How Are Returns Distributed in Prediction Markets?

Identifying Traits and Strategies of Consistently Successful Predictors

J. Vallejo J. Kollenberg
London School of Economics and Political Science (LSE)

Abstract

In the last three years, prediction markets have gone from fringe betting platforms to being routinely cited by major news outlets as real-time indicators of event likelihood. However, little is known about who actually profits on these platforms or which behaviours distinguish successful traders from those who lose money. Using transaction data for Polymarket users, the largest prediction market, we attempt to answer these questions.

We find that returns are extremely unequal: the distribution is heavy-tailed, with a small fraction of users capturing outsized gains while nearly two-thirds of participants lose money. To identify the causes of positive performance, we evaluate linear regression alongside three classification approaches: logistic regression, k-nearest neighbours, and random forests. All three models suggest that risk management is the dominant predictor of success. Profitable users diversify across markets and are disciplined in their exit strategies, whereas the majority of traders hold concentrated positions that ultimately decrease in value.

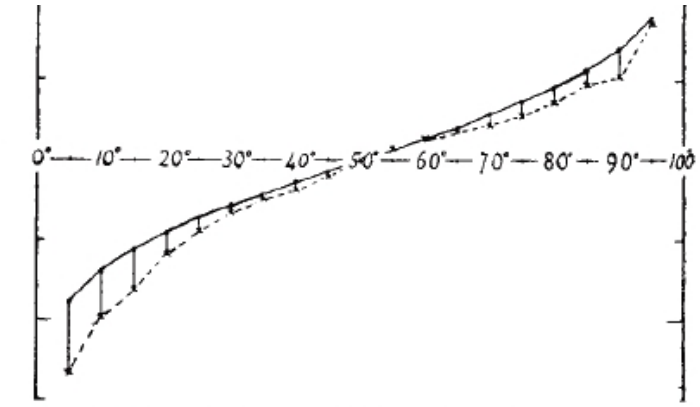
These findings highlight the effectiveness of different strategies and carry practical implications for traders and regulators.

Contents

1 Introduction

1.1 Background and Motivation

Francis Galton was a Victorian polymath known for contribution to the field of statistics, his inventions, and his deep skepticism of democracy. In 1906, he found himself at a livestock fair where more than 800 spectators were attempting to guess the weight of an ox. Eager to demonstrate what he called, "ignorance of the masses", he recorded the guesses and produced summary statistics. However, his results were surprising. Averaging the guesses, the crowd was off by less than 0.1%. Galton wrote a paper, published in 1907, titled The Voice of the People describing the "Power of Aggregated Knowledge".



The continuous line is the normal curve with p.e.=37.
The broken line is drawn from the observations.
The lines connecting them show the differences between the observed and the normal.

Figure 1: Galton's 1907 analysis. The broken line represents the cumulative distribution of observed guesses, the solid line represents the expected normal distribution.

It is difficult to quantify the vast impact this idea has had over the last century, but many designate Prediction Markets as a direct consequence. A prediction market is a platform that allows users to buy and sell contracts based on the outcome of future events - for example, whether a political candidate will win an election. Each contract typically pays $1 if the event occurs and $0 otherwise. As a result, the price of the contract reflects the probability estimate of that event happening. [1]. Thus, if a contract trades at $0.68, the market believes there is a 68% chance the event will occur.

[1] Strictly speaking, the price of a prediction market contract does not always equal its implied probability, as factors such as liquidity, bid–ask spreads, and certain market designs can cause deviations. In liquid binary markets, however, price is commonly used as a close approximation.

Many use these platforms as sources of information rather than speculation. As such, some argue against treating and regulating these platforms as "betting" houses, but instead argue they should be treated as news providers. They argue that people are not vulnerable, as all guesses are simply unbiased approximations (Wolfers and Zitzewitz (2006)). Prediction markets have shown forecasting capabilities across various domains, (Berg et al. (2008)) but their efficiency remains an empirical question. Return analysis offers a practical window into these markets' functioning. Economic theory suggests efficient markets should primarily compensate for risk, with persistent extraordinary profits being rare (Brown et al. (2019)).Patterns of extreme returns or consistently successful strategies may indicate inefficiencies, information asymmetries, or manipulation. Our research on return distributions and successful participant characteristics contributes to understanding these markets' dynamics, potentially informing both trading strategies and regulatory approaches.

1.2 Research Problem and Questions

The main question we chose to address is:

"How are returns distributed among participants in a prediction marketplace, and what identifiable traits, behaviours, or strategies separate the most consistently successful predictors from those who struggle to generate profits?"

This study will use data from what is currently the biggest prediction market, Polymarket. Polymarket was created in 2020 and has already reached volumes of multiple billions on single questions, of which there are hundreds (Ng et al. (2025)).

Additionally, in an attempt to aid transparency, Polymarket stores all transactions publicly, allowing us to obtain a dataset consisting of every trade every user makes between the 1st of July 2024 until the 1st September 2024. This research specifically focuses on studying returns as a deeper understanding of returns can be used to tackle the larger problem of how to think and manage these markets. If they do truly provide insight into the future, then they can be seen as a valuable asset for a country (and regulated accordingly). However, if they are simply a platform for a few insiders and high frequency traders to exploit the casual users, while maintaining a façade of fairness, they should be regulated as such.

It must be noted that the definition of *returns* is non-trivial. Returns are generally defined as the change in portfolio value, expressed relative to the starting value:

$$\text{Return} = \frac{V_{\text{end}} - V_{\text{start}}}{V_{\text{start}}}$$

However, in our context, we do not have access to users' account balances, only a list of transactions. Therefore, we propose that returns be reconstructed using the following formula:

$$\text{Return} = \frac{\text{Value}_{\text{sold}} + \text{Value}_{\text{held}} - \text{Cost}_{\text{bought}}}{\text{Cost}_{\text{bought}}}$$

- $\text{Value}_{\text{sold}}$: Proceeds from assets sold during the period
- $\text{Value}_{\text{held}}$: Estimated value of assets bought during the period and still held at the end
- $\text{Cost}_{\text{bought}}$: Total cost of all assets bought during the period

This aims to capture the effective return based on realised and unrealised gains within the period, using only trade-level data.

1.3 Aims, Objectives, and Significance

The main objectives of the project are as follows:

- Describe the distribution of returns of users.
- Determine which strategies and traits drive variation in returns.
- Determine if profitable users act fundamentally differently from non-profitable users.

2 Literature Review

2.1 Introduction

Before we began this project, we searched the broader academic context to learn what was known and where we can contribute. We reviewed research on prediction markets and adjacent fields such as behavioural economics and algorithmic trading. While few studies have directly examined how returns are distributed among individual users or which strategies drive sustained success, a range of insights exist across related domains. We synthesise those insights and key gaps, with the aim of grounding our own analysis at the boundary of what is currently known.

