Predicting Stock Movements Through Insider Trading Patterns

Bridging the Gap Between Retail Investors and Corporate Investors

Lexa Fu, Kidus Yohannes, Alex Lee, Mihir Argulkar
Machine Learning in Finance
Prof Karthik Krishnan
Group 5

MEET THE TEAM

Lexa Fu

Fifth year Brand Management and FinTech

LinkedIn

Alex Lee

Fifth year Data Science and FinTech

LinkedIn

Mihir Argulkar

Third year Data Science and FinTech

LinkedIn

Kidus Yohannes

Fifth year Computer Science and FinTech

LinkedIn

AGENDA

- Business Problem
- Business Context
- Who, What, Why
- Data
- Model: Ensemble
- Model: XGBoost
- Strategy: Insider Trade Signal
- Business Insights
- Limitations
- Demo

OUR BUSINESS QUESTION

Can publicly available insider trading data be used to build a trading strategy that helps retail investors achieve better performance than traditional investment approaches?

BUSINESS CONTEXT

- Corporate insiders often trade before major events, earning ~4–6% abnormal returns within six months
- Trades are publicly disclosed, (Form 4) but retail investors lack the tools to interpret these signals
- This creates **information asymmetry:** insiders act before news becomes public, while everyone else reacts after the market moves
- As a result, insiders consistently earn abnormal returns that **average investors cannot access**

BUSINESS CONTEXT

Legal insider trades reported on SEC Form 4 occur at massive scale, averaging 200,000 insider transactions every year

WHO, WHAT, WHY

WHO

- Retail investors

WHAT

- Build a machine learning model that analyzes insider trading patterns
- Identify abnormal trading signals
- Predict short-term stock direction following insider activity

WHY

- Insider trades often reflect non-public expectations about future events
- Detecting predictive patterns can help bridge the information gap

BUSINESS CONTEXT

But most importantly...
our model aims to demonstrate how data can turn insider activity into an accessible signal for anyone making investment decisions, allowing retail investors to also make profitable decisions

DATA OVERVIEW

- **Form 4 Public Filings (SEC EDGAR)**
  - OpenInsider.com
- **Yahoo Finance Return % Data**
- **Twelve years spanning four distinct market regimes:**
  - Post-crisis recovery (2013-2017)
  - Bull market (2018-2019)
  - COVID shock + recovery (2020)
  - Bear market + stabilization (2021-2025)
- **Removed delisted companies**
  - No return % data

Date Range	**2013 - 2025**
Total Insider Transactions	**189,484**
Unique Companies	**9,125**
Unique Insiders	**40,501**

MODEL- XGBOOST

MODEL 1 (30 DAY RETURN PREDICTION)

```
🚀 === Test Set Performance ===
Mean Squared Error: 0.0399
Root Mean Squared Error (RMSE): 0.1997
Mean Absolute Error (MAE): 0.1112
R² Score: 0.0907

Baseline MAE (predicting mean): 0.1145
Improvement over baseline: 2.91%
```

MODEL 2 (60 DAY RETURN PREDICTION)

