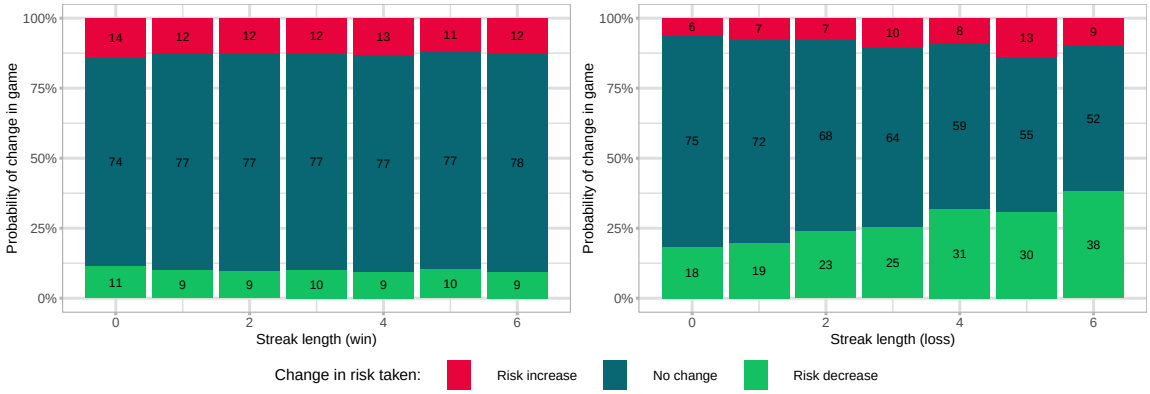

Figure 4.3: Population-wide probability of increasing/decreasing or holding odds after a bet with a winning or losing streak. The numbers inside the bars represent the (rounded) proportions of observations for each individual change in risk, following a certain streak length. The behavioral pattern for those in winning streaks is seemingly independent of the streak's length, while for those with losing streaks increasingly tend to change their odds as the streaks get longer.

However, as Figure 4.3 clearly shows gamblers do not change their odds as often as Xu and Harvey (2014) suggests. As this figure illustrates, the likelihood of changing the odds in consecutive games decreases marginally for gamblers who encounter a winning streak. Subjects tend to maintain their previous choice at a rate of at least 75% of the time. In contrast, gamblers experiencing a losing streak exhibit a higher tendency to alter their odds. However, this probability remains below 50%, indicating that the majority of players do not change their strategies for most games. It is worth mentioning that as the losing streaks prolong, there is an increasing proportion of players who opt for safer odds by reducing the risk they take. This results directly contradict the underlying mechanism proposed in Xu and Harvey (2014).

Furthermore, when examining cases where players do modify their choice of odds and take into account the magnitude of these changes, we once again observe results that contradict the findings of Xu and Harvey (2014). Figure 4.4 depicts the relative changes in winning probabilities when players do make a change in their choice after a winning or a losing streak of length n for each period analyzed. The observed pattern reveals that for winning streaks, changes in odds are clustered on the negative side. This suggests

that as the length of the winning streak increases, gamblers exhibit a consistent rise in their risk appetite. Importantly, the effect size of these changes appears to diminish as the streak lengthens. Conversely, for subjects experiencing losing streaks, there is a tendency to increase their winning probabilities or select less risky games. This effect becomes more pronounced with longer streak lengths, albeit inconsistently. It is worth noting that the mean relative change in winning probabilities for games without consecutive identical outcomes (i.e. streak length of zero) tends to cluster around zero.

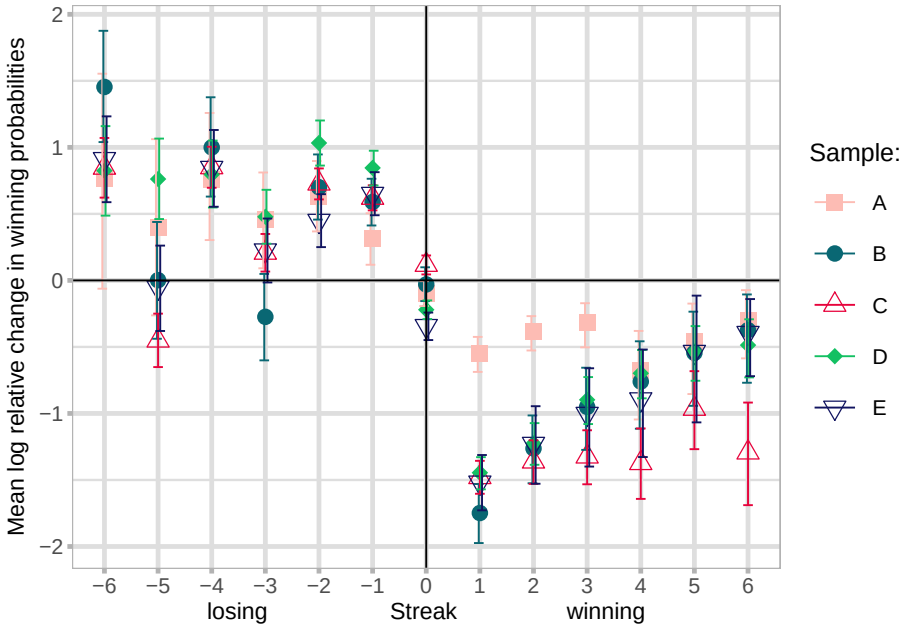

Figure 4.4: Means of the relative logarithmic changes in chosen winning probabilities ($E\left[\log_2 \frac{\rho_{n+1}}{\rho_n}\right]$) for different streak lengths over the observed periods. Error bars indicate 99% bootstrap confidence intervals (CIs) of the means. Positive changes indicate an increase, negative ones a decrease in the winning probability.

To summarize, the majority of players tend to maintain their behavior after experiencing a winning streak. However, a small group of players do adjust their risk levels in subsequent games, as shown in Figure 4.3. Interestingly, the number of players who choose to increase their risk outweighs those who decrease their risk for each length of a winning streak. This suggests that, overall, we should expect to see an increase in risk and as a consequence a decrease in the probability of winning as the winning streak becomes longer, which would be contradictory to what is observed in Figure 4.2. However, note that the group of players

who increase their risk levels is small for each streak length. Additionally, by choosing a higher risk, their probability of continuing a winning streak decreases. Consequently, they become increasingly underrepresented in the subset of players with longer winning streaks.

In simpler terms, players with a lower risk appetite have a higher chance of achieving longer winning streaks. Meanwhile, players with higher initial risk or those who choose to increase their risk level experience a decrease in their chances of continuing to win. As the streaks become longer, the sample becomes increasingly dominated by players with a low risk appetite, leading to higher and higher winning probabilities, as illustrated in Figure 4.1 and 4.2. The inverse pattern holds for players experiencing losing streaks.

To address the concern raised by Demaree et al. (2015) regarding potential selection bias influencing the results, Xu and Harvey (2015) investigated whether gamblers with longer winning streaks were generally more cautious players, using personal-level data. They assessed the risk propensity of each player who achieved a certain length of winning streak by calculating the average probability of winning across all their bets. Similarly, they performed the same analysis for gamblers with varying lengths of losing streaks. The authors ultimately concluded that the length of the streak do not have a significant impact on the betting choices made by gamblers, suggesting the absence of selection bias that could have affect the outcomes.

However, the mean winning probability of players used by Xu and Harvey (2015) is not a suitable measure for assessing the risk propensity when gamblers have different frequencies of play. One risky player might gamble significantly more times than another safer player, but the measure used by Xu and Harvey (2015) fails to account for the effect of varying gambling frequencies, leading to an over-representation of certain players in different streak brackets. By employing a selection criterion that includes gamblers who have achieved a consecutive win streak of at least n times, Xu and Harvey (2015) creates a highly restrictive measure that primarily focuses on the maximum streak length attained. As a result, when considering the case of $n = 0$, the selected group of players consists of individuals who have never won a single bet, including players who quit after a single loss and those who placed numerous highly risky bets without ever achieving a win.

To overcome this limitation and balance the effects of different gambling behaviors, an alternative approach can be implemented. Instead of measuring the risk propensity of

players, we can calculate the mean odds of each individual bet after categorizing gamblers based on their maximum winning streak. By separating gamblers according to their streak length, we can mitigate the bias introduced by the restrictive measure used by Xu and Harvey (2015). This approach provides a clearer distinction between players based on their actual betting choices. [9]

To compare the two approaches, we calculated the risk propensity of players based on both measures.

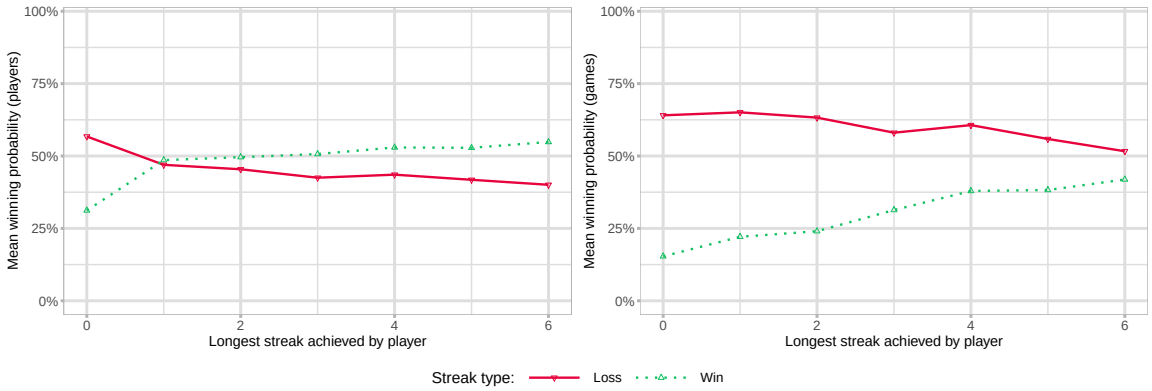

Figure 4.5: A (left): Mean winning probability for the longest winning/losing streaks experienced as in Xu and Harvey (2015). B (right): Mean winning probability for the longest winning/losing streaks experienced based on the mean winning probability of the games selected by players with a specific maximum length of winning/losing streak.

As one can see on Figure 4.5.A a similar result can be obtained with our dataset as in Xu and Harvey (2015) by calculating the mean risk for each player first. However, if we take into account that gamblers might play with different frequencies (as they indeed do, according to the data) and average over all individual bets after separating for maximum streak length achieved, the conclusions we can draw are very different. The results presented in Figure 4.5.B demonstrate a notable contrast with the findings of Xu and Harvey (2015). Specifically, when considering the influence of different gambling frequencies and averaging

[9] Formally, the measure in Figure 4.5.A can be given as $E\left[P_{win}(n)\right]_A = N_{i(n)}^{-1} \sum_{i(n)} N_{j_{i(n)}}^{-1} \sum_{j_{i(n)}} P_{win}\left(j_{i(n)}\right)$ while the measure in Figure 4.5.B as $E\left[P_{win}(n)\right]_B = N_{ij(n)}^{-1} \sum_{i(n)} \sum_{j_{i(n)}} P_{win}\left(j_{i(n)}\right)$, where $i(n)$ stands for player i with a maximum streak length of n, $j_{i(n)}$ denotes player $i(n)$'s bets, and $ij(n)$ the entirety of games in which players $i(n)$ takes part, while N represents the frequency with which the game occurs within the set of games involving player $i(n)$.

over individual bets while separating them based on maximum streak length achieved, our observations diverge significantly. It is evident that gamblers experiencing longer winning streaks tend to select bets with lower mean odds, indicating a preference for safer bets. Conversely, gamblers with longer losing streaks tend to opt for riskier bets. Notably, the disparity in the starting points of the two curves can be attributed to the betting behaviors of gamblers who have played numerous games but have never won a single game or have achieved only a few consecutive wins. These gamblers tend to place risky bets with low winning probabilities on average. Conversely, players who have never experienced a single loss or have only encountered a few consecutive losses tend to exhibit a tendency towards safer bets with higher winning probabilities on average.

4.4 Conclusion

According to Xu and Harvey (2014), there is evidence suggesting the presence of the hot hand phenomenon in sport betting. The proposed mechanism underlying their findings is simple: gamblers progressively adjust their risk-taking behavior by opting for safer odds in each consecutive game during a winning streak, while individuals experiencing a losing streak are inclined to choose riskier games in subsequent rounds, thereby increasing their chances of losing.

In this article we have argued that this reasoning, even though seemingly convincing and supported by aggregate-level data analysis, may be misleading. We have countered the claimed findings, demonstrating that while comparable patterns may emerge at the aggregate level, there are significant discrepancies when examining individual-level data. While it is worth acknowledging that in a subsequent paper Xu and Harvey (2015) presented new insights by utilizing personal-level data, however, the results may be as misleading as in Xu and Harvey (2014). Therefore, we have argued that the analysis presented in Xu and Harvey (2014) is not robust enough and their conclusion regarding the existence of the hot hand phenomenon appears to be premature and lacks a sufficient level of certainty.