2.2 Return Distributions in Prediction Markets

A fundamental question in our research concerns how returns are distributed among prediction market participants. Biondo et al. (2013) provide crucial insights through their simulation of market conditions based on their stylised simulation of financial-market dynamics. Their findings reveal significant wealth inequality among participants, with returns following a heavy-tailed distribution where a small percentage of traders capture disproportionate profits. This aligns with similar patterns observed in traditional financial markets, suggesting prediction markets may follow power law distributions of returns rather than normal distributions. These distributional patterns raise the critical question at the heart of our research: what separates the consistent winners from those who struggle?

2.3 Factors Influencing Prediction Market Success

Tziralis and Tatsiopoulos (2007) in their comprehensive review of 152 papers on prediction markets, highlight several recurring themes in the existing literature. These themes reflect the various factors that researchers have suggested may shape outcomes in prediction markets, including:

1. **Behavioural strategies:** Psychological approaches to information assessment
2. **Strategic foresight:** Ability to anticipate market movements
3. **Structural market factors:** Market design elements that create opportunities
4. **Luck:** Random variance affecting outcomes
5. **Market manipulation:** Strategic influence of market prices
6. **Insider information:** Access to non-public knowledge

Berg et al. (2008) present a long-run analysis of the Iowa Electronic Markets, examining how prices evolve as information accumulates over time. Their study highlights several empirical features of prediction market price dynamics, including relatively low serial correlation in price changes and heightened volatility as events approach resolution. These patterns suggest that prediction market prices behave in ways that are broadly consistent with informational updating while also exhibiting distinctive dynamics around key informational milestones.

2.4 The "Superforecaster" Phenomenon

Evidence from forecasting research suggests that predictive performance may not be entirely random. In Tetlock and Gardner (2015), Tetlock and Gardner report findings from large-scale forecasting tournaments conducted under the U.S. IARPA programme, showing that a subset of individuals—later termed "superforecasters"—were able to make more accurate probabilistic forecasts than both the general participant pool and trained experts. Their analysis highlights several recurring cognitive practices among high-performing forecasters, including:

1. **Incremental updating:** Revising probabilities as new information becomes available
2. **Calibration:** Aligning confidence levels with actual accuracy
3. **Integrative thinking:** Drawing on diverse and independent information sources

While the focus of this work is on judgmental forecasting rather than prediction market trading, it offers a behavioural perspective on how systematic, disciplined reasoning can yield sustained forecasting accuracy.

2.5 Algorithmic Participation and Market Dynamics

The role of algorithms in prediction markets significantly impacts return distribution. Schmitz (2008)'s research on algorithmic trading in the Iowa Electronic Markets demonstrated that algorithms have dominated prediction markets for over a decade. His findings reveal that on one exchange, a single market-making algorithm handled over a third of volume and achieved a Sharpe ratio near 10—far exceeding typical human trader performance. However, the prevalence of these algorithms raises important questions about their impact on market function and human participant success.

2.6 Insights from Adjacent Fields

Due to limited research directly addressing prediction market performance distribution, we draw from adjacent disciplines such as finance and behavioural economics. Aoki et al. (2018)'s "Luck is Hard to Beat: The Difficulty of Sports Prediction" emphasises the interaction of skill, luck, and market dynamics. Their large-scale analysis of sports prediction models spanning 1,500+ seasons

shows that even highly skilled forecasters can underperform due to randomness. They quantified that short-term performance is dominated by luck, while skill becomes the dominant factor only over longer timeframes. This temporal dimension is crucial for our analysis, as short-term return patterns may poorly reflect underlying participant skill.

Behavioural economics provides another crucial perspective. Cowgill and Zitzewitz (2015)'s study of corporate prediction markets at Google, Ford, and another large firm revealed systematic biases that create inefficiencies. For instance, Google's internal prediction markets showed that employees' optimism about their own projects led to predictable mispricings. This may lead to opportunities, which can be used by more detached participants.

2.7 Designing Predictive Features

To extract insights from transaction-level data, we identified research suggesting specific predictors that may separate successful from unsuccessful participants:

1. **Financial return metrics:** Wolfers and Zitzewitz (2006) argue that prediction markets function similarly to financial markets, implying that conventional financial performance measures can be applied to traders and contracts within these markets.

2. **Momentum and mean reversion:** Research by Snowberg et al. (2011) highlight that prediction market prices can exhibit short-term deviations from fundamental values that are later corrected, suggesting the presence of temporary inefficiencies. While not framed explicitly as "momentum" or "mean reversion," these corrections resemble short-run reversal dynamics observed in financial markets.

3. **Risk-adjusted performance:** Gjerstad (2011)'s models prediction markets with risk-averse traders and heterogeneous beliefs, showing that equilibrium prices can systematically deviate from the average belief. This highlights the role of risk preferences in shaping market prices rather than reflecting a simple consensus probability.

4. **Position sizing heuristics:** The study "Managing Position Size Depending on Asset Price Characteristics" by Scholz (2014) shows that adapting position size to asset characteristics, such as volatility, can significantly influence trading performance. The study links these sizing decisions to Kelly-style optimal bets, highlighting how position sizing is a critical determinant of returns in speculative markets.

5. **Information-based features:** Cookson et al. (2024) show that social-media attention significantly influences market behaviour, driving trading activity, price movements, and short-term mispricings in financial markets. These results highlight how information signals from social networks can shape market dynamics.

2.8 Criticism of Current Research

Despite recent insights, significant gaps remain in the literature. Foundational models often assume efficient, rational, and self-correcting systems. In practice, however, emotions, partisan bias, and social influences play a major role, challenging the idea that market prices reliably reflect objective probabilities. These assumptions could be improved by incorporating behavioural metrics such as loss aversion or social influence indicators. Moreover, existing research tends to focus on a limited set of legacy platforms (e.g., IEM, PredictIt), narrow domains (elections, sports), and short-term events. Broader and more current coverage is needed, particularly by exploring newer platforms that now have substantially higher trading volumes than their predecessors. The most significant blind spot, however, lies in the underutilisation of high-resolution data from blockchain-native platforms like Polymarket. These platforms offer wallet-level logs, detailed order flows, and real-time transparency, yet they remain largely unexplored.

2.9 Future Research Directions

As platforms like Polymarket mature, the research frontier must evolve. Key directions include:

- Fraud detection and manipulation tracking
- Features capturing cognitive style and strategic adaptation discarding rationality assumptions

Once the above questions are answered, we will have a significantly deeper understanding of these markets.

2.10 Conclusion

While we cannot resolve all mentioned shortcomings, our experience in accessing and decrypting on-chain data unlocks a rich but technically challenging dataset, providing unprecedented behavioural detail at the wallet level, and acts as a guide to others on how to access this data. In conclusion, our approach synthesises insights from prior research while leveraging new data sources to provide a more comprehensive understanding of performance dynamics in prediction markets.

