



Abstract

- SHAP is widely used to rank features for pruning, but its benefit is rarely benchmarked against the serving cost it saves or on model-agnostic alternatives
- On next-day equity prediction across five tickers under standard PnL and transaction-cost assumptions, no tested horizon clears its cost break-even
- SHAP-based pruning beats Spearman-correlation and PCA pruning only in a narrow band of feature counts and is unstable elsewhere
- Because the advantageous regime is narrow, its practical value as a pruning criterion is limited

Introduction and Motivation

- SHAP (SHapley Additive exPlanations) attributes a fitted model’s output, answering what features a model relied upon in its inference [1]
- SHAP can be used as a budgeting tool: rank features by attribution per unit of serving cost, then prune
- PCA projects correlated inputs onto a set of uncorrelated components, useful for feature compression [2]

Methodology

- Trained Tensoral Fusion Transformer (TFT) and XGBoost models on next-day returns.

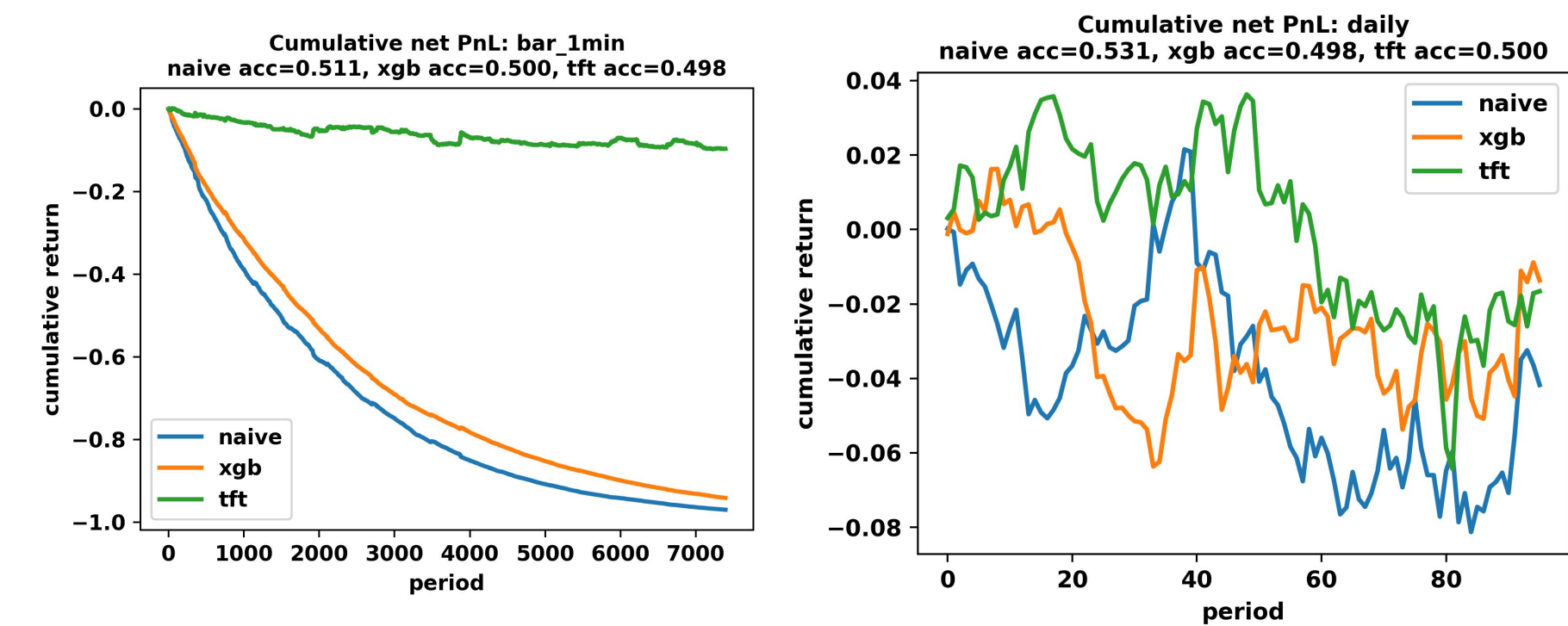


Figure A: XGBoost and TFT vs. naive baseline.