As mentioned above, Figure 4.1 can seem contradictory to Figure 4.5 at first, but let us consider again the difference between the streak categories used in Xu and Harvey (2014) (to create the former) versus the ones in Xu and Harvey (2015) (to create the latter figure):

when looking at winning probabilites in the context of streak length, our groups grew ever more exclusive, the group related to streak length n contains only those games that have reached this streak length and only include those who may even did reach an $n+1$ streak. To the contrary, the streak length categories of Xu and Harvey (2015) get more inclusive as the length n grows: the group who's longest streak experienced was n may have experienced a streak of $n-1$ as well, but never and $n+1$ long. This difference explains why the first figure has large difference between wins and losses at large n, while the second shows it at small ones.

We acknowledge that there are differences between the dataset used in our analysis and the one employed by Xu and Harvey (2014). While they analyze gambling data specifically on sports betting, our dataset is on online betting. It is important to note that while the outcomes in our analysis are intentionally designed to be independent and identically distributed, this may not hold true for other types of betting such as horse racing, football, or cricket. Furthermore, the bets in our case involve objective probabilities commonly known to the bettors. This is different from betting on events involving human action, which Xu and Harvey (2014) and Xu and Harvey (2015) studied. As Oskarsson et al. (2009) have convincingly shown, however, the nature of the data-generating process is critical in determining how humans respond to streaks. This might make the two analysis difficult to compare. However, irrespective of the specific data generating process, the finding of the "presence" of the hot hand phenomenon in our dataset highlights the potential weaknesses in their applied methodology. This suggests that their conclusions may be questionable and calls for further examination and scrutiny of their findings.

There is a compelling theoretical argument, detailed in the Appendix, that further contradicts Xu and Harvey (2014)'s reasoning. It is argued that as long as the set of games from which gamblers make their selections is not limited to a single option and gamblers do not modify their betting behavior during a streak, the observed probability of winning increases as the length of the winning streak grows. Additionally, the observed probability of winning converges to a streak-independent level in the absence of continuous winning or losing streaks.

In the Appendix, we demonstrate through numerical simulations that these theoretical results yield outcomes equivalent to the findings reported by Xu and Harvey (2014).

However, it is important to note that these simulations are based on the assumption that gamblers do not alter their risk behavior and consistently choose the same odds while placing identical bets over time. Consequently, the results obtained in Xu and Harvey (2014) may simply arise as a consequence of the specific analytical approach employed.

The situation is somewhat similar to what has been pointed out by Wardrop (1995) regarding Gilovich et al. (1985). By using the average results of a pool of players, one may arrive to conclusions that are not necessarily true on a player by player basis. Considering the performance of individual players, we can see that the winning probability is independent from the length of streak experienced. Yet, if all players are pooled up, risk seekers will be over-represented in the population with long losing streaks and vice versa for the risk averse. Moreover, with this aggregation we may also introduce a selection bias by gradually restricting the conditions for inclusion in the streak group.

In conclusion, the findings presented in this paper strongly suggest that the analysis employed by Xu and Harvey (2014) is inadequate for making any substantial claims regarding the existence of the hot hand phenomenon, let alone providing conclusive evidence of its presence in gamblers' behavior. All in all, we believe that the existence of the hot hand phenomenon in gambling remains an unresolved question that requires further investigation and analysis.

4.5 APPENDIX: The illusion of hot hand with persistent players

Generally, in any gambling situation there are two important factors that characterize a gamble. One is the wager or the bet amount, which is the basis of the payout. The relationship between the wager and the payout is always linear. The other important attribute for any gamble is the winning odds. This usually is represented as a multiplier applied on the wager for each outcome. These multipliers are constructed as $M = (\rho+\chi)^{-1}$, where ρ stands for the unbiased probability of the winning outcome and χ for the house cut. This characteristic of the games makes gambles comparable to each other, even if the events that the outcomes are derived from are totally different in nature. The key variables analyzed in Xu and Harvey (2014) are the observed winning probabilities and the mean

odds for different winning and losing streak lengths.

Xu and Harvey (2014) argue that gamblers by decreasing or increasing their odds after each lucky or unfortunate outcome create the circumstances for the hot hand to prevail. Even though seemingly convincing, we argue that the results presented in Xu and Harvey (2014) are not necessarily the consequences of gamblers' changing their risk behavior and as we will show in this section it can be obtained by assuming persistent players who do not change their behavior during subsequent games.

To demonstrate this, consider a continuum of players with uniformly distributed preferences who randomly select an implied winning probability between 0% and 100%. Assume that gamblers, after selecting their characteristic winning probabilities, do not change their behavior (the odds of the game or the bets they place) in any way. For the sake of simplicity, we assume that the house cut equals to zero.

Calculating the probability distribution for players with an initial choice of winning probability (denoted by ρ) having n sequential wins, we have that

$$\Phi_n^w(\rho) = (n+1)\rho^n \tag{4.1}$$

where ρ is drawn from the uniform distribution. The same can be derived for players who do not have a winning streak of length n

$$\Phi_n^{\bar{m}}(\rho) = \frac{(n+1)(1-\rho^n)}{n}. \tag{4.2}$$

Determining the mean of these distributions over the choice of ρ, which gives us the observed average winning probability of the participants, we get

$$\chi_n^w = \int_0^1 \rho\Phi_n^w(\rho)d\rho = \int_0^1 (n+1)\rho^{n+1}d\rho = \frac{n+1}{n+2} \tag{4.3}$$

for players characterized with a winning streak of length n, and for those who are not lucky enough to be part of such a lengthy winning streak, we have

$$\chi_n^{\bar{m}} = \int_0^1 \rho\Phi_n^{\bar{w}}(\rho)d\rho = \int_0^1 \frac{(n+1)\rho(1-\rho^n)}{n}d\rho = \frac{n+1}{2n} - \frac{n+1}{n(n+2)}. \tag{4.4}$$

Considering the limits for expressions (4.3) and (4.4) when $n \to \infty$, we have that the expected probability of winning tends to 1 conditioned on having winning streaks, whereas the probability of winning outside of these winning streaks will limit to 1/2.

The losing case is the exact opposite of the above. For the probability distribution of winning probabilities after a losing streak of length n we get

$$\Phi_n^l(\rho) = (n+1)(1-\rho)^n \tag{4.5}$$

with the mean of

$$\chi_n^l = \int_0^1 \rho\Phi_n^l(\rho)d\rho = \int_0^1 (n+1)\rho(1-\rho)^{n+1}d\rho = \frac{n+1}{(n+2)(n+3)} \tag{4.6}$$

which tends to 0 as $n \to \infty$. If the outcomes are drawn from an independent and identically distributed set, the observed winning probabilities converge to the implied ones, although the speed of the convergence will be slower for probabilities close to 0% and 100%.

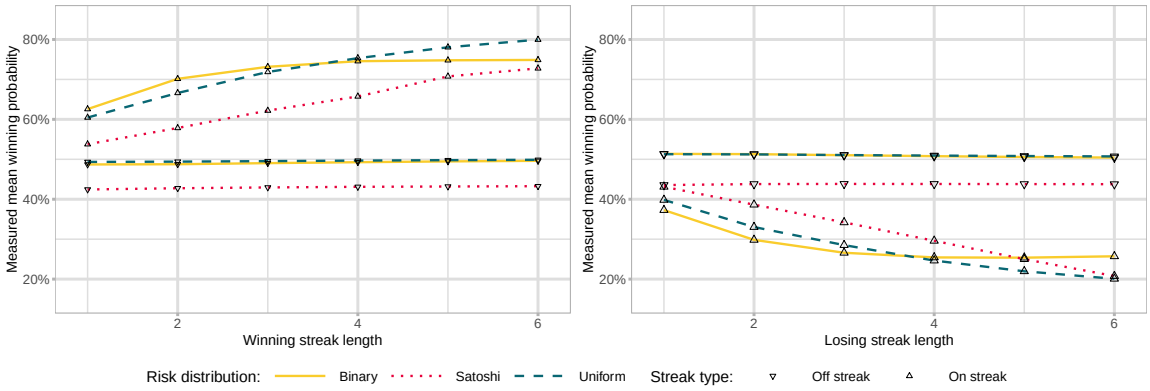

Figure 4.6: A (left): Observed winning probability of players with and without a winning streak of certain lengths. B (right): Observed winning probability of player with and without a losing streak of certain lengths. The probability is monotonically decreasing for players experiencing losing streaks while converging to 1/2 for others.

We should also consider that in reality, the distribution of the risk appetite might not necessarily be a uniform distribution. Even so, to achieve the convergence of the observed probabilities as in Xu and Harvey (2014) it is sufficient to have a choice set with at least two different risk levels, and it is not necessary to have a continuous and uniformly distributed game set. To demonstrate this, we present simulation results for two different scenarios and compare them to the uniform case presented above. One distribution (labeled as *Binary*) that is included in our simulations is a simple initial distribution of low risk (with 75% implied winning probability or, equivalently, with odds of 1.333x) and high risk (10% winning

probability or odds of 10x) choices with equal occurrences. Another distribution (labeled as *Satoshi*) that we present simulation results for, features the game choice distribution of gamblers in our dataset described in more details in section 4.2.

As one can observe in Figure 4.6 the probability of winning is monotonically increasing for players experiencing longer and longer winning streaks, while it is converging to 1/2 for gamblers without these lucky streaks. Moreover, as the figure clearly indicates, the trends regarding the winning probability after winning or losing streaks are independent of the actual size and distribution of the choice set. These results suggest that if gamblers can choose their gamble from a choice set with at least two games with different characteristics - which is usually the case in any real situation - the convergence of winning probabilities similar to the ones presented in Xu and Harvey (2014) will emerge without players ever changing their behaviour.

P_{win}	*Multiplier*	A	B	C	D	E
0.002%	64000	0.4%	2.2%	2.2%	2.7%	5.2%
0.003%	31982	0.1%	0.1%	0.0%	0.2%	0.2%
0.006%	15991	0.5%	0.1%	0.1%	0.2%	0.1%
0.01%	7995	0.1%	0.1%	0.1%	0.2%	0.1%
0.02%	3998	0.1%	0.1%	0.2%	0.1%	0.2%
0.05%	1999	0.1%	0.3%	0.2%	0.2%	0.3%
0.1%	999.4	0.3%	0.2%	0.2%	0.2%	0.3%
0.2%	499.7	0.4%	0.2%	0.2%	0.2%	0.2%
0.4%	249.8	0.2%	0.4%	0.2%	0.2%	0.2%
0.8%	124.9	0.4%	0.8%	0.9%	0.4%	0.4%
1.5%	63.97	0.4%	3.0%	4.3%	3.4%	6.0%
2.3%	42.64	0.7%	0.7%	0.2%	0.4%	0.3%
3.1%	31.98	0.3%	0.6%	3.0%	2.1%	8.0%
4.6%	21.32	0.3%	0.7%	0.2%	0.4%	0.4%
6.1%	15.99	0.3%	0.6%	0.6%	0.3%	3.0%
9.2%	10.66	0.9%	0.8%	0.8%	1.5%	1.2%
12%	8.000	0.7%	3.8%	5.1%	5.3%	7.4%
18%	5.335	1.4%	0.9%	3.6%	2.9%	6.1%
24%	4.003	2.0%	8.5%	5.7%	6.5%	7.9%
37%	2.670	5.5%	4.8%	7.8%	5.6%	5.3%
49%	2.004	24%	45%	27%	23%	10%
50%	1.957	47%	14%	16%	25%	22%
73%	1.338	12%	7.3%	11%	9.2%	7.4%
79%	1.235	NA	1.6%	3.6%	2.5%	2.0%
85%	1.147	NA	1.6%	1.8%	2.6%	1.2%
92%	1.071	NA	0.7%	4.4%	3.3%	3.6%
98%	1.004	1.3%	0.2%	0.2%	0.3%	0.3%

Table 4.2: Observed distribution of risk choice in the analyzed periods. P_{win} stands for the probability of a winning outcome and *Multiplier* is applied on the bet amount on a positive outcome. The other letters stand for the periods described in Table 4.1

Chapter 5

Unmasking Risky Habits: Identifying and Predicting Problem Gamblers Through Machine Learning Techniques

Abstract

The use of machine learning techniques to identify problem gamblers has been widely established. However, existing methods often rely on self-reported labeling, such as temporary self-exclusion or account closure. In this study, we propose a novel approach that combines two documented methods. First we create labels for problem gamblers in an unsupervised manner. Subsequently, we develop prediction models to identify these users in real-time. The methods presented in this study offer useful insights that can be leveraged to implement interventions aimed at guiding or discouraging players from engaging in disordered gambling behaviors. This has potential implications for promoting responsible gambling and fostering healthier player habits.[1]

[1]This work has been accepted and published as Sándor and Bakó (2024) by the Journal of Gambling Studies.