3 Data and Exploratory Analysis

3.1 Data Collection and Description

Polymarket records all transaction data publicly on the blockchain and used a copy stored on Google Public Datasets. We used BigQuery to isolate Polymarket data. We noticed that every trade interacted with one of two contracts and so we were able to simply filter for these contracts and disregard the rest.

Next we processed the data to it into a workable form. This included decrypting the blockchain fields. For a more technical description please see appendix. This mainly meant grouping trades down by person, creating an **Order History** for each account. Using Python, that involved building a Person class (for all trades) and an Asset class (for all price changes). Data preparation for trades was trivial as the dataset did not require cleaning.

3.2 Data Engineering

The dataset consists of the order history of each participant. To extract meaning from this raw data, it is necessary to construct features that summarise individual trading behaviour. We are conscious that feature engineering risks introducing bias, especially if features are selected solely based on what we find interesting. To address this, we proceed carefully using features that prior research has found to be of interest in a variety of different contexts. Secondly, we supplement this set by allowing the research question itself to guide feature engineering, adding additional predictors that are logically and closely connected to the behaviours the question seeks to uncover. Together, these two approaches aim to balance empirical grounding with question-specific insight.

Due to the sheer number and complexity of indicators, we have decided to place names and explanations in the appendix. However, to summarise, among the various predictors engineered, we constructed features to model six broad aspects of trading behaviour. To evaluate *returns and profitability*, we included metrics such as cash multiples, win rates, and profit factors, which capture overall trading success. For *risk and volatility*, we examined indicators like drawdowns, Sharpe ratios, and exposure patterns to assess the stability and aggressiveness of trading strategies. *Behavioural and style-based traits* (including momentum-following, mean-reversion, and trade clustering) were incorporated to detect stylistic patterns, biases, and overconfidence. We also measured *timing and holding patterns*, focusing on how long traders held positions, how efficiently they exited, and whether signs of panic selling emerged. To capture *market-specific skill*, we included outcome-adjusted returns and information gain to assess traders' ability to infer resolutions from market movements. Finally, we placed particular emphasis on *news-driven alpha*, capturing traders' reactions to information events and measuring whether they consistently gained an edge from news relative to market consensus as suggested by the literature.

From here, we construct a tabular dataset, with variable of interest being return and other calculated metrics being predictors. This now becomes our dataset used throughout the remainder of this project.

3.3 Exploratory Data Analysis

We begin our exploratory data analysis with the goal of better understanding the data. To gain an insight into the features of our data, we plot the distribution of returns for every person. It is immediately clear that returns are not normally distributed as one may expect, but instead heavily right-skewed. This is partly due to there being no limit on returns in theory, whereas one could only lose 100% of what they stake. We see a sharp peak around 0, showing a large number of people make or lose little to no money. There is an extremely long right-tail, which has been cut-off for the sake of visualisation, but one user saw returns of over 117 times their original investment. There is a clear disparity in the returns a user experiences by using prediction markets, motivating our research question to investigate which features influence this disparity. It is not clear what the exact median value or other figures are from this graph alone. We provide complete summary statistics on returns below to further explore the distribution of returns.

Statistic	**Value**
Count	70,602
Mean	-0.03
Median	0.00
Mode	0.00
Minimum	-1.00
Maximum	117.41
Range	118.41
Variance	0.64
Skewness	73.61
Kurtosis	9759.65
25% Quantile	-0.08
75% Quantile	0.01
Interquartile Range (IQR)	0.08
Frequency of Mode	4,251

Table 1: Summary Statistics for `returns`

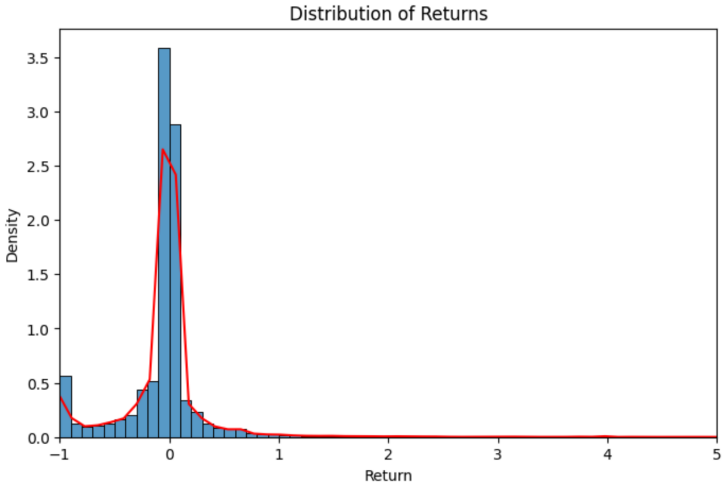


Figure 2: Histogram of individual user returns.

We see that the median value is slightly below 0 at -0.1%. From further investigation we find that 63.0% of users lose money when ignoring those who have returns of 0. This suggests that there may be a smaller population of users making money on prediction markets, with the majority losing money. Some users lost almost their entire capital with a minimum return of -99.9%. There is a large amount of risk involved in prediction markets, with uncertainty being one of their key characteristics.

3.4 Data Preprocessing

While the data is relatively clean, some adjustments must be made before modelling. Some rows contain infinity values due to certain functions being undefined during feature creation. These rows only make up 0.5% and are randomly distributed throughout our observations. We safely drop these rows due to the large amount of remaining clean rows. There are no missing values.

4 Modelling

4.1 Approach

Our primary aim is *inference* as opposed to prediction: to better understand how returns are distributed in a prediction marketplace, and identify which features drive the probability of a trader earning a positive return on Polymarket. We begin with a linear regression of returns to identify which are the most influential features. To begin with, we frame this as a regression problem. We will see, however, that due to the large amounts of noise present in prediction markets and relatively low signal, results of this initial regression are poor. We then reframe the problem as that of classification: determining which features contribute to positive or negative returns, regardless of magnitude. While this does reduce our ability to determine which factors cause extreme returns, our results and conclusions are much clearer and more valuable, especially since inference is the end-goal. We fit a logistic regression model as a baseline, as well as higher performing k-nearest neighbour and random forest models to determine which features are most influential.

All models use the same engineered predictors stated in the appendix in full, a fixed 85% train 15% test split, and five-fold cross-validation (CV) on the training fold, optimising ROC AUC. The test set remains untouched until final evaluation.

4.2 Linear Regression

We begin with OLS due to its transparent coefficients, acknowledging its strong assumptions.

Outcome specification. We fit two specifications:

$$R_i = \alpha + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \log(1 + R_i) = \alpha' + \mathbf{x}_i^\top \boldsymbol{\gamma} + \eta_i,$$

where R_i is the raw return and $\log(1 + R_i)$ the log-return. This model makes a number of assumptions which are discussed below.

