Predicting Post-Insider-Trade Stock Direction Using Supervised Learning

A Machine Learning Approach to Market Signals

Nathan Garza

CSCA 5622 - Supervised Learning Final Project

The Problem

Research Question

After an insider open-market purchase (SEC code `P`), can I predict whether the stock's short-horizon forward return (e.g., 20 trading days) will be non-negative?

Why This Matters

Insider trades are legally required disclosures that may signal private information about company prospects. If insiders consistently buy before price increases, I can potentially identify a profitable trading signal.

Task Type: Binary classification (supervised learning)

Target: Predict if 20-day forward return after insider purchase (code P) is positive or negative

Data Sources

OHLC Daily Data

- Full U.S. equity market
- 2.25M daily bars
- Open, High, Low, Close, Volume
- 1-year: Aug 2024 - Aug 2025

Source: Unusual Whales

Insider Trades (Form 4)

- 407K insider transactions
- Role flags: officer, director, 10% owner
- Transaction codes, prices, roles
- 1-year: Aug 2024 - Aug 2025

Source: SEC via Unusual Whales

Feature Engineering

Engineered Features (leakage-safe):

1. Technical: momentum (5-day, 10-day, 20-day returns), volatility, ATR, overnight gaps, drawdowns
2. Insider: transaction size, officer title, ownership type, filling delay, liquidity terciles
3. Regime: time-based market conditions

Critical Design Choice

- Leakage prevention: All features use data prior to the insider purchase event to avoid look-ahead bias.

ML Approach & Methods

Models Tested

1. HistGradientBoosting
2. RandomForest
3. Stacking Ensemble
4. Logistic Regression

Evaluation Strategy

- Walk-forward cross-validation: Respects temporal ordering of financial data
- Multiple metrics: ROC-AUC, PR-AUC, Brier score, F1 score
- Isotonic calibration: Ensures predicted probabilities are reliable
- Hold-out test set: Final performance on unseen data

Key Iterations

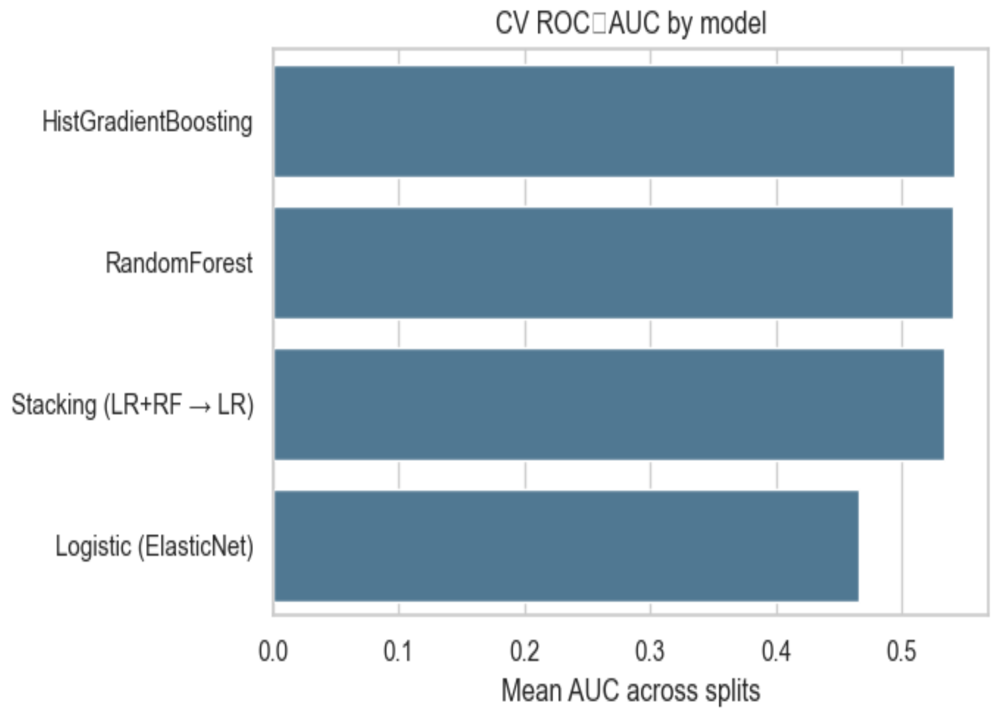
- Baseline with core features and leakage-safe labels
- Added heterogeneity features (liquidity tercile, market regime)
- Enhanced technical indicators (ATR, overnight gaps, drawdowns)
- Feature ablation studies to validate signal sources

Key Results: Model Performance

HistGradientBoosting
Cross-validated ROC-AUC: ~0.542
Hold-out ROC-AUC: ~0.543

RandomForest
Cross-validated ROC-AUC: ~0.540
Consistent across splits

Interpretation
Models detect a modest but reliable signal. Performance is significantly better than random (0.50), indicating insider purchases carry predictive information.

