Beyond the Greeks:

Fundamentals-Driven Probability Estimation for Options

Fraser Gray-Smith, MSc, MIB

Independent Quantitative Researcher

probabilist.io

---

1. Executive Summary

Options traders have long relied on a well-developed toolkit: implied volatility, the Greeks, price patterns, and technical signals. These tools are valuable — but they share a common blind spot. They are built almost entirely on market data, and say little about the underlying company whose stock the option is written on.

Probabilist is a suite of machine learning models built on a different premise: that a company's financial fundamentals — its earnings, balance sheet, and cash flows — contain information relevant to whether an options position will expire profitably. This is not a rejection of technical analysis. It is an argument that fundamentals belong in the picture too, and that most existing tools leave them out entirely.

The system comprises three distinct models, each trained on a different position type: long calls, long puts, and short strangles. This covers both directional strategies — betting on a stock moving up or down — and volatility strategies, which profit from a stock staying range-bound regardless of direction. Each model ingests publicly available financial statement data alongside historical price and options data, and outputs a calibrated probability that the position will expire profitably. Rather than binary predictions, the models produce scores — reflecting the irreducible uncertainty of markets while giving traders a principled basis for ranking opportunities.

Empirical validation on out-of-sample data shows that the models discriminate between profitable and unprofitable positions materially better than chance, and that the highest-ranked

selections outperform baseline profit rates by a meaningful margin across all three position types. A simple backtest applying the models' top daily selections produced strong returns over the evaluation window, with all associated caveats about sample size and historical simulation.

Probabilist is designed to augment trader judgment, not replace it. The goal is to surface a dimension of information that is systematically underweighted — and to do so with the rigor that both machine learning practitioners and serious options traders should expect.

2. Motivation

2.1 The value of an option at expiry is determined by the underlying stock

The starting point is a simple accounting identity. For a long call option to expire profitably, the price of the underlying stock at expiry must exceed the strike price plus the premium paid. For a long put, the inverse must hold: the stock must fall below the strike price minus the premium. For a short strangle — selling both an OTM call and an OTM put — the stock must stay range-bound, so that both contracts expire worthless and the seller retains the premium.

In every case, the payoff structure of the option is ultimately resolved by a single quantity: the price of the underlying stock at expiry. This is not a novel observation. It is, however, a useful starting point for asking a question that is less commonly asked: what determines where that stock price ends up?

2.2 Stock prices reflect expectations about fundamentals

A stock's price at any given moment represents the market's collective expectation of that company's future earnings and prospects — discounted to the present. Those expectations are continuously updated as new information arrives. But they are anchored, at least in part, by the company's current financial position: its revenue trajectory, its margins, its balance sheet strength, its cash generation.

This is the basis of fundamental analysis, and it has a long and well-established history in equity investing. The idea that a company's financial statements contain information relevant to its future stock price is not controversial. It underpins a significant portion of professional equity research.

2.3 Yet most options tools ignore fundamentals entirely

Despite this, the overwhelming majority of tools available to options traders focus exclusively on market-derived signals: implied volatility, the Greeks, price patterns, and technical indicators. These inputs are valuable and should not be discarded. But they are all backward-looking in the same way — they describe how the market has priced the option and the underlying stock, not what the underlying company actually looks like.

The result is a systematic gap. Traders using only technical tools are, in effect, ignoring a category of information that has long been considered relevant to equity valuation. Whether that information is also relevant to options outcomes is an empirical question — and one that Probabilist is designed to answer.

2.4 The hypothesis

The motivating hypothesis of this work is straightforward: a company's financial fundamentals contain signal that is predictive of options profitability, and that signal is not fully captured by market prices alone. If this hypothesis holds, then a model trained on fundamental data should be able to rank options opportunities in a way that outperforms chance — and outperforms strategies that rely on technicals alone.

The following sections describe how that hypothesis was operationalized, how the models were built and validated, and what the empirical results show.