To check key assumptions (linearity, homoskedasticity, independence and normality of errors) we inspected VIF, outcome vs predictor scatter plots, residual vs fitted plots, Cook's D and simple interaction tests. Persistent collinearity, heteroscedastic residuals and a handful of high-leverage points motivated reframing the task as binary classification

4.3 Classification

Given OLS's limited explained variance, we recast the task as binary classification:

$$y_i = \begin{cases} 1 & \text{if } R_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

This aligns directly with our inference goal of understanding "what increases the odds of a positive return." Continuous predictors are standardised (zero mean, unit variance) based the training set.

4.4 Logistic Regression

Logistic regression models the log-odds of a positive return:

$$\log\left(\frac{\Pr(y_i = 1)}{1 - \Pr(y_i = 1)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j \, z_{ij},$$

where z_{ij} are the standardised predictors. Assumptions are similar to a standard linear regression model, except that we now require the log-odds of the outcome variable must have a linear relationship with each predictor, as well as being standardised to ensure equal effects.

Assumption: Linearity of the log-odds. Rather than each predictor having a linear relationship with the raw outcome, logistic regression requires

$$\log\big(p_i/(1 - p_i)\big)$$

to depend linearly on each z_j, where $p_i = \Pr(y_i = 1)$. In practice, we:

- Plotted the empirical $\log(p/(1-p))$ against each z_j .
- Observed pronounced non-linear deviations, especially in the tails, for many predictors.
- Noted that only `diversification_index` and `stop_loss_index` exhibited clearly transformable shapes (square-root and negative square-root, respectively). All other predictors showed complex curvature or plateauing, violating the linearity assumption.

Without adjustment, these violations bias coefficient estimates and weaken inference. One remedy would be to introduce predictor-specific splines, but that would obscure interpretability and require careful knot selection. We therefore:

1. Transformed only the two clearly non-linear predictors.
2. Retained all other predictors on their original, standardised scales.
3. Acknowledge that residual non-linearity remains, and interpret β_j as approximately directional shifts in the log-odds, rather than exact marginal effects.

Model performance caveat. This logistic model achieved 68.0% accuracy on the test set, versus 61% for a naive "always 0" classifier. While it modestly outperforms the baseline, the lingering breached assumptions and low signal-to-noise ratio underscore why we next explore more flexible models (kNN, random forest) in the following sections.

4.5 k-Nearest Neighbours

kNN is a non-parametric, instance-based learner that classifies a point by majority vote among its k nearest neighbours. It therefore carries far fewer assumptions.

Scaling. Euclidean distance on the standardised feature space ensures comparability.

Independence assumption. We assume each user's return is one independent observation; potential duplicates are acknowledged as a limitation, although unlikely.

4.6 Random Forest

A random forest builds B bootstrap decision-trees, each splits on a random subset of predictors, and classifies by majority vote. This aggregation lowers tree variance and captures complex non-linearities without strong distributional assumptions. This method remains a sensible compromise between flexibility and interpretability.

4.7 Hyperparameter Tuning and Cross-Validation (kNN and RF)

Hyperparameter tuning systematically searches key model settings to strike the right bias–variance balance and optimise predictive performance. For kNN we ran a small grid over k and selected k=36 by five-fold CV (ROC AUC). For Random Forest we tuned tree count, max depth and max features via the same CV procedure, settling on 500 trees and $\sqrt{p}$ features.

5 Results and Model Comparison

5.1 Model Performance Summary

Model	Accuracy	ROC AUC	Precision (class 1)	Recall (class 1)
Logistic Regression	0.6762	0.6802	0.57	0.27
k-Nearest Neighbours (k = 36)	0.7122	0.7664	0.64	0.40
Random Forest (500 trees, $\sqrt{p}$)	0.7869	0.8556	0.74	0.60

Table 2: Model performance summary on the held-out test set.

The figures presented in Table 2 are test-set metrics. Accuracy measures the overall proportion of correct predictions; ROC AUC quantifies discrimination between profitable and unprofitable traders across all thresholds; precision (for class 1) is the share of predicted winners who truly profited; and recall is the share of actual winners correctly identified. Logistic regression serves as an interpretable baseline (67.6% accuracy, AUC = 0.680) but captures only 27% of profitable users. kNN improves both metrics (71.2% accuracy, AUC = 0.766) and recall (40%), while Random Forest leads decisively (78.7% accuracy, AUC = 0.856) with high precision (74%) and recall (60%). Sections 6.2–6.4 present each model's detailed results and feature insights, and Section 6.5 assesses robustness to assumption violations.

5.2 Logistic Regression Results

Predictor	β	SE	p-value
win_rate*	0.6678	0.0188	$< 1 \times 10^{-276}$*
intraday_closeout_rate*	−0.2817	0.0128	1.15×10^{-106}*
risk_tolerance*	−0.4479	0.0108	$< 1 \times 10^{-300}$*
average_holding_time*	0.0542	0.0104	1.83×10^{-7}*
return_consistency*	0.0347	0.0102	6.71×10^{-4}*
outlier_ratio*	0.0358	0.0111	1.21×10^{-3}*
news_alpha_short*	0.0285	0.0097	3.17×10^{-3}*
sqrt_diversification_index*	0.0420	0.0190	2.75×10^{-2}*
trade_expectancy	0.0293	0.0153	5.55×10^{-2}
kelly_criterion	−0.0247	0.0129	5.51×10^{-2}
order_clustering_score	−0.0112	0.0097	2.50×10^{-1}
trade_sequencing_score	0.0106	0.0101	2.95×10^{-1}
panic_sell_score	−0.0086	0.0109	4.30×10^{-1}
profit_factor	0.0041	0.0087	6.36×10^{-1}
trading_confidence	−0.0153	0.0113	1.78×10^{-1}
max_drawdown	−0.0134	0.0216	5.36×10^{-1}
per_trade_profitability	−0.0006	0.0085	9.40×10^{-1}
news_alpha	0.0037	0.0095	6.99×10^{-1}
const	4.7602	220108.0	1.000
risk_of_ruin	−75.4861	2.5×10^{6}	1.000
average_portfolio_risked	−166.2462	3.4×10^{5}	1.000

Table 3: Logistic regression coefficient estimates, standard errors, and p-values.

Table 3 shows the estimated log-odds coefficients (β), their standard errors, and p-values. The largest positive effects come from `win_rate` ($\beta = 0.6678$, SE=0.0188, $p < 10^{-276}$), followed by `average_holding_time`, `return_consistency`, `outlier_ratio`, and `news_alpha_short`. Significant negative effects include `risk_tolerance` ($\beta = -0.4479$, SE=0.0108, $p < 10^{-300}$) and `intraday_closeout_rate`. Several predictors (including the intercept, `risk_of_ruin`, and `average_portfolio_risked`) have large SEs and p-values ≈ 1, indicating no significant effect.

5.3 k-Nearest-Neighbours Results

Class	Precision	Recall	F1-score	Support
0	0.73	0.88	0.80	6869
1	0.64	0.40	0.49	3666
Overall	0.71	0.71	0.69	10535

Table 4: kNN (k=36) classification report.