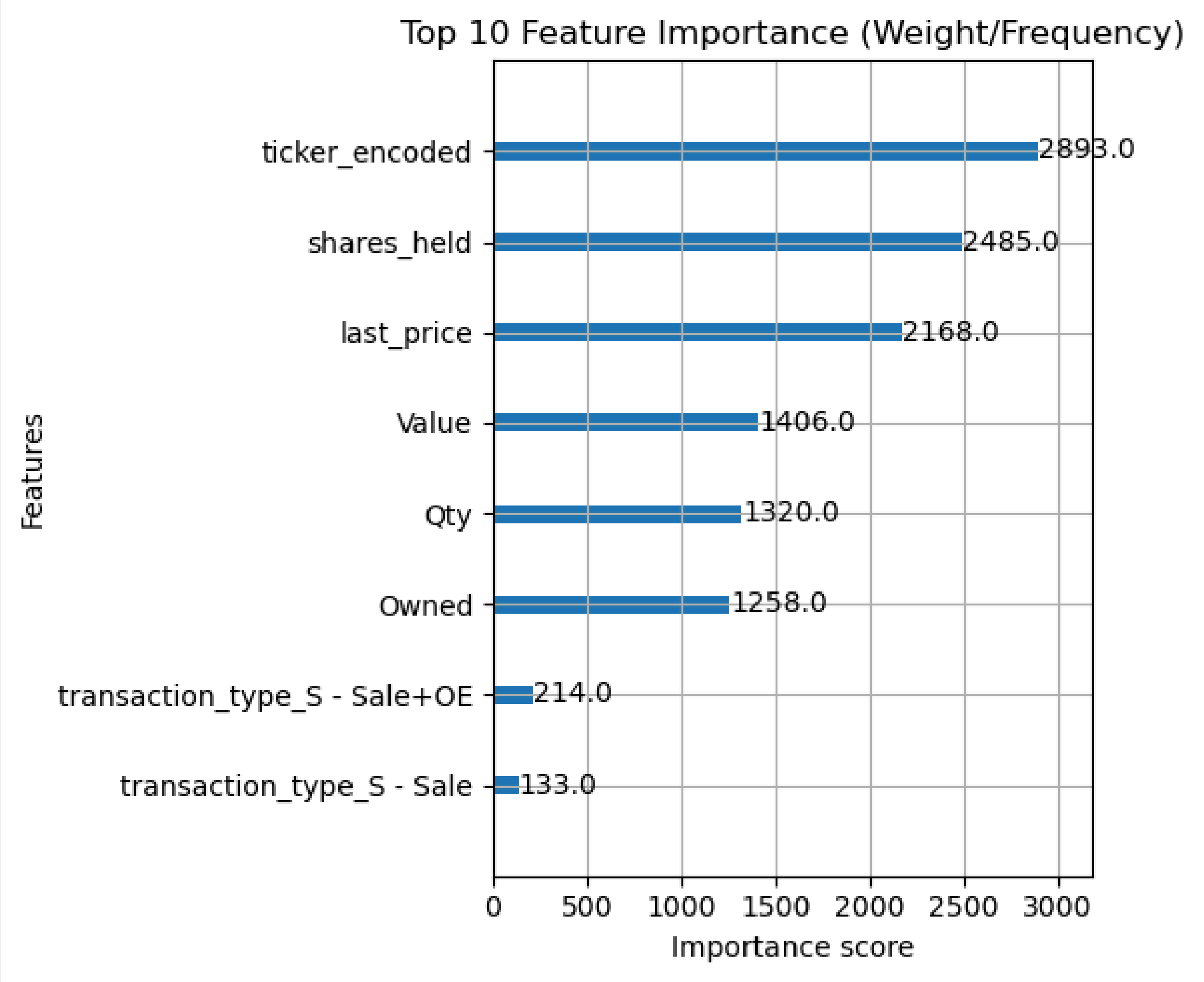
```
🚀 === Test Set Performance ===
Mean Squared Error: 0.0826
Root Mean Squared Error (RMSE): 0.2874
Mean Absolute Error (MAE): 0.1531
R² Score: 0.1403

Baseline MAE (predicting mean): 0.1605
Improvement over baseline: 4.59%
```

MODEL 3 (90 DAY RETURN PREDICTION)

```
🚀 === Test Set Performance ===
Mean Squared Error: 0.1214
Root Mean Squared Error (RMSE): 0.3484
Mean Absolute Error (MAE): 0.1866
R² Score: 0.1493

Baseline MAE (predicting mean): 0.2006
Improvement over baseline: 6.98%
```



MODEL: ENSEMBLE

ENSEMBLE STRATEGIES

Strategy	Avg Return	Win Rate	Trades
Smart Ensemble	5.78%	63.6%	4,231
High Conviction	7.17%	66.1%	1,034
Ultra Ensemble	9.27%	88.9%	18

```
=================================================
ENSEMBLE CONSTRUCTION
=================================================
Ensemble weights (by inverse MAE):
  XGBoost: 0.251
  LightGBM: 0.250
  Ridge: 0.248
  Quantile: 0.251

Simple Ensemble MAE: 0.1207
```

```
--- REGRESSOR PERFORMANCE ---
XGBoost MAE:      0.1207
LightGBM MAE:     0.1208
Ridge MAE:        0.1220
Quantile MAE:     0.1207
Ensemble MAE:     0.1207
Weighted MAE:     0.1207

--- CLASSIFIER PERFORMANCE ---
CatBoost AUC:     0.5577
CatBoost Acc:     0.5632
```

ENSEMBLE MODEL

Why XGBoost Over Ensembles?