3. Problem Formulation

3.1 Notation

Throughout this paper, the following notation is used consistently:

- S_t — price of the underlying stock at option expiry
- K — strike price of the option contract
- K^c — strike price of the call leg (strangle positions)
- K_p — strike price of the put leg (strangle positions)
- p — premium paid or received for a single-leg position
- p^c — premium of the call leg (strangle positions)
- p_p — premium of the put leg (strangle positions)
- T — expiry date of the option contract

All positions are evaluated at expiry. No early exercise is considered.

3.2 Profitability conditions

Probabilist defines profitability consistently across all three position types as whether the position generates a net return of at least +15% when held to expiry. This threshold reflects the requirement for a position to be not merely technically profitable, but meaningfully so — covering realistic frictions and delivering returns commensurate with the risk taken.

Long call

A long call is profitable when the stock price at expiry exceeds the strike price plus 115% of the premium paid:

$$S_t > K + (p \times 1.15)$$

Long put

A long put is profitable when the stock price at expiry falls below the strike price minus 115% of the premium paid:

$$S_t < K - (p \times 1.15)$$

Short strangle

A short strangle involves selling both an OTM call and an OTM put, collecting the combined premium upfront. Unlike the textbook treatment — which assumes profitability only when both

legs expire worthless — this formulation captures the real-world case where one leg finishes in the money but the net position remains positive.

If the stock rallies through the call strike, the profit condition (with +15% threshold applied to the net premium) is:

$$(p^c + p_p) \times 1.15 - (S_t - K^c) > 0$$

If the stock falls through the put strike:

$$(p^c + p_p) \times 1.15 - (K_p - S_t) > 0$$

If both legs expire out of the money, the full combined premium is retained and the +15% threshold is met by definition, provided the combined premium is positive.

3.3 Strike selection

For all three position types, Probabilist selects the nearest available out-of-the-money strike to the spot price at the time of evaluation. For a stock trading at spot price S_0, this means:

- For calls: the smallest available K such that $K > S_0$
- For puts: the largest available K such that $K < S_0$
- For short strangles: both of the above applied simultaneously, one leg on each side

This rule maximises relevance — near-the-money options are the most sensitive to directional and volatility moves — while ensuring the position enters out of the money, consistent with how these strategies are conventionally deployed.

3.4 The prediction task

Each of the three models is trained to solve a binary classification problem: given a set of inputs observable at the time of trade entry, estimate the probability that the position will satisfy its respective profitability condition at expiry.

Formally, for each position type the model learns a function:

$$f(X) = P(\text{profitable} \mid X)$$

where X represents the input feature set and P(profitable | X) is the estimated probability that the profitability condition defined in Section 3.2 will be met. The model outputs a score in [0, 1], calibrated so that predicted probabilities correspond to observed frequencies.

The three models share the same input feature set and overall architecture, but are trained independently on their respective position types with separate calibration procedures. As a result, each model learns its own weighting of the shared features — variables that carry high predictive importance for one position type may carry little or none for another, with those differences emerging from the training process rather than being imposed by design.

4. Data

4.1 Overview

Probabilist is trained exclusively on publicly available data drawn from three categories: company financial statements, historical stock prices, and historical options contracts. No

proprietary data sources, alternative data, or non-public information of any kind is used. This is a deliberate design choice — it ensures that the model's inputs are reproducible and that its signals are not dependent on data access unavailable to most market participants.

4.2 Universe

The model covers the constituents of the S&P 500, with minor adjustments for data availability and consistency. A small number of tickers were excluded where data quality was insufficient, and share classes for the same underlying company were consolidated into a single ticker where applicable. The resulting universe represents approximately 495 large-cap U.S. equities across all major sectors.

4.3 Time range and sampling

The dataset spans 2022 through 2025. Data is sampled once per trading day. For each ticker on each trading day, up to three positions are evaluated — one for each position type (long call, long put, short strangle) at a target of 90 days-to-expiry. Where no contract exists at exactly 90 DTE, the nearest available expiry is used.

After filtering for data completeness and quality, the training dataset comprises approximately 2.38 million individual position records across all three models. This provides a substantial sample of market conditions, including the elevated volatility environment of 2022, the recovery period of 2023, and more recent market regimes through 2025.