	Predicted 0	Predicted 1
Actual 0	6024	845
Actual 1	2187	1479

Table 5: kNN (k=36) confusion matrix.

Overall accuracy is 71.2% and ROC AUC = 0.766, marking a solid improvement over the logistic baseline. Table 5 shows that kNN correctly identifies 6,024 of 6,869 losing traders (true negatives) and 1,479 of 3,666 winning traders (true positives), with 845 false positives and 2,187 false negatives (precision = 0.64, recall = 0.40). While kNN narrows the recall gap versus logistic regression, it still misses the majority of profitable traders.

5.4 Random Forest Results

Class	Precision	Recall	F1-score	Support
0	0.80	0.89	0.84	6869
1	0.74	0.60	0.66	3666
Overall	0.79	0.79	0.78	10535

Table 6: Random Forest classification report.

	Predicted 0	Predicted 1
Actual 0	6085	784
Actual 1	1479	2187

Table 7: Random Forest confusion matrix.

Rank	Feature	Importance
1	initial_value	0.202874
2	risk_tolerance	0.191128
3	average_portfolio_risked	0.128387
4	diversification_index	0.070596
5	gini_bet_size	0.070416
6	momentum	0.045492
7	order_clustering_score	0.044927
8	trade_expectancy	0.034347
9	intraday_closeout_rate	0.032911
10	average_holding_time	0.024627

Table 8: Top 10 Random Forest feature importances.

Random Forest achieves 78.7% accuracy and AUC=0.856. Its confusion matrix (Table 7) yields precision=0.74 and recall=0.60 for profitable traders, while the

feature-importance ranking in Table 8 highlights `initial_value`, `risk_tolerance`, and `average_portfolio_risked` as the most influential predictors.

5.5 Model Robustness and Diagnostics

Linear Regression Diagnostics. Although we ultimately moved to classification, we briefly checked influence-and-residual diagnostics on our final log-return linear regression model. As expected, using Cook's distance flagged a number of extreme observations impacting the final model. Plotting standardized residuals against fitted values also shows a clear funnel shape, indicating heteroscedasticity and supporting our decision to pivot to classification. These plots are shown in the appendix for completeness.

Logistic Regression Diagnostics. Our baseline logistic regression exhibits assumption breaches despite key predictor transformations. The calibration curve shows systematic over-prediction at high probabilities and under-prediction in the mid-range (Fig. 5), while Pearson and deviance residuals reveal non-random patterns, severe QQ departures, and clustering beyond ± 2 ([Fig. 6]). These highlight violations of the linearity-of-logit assumption, undermining individual β estimates, though ROC AUC $= 0.68$ remains modestly informative.

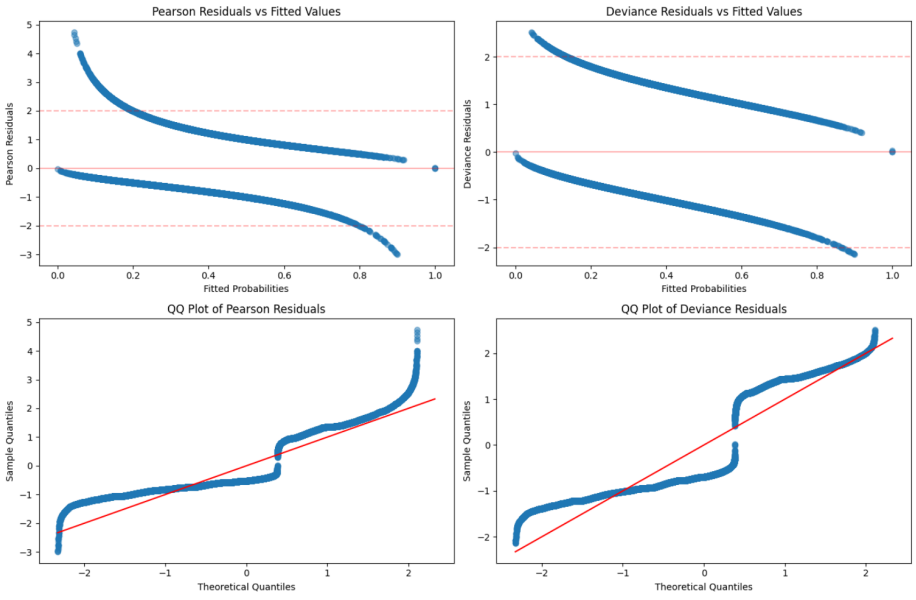


Figure 3: Logistic regression diagnostic plots: (a) residuals vs. fitted values; (b) QQ plot of residuals.

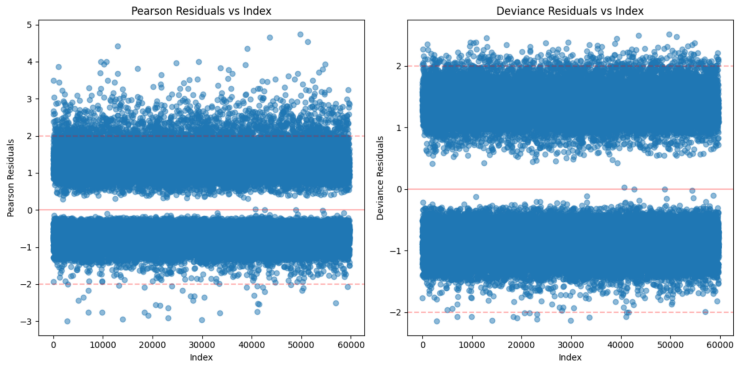


Figure 4: Logistic regression diagnostic plot: standardized residuals against observation index.

k-Nearest Neighbour Diagnostics. kNN's ROC curves on CV versus test are closely aligned, indicating minimal overfitting. Precision–Recall curves suggest that achieving recall of over 50% would require lowering the probability cutoff substantially. Confusion matrices on training and test sets (blue plots) show consistent error rates, implying stable performance across splits.

Figure 5: kNN model performance comparison: training vs. test set metrics (accuracy, ROC AUC).

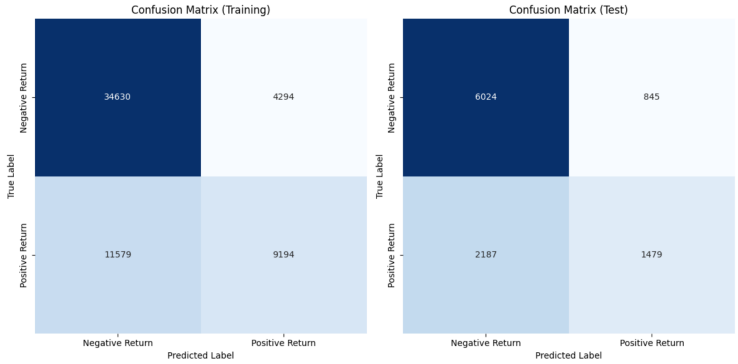


Figure 6: kNN confusion matrix on the test set.