Does SHAP-Based Feature Pruning Pay for Itself? A
Cost-Benefit Benchmark on Next-Day Equity Prediction

Ayaan Qayyum¹

¹Electrical Engineering Department, Columbia University

✉ aaq2109@columbia.edu

Results

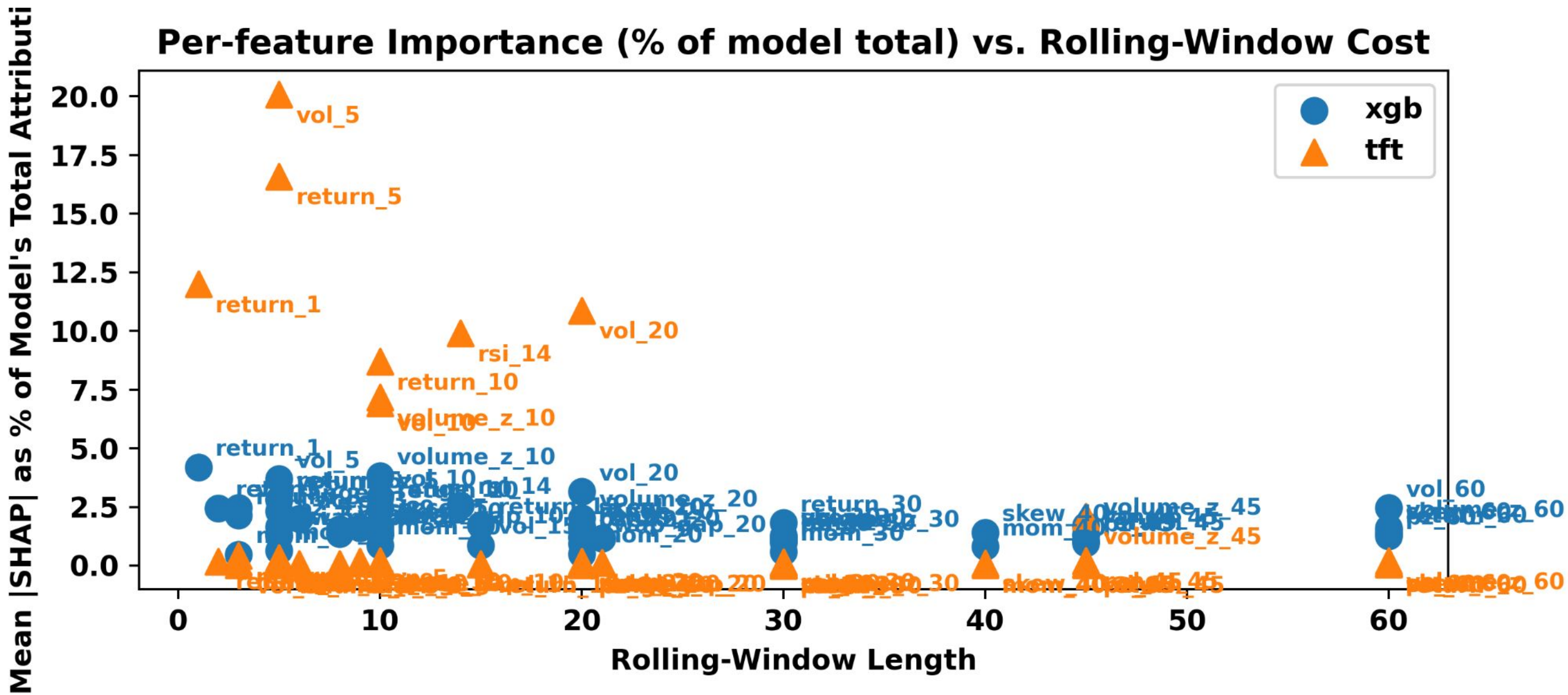


Figure B: SHAP attribution vs. each feature’s required history window on 1-min data.

- A majority of the rolling window features, computed from the same input data, have very low utilization according to the mean |SHAP| score.
- Perhaps there is a number of features to include to optimize the Sharpe score.