5.1 Introduction

The gambling industry's rapid technological transformation has led to unprecedented accessibility, contributing to a concerning rise in problem gambling cases (Potenza et al., 2011; Chagas and Gomes, 2017). Although the recent pandemic initially reduced overall gambling participation, it triggered a surge in online and problem gambling, with younger individuals disproportionately affected (Wardle et al., 2021; Hodgins and Stevens, 2021). The societal costs associated with problem gambling are projected to have a profound impact on the economy (Hofmarcher et al., 2020). Notably, online gambling platforms employ persuasive tactics called "sludges" to entice users to engage in longer and riskier betting practices (Newall, 2019; Newall et al., 2020). Moreover, the industry utilizes industrial machine learning solutions to support these practices (Coussement and De Bock, 2013), and the utilization of dark patterns has demonstrated significant effects on consumer manipulation (Bogliacino et al., 2023). Consequently, regulatory bodies have initiated investigations into the adverse implications of online choice architecture.

To address the issue of problem gambling, various studies have examined the effectiveness of nudges, such as implementing loss limits and providing personalized feedback, in discouraging addictive behaviors (Brodeur, 2019; Auer et al., 2018; Auer and Griffiths, 2020). Promising results have been observed in brick-and-mortar casino gambling through the introduction or promotion of self- and forced exclusion periods (Kotter et al., 2018). In the case of online gambling, interventions that disrupt the gambling flow, such as fixed or self-defined monetary limits, have shown effectiveness (Folkvord et al., 2019). However, results by Caillon et al. (2019) and Giroux et al. (2017) suggest that the effectiveness of these measures in online gambling remains unclear.

Unsupervised machine learning techniques have been successfully used to identify vulnerable user groups in gambling (Deng et al., 2019; Braverman and Shaffer, 2012; Xuan and Shaffer, 2009). Machine learning algorithms can also predict the development of addictive patterns (Mak et al., 2019). Previous studies on gambling data have effectively predicted self-exclusion using supervised learning techniques like logit regression, gradient boosting, and neural networks (Percy et al., 2016; Ukhov et al., 2021; Buttigieg et al., 2022; Finkenwirth et al., 2021), relying on observed behavioral markers like frequency of

play, risk-taking behavior, and bet sizes. However, one limitation of previous studies is their reliance on rule-of-thumb measures to select the target subgroup of gamblers. This approach may introduce researcher bias and hinder the transferability and general efficacy of the results across different game types and designs.

In this analysis, we propose a new approach that avoids using pre-observed labeling information, thereby reducing potential bias towards self-aware gamblers. Instead, we combine unsupervised machine learning techniques to create labels for problem gamblers. Once the target categories are established, we simplify the process of selecting specific prediction algorithms using automatic machine learning (autoML) algorithms. This approach ensures a more objective and robust method for identifying problem gamblers and predicting their behavior.

5.2 Methods

In this study, our main goal is to demonstrate the effectiveness and ease of predicting problem gambling. To achieve this, we adopt a dual approach. First, we employ k-means clustering to categorize our target users based on their gambling behavior over a 7-day period, following an initial 3-day period. This clustering process helps us assign labels to our problem gambling group. Next, we develop predictive models that can forecast the cluster label of each player based on their behavior during the initial 3-day period. To accomplish this, we utilize a large dataset of betting transactions extracted from publicly available data sources.[2]

5.2.1 Dataset

Among the early use cases of Bitcoin, the pioneering decentralized digital currency, online gambling emerged as a prominent application. Bitcoin's innovative system provided an ideal environment for experimentation, and due to its unregulated nature, numerous online gambling sites have sprung up since 2012, leveraging the Bitcoin ecosystem. One of the most

[2]Scripts used for data preparation and analysis are made publicly available at github.com/sampaat/problem_gambler_prediction.

successful ventures within the cryptocurrency community was Satoshidice.[3] This platform implemented a simple yet fair gambling system, offering games to players with varying odds or levels of risk. The fairness of the games was ensured through two mechanisms: the expected return for each game was fixed, thereby creating a house cut that remained independent of the risk level. Additionally, the game outcomes were determined by a "dice roll" generated by combining information from the Bitcoin ledger related to the bet itself and a pre-set secret, which could be independently verified by the players.

The game process was straightforward. Players selected their desired level of risk by choosing a specific game from a predefined list, which presented various winning probabilities (inversely proportional to the odds) alongside a unique wallet address. By initiating a transaction to one of these addresses, the player placed a bet with the sent amount (within specific bet limits). The site assessed the bet based on transaction details and the secret key, promptly sending a return transaction reflecting the outcome. Although blockchain confirmation times in 2013 typically ranged from 5-7 minutes, most bets received instantaneous responses from the site.

Given the blockchain's public nature, it is possible to extract a comprehensive history of all incoming and outgoing transactions associated with any address on the network. We collected all bets placed at and return transactions sent by SatoshiDice during its operational period in the specified form (the site transitioned to a prepay system in 2014). Our dataset comprises a complete longitudinal observation set of betting transactions, with five 21-day periods used to assess the robustness of our procedure over different samples and time frames. For detailed information on the data gathering methodology and resources, see Bakó and Sándor (2021).[4].

From the transaction details, we can directly observe the following descriptors:

- **Player ID**: User identification label created based on the dataset of Kondor et al. (2014). The ID links transactions associated with the Bitcoin addresses controlled by the same entity. However, it does not provide any personal or location information about the player in question.

- **Time of bet**: Timestamp given to the Bitcoin transaction of the bet placed.

[3]see https://web.archive.org/web/20121103121459/http://www.satoshidice.com/

[4]The dataset used for the analysis can be accessed at DOI: 10.5281/zenodo.5600259.

	Start date	Number of bets	Number of users	Total bets placed (BTC)	Median bet size (BTC)	Mean daily price (USD/BTC)
A	2012-05-02	119,399	1002	46,649	0.04	5
B	2012-09-17	129,265	2114	85,280	0.06	12
C	2012-12-17	252,301	3405	407,140	0.04	13
D	2013-05-04	329,155	3432	100,430	0.02	111
E	2013-09-11	86,520	1400	62,920	0.03	123

Table 5.1: Summary statistics of the subsets of the observed gambling history used. The exchange rates have been sourced from the public historical data published by the online cryptocurrency exchange aggregator BitCoinCharts (see http://www.bitcoincharts.com)

- **Time of answer**: The timestamp is assigned to the answering Bitcoin transaction, which we have paired with the bet.
- **Game ID**: The game selected by the player is determined by the target of the betting transaction. Directly linked to this target, we can assign a fixed winning probability and odds to the respective bet. This enables us to determine the specific game being played and the associated chances of winning for each betting transaction.
- **Bet amount**: The part of the bet transaction that has been directed towards the selected game address.
- **Answer amount**: The amount of Bitcoin directed back to the betting addresses from the SatoshiDice wallet determines the outcome of the gamble. This return transaction reflects the winnings or losses of the bet and makes it possible to determine the final result of the gambling activity.

From the variables mentioned above, we can derive several informative descriptive measures of the gambling process. While one approach could involve treating this data as a time series, as demonstrated in Peres et al. (2021), we find that producing daily aggregates achieves similar clustering outcomes without the computational complexity associated with the former method.

To facilitate both the labeling and predicting exercises, we have derived the following aggregates. It's worth noting that these aggregates largely align with the observed behavioral markers used in previous studies (Deng et al., 2019). As the last step of preparation, all of the measures are standardized. By employing these derived measures, we can effectively capture important aspects of the gambling behavior and use them to categorize and predict the behavior of interest.

- **Number of games**: Number of bets placed on the given period, transformed to a logarithmic scale.
- **Number of days active**: Number of calendar days that the player placed bets from the observed period (only used for labeling).
- **Number of sessions per days active**: Number of game sessions played defined by successive bet chains where no more than 1 hour has been spent inactive by the user, divided by the number of active days.
- **Median winning probability**: Median of the implied winning probability of the bets placed during the period. This describes the risk appetite of the player.
- **Range of winning probability**: Distance of the smallest and largest implied winning probability of the bets placed during the period. This describes the variability in risk taken by the player.
- **Mean bet**: Mean of the bet amounts placed during the period (in BTC). A logarithmic transformation has been applied.
- **Maximum bet**: Maximum of the bet amounts placed during the period (in BTC). A logarithmic transformation has been applied.
- **Total payout**: The aggregated amount of bets placed and answers received by players during the period (in BTC) resulting in the total gains/losses.

Our analysis consists of two main steps, with the second step involving the prediction of labels created in the first step. To facilitate this process, we establish two distinct subsets from each of our samples. What sets our approach apart is that we use shorter

sample durations for both clustering and prediction compared to previous studies such as Braverman and Shaffer (2012) or Xuan and Shaffer (2009), which typically relied on 30-day to full history samples. For each gambler in each sample, we identified a 10-day period starting from their first betting day in the given sample. This window was then divided into the first 3 days and the last 7 days. The last 7-day window was utilized to identify emergent behavioral patterns indicative of problem gambling tendencies. On the other hand, the first 3-day window served as the basis for predicting the labeling of problem gambling behavior.

To create the clustering dataset, we aggregated relevant variables over the week-long window. Conversely, for the prediction dataset, we aggregated the data on a daily basis. Additionally, we introduced additional variables representing the change over days in the number of daily bets and the mean bet size. These features are crucial for predicting problem gambling labels effectively. By employing shorter sample duration and employing different aggregation methods for clustering and prediction, we demonstrate the robustness and efficiency of our approach. This allows us to effectively identify and predict problem gambling behaviors with improved accuracy and computational efficiency compared to previous studies.

5.2.2 Labeling problem gamblers: unsupervised learning

k-means clustering is a widely used method in behavioral profiling, employed in marketing (Arumawadu et al., 2015), psychological settings (Stegmann et al., 2019), and specifically in analyzing gambling behavior (Braverman and Shaffer, 2012; Xuan and Shaffer, 2009). The key advantage of this unsupervised method is that it provides an unbiased separation of players based solely on their gambling profiles, devoid of any influence from researchers or regulators.

In our analysis, we use one week-long aggregates of the measures presented in Section 5.2.1. This observation period begins three days after the players' first observed betting day. It's worth noting that inclusion in this set indicates that players placed bets between the third and tenth day after their first bet in the sample. The user retention rate, as observed in this manner, varies between 14% (sample C) and 35% (sample E). Spearman correlations between the input variables generally stay below $r < .6$. Slightly higher correlations ($.6 <$

$r < .8$) are observed between the mean logarithmic bet versus the maximum bet and the number of active days versus the number of games played. However, deviations from this linear trend are significant in both separation and later prediction, indicating substantial variations in bet amounts and activity levels. We acknowledge the presence of outliers in our dataset (e.g., extreme number of bets or extremely large maximum bets), which can impact the robustness of k-means clusters. To address this, we employ the method of trimmed k-means clustering (Cuesta-Albertos et al., 1997; Hennig, 2020), allowing for a 1% trimming factor, ensuring high stability for the specific separation we are focusing on.[5] Based on measurements of both the Silhouette and Dunn indices, an optimal common choice for number of clusters to be two (see Figure 5.1).

5.2.3 Predicting gambling behavior: autoML

Our prediction process involves categorization, where various techniques can be used, such as generalized linear models, random forests, gradient boosting, and deep learning algorithms. However, manually detailing and fine-tuning these methods to find the optimal one (or ensemble) for the given problem can be cumbersome. Instead, we demonstrate a more efficient approach using automatic machine learning techniques, specifically H2O's autoML package (LeDell and Poirier, 2020; LeDell et al., 2022). This approach allows us to find the optimal model (or combination) by leveraging goodness-of-fit measures. Using autoML, we streamline the model selection process, producing robust and cross-validated models. This automation not only saves time, but also ensures a reproducible process that can be easily deployed into production and archived for future reference or investigation.

In our prediction process, we have two targets: user retention, which predicts whether the observed player will continue placing bets in the target period or leave the game, and identification of players belonging to the group labeled as "intensive" during the clustering phase, indicating signs of problem gambling. To train the model, we use the variables described in Section 5.2.1, aggregated on a daily basis for the 3-day gambling period of

[5] We excluded the trimmed "cluster" of users from the comparisons and the modeling exercise. Our case to do so is twofold: because of the choice of the trimming parameter, the size of this group is relatively small for each clusters ($10 - 35$ players) and even in these tiny groups, their gambling style being very diffused and extreme even compared to each other.