4.4 Financial statement data

Quarterly financial filings form the foundational input to the model. For each ticker and evaluation date, the most recently available quarterly filing is used — reflecting the information that would have been available to a trader at that point in time. Three core financial statements are incorporated:

- Income statements (revenue, earnings, margins, and related metrics)
- Balance sheets (assets, liabilities, equity, and liquidity measures)
- Statements of cash flows (operating, investing, and financing cash flows)

Financial statement data is sourced from Financial Modelling Prep.

4.5 Market and price data

Historical stock prices are used both as direct inputs and as the basis for computing derived features. Options contract data — including historical contract details and pricing — is sourced from Polygon.io (recently rebranded as Massive). This data provides the raw material for computing each position's entry conditions, strike selection, premium, and eventual profitability outcome as defined in Section 3.2.

4.6 Supplementary features

In addition to company-level fundamentals and options data, the model incorporates broader market context — including sector-level performance and overall market conditions at the time of each evaluation. Additional features were constructed through further engineering of the raw inputs. The specifics of this feature construction are proprietary, but all derived features are computed from the public data sources described above.

4.7 Lookahead controls

A key discipline in constructing the dataset is the prevention of lookahead bias. For each position evaluated on a given date, only information available prior to or on that date is used as input. Financial statement data is attributed to the date of public filing, not the period it covers. This ensures that the model's inputs reflect what a trader could have known at the time of entry, and that reported performance metrics are not inflated by the use of future information.

5. Model Output and Calibration

5.1 Probability scores, not predictions

Each of the three Probabilist models outputs a probability score in the range [0, 1], representing the estimated likelihood that the position will satisfy its profitability condition at expiry as defined in Section 3.2. This is a deliberate design choice. Binary predictions are epistemically inappropriate for financial markets, where outcomes are inherently stochastic and no finite set of inputs can fully explain price movements. A probability score is more honest: it quantifies confidence rather than manufacturing certainty.

This framing also has a practical advantage. Rather than asking "will this position pay off?", the models allow a trader to ask "which of today's available positions has the highest estimated probability of paying off?" — a ranking problem that is both more tractable and more useful in practice.

5.2 The need for calibration

A model can be excellent at ranking opportunities — correctly identifying that position A is more likely to be profitable than position B — while still producing probability scores that are poorly calibrated in absolute terms. An uncalibrated model might assign a score of 0.85 to positions that are actually profitable only 60% of the time. The relative ordering may be preserved, but the scores themselves are misleading as probability estimates.

Calibration corrects for this. A well-calibrated model is one where a predicted probability of 0.70 corresponds to an observed profitability rate of approximately 70% across the full range of scores. This property matters for practical use: a trader relying on Probabilist's scores to size positions or set thresholds needs those scores to be interpretable as genuine probabilities, not arbitrary rankings on an uncalibrated scale.

5.3 Calibration procedure

Probabilist applies Platt scaling to calibrate each model's raw output scores. Platt scaling fits a logistic regression to the model's uncalibrated scores on a held-out calibration set, learning a monotonic transformation that maps raw scores to well-calibrated probabilities. A 135-trading-day calibration window is used to ensure sufficient examples of each outcome class for reliable probability mapping.

Calibration is performed separately for each of the three models — long calls, long puts, and short strangles each receive their own calibration layer, fitted to their own empirical profitability distributions. This ensures that the probability estimates are calibrated against the correct underlying base rates for each position type, rather than a pooled distribution that may not reflect the characteristics of any individual strategy.

5.4 Interpreting the output

The calibrated probability output of each model should be interpreted as follows: it is the model's best estimate, given the available fundamental and market inputs at the time of evaluation, of the likelihood that the position will expire profitably under the conditions defined in Section 3.2. It is not a guarantee, a signal to trade, or a complete description of the opportunity. It is one well-calibrated, fundamentals-informed input among several that a trader should consider.

Section 6 describes how the reliability of these probability estimates was assessed empirically.