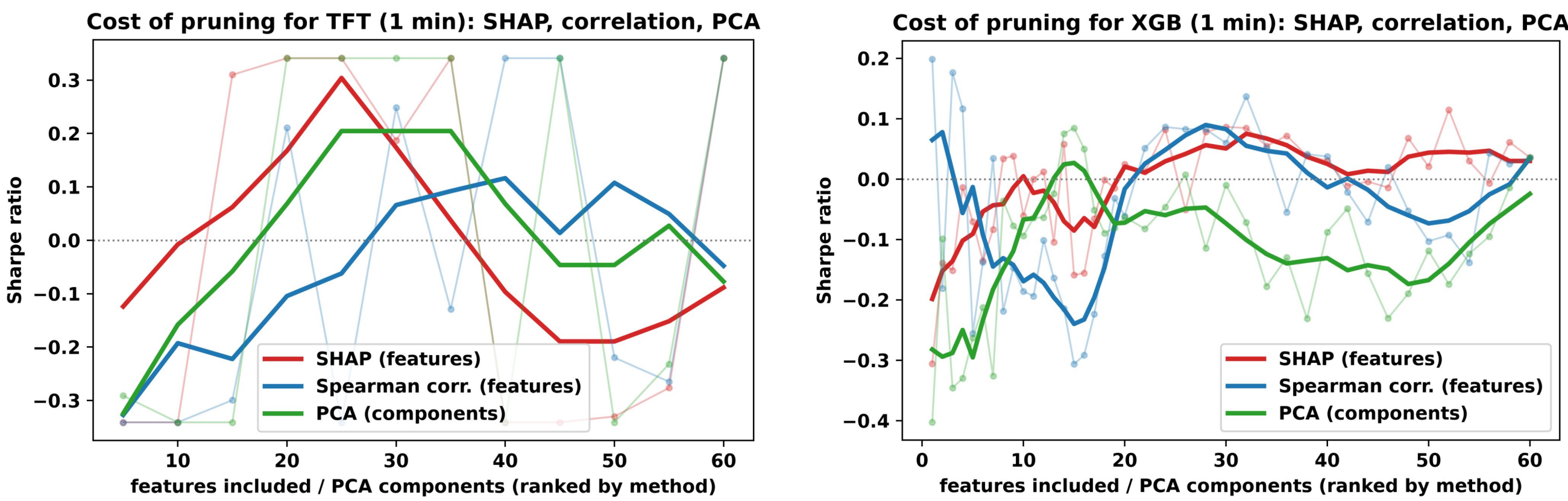


Figure C: Test of feature pruning with three methods: lowest mean SHAP score, lowest PCA component strength, or lowest Spearman correlation. Mean SHAP is the only pruning technique tested that is not model agnostic.

Data Collection Setup

- Collected Bloomberg data to test whether short-horizon equity return prediction can clear its transaction-cost break-even across multiple timescales
- Tested on five tickers: AAPL, MSFT, SPY, QQQ, TSLA

Horizon	Date Range	Ticker Rows	Train	Val	Test	Break-Ev en Accuracy
10s ticks	4/7 -> 4/20	23,421	13,437	4,479	4,479	54.20%
1-min bars	12/2 -> 4/20	37,078	21,471	7,157	7,157	51.40%
5-min bars	3/23 -> 4/20	3,745	2,150	717	717	50.80%
Daily bars	4/22/24 -> 4/20/26	500	245	82	82	50.00%

Table A: Time horizons benchmarked.

Ticker	Median Quoted Spread	One-Way Cost	Timestamps
SPY	0.14 bps	0.07 bps	70
QQQ	0.15 bps	0.08 bps	9
AAPL	0.73 bps	0.37 bps	740
TSLA	0.99 bps	0.49 bps	433
MSFT	1.19 bps	0.59 bps	928

Table B: Measured Transaction Costs. One way cost is measured with half spread.

Future Direction

- Extend to online analysis, larger timescales and horizons
- Integrate sparse, high-dimensional data from external sources, such as LLM headline text embeddings

References

[1] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 4765–4774.
[2] Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.
[3] Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748–1764.
[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.