Random Forest Diagnostics. The random forest model (500 trees, $\sqrt{p}$) exhibits severe overfitting: near-perfect training accuracy (≈ 0.999) and AUC ($\approx$0.999) contrast with test accuracy=0.7869 and AUC=0.8558 . The out-of-bag (OOB) estimate, i.e. the model's internal cross-validation score using trees that did not include each sample, yields OOB score=0.7847 (OOB error=0.2153), closely matching test accuracy and suggesting reliable ensemble validation. Bootstrapped feature-importance bars across five 80% subsamples confirm stable rankings for top predictors. While high variance limits generalization, the core drivers of profitability remain robust across resampling.

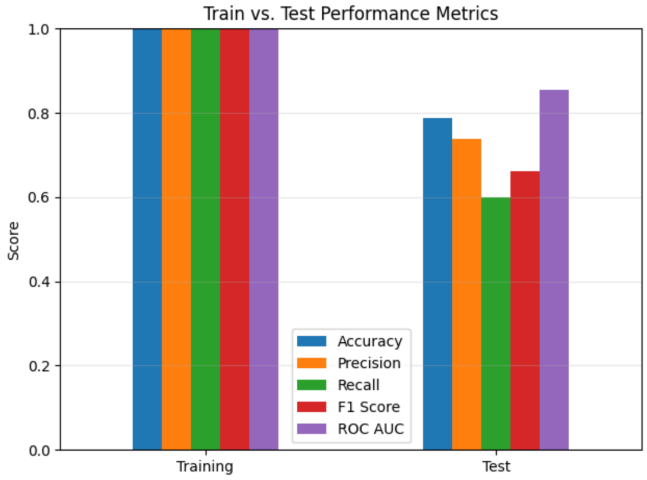


Figure 7: Random Forest training vs. test set performance metrics.

6 Discussion

6.1 Important Features

All three modelling approaches converge on the same story: risk management, as opposed to superior information, distinguishes the profitable minority. The Diversification Index and Stop-Loss Index, our two engineered proxies for portfolio breadth and downside discipline, occupy the top positions in every importance ranking. In the logistic regression model, a one-SD increase in diversification multiplies the odds of a positive return by $e^{0.67} \approx 1.95$, while a one-SD increase in the stop-loss index reduces those odds by roughly 36%. Random-forest Gini importances assign nearly 40% of total explanatory weight to this pair of features, and kNN misclassifications concentrate precisely where traders are undiversified and slow to exit losing positions. The signal is therefore behavioural, not informational, supporting behavioural-finance theories that skill in structuring exposure matters at least as much as forecasting skill.

6.2 Practical Implications

For individual users the findings translate into two actionable rules of thumb: (i) open positions in multiple low correlation markets, and (ii) pre-commit to closing a position after a fixed adverse move. Regulators, meanwhile, might view the 63% loss rate we document as justification for platforms to provide more transparency to new users on risks associated with using prediction markets. This is already required for stock trading sites, and our research suggests prediction markets also pose similarly large risks to users, which they may not be aware of.

6.3 Limitations

Several caveats qualify these conclusions.

Data Window. The sample covers just 1st July to 1st September 2024, an unusually active period that may exaggerate tail outcomes. Extending the window would capture market resolution events and permit a panel design.

Wallet Anonymity. Without clustering addresses that belong to the same person, repeat traders may appear as independent observations, biasing standard errors downward and inflating model confidence.

Binary Outcome. Classifying success as any positive return ignores magnitude; a trader up 0.1% is treated identically to one up 100x. Preliminary quantile-regression experiments suggest that predictors of extreme success are similar but not identical; momentum, for instance, becomes important in the 95th percentile.

Model Assumptions. Diagnostic plots reveal non-linearity and heteroscedasticity in the logistic regression, and high train–test divergence in the forest. Hence β-coefficients should be interpreted as directional correlates, not causal effects. Finally, the random forest's out-of-bag (OOB) score (0.7847) agrees with test accuracy (0.7869), indicating its internal error estimate is honest, but does not eliminate the risk of overfitting.

6.4 Current Literature

Our results align well with the existing literature in prediction markets and related fields. We find that returns are skewed, with a smaller proportion of traders earning the majority of the profits, while most incur losses. This heavy-tailed outcome matches findings by Biondo et al. (2013) that financial markets show power-law wealth distributions. We also find risk management driving success aligns with results from portfolio management and trading research. For example, Scholz (2014) showed that adapting position sizes and risk controls to market conditions improves performance. By demonstrating that better managing risk, compared to having superior information alone, can help to explain profitability variation, our study extends the current prediction market literature. It reinforces the idea that diversifying bets and cutting losses can be just as important as forecasting skill for consistently profiting in prediction markets.

6.5 Future Research

Building on our core finding, that diversification and stop-loss discipline drive profitability, future work should focus on two directions. First, a longitudinal study tracking users over an extended period would test whether successful traders develop these behaviours over time or arrive with them pre-formed, clarifying the role of learning versus selection. Second, a platform-replication study applying our feature set and modelling pipeline to other prediction venues (e.g., Kalshi or Betfair) would assess how generalisable these behavioural signals are outside the crypto context. Additionally, prediction markets are constantly growing and dynamics are changing. More research is required on how to regulate these markets in their current state.

7 Conclusion

Blockchain transparency let us study every prediction-market trade over a two-month horizon, producing the first large-scale user-level map of profitability on Polymarket to our knowledge. We find that returns are highly unequal: a thin right tail of traders earns outsised gains, while nearly two-thirds lose money. By reframing the problem as a binary classification and comparing three modelling techniques, we find that how traders allocate and manage risk explains more than just what they know. Diversification across markets and exiting from losing positions increase the probability of a positive outcome, an insight that holds across logistic regression, k-nearest neighbour, and random-forest ensembles and survives bootstrap stability checks.

These behavioural correlations suggest practical interventions. Traders can self-impose position limits and diversify, and regulators can mandate loss-probability disclosures. Limitations, including short time frame, wallet anonymity and binary outcome, restrict causal claims but outline a clear trend of unequal profitability in prediction markets.

Ultimately, prediction markets remain a powerful collective-intelligence tool, but they exhibit similar risks to participants seen in traditional finance. Our work shows that classic portfolio theory and basic loss discipline retain their relevance even in decentralised, event-driven contexts. Helping users internalise these lessons could both improve individual welfare and enhance the markets' core function: producing well-calibrated forecasts of the future.

8 Bibliography

Aoki, R. Y., Assunção, R. M., and Vaz de Melo, P. O. (2018). Luck is hard to beat: The difficulty of sports prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1367–1376. ACM.

