

Machine-Learned Conditional Trading of TP2 Violations in SPY American Call Options

Theory-guided event detection, American exercise diagnostics, and walk-forward portfolio selection

Jyotir Patel • Omkar Lashkare • Pawel Polak
Stony Brook University



Research question. When a TP2 violation is executable across the NBBO, which episodes retain positive net payoff after American exercise effects and transaction costs?

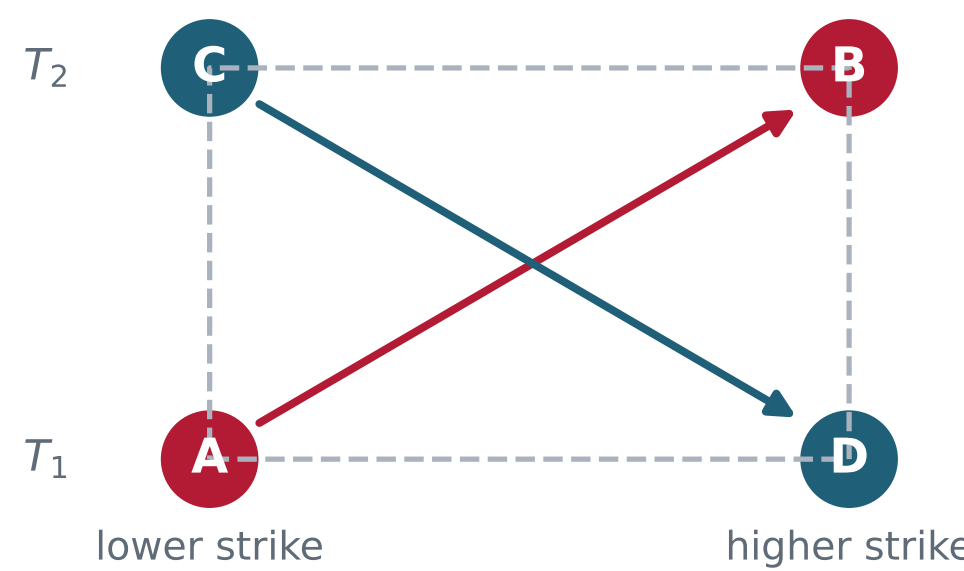
1 Shape restriction

$$\Delta_{\text{TP2}} = AB - CD \geq 0$$

European call-price surface restriction

Executable strong violation

$A^a B^a < C^b D^b$ across quoted spreads



Forward-moneyness alignment
using the parity-implied forward
A TP2 violation is not an arbitrage.

2 American exercise

$$C^{\text{am}}(K, T) = C^{\text{eu}}(K, T) + \pi(K, T)$$

Contract-level tests separate the European benchmark from the American early-exercise wedge

Prop. 2.1
zero premium: $\pi = 0$

Prop. 2.2
dividend bound: $V^q > M$



2021-2026 theory categories

61.0% no distribution 10.0% dividend-spanning
0.7% dividend-bound 28.2% unresolved

71.0% satisfy Proposition 2.1.

Models are trained and evaluated on all four groups.

3 Executable trade

Two candidate actions per episode

$$s \in \{T_1, K_2\}$$

BUY
underpriced pair
at the ask

SHORT
overpriced pair
at the bid

$$Y_{i,s} = \mathbf{1}\{R_{i,s}^{\text{net}} > 0\}$$

Hold to expiration | \$1,000-per-trade normalization

Commission: \$1.30 per contract-leg
(two \$0.65 execution charges)

Quoted spreads are crossed before the commission is applied.

4 Walk-forward selection

$$\hat{p}_{i,s} = \Pr(Y_{i,s} = 1 | X_{i,s})$$

Six rolling folds: every test year is untouched

TRAIN 3Y

CALIBRATE 1Y

TEST 1Y

2021 | 2022 | 2023 | 2024 | 2025 | 2026 (partial)

LightGBM | MLP | Transformer | MV-ALiBi | Logistic

149 point-in-time features

fold-specific top-48 subset for neural and linear models

Trade the higher-probability action if
 $\hat{p} \geq 0.5$; **otherwise abstain**

Date-block bootstrap | calibration | matched random-subset null

All five specifications finish with positive
aggregate out-of-sample P&L

Figure 1. Theory-guided conditional trading architecture. TP2 defines the event set; Propositions 2.1–2.2 diagnose the American early-exercise wedge; a leak-free classifier estimates after-cost profitability; portfolio statistics evaluate the selected trades. The propositions are stated in the lower-left panel.