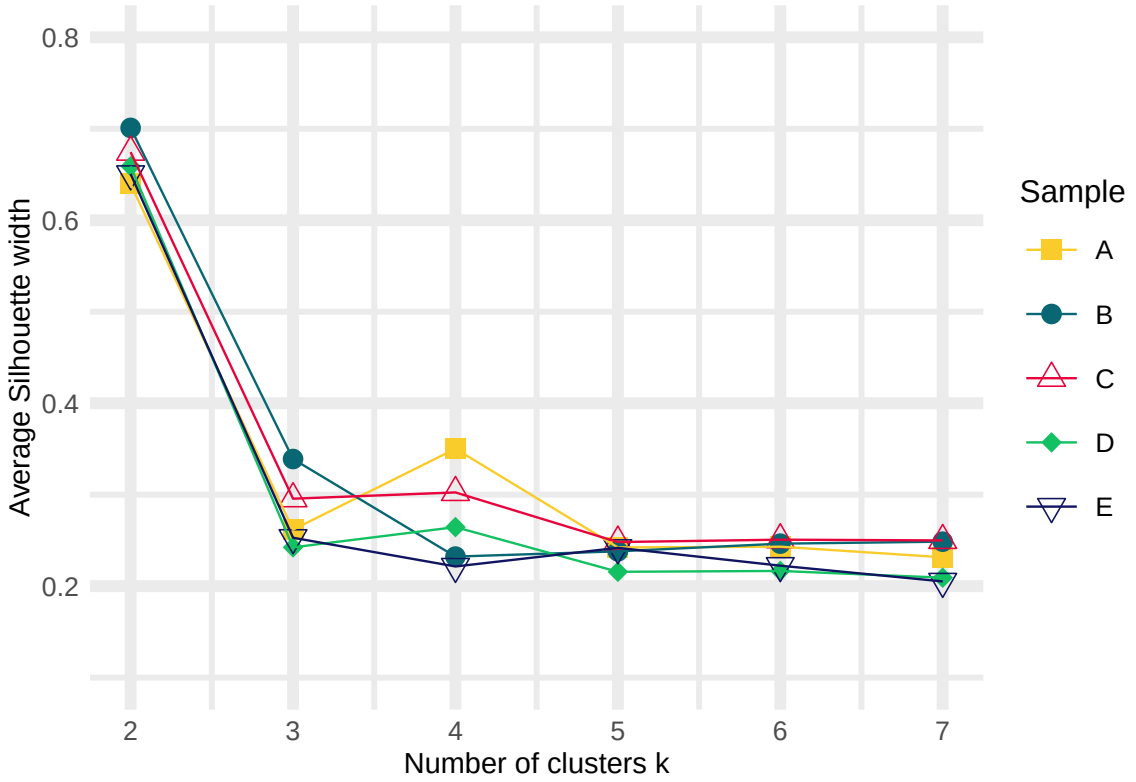

Figure 5.1: Average of silhouette widths over different values of k in the samples (see Section 2.3.3 for definition). A uniform choice of $k = 2$ seems adequate for all samples.

our users starting from their first betting day in our samples. After our above detailed steps of data preparation and plentiful variable selection, we ran the autoML algorithm with default settings, including 5-fold internal cross-validation, creating 10 model sets, and a computational limit of 300 seconds. [6] The calibrations are performed on a desktop computer without GPU support. This automated approach ensures an optimized model selection process and facilitates efficient and accurate predictions for both user retention and problem gambling identification.

[6]Despite our best efforts for setting up individual optimization methods using various implementations detailed in Section 2.4, we have found that the autoML approach of combined models prevail in performance every time.

5.3 Results

5.3.1 Labeling

Table 5.2 presents the median values of the input variables for the identified clusters. A clear contrast is evident for most of these measures between the casual (-) and intensive (+) groups. The most notable difference lies in the dimensions of gambling frequency: the intensive group places significantly more bets (ranging from 62 to 303) compared to the casual group (ranging from 6 to 7). Furthermore, members of the intensive group engage in gambling almost every day during the observation period, while casual players only participate for 1 to 2 days. Furthermore, the intensive group returns to betting multiple times a day, with the number of daily sessions exceeding 2.

Analyzing risk-taking behavior, we observe that both groups often opt for "balanced" bets, offering approximately 50% probability of winning (or a multiplier of 2). However, the intensive group displays a much wider variation in risk-taking compared to the casual group. A similar pattern is noticeable for bet sizes. Although the average bet sizes might not differ significantly, the maximum bets placed by players in the intensive group tend to be approximately an order of magnitude higher on average. The difference in expected losses (total payout) is a direct consequence of the aforementioned observations. Since the game is implemented fairly, with the house cut independent of the wager's risk level, players in the intensive group, who engage in more frequent and higher-risk betting, can expect to accumulate larger losses on average.

The identified clusters exhibit distinct behavioral patterns, with the intensive group demonstrating a higher frequency of gambling, risk taking, and bet sizes, resulting in higher expected losses due to the nature of the game's fairness. These patterns overlap heavily with the behavioral markers referenced in the DSM-5 guidelines (American Psychiatric Association et al., 2013): a tendency to play more frequently over the days, with multiple sessions (bets placed within a day with at least a 4-hour pause between) results in orders of magnitudes more total bets placed. This combined with a distinctly more risk taking behavior in choosing scenarios and occasionally placing much larger bets results in higher overall losses and volatility in payout. Our methodology still misses a list of criteria required

	Group	Games	Days	Sessions	Median risk	Risk range	Mean bet	Max bet	Total payout
A	-	6 (4)	2 (2)	1.5 (0.8)	48.8 (25.0)	0.0 (36.6)	0.07 (13.53)	0.18 (10.00)	-0.10 (0.44)
	+	89 (5)*	6 (3)*	2.8 (1.0)*	48.8 (24.4)	42.5 (52.6)*	0.13 (18.52)†	2.00 (7.98)*	-0.33 (3.67)
B	-	7 (6)	1 (2)	1.5 (1.0)	48.8 (28.1)	0.0 (24.4)	0.05 (12.81)	0.20 (10.00)	-0.02 (0.42)
	+	99 (6)*	6 (3)*	2.1 (1.4)*	48.8 (36.6)	48.8 (48.8)*	0.15 (10.08)†	2.48 (9.87)*	-0.57 (6.78)
C	-	9 (7)	2 (2)	1.5 (1.0)	50.0 (24.4)	1.2 (32.6)	0.07 (17.79)	0.37 (40.00)	-0.04 (0.68)
	+	303 (8)*	6 (3)*	2.3 (1.2)*	48.8 (13.4)	71.7 (47.3)*	0.05 (9.12)	2.56 (9.43)*	-0.70 (7.79)†
D	-	7 (7)	2 (3)	1.5 (1.3)	48.8 (13.4)	6.1 (67.1)	0.02 (5.80)	0.04 (9.29)	-0.03 (2.45)
	+	105 (5)*	6 (2)*	2.6 (1.0)*	48.8 (37.8)†	48.7 (36.6)*	0.04 (3.33)†	0.85 (9.05)*	-0.32 (0.12)
E	-	6 (5)	2 (2)	1.6 (0.8)	48.8 (54.9)	0.0 (25.6)	0.02 (7.33)	0.04 (12.31)	-0.02 (0.18)
	+	62 (5)*	6 (3)*	2.5 (1.4)*	48.8 (31.7)	50.0 (40.0)*	0.03 (5.28)†	0.47 (14.27)*	-0.15 (1.81)

Table 5.2: Median (IQR) statistics of the clusters identified in the data samples (described in Table 5.1.). The groups labeled to be the intensive gamblers are signed with a + in the group column and highlighted with gray background. The results of Kruskal-Wallis rank sum tests for difference between the group descriptors are signed on the intensive group values (P-levels: $\dagger.05, *.001$). The separation of the groups is consistent in the dimensions of game frequency (games, days active and sessions per day) and also the wider risk and bet size range.

for diagnosis by design: restlessness, irritability, and preoccupation can take many forms that are unmeasurable in our context, as well as complex patterns like lying to conceal gambling intent and the degradation of the social fabric around gamblers. Still, as much as our labeling can fit someone who just enjoys playing heavily and consistently over time - albeit them not necessarily being problem gamblers - it creates the environment for the surrounding symptoms to emerge.

5.3.2 Prediction

The top section of Table 5.3 displays the predictive performance of the best models generated by the autoML algorithm for all our samples. The results reveal remarkably high area under the curve (AUC) measures and low errors, alongside satisfactory log loss compared to the target prevalence. These findings indicate that, on average, we can accurately predict whether a player will or will not place a bet in the 4th to 10th day following their initial betting day, based on the optimal probability level set. This high accuracy in predictability

Retention prediction							
	Sample size	Prevalence	Best model	Models in ensemble	AUPCR	Logloss	RMSE
A	445	34%	GBM		0.92	0.20	0.25
B	1200	18%	Ensemble	DL/DRF/GBM/GLM	0.88	0.11	0.18
C	1837	14%	Ensemble	DRF/GBM/GLM	0.91	0.09	0.16
D	2263	30%	Ensemble	DL/DRF/GBM/GLM	0.91	0.16	0.22
E	741	35%	GBM		0.90	0.18	0.23
Intensive period prediction							
	Sample size	Prevalence	Best model	Models in ensemble	AUPCR	Logloss	RMSE
A	445	17%	GBM		0.70	0.23	0.27
B	1200	10%	Ensemble	DL/DRF/GBM/GLM	0.63	0.15	0.22
C	1837	6%	GBM		0.58	0.11	0.19
D	2263	14%	GBM		0.74	0.19	0.25
E	741	16%	GBM		0.62	0.25	0.29

Table 5.3: Descriptors of prediction performance of top models found using the autoML method. Sample size shows the number of players who have placed a bet in the first 12 days of the sample. Prevalence refers to the relative size of the target group compared to the sample size. The models used are gradient boosting (GBM), deep learning (DL), distributed random forest (DRF) and generalized linear (logit) model (GLM). Submodels are only detailed for ensembles.

of user retention is not surprising since modeling this metric has already become an industry standard, hence yielding expectedly strong results.

Looking at the lower part of Table 5.3, we observe the same statistics for predicting player inclusion in the intensive clusters, as described in Section 5.3.1. Comparing this prediction to the user retention case, we notice a slightly weaker predictive strength, but the metrics still demonstrate good predictive quality. The area under the curve metrics remain very high, and the log losses are significantly below trivial levels. With the optimal probability threshold, these models provide categorization with only a few instances of mislabeling for each sample. These models exhibit explanatory power similar to recent analyses, as seen in Finkenwirth et al. (2021). In most cases, gradient boosting models performed the best as standalone models, and ensembles of gradient boosting and other models were used in

	1#	2#	3#	Importance
A	Max bet (1)	Risk range (3)	Games (3)	29.7%
B	Games (3)	Median risk (2)	Sessions (3)	32.5%
C	Sessions (1)	Risk range (3)	Median risk (2)	30.7%
D	Risk range (3)	Median risk (1)	Total payout (1)	6.1%
E	Mean bet (3)	Total payout (1)	Total payout (2)	29.1%

Table 5.4: Top 3 most important variables from the 3 day observation periods and their total of respective relative importance relative to all other variables. The day of observation for each variable is provided in the parenthesis following them. For sample B the best single model (GLM) was used.

other instances. It is worth noting that during the autoML training, a set of alternative methods (both standalone and ensemble) were provided, and they exhibited comparable performance levels. The high predictive quality of these models, even in standalone configurations, highlights their robustness and effectiveness in identifying players likely to belong to the intensive gambling clusters.

If we look at the list of top explanatory variables in Table 5.4, we can see that while indicators of frequency (number of sessions and games) do show up as top explanators in the first three samples, the landscape is not at all dominated by these variables, that would point to trivialized solutions found. We can see that the models are rather complex, as the top 3 variables together do not explain the majority of the model (with the model for sample D being even more dispersed in this sense), but also that they are not dominated by descriptors of the last day of observation. These signs hint at the process behind our predictive algorithm not only being complex but non-stationary over the lifetime of the game, highlighting the need for a methodology that can be generally applied and retrained over the evolution of unobserved variables.

5.4 Conclusion

The successful demonstration of the effectiveness of unsupervised learning methods in separating players exhibiting signs of problem gambling has significant implications for the field

of responsible gambling and player protection. By identifying key variables that measure the intensity of gambling, such as the number of bets placed and the frequency of betting sessions, we can easily detect the group displaying problem gambling attitudes. This separation process has proven to be robust and reliable across various observation periods, even when dealing with varying sample sizes, making it a valuable and adaptable tool for early identification of problem gambling behaviors.