- **More Trades**: 16,341 signals → stronger statistical validity
- **Significance**: 9.27% return on 18 trades = too small a sample
- **Comparable Error**: MAE 0.1207 ≈ ensemble performance
- **Simpler**: Easier to explain, debug, and deploy
- **More Robust**: Larger trade volume → more reliable results

Higher returns from ensembles come at the cost of drastically reduced trade volume, making results statistically unreliable

STRATEGY: INSIDER TRADE SIGNAL MODEL

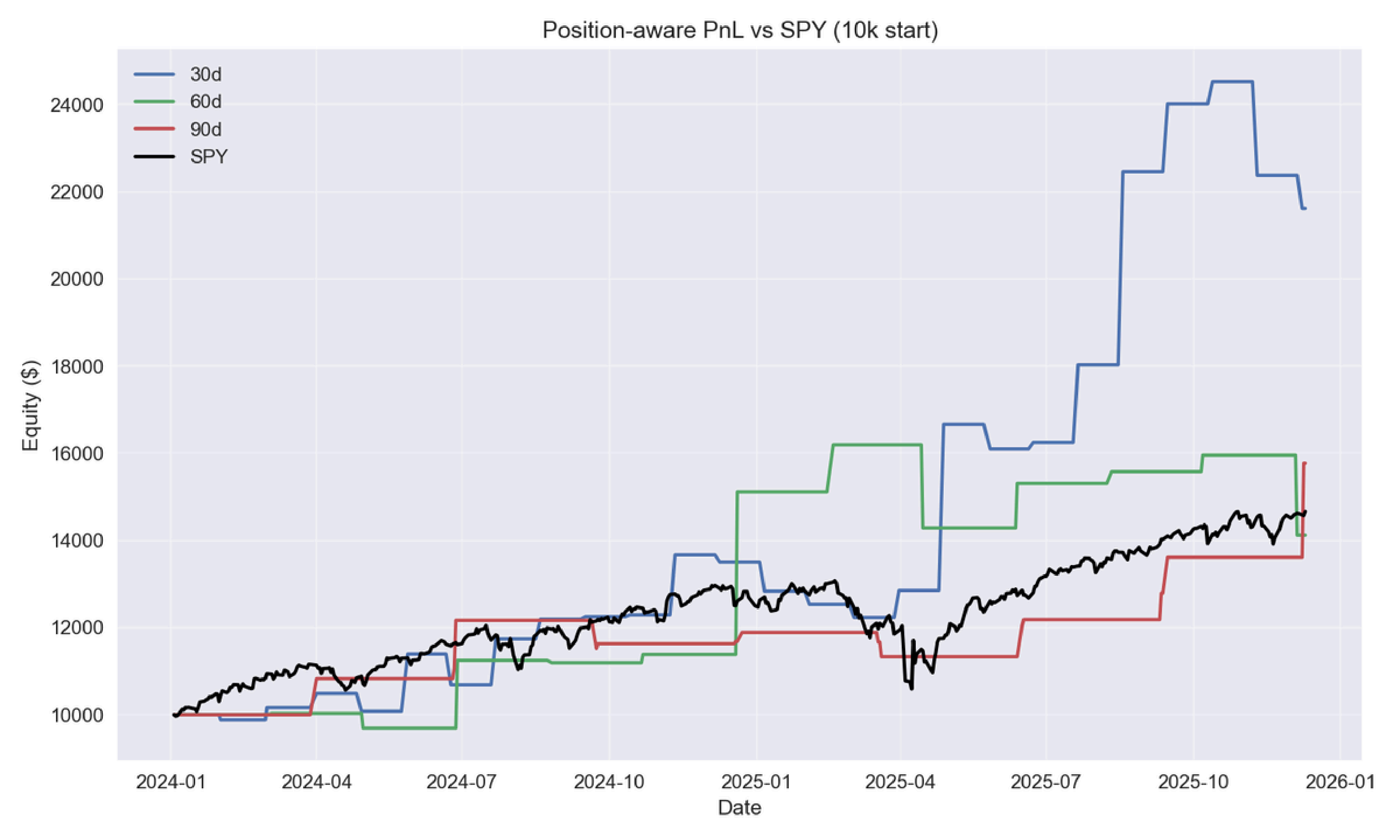
HOW IT WORKS

- Signal Generation
  - Model predicts movement of insider-trade events
  - POSITIVE → LONG BUY SIGNAL
  - NEGATIVE → SHORT SELL SIGNAL

- Execution
  - Enter next trading day
  - Hold for a fixed period (30/60/90 days)
  - Recycle capital