Theory, data, and identification

177,487

TP2 episodes
2014–2026

125,002

test episodes
2021–2026

250,004

candidate trades
T1 and K2 actions

American-option propositions used in the analysis

Proposition 2.1 (sufficient condition for zero early-exercise premium).

For an American call with strike K and expiry T , suppose either (i) no ex-dividend date lies in $(0, T]$, or (ii), at every ex-date $t_i \leq T$,

$$\delta_i + \sum_{j \geq i} \delta_j e^{-r(t_j - t_i)} \leq K \left(1 - e^{-r(T - t_i)}\right).$$

Then $\pi(K, T) = 0$, so $C^{\text{am}}(K, T) = C^{\text{eu}}(K, T)$. If the condition holds for all four contracts, the rectangle can be evaluated against the European TP2 benchmark.

Proposition 2.2 (dividend-bound certificate).

Let $\bar{D}_1 = \bar{D}(T_1)$ and $V^q = C^b D^b - A^a B^a$. If

$$V^q > M := \bar{D}_2 D^b + \bar{D}_1 C^b, \quad C^{\text{bid}} \leq C^{\text{am},*} \leq C^{\text{ask}} \quad \text{for each leg,}$$

then the corresponding European values also violate TP2: the American early-exercise premium cannot explain the observed inequality.

Assumptions at each signal date: deterministic cash distributions and discount factors, frictionless exercise, and no ownership benefit beyond distributions. Here δ_i is the announced cash distribution at t_i , r is the discount rate, and $\bar{D}(T)$ is the present value of remaining distributions through T .

Executable label and chronological evaluation

Closing OPRA NBBO quotes are sampled at or before 16:00 ET. Strong signals satisfy $A^a B^a < C^b D^b$. For $s \in \{T_1, K_2\}$, long legs enter at the ask, short legs at the bid, positions are held to expiration, and the backtest charges **\$1.30 per contract per leg**.

$$R_{i,s}^{\text{net}} = \frac{\text{premium}_{i,s} + \Pi_{i,s} - \text{cost}_{i,s}}{G_{i,s}}, \quad Y_{i,s} = \mathbf{1}\{R_{i,s}^{\text{net}} > 0\}.$$