The ability to apply the chosen behavioral descriptors to different types of gambling, regardless of their specific structures, highlights the potential universality of this approach. This flexibility allows for the assessment of problem gambling tendencies in various gambling contexts, providing valuable insights for policymakers, regulators, and gambling operators. However, there are certain manual steps involved in the process, which may vary when dealing with other types of gambles. Determining the optimal number of groups for separation and subsequent labeling requires careful consideration and domain-specific knowledge. Additionally, the lack of a follow-up measure to validate whether the identified players are indeed problem gamblers may lead to lower labeling accuracy for true problem gamblers. Future research should focus on incorporating follow-up measures to enhance the accuracy and reliability of player categorization.

Machine learning approaches, such as the ones used in this study, offer an easy-to-implement monitoring tool for gambling platforms. These models can serve as a foundation for implementing proactive measures, such as nudging or forced exemptions, to deter at-risk gamblers from developing or continuing problem gambling behaviors. By identifying players early on who show signs of problematic gambling, operators can provide targeted interventions and support to promote responsible gambling practices and minimize harm. It is essential to recognize that the effectiveness of forced exemptions hinges on their widespread application on a market-wide scale. This measure prevents problem gamblers from simply shifting to other gambling venues or online sites, ensuring a more comprehensive and effective approach to player protection.

While the results of this study are promising, further replication and validation on other forms of gambling, such as online versions of classical casino games and sports betting, are necessary to assess the generalizability of the findings.[7] Conducting a control group study

[7]Another limitation of our analysis is that the identification of risky behavior relies on specific past

with real gamblers, along with follow-ups and psychological profiling, would provide valuable data to compare the effectiveness of player selection and the optimal combination of nudging or forced deterring techniques. This comprehensive investigation would yield deeper insights into the potential impact of these interventions on curbing problem gambling and fostering responsible gambling practices on a broader scale.

behavior which may not be available or observable. In such cases, a method presented in Codagnone et al. (2020) could provide promising results.

Chapter 6

How Bitcoin's Ups and Downs Are Changing the Way You Bet

Abstract

We investigate the relationship between Bitcoin price fluctuations and gambling behavior on the blockchain-based LuckyBit platform. Using transaction data spanning four years, we analyze how Bitcoin's exchange rate, price movements and volatility impact key gambling metrics, including bet size, gambling frequency, and risk appetite. Employing regression models and clustering techniques, we identify distinct behavioral responses among different gambler cohorts. Our findings reveal that higher Bitcoin price levels are associated with increased risk-taking but reduced gambling frequency, whereas higher weekly price volatility leads to a significant reduction in average bet sizes across all player groups. These results suggest that gamblers perceive Bitcoin's price movements as a psychological reference point, influencing their betting decisions in a manner similar to traditional financial decision-making. [1]

[1] This work has recently been accepted and published as Bakó and Sándor (2025) by Economics Letters.

6.1 Introduction

Imagine a casino where, instead of traditional chips, players use a special digital coin that displays its real-time monetary value in USD. As you hold this coin, its value fluctuates – sometimes increasing unexpectedly, making you feel wealthier and more inclined to place larger bets, and sometimes decreasing sharply, prompting caution or perhaps riskier bets in an effort to recover losses. This digital coin, with its constantly shifting value, creates a unique challenge for players as they weigh their decisions at the gambling table. This scenario encapsulates a key question: how does perceived value, especially when volatile, shape an individual's appetite for risk? This question has significant implications for understanding behavior in both gambling and financial decision-making, two areas that exhibit strikingly similar patterns and can often be analyzed using comparable metrics. The emergence of cryptocurrencies like Bitcoin, with their extreme price volatility, offers a unique lens through which to explore this dynamic.

In this study we investigate how Bitcoin's price fluctuations affect gambling behavior, focusing on three key dimensions: i) risk appetite, ii) bet size, and iii) gambling intensity. To address these questions, we analyze transaction data from the Bitcoin-based gambling platform LuckyBit, linking daily Bitcoin price movements to gambling behaviors across different cohorts of players. While prior studies have explored the effects of gambling activity on financial markets, such as Conlon and McGee (2020) analysis of LuckyBit, which found that betting activity explained 32% of BTC/USD exchange rate movements over six months, less attention has been paid to the reverse relationship: how Bitcoin's price volatility shapes gambling behavior.

Research suggests that volatility significantly affects risk-taking behavior. Typically, higher volatility leads to more cautious behavior, as people become more risk-averse in uncertain situations. In finance, for instance, increased volatility often pushes investors to favor safer assets (see Huber et al. (2022)), which aligns with the concept of loss aversion, where the fear of losses outweighs the potential for gains (see Kahneman and Tversky (1979); Barberis and Huang (2001); Schmidt and Zank (2005); or Bakó and Neszveda (2024)). However, volatility can also encourage risk-seeking behavior, especially among traders or gamblers who see opportunities to profit from price fluctuations (see Kuhle (2020)). This

dual nature of volatility, influencing both risk aversion and risk-seeking behavior, is further shaped by psychological factors like the 'house money effect' and the 'break-even effect' (see Richard and Eric (1990)), where past outcomes influence future risk-taking decisions. Given these insights, it would be expected that volatility would influence gamblers' behavior, particularly their risk appetite. Volatility creates an environment of uncertainty, which could cause gamblers to adjust their betting strategies based on the potential for gains or losses. However, our results indicate that while volatility does influence the size of bets placed, it does not significantly affect the level of risk taken. This suggests that volatility impacts how much gamblers are willing to stake, but not the riskiness of their bets themselves. Instead, volatility seems to affect the amount gamblers bet rather than the type of bet they choose.

Additionally, changes in the exchange rate between Bitcoin and USD could be expected to affect bet sizes. If gamblers are accounting for bets in USD, they might perceive their holdings as more valuable when Bitcoin increases in price, leading them to place smaller bets. However, our results show that exchange rate changes between Bitcoin and USD do not significantly affect bet sizes. Instead, they influence gamblers' risk appetite. This suggests that while Bitcoin's perceived value affects how much gamblers stake, it is their emotional response to risk, rather than the absolute amount they bet, that drives their behavior.

6.2 Data & Methods

As described by Conlon and McGee (2020), the LuckyBit gambling platform operates a probabilistic game that simulates a Galton board with 17 distinct outcomes.[2] This setup defines a payout function following a binomial distribution, where each possible outcome is associated with a specific multiplier. The platform offered four different game variations, each with its own set of multipliers. Table 6.1 summarizes the key characteristics of these games. While the expected returns (EP_i) are nearly identical, they exhibit significant differences in overall winning probabilities and maximum multipliers, ranging from 2 to

[2] A representative archive version of the website containing most graphical elements and the gambling rules is available at the web archives https://web.archive.org/web/20150314200358/http://luckyb.it/.

998.

Game (i)	$\max_i M_{i,j}$	$\min_j M_{i,j}$	$\sum_{M_{i,j}\leq 1} P_{i,j}$	EP_i	CV_i
blue	2	0.2	86.7%	98.27%	0.31
green	21	0.4	45.4%	98.25%	0.38
yellow	110	0.3	21.0%	98.24%	1.50
red	998	0.2	21.0%	98.24%	6.65

Table 6.1: Summary statistics of LuckyBit games. Column descriptions are detailed in Section 6.2.1. While expected returns are nearly identical, the coefficient of variation (CV_i) varies significantly, offering different levels of risk exposure.

To place a bet on the site, users would send a Bitcoin transaction to a static wallet address assigned to a specific game. The amount wagered was encoded in the transaction, and the outcome was determined by the transaction ID recorded on the blockchain, ensuring provable randomness. Once evaluated, the site returned a transaction to the initiating player. This publicly recorded process makes it possible to extract and analyze gambling behavior from blockchain data, as done by Conlon and McGee (2020) and Scholten et al. (2020).

6.2.1 LuckyBit data

To gather Bitcoin transaction records related to LuckyBit, we utilized the Bitcoin ledger dataset curated by Kondor et al. (2014), which has been extended through February 7, 2018.[3,4] Unlike in Conlon and McGee (2020), our dataset includes entity-level approxima-

[3]See the shared dataset `luckybit_names_address.csv` for the map of addresses.

[4]Our dataset concludes in February 2018 because in early 2018 the platform introduced a a registration-based, closed-account system that swiftly overtake its publicly visible "direct-betting" interface in popularity, due to decreased transaction costs and an increase in evaluation speed. This change effectively removed public access to wager information, leaving only a handful of observable bets during the transition and none thereafter, so systematic data collection beyond 2018 is not feasible.

tions[5] , allowing us to examine user-specific behavior rather than only individual bets.[6]

One limitation, however, of blockchain-based timing is that transaction timestamps correspond to the block creation time, which fluctuates between 10 and 20 minutes. Therefore, exact daily assignments, particularly around GMT midnight, may introduce minor inaccuracies. We use GMT for all timestamps, including exchange rate data.[7]

To capture key aspects of gambling behavior, we define three daily-level measures.

- *Gambling intensity:* the average number of bets per player per day:

$$\widetilde{N_t} = N_t/n_t, \tag{6.1}$$

  where N_t is the total number of bets on day t, and n_t represents the number of active players that day.

- *Bet size:* since bet amounts vary from 0.001 BTC to 50 BTC, we apply a logarithmic transformation:

$$\widetilde{\log_{10} B_t} = \sum_{i \in t} \log_{10} B_i / N_t. \tag{6.2}$$

- *Risk appetite (CV_t):* The coefficient of variation [8] for game i is calculated as:

$$CV_i = \frac{\sum_j P_{i,j} \left(M_{i,j} - EP_i\right)}{EP_i}, \tag{6.3}$$

  where $M_{i,j}$ is the payout multiplier, $P_{i,j}$ is the probability of outcome j, and $EP_i = \sum_j P_{i,j} M_{i,j}$ is the expected payout. The daily risk appetite for the population is given by:

$$\widetilde{CV_t} = \sum_{i \in t} CV_i / N_t. \tag{6.4}$$

[5] Although the original source of transaction data at www.blockchain.com is inevitably still the blockchain itself, the site turned its data sharing model to include transaction costs, thus making big data requests like ours, quite costly.

[6] The dataset used for the analysis can be publicly accessed at DOI: 10.5281/zenodo.14926295.

[7] The dataset does not include nationality or socioeconomic details about users, preserving anonymity and ensuring ethical use of the data.

[8] Compared to the standard statistical definition containing the main return or payout (that is multiplier times bet size), we have used the multipliers in the equation instead to keep the wager and variance risk dimensions separate in the analysis. A combined measure (that is weighting with bets sizes) brings almost identical overall conclusions to the study.

To control for days with very low participation, we exclude days with fewer than 10 bets, resulting in 1,486 days of observations between October 20, 2013, and February 7, 2018, covering 2,060,601 bets from 18,220 unique players. This period includes Bitcoin price fluctuations from a low of $171 to a high of $1280 per BTC, featuring extreme price jumps and volatility (see Figure 6.1).[9]

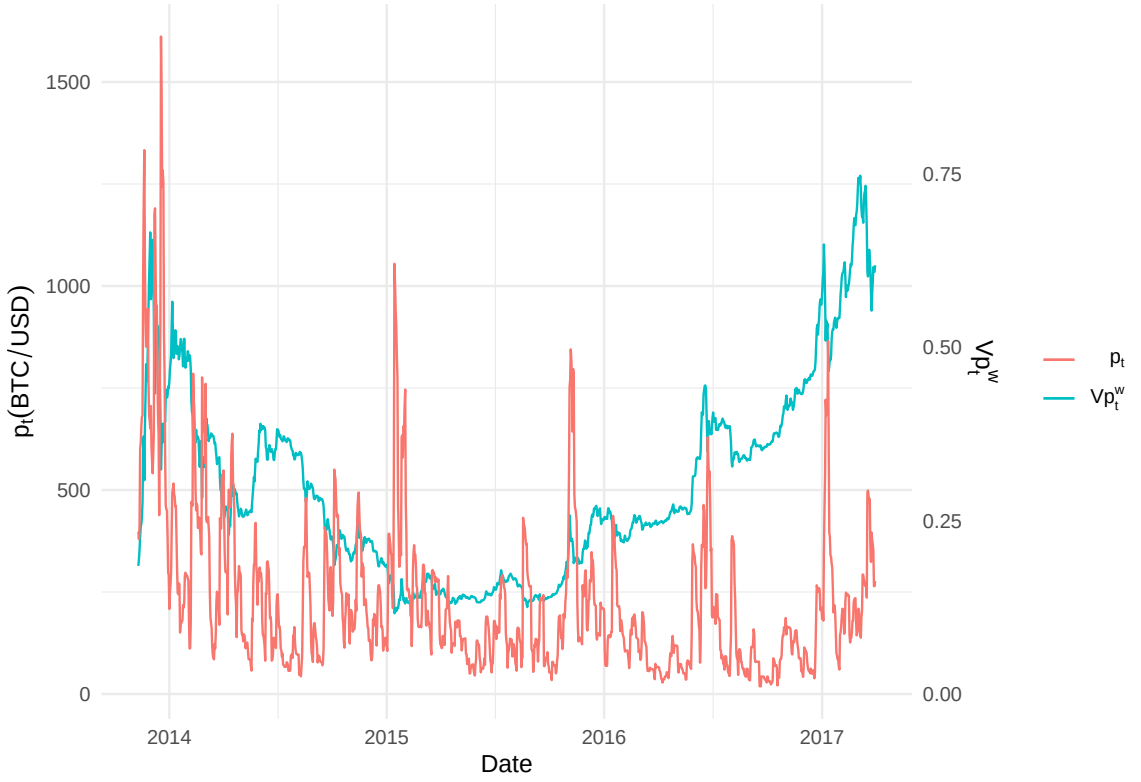

Figure 6.1: Bitcoin exchange price p_t and weekly volatility Vp_t^w over the observed period. We can see that over the 4 years of the discussed timeline both the price and the volatility had similar high and low values and clusters over time providing a great variability of effect combinations from these two aspects of the exchange rate. Data shown is gathered from the publicly available historic records of Coinmarketcap (see Section 6.3).

