Forecasting S&P 500 Volatility with Machine Learning Models

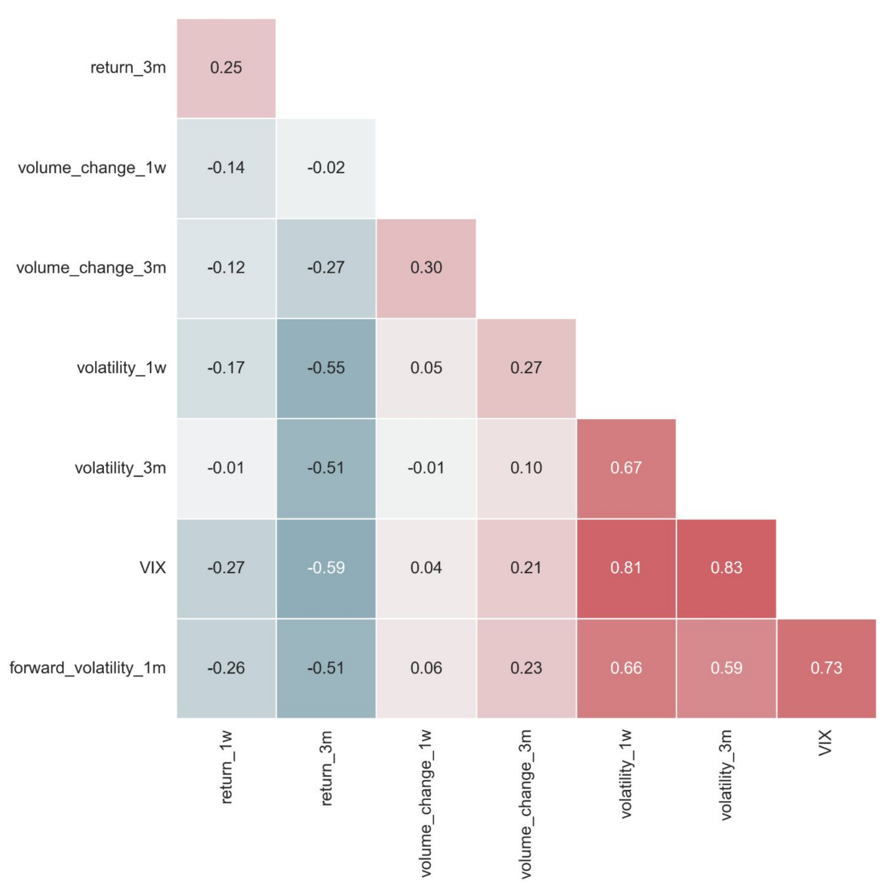
Jonah Singer

Abstract

- Task: predict **1-month ahead SPX realized volatility**.
- Features: **VIX**, realized vol, returns, **volume Δ**
- Methods: **OLS/Ridge/Lasso**, **KNN**, **FFNN**; **time-aware CV**.
- Outcome: **OLS** generalizes best and **beats VIX-only** baseline.

Data Exploration & Data Split

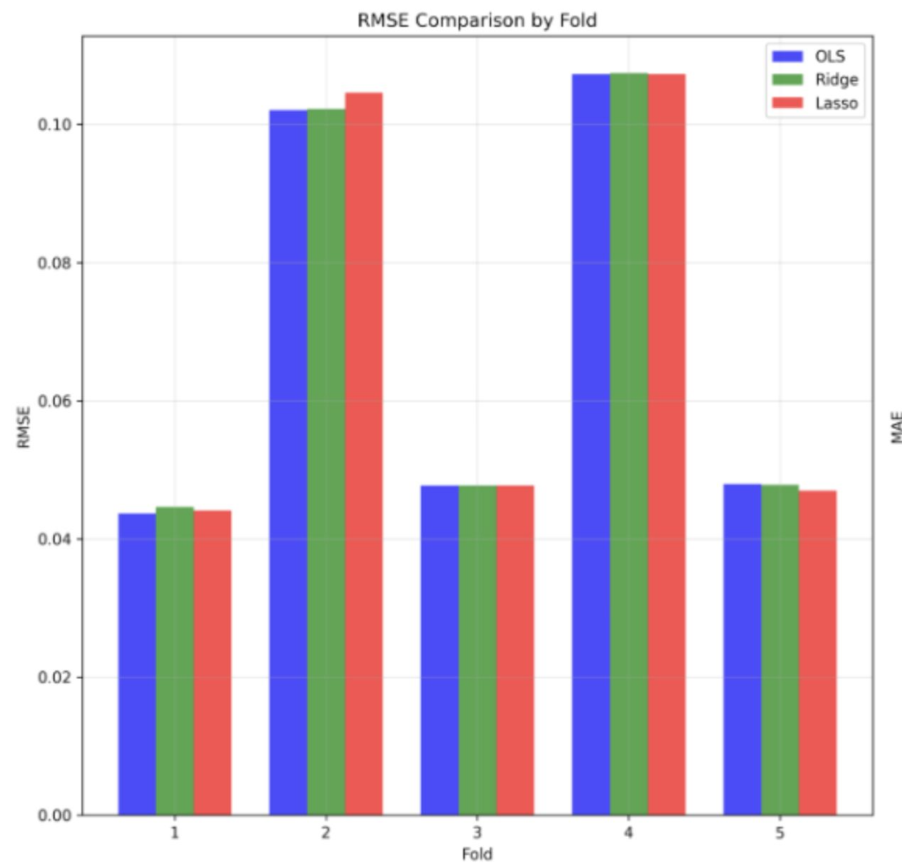
- **Stylized facts: volatility clustering**; **leverage effect** (neg. returns → higher future vol).
- **Split: forward-chaining** time-series CV (train past → validate future) to avoid regime leakage.