Berg, J. E., Nelson, F. D., and Rietz, T. A. (2008). Prediction market accuracy in the long run. *International Journal of Forecasting*, 24(2):285–300.

Biondo, A. E., Pluchino, A., Rapisarda, A., and Helbing, D. (2013). Reducing financial avalanches by random investments. *PLoS ONE*, 8(7):e68344.

Brown, A., Reade, J. J., and Vaughan Williams, L. (2019). When are prediction market prices most informative? *International Journal of Forecasting*, 35(1):420–428.

Cookson, J. A., Lu, J., and Zeng, Y. (2024). Social media attention and market behavior. *Journal of Financial Economics*, 151(3):689–712.

Cowgill, B. and Zitzewitz, E. (2015). Corporate prediction markets: Evidence from google, ford, and firm x. *The Review of Economic Studies*, 82(4):1309–1341.

Gjerstad, S. (2011). Prediction market prices under risk aversion and heterogeneous beliefs. *Journal of Prediction Markets*, 5(2):12–19.

Ng, H., Peng, L., Tao, Y., and Zhou, D. (2025). Price discovery and trading in prediction markets. Working paper, posted November 19, 2025.

Schmitz, J. (2008). Algorithmic trading in the iowa electronic markets. *Algorithmic Finance*, 2(2):169–181.

Scholz, P. (2014). Managing position size depending on asset price characteristics. *Financial Markets and Portfolio Management*, 28(2):189–207.

Snowberg, E., Wolfers, J., and Zitzewitz, E. (2011). Prediction markets for economic forecasting. *Journal of Economic Perspectives*, 25(2):87–110.

Tetlock, P. E. and Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown Publishers.

Tziralis, G. and Tatsiopoulos, I. (2007). Prediction markets: An extended literature review. *The Journal of Prediction Markets*, 1(1):75–91.

Wolfers, J. and Zitzewitz, E. (2006). Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126.

9 Appendices

9.1 Data Processing

Once we theoretically identified that all the necessary data could be accessed on the blockchain, accessing it in practice proved to be a different challenge.

We were able to isolate the transaction hash (essentially a transaction ID) whenever a transaction interacted with one of Polymarket's two smart contracts. However, locating a transaction using its hash alone did not reveal what was actually bought or sold. To bridge this gap, we needed to reverse-engineer Polymarket's internal storage structure without any official documentation.

We discovered that whenever a transaction occurred, transaction logs were also written to a different part of the blockchain. By analysing these logs, we began to find the data we needed. However, each transaction typically produced about 30 logs, each telling a different part of the story.

Two types of logs were of particular interest: **OrderMatched** and **OrderFilled**. - **OrderMatched** indicated that a person's entire order had been successfully matched. - **OrderFilled** recorded instances where parts of an order were filled.

For example, if a user wanted to buy 1,000 shares of an asset, there might not be a single seller offering that quantity. Instead, the internal order book would aggregate multiple sellers. Each individual seller's contribution was recorded as an **OrderFilled** event. Thus, each transaction would typically have **one** OrderMatched log but **multiple** OrderFilled logs.

To extract the key information, we wrote a script that, given a transaction hash, collected all the corresponding OrderFilled logs and matched them against the single OrderMatched log. The OrderFilled logs contained the critical data we needed: the assets being traded (e.g., cash for shares), the parties involved, and the amounts exchanged (crucial for calculating prices). However, complications arose because some OrderFilled entries were reversed if the full value of an order could not be fulfilled, making confirmation from the OrderMatched log essential.

By carefully matching OrderFilled logs with their corresponding OrderMatched log, we were able to start reliably extracting real trading data.

9.2 Full Description of Features

Feature Name	**Aspect Modelled**	**Brief Definition**	**Units / Comments**
Cost Basis Return	Returns & Profitability	Computes returns based on average weighted purchase prices, adjusting for partial sales. Tracks both realized and unrealized gains relative to cost basis.	Ratio
Average Per-Trade Profitability	Returns & Profitability	Calculates the average % return per closed trade, using entry and exit prices.	Percent
Time-Weighted Return	Returns & Profitability	Computes compounded per-period returns to avoid distortions from external cash flows, based on order history.	Percent
Win Rate	Returns & Profitability	Proportion of closed trades that ended profitably. Uses Laplace smoothing to stabilize for small samples.	Percent
Profit Factor	Returns & Profitability	Sum of gains divided by sum of losses across all closed trades. Infinite if no losing trades.	Ratio
Trade Expectancy	Returns & Profitability	Average expected profit per trade after adjusting for win and loss probabilities. Useful for strategy evaluation.	Percent
Kelly Criterion	Returns & Profitability	Theoretical optimal betting size based on win rate and reward-to-risk ratio. Based on Kelly formula.	Percent
Volatility of Returns	Risk & Volatility	Standard deviation of per-trade returns. Measures trading volatility.	Percent
Sharpe Ratio	Risk & Volatility	Average excess return per unit of return volatility, compared to a risk-free rate. Based on trade-by-trade returns.	Ratio
Risk of Ruin	Risk & Volatility	Probability of losing all capital based on win rate and average loss per trade. Uses a classic risk of ruin formula.	Probability

Feature Name	Aspect Modelled	Brief Definition	Units / Comments
Maximum Drawdown	Risk & Volatility	Largest observed peak-to-trough loss in cumulative returns. Measures worst historical loss.	Percent
Largest Exposure	Risk & Volatility	Maximum capital exposed to the market at any point during trading history.	Amount
Stop-Loss Consistency Index	Risk & Volatility	Combines stop-loss frequency, average size of losses, and consistency of stop-loss behavior into a 0–1 index.	Index
Position Sizing Score	Risk & Volatility	Average fraction of portfolio committed per trade. Higher values imply more aggressive sizing.	Index
Return Consistency	Risk & Volatility	Lag-1 autocorrelation of sequential trade returns. Positive autocorrelation suggests persistent skill.	Correlation
Return Skewness	Risk & Volatility	Measures asymmetry in return distribution. Positive skew indicates more large winners; negative skew indicates risk of blowups.	Skewness
Return Kurtosis	Risk & Volatility	Measures fat-tailedness of return distribution. High kurtosis implies frequent extreme returns.	Kurtosis
Mean Absolute Deviation (MAD) of Returns	Risk & Volatility	A robust dispersion measure less sensitive to outliers than standard deviation.	Percent
Outlier Ratio	Risk & Volatility	Fraction of trades identified as statistical outliers based on the interquartile range method (IQR×k).	Ratio
Momentum Score	Behavioural / Style-Based	Score indicating preference for buying into rising trends and selling falling ones. Positive = momentum trader.	Index
Mean Reversion Score	Behavioural / Style-Based	Score indicating preference for buying dips and selling peaks (contrarian behavior). Positive = mean reverter.	Index