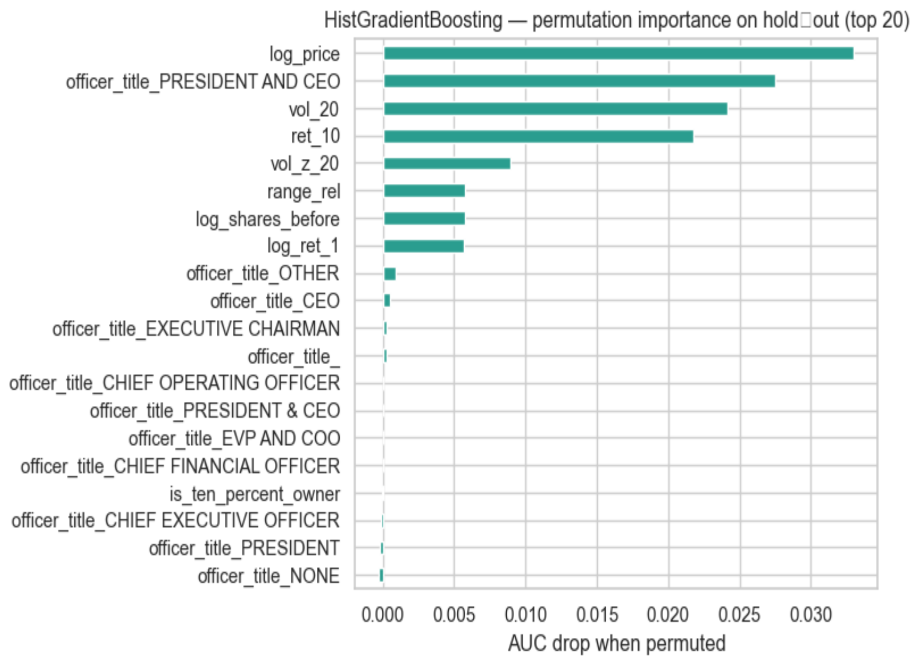


Feature Importance Analysis

Most Predictive Features:

1. Stock price (log_price): Highest importance - larger companies may show different patterns
2. Officer title (President/CEO): Trades by top executives carry stronger signals
3. Recent volatility (vol_20): Market conditions matter
4. Recent returns (ret_10, ret_5): Momentum effects

Key Insight: The model relies on a combination of stock characteristics, officer seniority, and technical indicators rather than a single dominant feature.



Model Calibration & Reliability

Probability Calibration

I applied isotonic calibration to ensure predicted probabilities match observed frequencies.

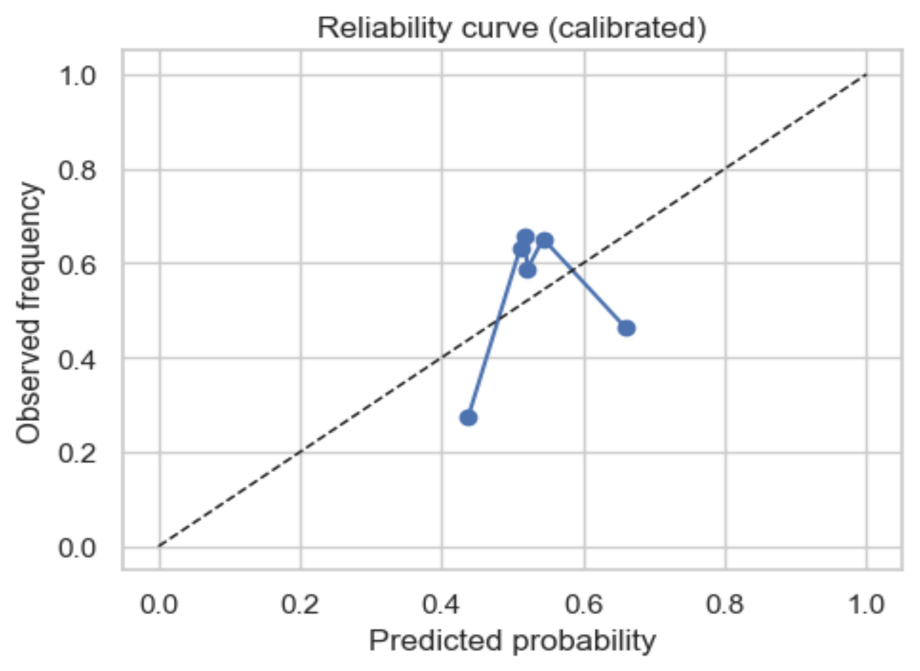
Why it matters: Calibrated probabilities enable decision-making based on confidence levels.

Additional Metrics:

ROC-AUC: ~0.542

PR-AUC: ~0.576

Bootstrap 95% CI confirms hold-out AUC is reliable



Conclusions & Future Work

What I Learned

- Insider purchases carry a small but reliable signal for short-term stock direction
- Tree-based models (HistGradientBoosting, RandomForest) outperform linear models
- Proper temporal validation and leakage prevention are critical for financial ML
- Feature engineering matters: combining price, volatility, and insider characteristics improves performance

Limitations & Future Improvements

- Limited to 1 year of data - need more regime coverage for robustness
- Could incorporate additional features: sentiment analysis, sector trends, macroeconomic indicators
- Explore shorter and longer prediction horizons (2 to 4 day, 21 to 90 day returns)

Supervised learning successfully identified predictive patterns in insider trading data, demonstrating the value of ML in financial forecasting.