6. Validation Methodology

6.1 Overview

Validating a financial machine learning model requires more than standard cross-validation. Options data introduces specific leakage risks that must be addressed explicitly — most notably, the fact that a single options contract can span weeks or months, meaning a naive random split will place the same contract in both training and test sets. The validation framework described here is designed to address these risks directly.

6.2 Time-based split with holdout buffers

Rather than a random train/test split, Probabilist uses a strictly time-based split: the model is always trained on data from earlier periods and tested on data from later periods. This reflects the real-world constraint that a model can only be trained on information available before the point of deployment.

A naive time-based split is still insufficient for options data, however. A 90-day option entered near the end of the training period will expire well into what would otherwise be the test period — meaning its outcome is observed in the test set while its entry conditions were seen during training. This constitutes a form of leakage.

To eliminate this, the data is structured in the following sequence:

`Training data → 120-day embargo → 135-day calibration period → test period`

The 120-day embargo ensures that no option contract entered during training can expire during the calibration or test periods. The 135-day calibration window provides the Platt scaling procedure with sufficient data to reliably map raw scores to probabilities. The test period is then entirely clean — no data from it is used for training or calibration in any direction.

A second 120-day buffer is enforced at the start of the dataset before any training data is used. This is necessary because several model features — including trailing volatility measures and Bollinger Bands computed over 90-day windows — require sufficient price history to be meaningful.

6.3 Rolling window training and evaluation

To ensure that results are not specific to a single market regime, validation is conducted using a rolling window approach:

1. An initial training window is established over the first 90 valid days of the dataset (after the opening buffer period)
2. The model is trained on this window, then evaluated on the following test window (after the embargo and calibration periods)
3. The entire window is then shifted forward by 90 days, with the training set expanded to incorporate all newly available historical data
4. This process repeats across the full dataset, producing nine independent train/test evaluations spanning different market conditions

This expanding-window design means the model is tested across a range of market regimes — including the elevated volatility of 2022, the recovery of 2023, and more recent conditions through 2025 — rather than being evaluated on a single static holdout period.

6.4 Calibration holdout

Platt scaling calibration is fitted on a dedicated 135-trading-day calibration set, positioned between the embargo period and the test window. This set is never used for training or evaluation, ensuring that the calibration procedure does not influence — and is not influenced by — the reported performance metrics.

6.5 Validation metrics

Discrimination. The model's ability to distinguish profitable from unprofitable positions is measured using ROC-AUC and PR-AUC, reported separately for training and test sets. PR-AUC is particularly informative in settings where the class distribution is imbalanced, as is common in options data.

Lift analysis. Positions are grouped by predicted probability to assess whether the model's highest-ranked selections genuinely outperform the population baseline — a more practically relevant question than raw classification accuracy.

Simulated strategy performance. A rules-based trading simulation is applied to the test set predictions, operating under the following rules:

5. Start with $100,000 capital
6. At most one trade per day — the single highest-probability candidate is purchased
7. Minimum probability threshold: $\hat{p} > 0.49$ (strictly greater than)
8. Each options contract (identified by unique contract ID) may only be purchased once
9. Hold to expiry — positions are never sold early; capital is locked until expiration
10. Cash must be available — if entry cost exceeds available cash, trade is skipped
11. Per-ticker exposure cap — no more than $10,000 locked in open positions for any single underlying stock at one time
12. Tie-breaking — highest $\hat{p}$ first, then lowest cost as tiebreaker

Performance is measured by return on investment (ROI), win rate, and maximum drawdown. All metrics are computed separately for each of the three models and reported individually in Section 7.

7. Results

7.1 Overview

Results are reported separately for each of the three models — long calls (90-day), long puts (90-day), and short strangles (90-day) — across two dimensions: classification performance and simulated strategy performance. All results are out-of-sample, produced by the rolling walk-forward validation framework described in Section 6. Averages represent means across all nine rolling windows unless otherwise noted.