Feature correlations

Linear Models

- **OLS:** RMSE **0.0689**, MAE **0.045**.
- **Takeaway:** features are well-behaved; **regularization adds little** beyond OLS.



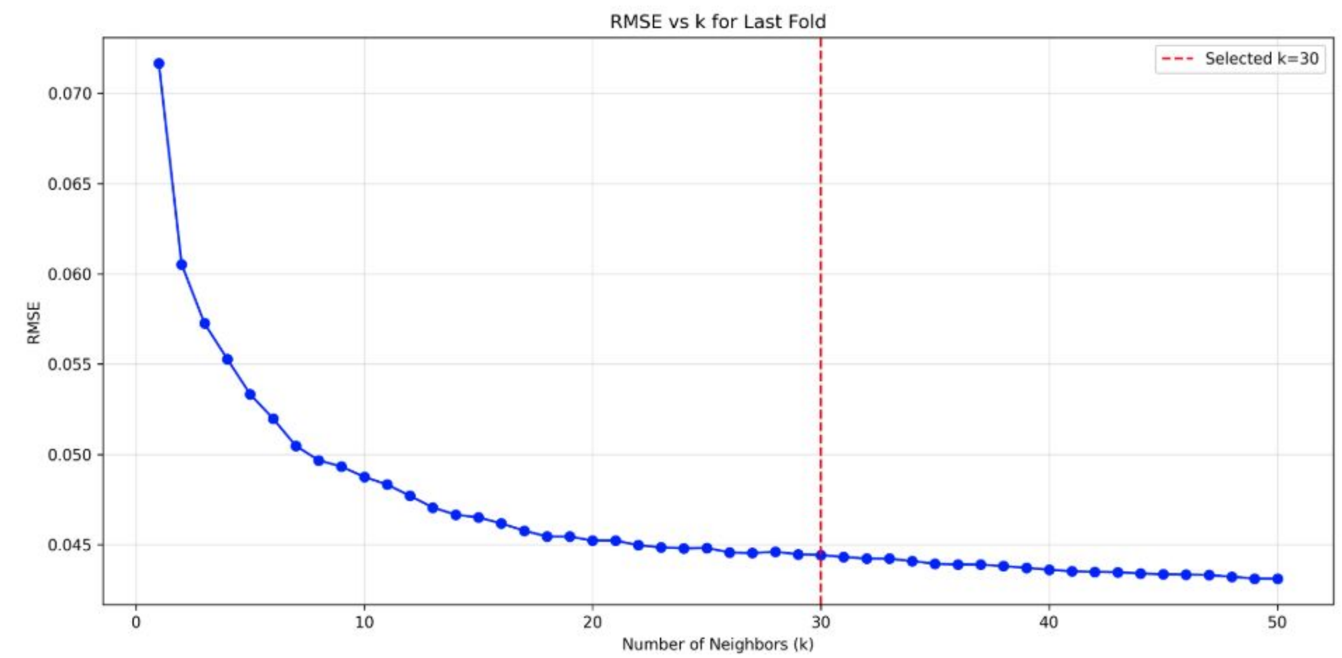
RMSE linear models RMSE

Neural Net

- **Architecture:** FFNN **64→32→1**, ReLU; weight decay **α=0.1**; **constant LR**; early stopping.
- **Metrics:** RMSE **0.0745**, MAE **0.0473**.
- **Result:** better than KNN but **worse than linear**; added complexity without gains.

KNN

- **Setup:** distance-weighted; standardized inputs; grid **k=1–100**.
- **Model selection:** elbow **~k≈30**; average selected **k=47**
- **Metrics:** RMSE **0.0749**, MAE **0.0481**.
- **Note:** smooths locally but **underperforms OLS**; sensitive to regime shifts.



Conclusion

- **What was done:** literature-driven features → leakage-safe pipeline → **time-aware CV** → compare **linear / KNN / NN**.
- **What we found: VIX strongest single input**; **returns (−)** & **realized vol (+)** matter; **volume** adds little.
- **Best model: OLS** balances bias–variance and **wins on RMSE** out-of-sample.
- **Linear OLS reduces RMSE by 11%** vs VIX alone
- **Implication: Parsimonious, interpretable** models can beat implied-volatility baselines for 1-month SPX vol.
- **Next:** semivariances/jumps, regime-switching linear models, alternate horizons.

=== Model Comparison ===

Model	Avg RMSE	Avg MAE
OLS	0.0698	0.0451
Ridge	0.0700	0.0453
Lasso	0.0701	0.0453