[9]Scripts used for data preparation and analysis are made publicly available at github.com/sampaat/luckybit_gambling.

6.2.2 Player cohorts

Using daily averages introduces potential bias by oversampling highly active players. Additionally, online gambling, particularly in the crypto space, may involve automated bots. To address these concerns, we apply clustering techniques to segment gamblers based on aggregated behavioral patterns.

As shown by Braverman and Shaffer (2012), k-means clustering can effectively distinguish casual players from problem gamblers. Furthermore, Sándor and Bakó (2024) demonstrate that trimmed k-means clustering is particularly useful for identifying players exhibiting extreme behavior, which may include automated bettors or bots.[10]

To segment players, we compute several behavioral metrics: the total number of bets placed (N_{total}), the number of active calendar days (D), the coefficient of variation of their bet portfolio weighted by bets placed ($CV_p = \sum_i CV_i \cdot B_{i,p} / \sum_i B_{i,p}$), and both the total ($log_{10}\left(\sum_i B_{i,p}\right)$) and maximum ($\max\left[log_{10}\left(B_p\right)\right]$) wager placed, measured on a logarithmic scale. A logarithmic transformation is applied to wagers beforehand for normalization.

For clustering, we set $k = 2$ cohorts with a 1% trimming limit. As shown in Table 6.2, the primary distinguishing factors are the total number of bets placed and the average bet size. While some overlap exists, other dimensions also offer insights into group characteristics. For instance, more committed players tend to place larger but less risky bets compared to those who gamble only occasionally, whom we label as casual players. The trimmed cohort consists of clear outliers, either pathological gamblers or bots executing a specific strategy.

6.3 Influence of the price of BitCoin

We use the USD/BTC exchange rate to capture the risk associated with Bitcoin's value.[11] To further break down the risk within the exchange rate, we derive the following price

[10] Our classification of the trimmed group as potentially containing bots is based on the assumption that bots place a high number of bets over either a short or extended period. As a result, they are separated from the rest of the player base based on the total number of bets or the days played descriptor used in the clustering process.

[11] We used the publicly available historic data of Coinmarketcap at https://coinmarketcap.com/currencies/bitcoin/historical-data/

Cluster	$N_{players}$	N_{total}	D	CV_p	$\max[log_{10}(B_p)]$	$log_{10}(\sum_i B_{i,p})$
All players	18220	4(21)	1(1)	1.47(4.57)	-2.43(-7.28)	-2(-1.38)
Casual	11329	2(6)	1(0)	1.5(6.27)	-2.67(-7.44)	-2.52(-2.27)
Committed	6709	24(99)	2(3)	1.03(2.32)	-1.85(-7.03)	-1.16(-0.91)
Extreme	182	1432(4209)	63(85)	1.68(2.96)	-2.47(-7.15)	-0.89(0)

Table 6.2: Corhort sizes ($N_{players}$) and median (IQR) descriptors of the whole player base and the three cohorts separated by the clustering (Casual and Committed) and trimming (Extreme). The bet values are represented on a logarithmic base of 10 in BTC.

markers:

- Daily mid rate: p_t
- Relative daily return: $\delta p_t = \frac{p_t}{p_{t-1}}$
- A proxy for the relative intra-day price volatility: $Vp_t = \frac{p_t^{high} - p_t^{low}}{p_t}$

To account for the fact that while many Bitcoin holders are active investors, others may react to price information more slowly, we also included:

- Weekly return: $\delta p_t^w = \frac{p_t}{p_{t-7}}$
- Weekly volatility: $Vp_t^w = \frac{\max_{t\ldots t-7} p_\tau^{high} - \min_{t\ldots t-7} p_\tau^{low}}{p_t}$.

Before estimating the regressions, all dependent and explanatory variables were standardized to mean zero and unit variance, which makes coefficients directly comparable across specifications. To assess collinearity among the five price-related regressors, we conducted a principal component analysis (PCA): the first component explains only 36% of the total variance, and three components are required to exceed 90%. This pattern indicates overlap in information but not a single dominant (i.e., highly collinear) dimension. Because pure LASSO can drop one variable from a correlated set, we employed Elastic Net regularization to retain its selection ability while adding ridge-type shrinkage.[12] In applications with

[12]Elastic Net, introduced by Zou and Hastie (2005), combines the variable-selection power of LASSO with the grouping effect of ridge regression, promoting both sparsity and the joint retention of correlated predictors.

strong collinearity, a balanced mix of the two penalties (e.g., $\alpha = 0.5$) is often used; in our case collinearity is modest, so we set $\alpha = 0.7$ to preserve some grouping while favoring sparsity. The penalty parameter λ was chosen by ten-fold cross-validation, and predictors with non-zero coefficients were then entered into the ordinary least squares models reported in Table 6.3.[13]

Behavior	Cluster	p_t	δp_t	Vp_t	δp_t^w	Vp_t^w	$R^2_{adj.}$
$\widetilde{log_{10}B_t}$	All players			-0.07*		-0.19***	5.6%
	Casual	0.08*	0.01	-0.03**	0.06**	-0.28***	9.2%
	Committed				0.06**	-0.2***	4.1%
	Extreme			-0.04**		-0.13***	2.7%
$\widetilde{N_t}$	All players	-0.39***			-0.06**		16.3%
	Casual	-0.12***	-0.06*	-0.04	0.06*	0.17***	2.5%
	Committed	-0.24***			-0.05*	-0.04	6.6%
	Extreme	-0.1***	-0.06*	0.11***	-0.05*	0.18***	8.7%
$\widetilde{CV_t}$	All players	0.29***	0.02	-0.04	-0.04	-0.05	7.6%
	Casual	0.32***	0.07**		-0.1***	-0.06**	9.6%
	Committed	0.34***			-0.03	-0.12***	11.0%
	Extreme	0.21***		-0.05		-0.05	4.5%

Table 6.3: OLS regression coefficients and adjusted R^2 of the models organized by the targeted population behavior and the clusters created. Only those coefficients are shown with values that were pre-selected using Elastic Net regression (P-levels: *.1;** .05;*** .001).

The results indicate that price factors can explain a small but significant portion of betting behavior. Notably, the mean daily bet size is negatively affected by weekly volatility across all cohorts. A 10% increase in weekly volatility corresponds to a 1.9% decrease in average bet size, suggesting that gamblers adopt a more cautious approach during periods of high volatility. This effect is particularly strong for casual players, who also tend to place larger bets when Bitcoin prices are higher. For daily betting frequency, there is a general negative correlation with price levels, with the strongest effect observed among committed

[13]Using $\alpha = 0.5$ retained a few additional regressors, but none was statistically significant nor did they improve model fit, and the resulting selection was identical to that obtained with pure LASSO ($\alpha = 1$).

players. However, weekly volatility appears to have a positive impact on betting frequency for both casual and extreme players. When examining risk-taking behavior, we find that higher price levels consistently correspond to greater risk propensity across all cohorts. Some counteracting effects emerge from weekly returns and volatility, though these influences are neither as strong nor as consistent across different player groups.

6.4 Conclusions

Our results demonstrate that the USD/BTC exchange rate - given all other factors unchanged - plays a partial but meaningful role in shaping the behavior of Bitcoin gamblers, with consistent effects observed across different player segments, including casual users, highly addicted individuals, and potentially automated betting programs. This highlights the influence of cryptocurrency market dynamics on gambling patterns, reinforcing the idea that digital asset volatility can significantly impact financial decision-making in high-risk environments.

One of the key findings is that higher Bitcoin price levels tend to correlate with an increase in risk propensity while simultaneous decrease in betting frequency. This suggests that when Bitcoin prices rise, players may feel more confident, leading them to take greater risks with individual wagers. However, the reduction in betting frequency indicates a more selective approach, where players place fewer but potentially larger bets. These effects are more pronounced among committed gamblers than casual ones, implying that engagement level plays a crucial role in shaping how individuals respond to price fluctuations.

Regarding the impact of directional price changes and volatility, our findings suggest that gamblers respond to price movements on a weekly rather than daily basis. This indicates a longer memory in their behavior, meaning that players do not react instantaneously to market changes but instead adjust their gambling patterns based on sustained trends. Notably, higher weekly price volatility coincides with reduced wager sizes across all cohorts, with the strongest effect observed among casual players. This suggests that uncertainty in Bitcoin's value leads to more cautious betting behavior, particularly among those who are less engaged in gambling.

These changes in behavior can also be interpreted through the lens of perceived wealth by

viewing USD/BTC price movements as proxies for wealth shocks. When prices rise, players may feel wealthier and become more willing to take risks, consistent with the house money effect, where recent gains encourage bolder decisions (Richard and Eric, 1990). Conversely, falling prices create a sense of financial loss. While this does not appear to reduce players' overall risk appetite, it is associated with a clear reduction in wager size, particularly under conditions of high volatility. This may reflect the break-even effect (Richard and Eric, 1990), where individuals continue to pursue risky outcomes to recover earlier losses while limiting the size of their financial exposure.

Our results suggest that perceived changes in wealth affect different dimensions of gambling behavior. Risk appetite is more responsive to the level of Bitcoin prices than to volatility or recent declines, indicating that perceived gains influence the types of bets players choose. Bet size, on the other hand, moves more strongly alongside price uncertainty, especially during downward trends, suggesting that players reduce their stakes when they feel financially less secure, even if their risk preferences remain unchanged. Differences across player cohorts support this interpretation. Casual players show the most pronounced reduction in bet size during volatile or declining periods, suggesting greater sensitivity to perceived losses. Committed players, by contrast, respond more to rising prices, consistent with a stronger house money effect and a greater responsiveness to perceived gains. Gamblers classified as Extreme show a more erratic response, likely reflecting a mix of behavioral extremes, including automated betting or persistent loss-chasing. While they, too, reduce bet sizes in volatile conditions, their risk-taking appears less systematically tied to perceived wealth, suggesting that other factors, such as algorithmic strategies or compulsive tendencies, may dominate. Framing these results in terms of financial perception helps explain not only the average effects of market conditions on behavior, but also the asymmetries across different types of players.

Interpreting the results, we should not forget the limitation of the applied methodology and treat the relationships as associations. We cannot rule out the case of endogeneity by factors like the Bitcoin's popularity or the state of the world economy.There could also be omitted variables influencing both terms, like external shock events, of which the crypto-markets have many of.

These findings still have several important implications. First, they highlight the broader

psychological and financial risks associated with cryptocurrency-based gambling. Unlike traditional gambling environments where currency values remain stable, crypto-gamblers must navigate not only their own risk preferences but also external financial volatility, which may amplify risky decision-making or trigger more conservative betting patterns depending on the context. The stronger behavioral shifts observed among casual and committed players suggest that different types of gamblers are uniquely vulnerable to market-induced pressures, warranting a differentiated approach in regulatory and harm reduction strategies.

Furthermore, our study underscores the growing overlap between online gambling, speculative trading, and digital asset markets as also discussed by Delfabbro et al. (2021). As the boundaries between these activities become less distinct, understanding how market trends influence betting behavior is increasingly important. The fact that gamblers appear to react more to weekly trends than daily fluctuations suggests that exposure to cryptocurrency markets may shape their decision-making processes. This raises concerns about the potential for crypto volatility to exacerbate gambling addiction or financial losses, particularly among players who lack the experience or knowledge to navigate these risks effectively.