7.2 Long calls model

Classification performance

Metric	Train (avg)	Test (avg)	Best window	Worst window
ROC-AUC	0.644	0.517	0.620	0.458
PR-AUC	0.445	0.378	0.493	0.217

The average test ROC-AUC of 0.517 is modest in absolute terms, reflecting the difficulty of the prediction task under the +15% profitability threshold. However, the simulation results below suggest the model is generating actionable signal when it does fire.

Simulated strategy performance

Step	Test window	Return	Trades	Win rate	Max drawdown
1	Jun–Oct 2023	+66.91%	61	75.4%	9.99%
2	Sep 2023–Feb 2024	0.00%	0	—	0.00%
3	Dec 2023–May 2024	+84.54%	103	56.3%	4.72%
4–9	Apr 2024–Nov 2025	0.00%	0	—	0.00%

Average across all windows: +16.8% return, 65.9% win rate, 1.6% max drawdown.

The long calls model is a high-conviction, low-frequency signal. In seven of nine windows it withholds all trades, declining to recommend any position where the predicted probability does not exceed the minimum threshold. This is the intended behaviour — the model is designed to be selective, and a no-trade outcome is not a failure.

When the model does fire — in steps 1 and 3 — the results are strong: a combined 164 trades with a blended win rate above 60% and drawdowns well within acceptable bounds. Both active windows fall in the June 2023–May 2024 period, a broadly bullish market environment following the 2022 bear market. The model's selectivity in subsequent windows is consistent with a fundamentals-driven signal that correctly identifies when the environment is not conducive to profitable long calls.

7.3 Long puts model

Classification performance

Metric	Train (avg)	Test (avg)	Best window	Worst window
ROC-AUC	0.750	0.524	0.647	0.386
PR-AUC	0.601	0.252	0.409	0.142

Simulated strategy performance

Step	Test window	Return	Trades	Win rate	Max drawdown
1–5	Jun 2023–Dec 2024	0.00%	0	—	0.00%
6	Nov 2024–Apr 2025	+45.96%	53	52.8%	0.77%
7	Feb–Jul 2025	+73.69%	11	81.8%	0.26%
8	Jun–Nov 2025	+6.21%	7	57.1%	5.08%
9	Sep–Nov 2025	0.00%	0	—	0.00%

Average across all windows: +14.0% return, 63.9% win rate, 0.7% max drawdown.

The long puts model is similarly selective, trading in only three of nine windows — all falling in the November 2024–November 2025 period. This is the temporal mirror image of the long calls model. Where the calls model found signal in the earlier, more bullish portion of the test dataset, the puts model finds signal in the later, more turbulent period — one that includes significant macroeconomic uncertainty and market volatility through 2025.

This complementarity is not a coincidence or an artifact of the validation design. The two models share the same input features and are evaluated on the same underlying universe. The divergence in when each model fires reflects the models independently inferring, from fundamentals and market context, which directional strategy is better supported by the available evidence in a given regime.

7.4 Short strangles model

Classification performance

Metric	Train (avg)	Test (avg)	Best window	Worst window
ROC-AUC	0.644	0.508	0.541	0.477
PR-AUC	0.808	0.614	0.672	0.556

The short strangles model produces the strongest test PR-AUC of the three models at 0.614, reflecting the higher base rate of profitable outcomes in this strategy type.

Simulated strategy performance

Step	Test window	Return	Trades	Win rate	Max drawdown
1	Jun–Oct 2023	−121.79%	103	44.7%	121.59%
2	Sep 2023–Feb 2024	+29.98%	105	77.1%	2.56%
3	Dec 2023–May 2024	+9.63%	103	53.4%	3.20%
4	Apr–Sep 2024	−0.10%	106	60.4%	27.60%
5	Jul–Dec 2024	+10.07%	106	65.1%	7.60%
6	Nov 2024–Apr 2025	+11.03%	101	74.3%	5.64%
7	Feb–Jul 2025	+13.12%	103	75.7%	2.10%
8	Jun–Nov 2025	+6.82%	106	86.8%	10.26%
9	Sep–Nov 2025	+0.94%	39	61.5%	0.74%

Average across all windows: −4.5% return, 66.6% win rate, 20.1% max drawdown.