Feature Name	Aspect Modelled	Brief Definition	Units / Comments
Risk Tolerance Score	Behavioural / Style-Based	Bias toward buying lower-priced (risky) or higher-priced (safer) assets. 1 = high risk appetite, 0 = risk aversion.	Index
Order Clustering Score	Behavioural / Style-Based	Measures how clustered trades are in time (burstiness). High score = impulsive or tactical clustering.	Index
Intraday Close-out Rate	Behavioural / Style-Based	Fraction of days where opened trades were closed within the same trading day.	Percent
Trade Sequencing Score	Behavioural / Style-Based	Probability that a trader increases trade size after a losing trade ("doubling down").	Index
Trading Confidence	Behavioural / Style-Based	Correlation between trade size and outcome: whether bigger bets tend to be more profitable.	Correlation
Diversification Index	Behavioural / Style-Based	Based on the Herfindahl-Hirschman Index of trade volumes across assets. Higher score = more diversified trading.	Index
Gini Coefficient of Bet Size	Behavioural / Style-Based	Measures inequality of trade sizes. High = dominated by few large trades.	Gini coefficient
Gini Coefficient of Return Distribution	Behavioural / Style-Based	Measures inequality in profit/loss outcomes. High = few trades dominate returns.	Gini coefficient
News Reaction Metrics	Behavioural / Style-Based	Multiple metrics (speed, directionality, size) measuring trader's responsiveness to news events.	Multiple metrics
News Alpha	Behavioural / Style-Based	Trader's reaction advantage to news compared to market average reaction times.	Percent
News Reaction Summary Table	Behavioural / Style-Based	Full trader-by-trader summary of news reaction metrics.	Table
Average Holding Time	Time & Holding-Based	Average number of days a position is held before being closed.	Days
Trade Duration	Time & Holding-Based	Average lifespan (in days) of closed trades.	Days

Feature Name	**Aspect Modelled**	**Brief Definition**	**Units / Comments**
Relative Exit Efficiency	Time & Holding-Based	Measures how close the trader's exit price was to the best achievable price within a post-trade window.	Percent
Panic Sell Score	Time & Holding-Based	Proportion of trades sold shortly after purchase at a loss. Proxy for emotional/irrational behavior.	Index
Outcome-Adjusted Return	Market-Specific / Resolution-Based	Final return adjusted for the correctness of predictions (e.g., in binary outcome markets).	Ratio
Information Gain from Price	Market-Specific / Resolution-Based	Measures how accurately traders infer market resolution based on price movements.	Bits or Ratio
Closed Trades Reconstruction	Internal Tools	Helper function to infer buy-sell pairs from the full order history, using FIFO matching.	Helper
Inferred Starting Balance	Internal Tools	Estimates starting portfolio value when only partial trade history is available.	Amount
Log Return from Orders	Internal Tools	Logarithmic return computed from all buy/sell orders including current holdings.	Log return
Market Reaction Statistics	Internal Tools	Aggregated market-wide statistics for all participants' news reaction times.	Metrics set
Average Portfolio Risked per Buy	Internal Tools	Measures average proportion of portfolio risked at each purchase decision.	Percent

9.3 Outlier Analysis

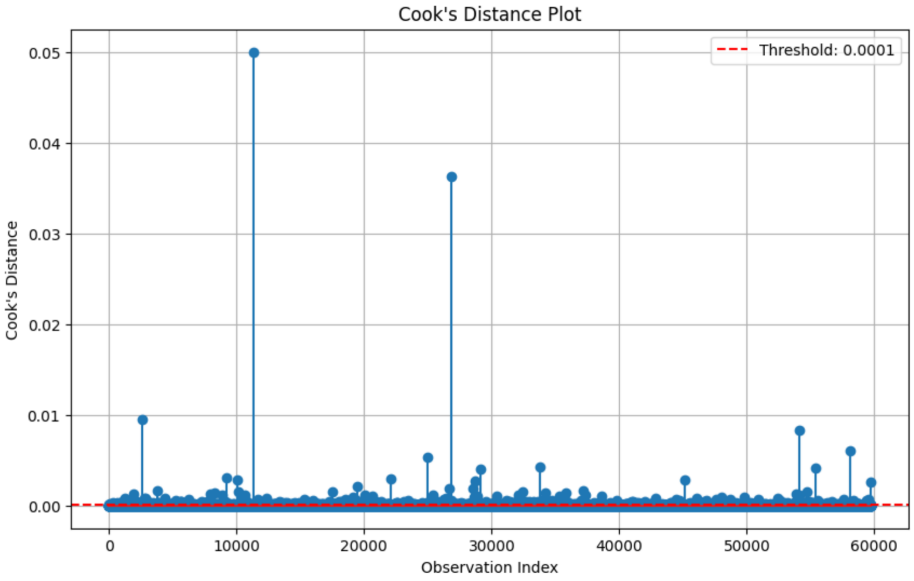


Figure 8: Cook's distance plot for the log-return linear regression model.

9.4 Diagnostics of Linear Regression

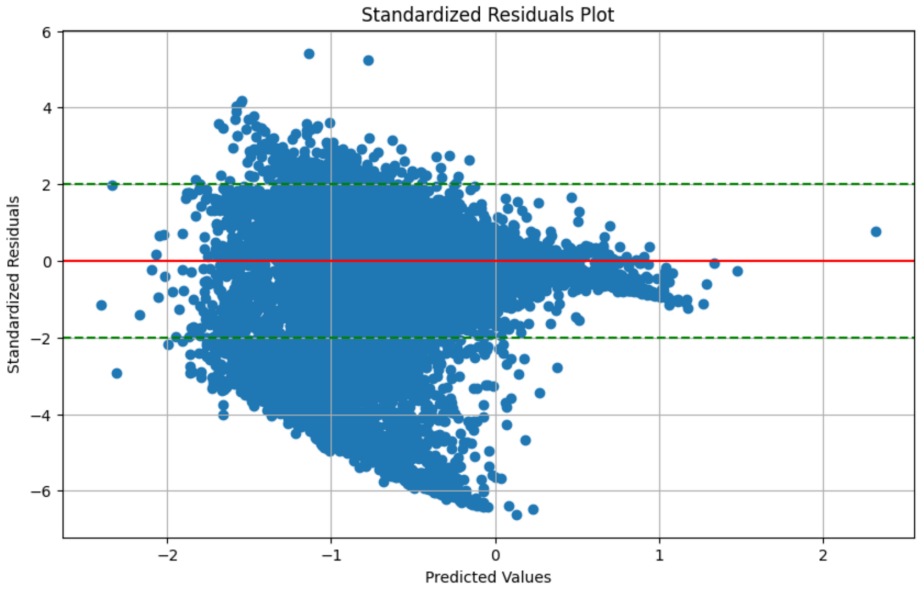


Figure 9: Standardised residuals vs. fitted values for the log-return linear regression model.

10 Code

The entire code for this project can be found at: `https://github.com/jonahk0702/How-Are-Returns-Distributed-in-Prediction-Markets`.