In the context of a global online gambling surge and heightened economic uncertainty, our findings emphasize the urgent need for regulatory interventions. Policymakers must consider the role of digital currencies in facilitating new forms of gambling, particularly those that appeal to younger or more vulnerable populations.[14]

Stronger consumer protections, responsible gambling initiatives, and public awareness campaigns could help mitigate the risks posed by the intersection of digital finance and gambling. As cryptocurrencies and blockchain-based betting platforms continue to evolve, understanding their psychological and economic effects on gamblers will be essential for developing effective policy responses and harm reduction strategies.

[14]This is especially relevant given the increasing use of digital currencies to enable gambling-like reward systems in video games targeted at teenagers and young adults, as discussed by Kim et al. (2023).

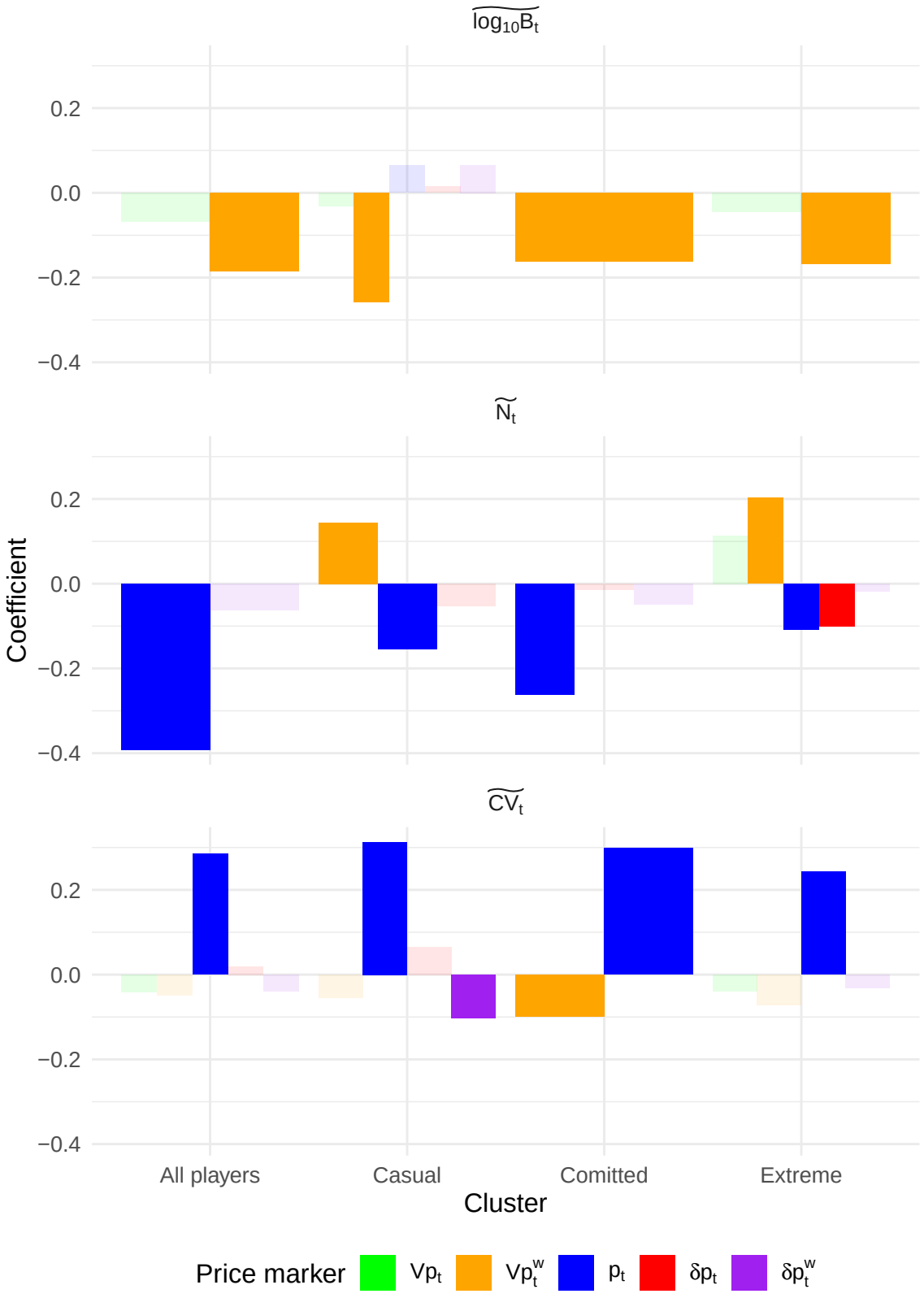

Figure 6.2: Visual comparison of OLS regression coefficients. Only those coefficients are shown with values that were pre-selected through LASSO regression, those with P-value below .1% are dimmed.

Chapter 7

Conclusion

This dissertation presents three works of research, that contributes to the understanding of gambling behavior through a multidisciplinary, data-intensive approach, integrating methodologies from behavioral economics and machine learning with some influence from finance. I have critically examined published behavioral models, introduced novel methodological frameworks, and provided empirical insights into the mechanisms underlying gambling decisions.

The three research projects are interconnected by their focus on understanding gambling behavior, yet they approach the problem from different angles. In the first study, we have provided a critical re-assessment of the hot hand fallacy, as presented by Xu and Harvey (2014), demonstrating that their prior findings on streak-dependent betting behavior may be artifacts of methodological biases rather than genuine cognitive distortions. In the second study, we have discussed the detection of problem gambling, by proposing a machine learning framework that classifies gambling behaviors without relying on self-reported measures. In this way, combining methodologies from different branches of earlier work, we offered a more objective alternative to approach this issue in a generalized and reproducible manner. The third study explores how the price volatility of Bitcoin affects gambling decisions, linking financial market conditions to shifts in betting behavior.

The key connection between these studies is that they all rely on newly collected datasets that are produced from publicly available data in a reproducible manner and also made available publicly in their organized and collected form, alongside the computer codes required to produce the published results. In addition, the papers demonstrate that gambling behavior

cannot be understood solely through psychological heuristics or economic rationality, but must be examined through the interplay of market dynamics and personal decision making. This integrated perspective advances the literature by providing an open, reproducible, data-driven approach to behavioral gambling research.

From a policy perspective, our findings have implications for market regulation, pointing to the fact that facilitators could and should play a larger role in responsible gambling initiatives. The machine learning framework developed for detecting early signs of problem gambling behavior should be used by regulators to enforce the application of early warning systems, reducing harm without overly relying on self-exclusion mechanisms.

While this research provides significant insights, it also opens the way for some new research directions and leaves some questions open. While we show, that a certain approach in identifying the hot hand fallacy is inadequate, a deeper analysis on how streaks could influence the overall behavior of gamblers - when properly controlled for other factors - could be systematically assessed using our data. Extending the classification study, the effectiveness of adding personalized interventions could be tested. Further studies could also refine the implementation and perhaps a case study on how to implement and design a regulatory framework around it. Finally, to expand the analysis of financial market influences beyond Bitcoin and integrate other external factor's influences over risk taking in gambling and the overlap with investment behavior.

Throughout the years of my PhD thesis, I have presented my works in university forums as well as international conferences in forms of posters and presentations as well as some revisions and rejections with valuable insights - now embedded into the works - from the quality journals of the research area. While Bakó and Sándor (2021) is currently under revision at a target journals of reproduction studies, Sándor and Bakó (2024) has been accepted and published by the Journal of Gambling Studies as well as Bakó and Sándor (2025) by Economics Letters.

In conclusion, this dissertation underscores the value of combining behavioral economics, machine learning and financial modeling to study gambling behavior. Through integrating these perspectives, this research improves our understanding of gambling behavior and offers practical tools and nudges policymakers and regulators to start using this toolkit. Hopefully, this work can help us to understand the rapidly evolving digital landscape of gambling and

to avoid the horrid fallout of this ancient leisure activity.

Bibliography

American Psychiatric Association, D., American Psychiatric Association, D., et al. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

Arumawadu, H. I., Rathnayaka, R., and Illangarathne, S. (2015). Mining profitability of telecommunication custormers using k-means clustering. *Scientific Research Publishing*.

Auer, M. and Griffiths, M. D. (2020). The use of personalized messages on wagering behavior of swedish online gamblers: An empirical study. *Computers in Human Behavior*, 110:106402.

Auer, M. and Griffiths, M. D. (2024). An empirical attempt to identify binge gambling utilizing account-based player tracking data. *Addiction Research & Theory*, 32(4):264–273.

Auer, M., Hopfgartner, N., and Griffiths, M. D. (2018). The effect of loss-limit reminders on gambling behavior: A real-world study of norwegian gamblers. *Journal of Behavioral Addictions*, 7(4):1056–1067.

Bacidore, J. M., Boquist, J. A., Milbourn, T. T., and Thakor, A. V. (1997). The search for the best financial performance measure. *Financial Analysts Journal*, 53(3):11–20.

Badev, A. I. and Chen, M. (2014). Bitcoin: Technical background and data analysis. *FEDS working paper*.

Bakó, B. and Neszveda, G. (2024). An aspirational perspective on the negative risk-return relationship. *Finance Research Letters*, 61:104977.

Bakó, B. and Sándor, M. C. (2021). Approaching the hot hand with a cool head. *Available at SSRN 3952051*.

Bakó, B. and Sándor, M. C. (2025). How bitcoin's ups and downs are changing the way you bet. *Economics Letters*, 255:112564.

Barberis, N. (2012). A model of casino gambling. *Management Science*, 58(1):35–51.

Barberis, N. and Huang, M. (2001). Mental accounting, loss aversion, and individual stock returns. *The Journal of Finance*, 56(4):1247–1292.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–96.

Baumöhl, E. and Výrostová, E. (2017). Do people gamble more in good times? evidence from 27 european countries. *Applied Economics Letters*, 24(18):1311–1314.

Behe, R. (2024). The industry looks back on 2024, and forward to 2025. *CDC Gaming Reports*. Accessed: 2025-02-24.

BetMGM, A. T. (2023). Machine learning in betmgm: Personalizing the player experience for enhanced retention. Accessed online 2025-05-16.

Bijker, R., Booth, N., Merkouris, S. S., Dowling, N. A., and Rodda, S. N. (2022). Global prevalence of help-seeking for problem gambling: A systematic review and meta-analysis. *Addiction*, 117(12):2972–2985.

Bogliacino, F., Pejsachowicz, L., Giovanni, L., and Francisco, L.-V. (2023). Testing for manipulation: Experimental evidence on dark patterns. *Available at SocArXiv sqt3j*.

Braverman, J. and Shaffer, H. J. (2012). How do gamblers start gambling: Identifying behavioural markers for high-risk internet gambling. *The European Journal of Public Health*, 22(2):273–278.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Brodeur, M. (2019). Public health and gambling: The potential of nudge policies. In *Harm Reduction for Gambling*, pages 112–119. Routledge.

Buttigieg, K. D., Caruana, M. A., and Suda, D. (2022). Identifying problematic gamblers using multiclass and two-stage binary neural network approaches.

Caillon, J., Grall-Bronnec, M., Perrot, B., Leboucher, J., Donnio, Y., Romo, L., and Challet-Bouju, G. (2019). Effectiveness of at-risk gamblers' temporary self-exclusion from internet gambling sites. *Journal of gambling studies*, 35(2):601–615.

Candel, A., Parmar, V., LeDell, E., and Arora, A. (2016). Deep learning with h2o. *H2O. ai Inc*, pages 1–21.

Carran, M. (2022). Monitoring gambling engagement and problem gambling prevalence within selected european jurisdictions. Technical report, City Law School, City, University of London, London, UK.

Chagas, B. T. and Gomes, J. F. (2017). Internet gambling: A critical review of behavioural tracking research. *Journal of Gambling Issues*, 36.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Codagnone, C., Bogliacino, F., Gómez, C., Charris, R., Montealegre, F., Liva, G., Lupiáñez-Villanueva, F., Folkvord, F., and Veltri, G. A. (2020). Assessing concerns for the economic consequence of the covid-19 response and mental health problems associated with economic vulnerability and negative economic shock in italy, spain, and the united kingdom. *PloS one*, 15(10):e0240876.

Conlisk, J. (1993). The utility of gambling. *Journal of risk and uncertainty*, 6(3):255–275.

Conlon, T. and McGee, R. J. (2020). Betting on bitcoin: Does gambling volume on the blockchain explain bitcoin price changes? *Economics Letters*, 191:108727.

Coussement, K. and De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, 66(9):1629–1636.

Croson, R. and Sundali, J. (2005). The gambler's fallacy and the hot hand: Empirical data from casinos. *Journal of risk and uncertainty*, 30(3):195–209.