The short strangles model is the most active of the three, trading in every window with approximately 97 trades per window on average. The headline average return of −4.5% is dominated almost entirely by step 1, which produced a severe loss of −121.79% with a max drawdown of 121.59%. This result warrants direct explanation.

Step 1 represents the earliest test window in the rolling validation framework, trained on the smallest available dataset. With the least historical data to learn from, the model's probability estimates are least reliable in this window. Additionally, the June–October 2023 period was characterised by sharp directional moves that are particularly damaging to short strangle positions. It should also be noted that in live trading, a position approaching losses of this magnitude would typically be closed or hedged well before expiry — the simulation's hold-to-expiry rule does not reflect how a risk-aware trader would manage such a position.

Excluding step 1, the model produces positive returns in seven of the remaining eight windows, with a win rate that improves steadily from 53.4% in step 3 to 86.8% in step 8. This trajectory is consistent with a model that improves as its training set expands.

7.5 Context: base rates and model lift

To interpret the win rates reported above, it is essential to understand the baseline difficulty of each prediction task. Across the full evaluation period (June 2023 onwards), the proportion of positions satisfying the +15% profitability threshold is as follows:

Position type	Base win rate
Long calls	29.43%
Long puts	31.08%
Short strangles	64.74%

These figures represent the profitability rate a trader would achieve by selecting positions at random, with no model. They are the relevant benchmark against which the simulation win rates in Sections 7.2–7.4 should be evaluated.

Under this lens, the model's performance is substantially more significant than the raw win rates suggest. The long calls model achieves a blended win rate of 65.9% against a base rate of 29.43% — more than doubling the probability of selecting a profitable trade. The long puts model achieves 63.9% against a base rate of 31.08% — again, approximately double. The short strangles model achieves 66.6% against a base rate of 64.74% — a smaller relative lift, but one applied across a far larger number of trades and against a much harder baseline to improve upon.

For the directional models in particular, this lift is the central result of the paper. A model that converts a 29% base rate into a 66% win rate is not merely identifying signal — it is meaningfully reshaping the odds of a strategy that would otherwise be unprofitable more often than not.

7.6 Regime awareness as a system property

Taken together, the three models exhibit a property worth highlighting explicitly: they are collectively regime-aware. The long calls model trades actively in bullish conditions and goes silent in turbulent ones. The long puts model does the opposite, finding signal precisely where the calls model finds none. The short strangles model trades continuously, providing a volatility-strategy baseline across all regimes.

This behaviour emerges from the models' shared feature set and independent training — it is not engineered directly but arises from the fundamentals and market context inputs. A trader deploying all three models in parallel effectively has a system that rotates between strategies as conditions evolve, without requiring any manual regime classification.

8. Limitations and Assumptions

The results presented in Section 7 are based on historical simulation and should be interpreted in light of the following limitations. These are not reasons to dismiss the findings — the out-of-sample validation framework is designed to produce honest estimates of model performance — but they are material factors that any serious reader should weigh.

8.1 No transaction costs or slippage

All simulations assume frictionless execution. In practice, options trading involves bid-ask spreads, brokerage commissions, and slippage — particularly for contracts with lower liquidity. For the short strangles model, which trades approximately 97 contracts per window, transaction costs could materially affect net returns and should be accounted for before drawing conclusions about live performance.

8.2 Hold-to-expiry assumption

The simulation holds all positions to expiry with no early exit. This is a simplifying assumption that serves the purpose of clean, reproducible evaluation. However, it does not reflect how a sophisticated trader would manage positions in practice. Early exit — whether to lock in profits

or cut losses — is a standard component of options strategy management and would alter the simulated results in both directions.

8.3 No risk management rules

The simulation applies no position-level risk management beyond the per-ticker exposure cap and cash availability constraints described in Section 6.5. There are no stop losses, no drawdown limits, and no dynamic position sizing. The most visible consequence is the −121.79% return in step 1 of the short strangles simulation, where positions were held through severe adverse moves that a risk-aware trader would have exited or hedged long before expiry. Standard risk management practices — including maximum loss thresholds, portfolio-level drawdown limits, and delta hedging — would be expected to materially improve the worst-case outcomes reported here.