- Portfolio Structure
  - $10k book per horizon
  - Equal-weight signals per day

- Includes trading costs and slippage



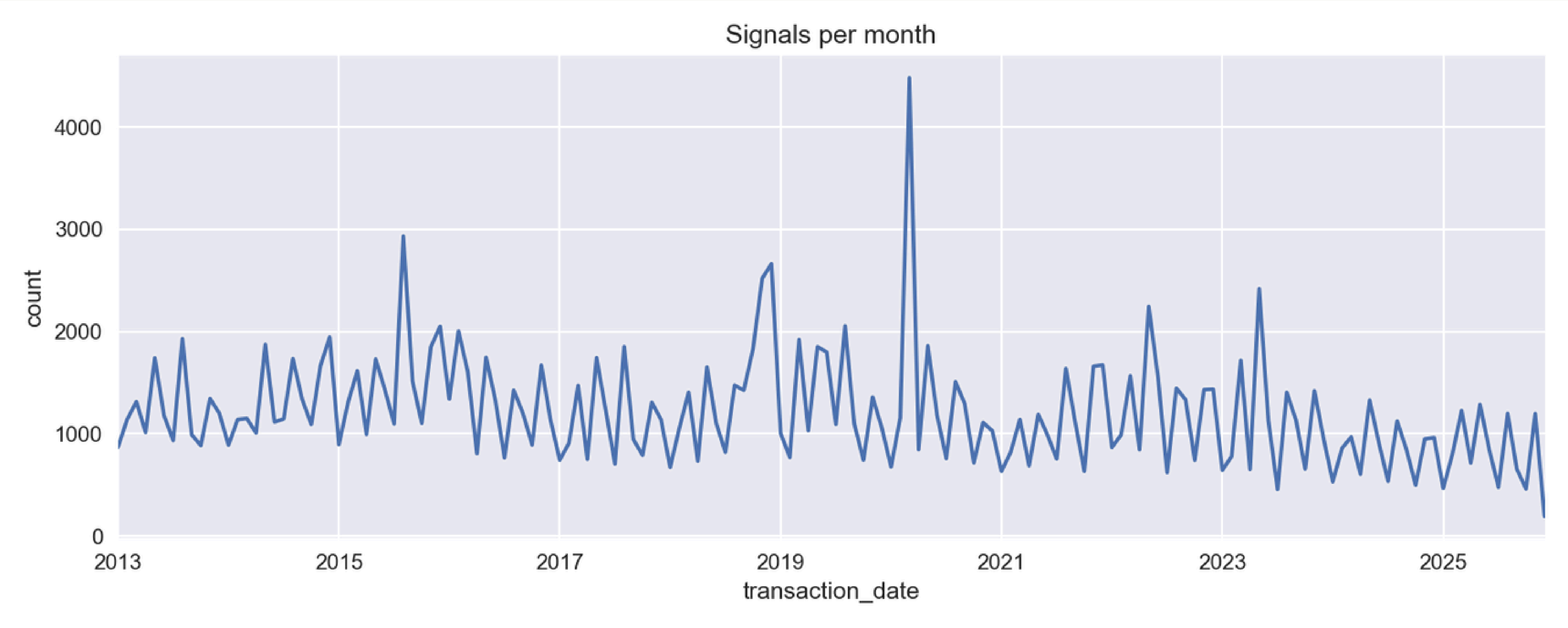
30-DAY: +116% (GREW TO $21.6K)
60-DAY: +41% **(GREW TO $14.1K)**
90-DAY: +58% (GREW TO $15.8K)
SPY ETF: +47% (GREW TO $14.7K)

SHARPE RATIO

1.31

BUSINESS INSIGHTS

Insider activity indicates a greater return on long positions relative to short positions.



The COVID period exhibited the highest volume of insider trading signals

LIMITATIONS

1. Insider motives carry noise
2. Indirect/proxy insider trading
3. Reporting lags in regulatory fillings
4. Unbalanced dataset

Demo

Portfolio-simulator

Thank you!

Any Questions?
GitHub

Lexa Fu, Kidus Yohannes, Alex Lee, Mihir Argulkar
Machine Learning in Finance
Prof Karthik Krishnan

References

Jeng, Leslie A., Andrew Metrick, and Richard Zeckhauser. Estimating the Returns to Insider Trading. The Rodney L. White Center for Financial Research, The Wharton School, University of Pennsylvania, July 1999. Web. https://rodneywhitecenter.wharton.upenn.edu/wp-content/uploads/2014/04/9919.pdf

"Insider Trading Statistics." SECForm4.com, https://www.secform4.com/training/insider-trading-statistics