TRAIN 3 YEARS

feature selection + fit

CALIBRATE 1 YEAR

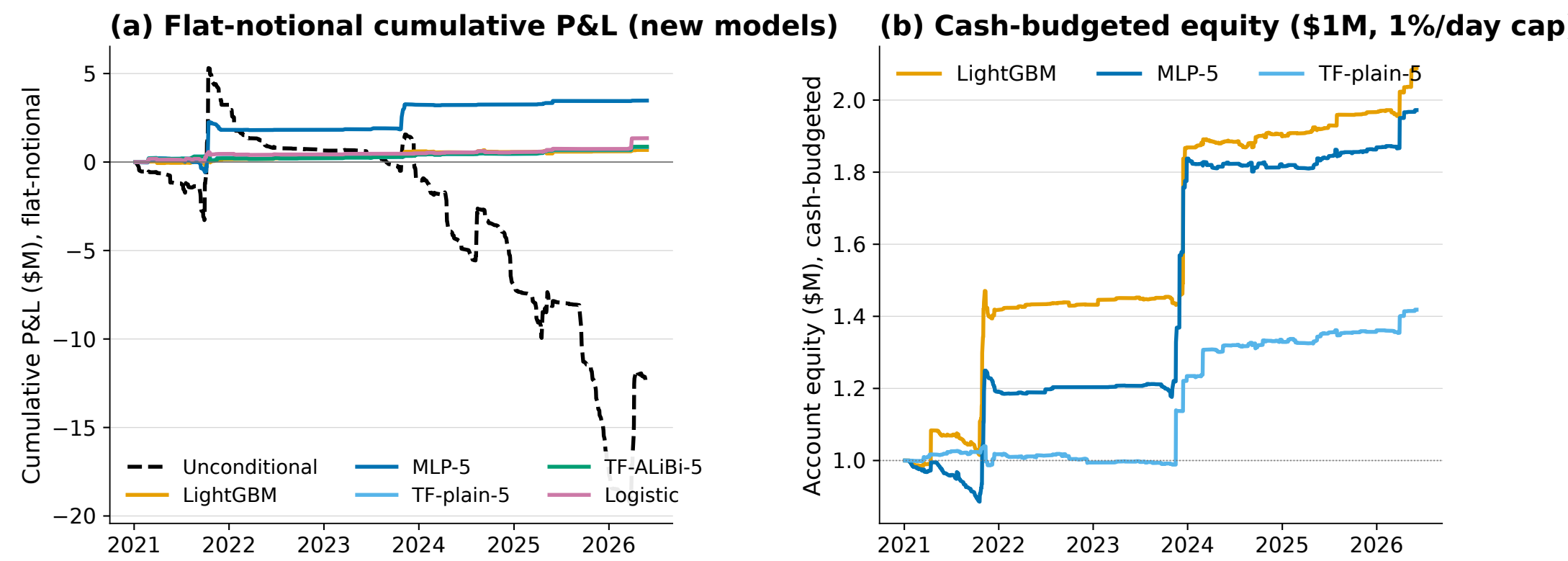
isotonic map

TEST 1 YEAR

untouched period

Six rolling folds test 2021–2026. The 149 point-in-time features cover surface location, short-leg Greeks, liquidity, quoted spreads, and violation history; feature selection, scaling, fitting, and calibration precede each test year. Dollar P&L uses the paper’s \$1,000-per-trade normalization.

Out-of-sample portfolio evidence



Cumulative evidence, 2021–2026. Left: normalized flat-notional P&L. Right: a separate \$1M premium-budget experiment with a 1% daily new-position limit.

Specification	Trades	P&L	Sharpe ¹	95% CI (\$M)	+yrs
Unconditional	250,004	−\$12.52M	−1.01	—	2/6
LightGBM	19,636	\$682K	1.25	[−0.14, 1.61]	5/6
MLP (5-seed)	19,398	\$3.47M	1.92	[−0.77, 10.11]	6/6
Transformer (plain)	15,931	\$840K	2.05	[0.06, 1.41]	5/6
Transformer (ALiBi)	15,542	\$866K	2.13	[0.07, 1.47]	6/6
Logistic (L_2)	12,803	\$1.35M	1.82	[0.05, 2.73]	5/6

¹Annualized active-day Sharpe. Confidence intervals use a date-block bootstrap of total P&L. 2026 is partial.

The criterion matters. The MLP has the largest normalized P&L, while the two attention models have the highest risk-adjusted performance. Logistic regression remains strongly positive, so the selection effect is not confined to high-capacity nonlinear models.

Economic decomposition and inference

Realized P&L by theory category (\$M)

	Prop. 2.1	Prop. 2.2	Unresolved
Unconditional	71.0%	0.7%	28.2%
MLP	3.24	−0.13	−10.08
Transformer (plain)	0.37	0.01	0.46
Transformer (ALiBi)	0.40	0.01	0.46
LightGBM	0.36	−0.05	0.38

The unconditional loss is concentrated in the unresolved American-option wedge. The MLP monetizes the zero-premium subset; the attention models and LightGBM obtain their highest category Sharpe in the unresolved group.

Statistical diagnostics

- Date-block 95% intervals exclude zero for both transformer variants and logistic regression.
- In an active-day/trade-count matched null, only 8% of random portfolios exceed the plain transformer’s P&L.
- The transformers are best calibrated (ECE 0.036/0.038); the MLP has the largest P&L despite ECE 0.065.

Capital interpretation and scope

- Premium-budget growth is 14.6% for LightGBM and 13.4% for the MLP; reserving Reg-T margin reduces all models to roughly 1–2%.
- Selected portfolios retain exposure to SPY and are not interpreted as market-neutral.
- One underlying, active-day Sharpe, partial 2026, and no full simulation of assignment, margin netting, or brokerage mechanics.

Selected references. [1] Glasserman, Li, and Pirjol (2025), *Quantitative Finance* 25(8), 1177–1198. [2] Glasserman and Pirjol (2026), *Mathematical Finance* 36(3), 571–594. [3] Press, Smith, and Lewis (2022), *ICLR*. [4] Künsch (1989), *Annals of Statistics* 17(3), 1217–1241.