Cuesta-Albertos, J. A., Gordaliza, A., and Matrán, C. (1997). Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576.

De Laplace, M. (1995). *A philosophical essay on probabilities*. Courier Corporation.

Delfabbro, P., King, D., Williams, J., and Georgiou, N. (2021). Cryptocurrency trading, gambling and problem gambling. *Addictive Behaviors*, 122:107021.

Demaree, H. A., Weaver, J. S., and Juergensen, J. (2015). A fallacious "gambler's fallacy"? commentary on. *Cognition*, 139:168–170.

Deng, X., Lesch, T., and Clark, L. (2019). Applying data science to behavioral analysis of online gambling. *Current Addiction Reports*, 6(3):159–164.

Eakins, J. (2016). Household gambling expenditures and the irish recession. *International Gambling Studies*, 16(2):211–230.

European Gaming Industry News (2023). Hungary regulator implements new gambling act amendments. `https://europeangaming.eu/portal/compliance-updates/2023/04/04/133132/hungary-regulator-implements-new-gambling-act-amendments/`.

Finkenwirth, S., MacDonald, K., Deng, X., Lesch, T., and Clark, L. (2021). Using machine learning to predict self-exclusion status in online gamblers on the playnow. com platform in british columbia. *International Gambling Studies*, 21(2):220–237.

Folkvord, F., Codagnone, C., Bogliacino, F., Veltri, G., Lupiánez-Villanueva, F., Ivchenko, A., and Gaskell, G. (2019). Experimental evidence on measures to protect consumers of online gambling services. *Journal of Behavioral Economics for Policy*, 3(1):20–29.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Friedman, M. and Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of political Economy*, 56(4):279–304.

Friedrich, S., Groll, A., Ickstadt, K., Kneib, T., Pauly, M., Rahnenführer, J., and Friede, T. (2023). Regularization approaches in clinical biostatistics: A review of methods and their applications. *Statistical Methods in Medical Research*, 32(2):425–440.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63:3–42.

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive psychology*, 17(3):295–314.

Giroux, I., Goulet, A., Mercier, J., Jacques, C., and Bouchard, S. (2017). Online and mobile interventions for problem gambling, alcohol, and drugs: A systematic review. *Frontiers in Psychology*, 8:954.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). The elements of statistical learning: Data mining, inference, and prediction.

Hennig, C. (2020). *trimcluster: Cluster analysis with trimming*. R package version 0.1-5.

Hodgins, D. C. and Stevens, R. M. (2021). The impact of covid-19 on gambling and gambling disorder: Emerging data. *Current opinion in psychiatry*, 34(4):332.

Hofmarcher, T., Romild, U., Spångberg, J., Persson, U., and Håkansson, A. (2020). The societal costs of problem gambling in sweden. *BMC public health*, 20(1):1–14.

Horváth, C., Günther, A., and Paap, R. (2010). Seasonal patterns in slot-machine gambling in germany. *International Gambling Studies*, 10(3):255–268.

Horváth, C. and Paap, R. (2012). The effect of recessions on gambling expenditures. *Journal of Gambling Studies*, 28:703–717.

Huber, C., Huber, J., and Kirchler, M. (2022). Volatility shocks and investment behavior. *Journal of Economic Behavior and Organization*, 194:55–70.

Ji, Q., Quan, X., Yin, H., and Yuan, Q. (2021). Gambling preferences and stock price crash risk: Evidence from china. *Journal of Banking & Finance*, 128:106158.

Johnson, J., Tellis, G. J., and MacInnis, D. J. (2005). Losers, winners, and biased trades. *Journal of Consumer Research*, 32(2):324–329.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Kim, H. S., Leslie, R. D., Stewart, S. H., King, D. L., Demetrovics, Z., Andrade, A. L. M., Choi, J.-S., Tavares, H., Almeida, B., and Hodgins, D. C. (2023). A scoping review of the association between loot boxes, esports, skin betting, and token wagering with gambling and video gaming behaviors. *Journal of Behavioral Addictions*, 12(2):309–351.

Kondor, D., Pósfai, M., Csabai, I., and Vattay, G. (2014). Do the rich get richer? an empirical analysis of the bitcoin transaction network. *PloS one*, 9(2):e86197.

Kotter, R., Kräplin, A., and Bühringer, G. (2018). Casino self-and forced excluders' gambling behavior before and after exclusion. *Journal of Gambling Studies*, 34(2):597–615.

Krčál, O., Kvasnička, M., and Staněk, R. (2016). External validity of prospect theory: The evidence from soccer betting. *Journal of Behavioral and Experimental Economics*, 65:121–127.

Kuhle, W. (2020). Thought viruses and asset prices. *Journal of Behavioral Finance*, 23(2):123–131.

LaBrie, R. A., LaPlante, D. A., Nelson, S. E., Schumann, A., and Shaffer, H. J. (2007). Assessing the playing field: A prospective longitudinal study of internet sports gambling behavior. *Journal of Gambling studies*, 23(3):347–362.

LaPlante, D. A. (2019). Replication is fundamental, but is it common? a call for scientific self-reflection and contemporary research practices in gambling-related research. *International Gambling Studies*, 19(3):362–368.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., and Malohlava, M. (2022). *h2o: R interface for the 'H2O' scalable machine learning platform*. R package version 3.36.0.4.

LeDell, E. and Poirier, S. (2020). H2O AutoML: scalable automatic machine learning. *7th ICML Workshop on Automated Machine Learning (AutoML)*.

LeDell, E. E. (2015). *Scalable ensemble learning and computationally efficient variance estimation*. University of California, Berkeley.

Lindner, P., Ramnerö, J., Ivanova, E., and Carlbring, P. (2021). Studying gambling behaviors and responsible gambling tools in a simulated online casino integrated with amazon mechanical turk: Development and initial validation of survey data and platform mechanics of the frescati online research casino. *Frontiers in Psychiatry*, 11:571954.

Luce, R. D., Ng, C., Marley, A., and Aczél, J. (2008). Utility of gambling ii: Risk, paradoxes, and data. *Economic Theory*, 36(2):165–187.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Mak, K. K., Lee, K., and Park, C. (2019). Applications of machine learning in addiction studies: A systematic review. *Psychiatry research*, 275:53–60.

Miller, J. B. and Sanjurjo, A. (2018). Surprised by the hot hand fallacy? a truth in the law of small numbers. *Econometrica*, 86(6):2019–2047.

Mills, D. J. and Nower, L. (2019). Preliminary findings on cryptocurrency trading among regular gamblers: A new risk for problem gambling? *Addictive behaviors*, 92:136–140.

Murphy, R. O. and ten Brincke, R. H. (2018). Hierarchical maximum likelihood parameter estimation for cumulative prospect theory: Improving the reliability of individual risk parameter estimates. *Management Science*, 64(1):308–326.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, page 21260.

National Council on Problem Gambling (2021). National survey on gambling attitudes and gambling experiences 2.0. Technical report, National Council on Problem Gambling.

Newall, P., Walasek, L., Ludvig, E., and Rockloff, M. (2020). Nudge versus sludge in gambling warning labels.

Newall, P. W. (2019). Dark nudges in gambling. *Addiction Research & Theory*, 27(2):65–67.

Oskarsson, A. T., Van Boven, L., McClelland, G. H., and Hastie, R. (2009). What's next? judging sequences of binary events. *Psychological bulletin*, 135(2):262.

Oyewole, G. J. and Thopil, G. A. (2023). Data clustering: application and trends. *Artificial Intelligence Review*, 56(7):6439–6475.

Pang, B., Nijkamp, E., and Wu, Y. N. (2020). Deep learning with tensorflow: A review. *Journal of Educational and Behavioral Statistics*, 45(2):227–248.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pearson, M. R., Schwebel, F. J., Richards, D. K., and Witkiewitz, K. (2022). Examining replicability in addictions research: How to assess and ways forward. *Psychology of Addictive Behaviors*, 36(3):260.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Percy, C., França, M., Dragičević, S., and d'Avila Garcez, A. (2016). Predicting online gambling self-exclusion: an analysis of the performance of supervised machine learning models. *International Gambling Studies*, 16(2):193–210.

Peres, F., Fallacara, E., Manzoni, L., Castelli, M., Popovič, A., Rodrigues, M., and Estevens, P. (2021). Time series clustering of online gambling activities for addicted users' detection. *Applied Sciences*, 11(5):2397.

Potenza, M. N., Wareham, J. D., Steinberg, M. A., Rugle, L., Cavallo, D. A., Krishnan-Sarin, S., and Desai, R. A. (2011). Correlates of at-risk/problem internet gambling in adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 50(2):150–159.

R Core Team (2013). *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Richard, T. and Eric, J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Science*, 36(6):643–660.

Rockloff, M. J. and Dyer, V. (2006). The four es of problem gambling: A psychological measure of risk. *Journal of Gambling Studies*, 22:101–120.

Salaghe, F., Sundali, J., Nichols, M. W., and Guerrero, F. (2020). An empirical investigation of wagering behavior in a large sample of slot machine gamblers. *Journal of Economic Behavior & Organization*, 169:369–388.

Sándor, M. C. and Bakó, B. (2024). Unmasking risky habits: Identifying and predicting problem gamblers through machine learning techniques. *Journal of Gambling Studies*, 40:1367–1377.

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference: festschrift in honor of vladimir N. Vapnik*, pages 37–52. Springer.

Schildberg-Hörisch, H. (2018). Are risk preferences stable? *Journal of Economic Perspectives*, 32(2):135–54.

Schmidt, U. and Zank, H. (2005). What is loss aversion? *Journal of Risk and Uncertainty*, 30(2):157–167.

Scholten, O. J., Hughes, N. G. J., Deterding, S., Drachen, A., Walker, J. A., and Zendle, D. (2019). Ethereum crypto-games: Mechanics, prevalence, and gambling similarities. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 379–389.

Scholten, O. J., Zendle, D., and Walker, J. A. (2020). Inside the decentralised casino: A longitudinal study of actual cryptocurrency gambling transactions. *PloS one*, 15(10):e0240693.

Solon, O. and Zuidijk, D. (2024). UK gambling operators face £100 million tax in harm reduction push. `https://www.bloomberg.com/news/articles/2024-11-27/uk-gambling-firms-face-100-million-tax-in-harm-reduction-push`.

Stegmann, Y., Schiele, M. A., Schümann, D., Lonsdorf, T. B., Zwanzger, P., Romanos, M., Reif, A., Domschke, K., Deckert, J., Gamer, M., et al. (2019). Individual differences in human fear generalization—pattern identification and implications for anxiety disorders. *Translational psychiatry*, 9(1):1–11.

Sándor, M. C. (2021). Satoshidice [dataset]. *Zenodo*, (5600259).

Sándor, M. C. (2025). Luckybit bets [dataset]. *Zenodo*, (14926295).

Team, S. R. et al. (2024). Social and economic impacts of casino introduction to massachusetts. *Amherst, MA: School of Public Health and Health Sciences, University of Massachusetts Amherst*.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323.

Ukhov, I., Bjurgert, J., Auer, M., and Griffiths, M. D. (2021). Online problem gambling: A comparison of casino players and sports bettors via predictive modeling using behavioral tracking data. *Journal of Gambling Studies*, 37(3):877–897.

Vali, I. (2024). Hungary – for a piece of the pie – yield sec. `https://g3newswire.com/hungary-for-a-piece-of-the-pie-yield-sec/`.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).

Wardle, H., Donnachie, C., Critchlow, N., Brown, A., Bunn, C., Dobbie, F., Gray, C., Mitchell, D., Purves, R., Reith, G., et al. (2021). The impact of the initial covid-19 lockdown upon regular sports bettors in britain: Findings from a cross-sectional online study. *Addictive Behaviors*, 118:106876.

Wardrop, R. L. (1995). Simpson's paradox and the hot hand in basketball. *The American Statistician*, 49(1):24–28.

World Health Organization (2024). Gambling. `https://www.who.int/news-room/fact-sheets/detail/gambling`.

Xu, J. and Harvey, N. (2014). Carry on winning: The gamblers' fallacy creates hot hand effects in online gambling. *Cognition*, 131(2):173–180.

Xu, J. and Harvey, N. (2015). Carry on winning: No selection effect. *Cognition*, 139:171–173.

Xuan, Z. and Shaffer, H. (2009). How do gamblers end gambling: Longitudinal analysis of internet gambling behaviors prior to account closure due to gambling related problems. *Journal of Gambling Studies*, 25(2):239–252.

Yaari, G. and Eisenmann, S. (2011). The hot (invisible?) hand: can time sequence patterns of success/failure in sports be modeled as repeated random independent trials? *PloS one*, 6(10):e24532.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.