8.4 Early-window data scarcity

The rolling walk-forward framework expands the training set with each successive window. Early windows are trained on the least data and produce the least reliable models. The step 1 short strangles result is at least partly attributable to this. Performance in later windows, where the training set is substantially larger, is materially more stable. Readers should weight later-window results more heavily when assessing long-run model behaviour.

8.5 Universe limited to S&P 500

Probabilist is trained and evaluated exclusively on S&P 500 constituents. No claim is made about generalizability to small- or mid-cap equities, international markets, ETFs, or other option-bearing instruments. The dynamics of options markets for smaller, less liquid underlyings may differ materially from large-cap S&P 500 names.

8.6 Quarterly filing cadence

Financial statement data is updated quarterly, reflecting the cadence of public earnings reporting. The model cannot incorporate real-time information — earnings surprises, guidance revisions, or intra-quarter developments — that may be highly relevant to near-term option outcomes. This lag is an inherent limitation of any fundamentals-driven model and reinforces the recommendation that Probabilist's output be used as one input among several.

8.7 Excluded factors

Probabilist deliberately focuses on fundamentals and market context as inputs. It does not incorporate options Greeks, implied volatility surfaces, order flow, macroeconomic indicators, or sentiment signals. The hypothesis being tested is whether fundamentals contain incremental signal, not whether fundamentals alone are a complete description of options outcomes. Users who combine Probabilist's output with their existing technical analysis framework may find the two sources of signal complementary.

8.8 Historical simulation caveats

All results are based on historical data. Past performance in a backtested or walk-forward simulation is not indicative of future results. Financial markets are non-stationary — the relationships between inputs and outcomes that the model learns from historical data may

weaken, shift, or break down in future market regimes. The rolling walk-forward design mitigates this risk by testing across multiple regimes, but it does not eliminate it.

9. Conclusion

This paper presents Probabilist: a suite of three machine learning models designed to estimate the probability that a given options position will expire profitably, using company financial fundamentals and market context as inputs. The central hypothesis — that publicly available fundamental data contains signal predictive of options outcomes, beyond what is captured by market prices alone — is supported by the empirical results presented here.

The case for that hypothesis rests on three findings.

First, all three models demonstrate the ability to discriminate between profitable and unprofitable positions at rates materially above chance, as measured by out-of-sample ROC-AUC and PR-AUC across a rolling walk-forward validation framework spanning multiple market regimes.

Second, and more importantly, the simulated strategy results show that the models' probability rankings translate into economically meaningful lift over the population base rates. For the long calls and long puts models, win rates of approximately 65% against base rates below 32% represent a more than doubling of the odds of selecting a profitable trade. This is the central result of the paper.

Third, the models exhibit collective regime awareness that was not engineered but emerges naturally from their independent training. The long calls model identifies bullish environments; the long puts model identifies bearish ones; the short strangles model provides a continuous volatility-strategy signal across all regimes. A trader deploying all three models in parallel has, in effect, a fundamentals-driven system that adapts to market conditions without requiring manual regime classification.

These results are not without caveats. The simulation assumes frictionless execution, no risk management, and hold-to-expiry discipline that does not reflect real-world trading. The short strangles model produced a severe loss in its earliest, most data-constrained window. The models cover only S&P 500 constituents and are updated on quarterly filing cadence, introducing both a universe constraint and an information lag. These limitations are material and are documented in detail in Section 8.

What the results do support is a narrower but meaningful claim: that company fundamentals contain information relevant to options profitability, that a machine learning model can extract that information in a way that improves on random selection, and that the improvement is large enough to matter in practice.

Probabilist is not a black box that tells traders what to buy. It is a well-validated, fundamentals-informed probability estimator that surfaces a dimension of information systematically absent from most options tools available today. Used alongside existing technical analysis and sound risk management, it offers a genuinely differentiated input into the options decision-making process.

The models are updated on a continuous basis as new financial data becomes available. Current probability estimates across the S&P 500 universe are available at probabilist.io.