Predicting Return and Volatility Distributions

Machine Learning Final Project

Jake Van Slyke & Jolie Walker – University of Oklahoma

12/7/2025

Problem Definition

- This project investigates whether machine learning can predict the distribution of short-term stock returns and volatility, not just point estimates, but calibrated confidence intervals that can quantify uncertainty.
- The key questions addressed are:
    1. Can we predict point estimates?
    2. Can we predict distributions?
    3. Do they generalize to truly unseen data?

Why It Matters

- Financial returns are noisy, and essentially random. Machine learning models rarely produce reliable point forecasts or direction predictions.
- The real area for opportunity lies in what is predictable, such as the uncertainty, volatility, and distribution of possible outcomes.
- If a model can provide well calibrated confidence intervals, we can manage the uncertainty and randomness rather than trying to predict it.

Dataset Overview

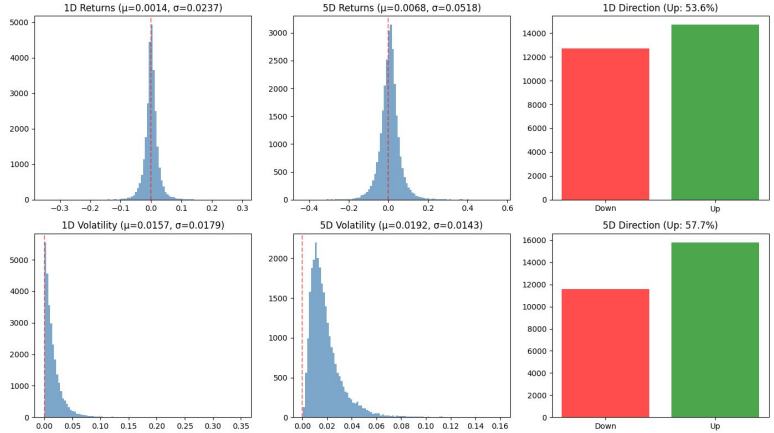
- Data source: Yahoo Finance.
- Symbols: QQQ + 10 major holdings (AAPL, NVDA, MSFT, AMZN, META, etc)
- Time Range: 2015-2025
- Size: ~27,412 samples, ~109 features
- Targets: future returns, price direction, and volatility for 1d/5d span.
- Why this data?
    - Sufficient history allows for extensive walk-forward validation, and the stocks selection are due to their prevalence and relationship to volatility, making them ideal for testing predictability.

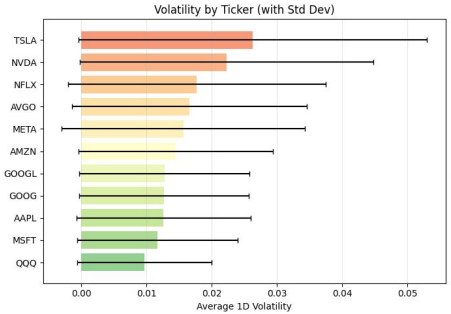
How We Prepared the Data

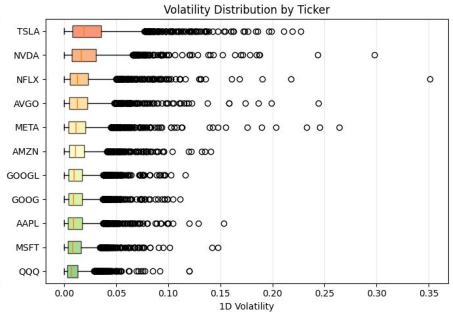
- Downloaded daily aggregate pricing data from Yahoo Finance API.
- Calculated relevant signals like price trends and volatility measures.
- Cleaned missing data (no missing data, just drop the warm up period for the rolling features like the moving averages).
- Time-based splits and walk-forward approach to avoid using future info leakage and retain time sequence.

EDA

- Returns:
  - Centered near zero - no consistent directional bias.
  - Fat tails - extreme moves happen more often than a normal distribution.
- Volatility:
  - Right-skewed - most days calm, but occasional large spikes.
- Per-Stock Volatility:
  - Wide spread between calmest and wildest stocks.
  - High-vol stocks also have more extreme outliers.
  - Model must handle very different volatility profiles.
  - Indicates that training on QQQ + QQQ top holdings then testing on QQQ could enable a model to learn the different volatility profiles.

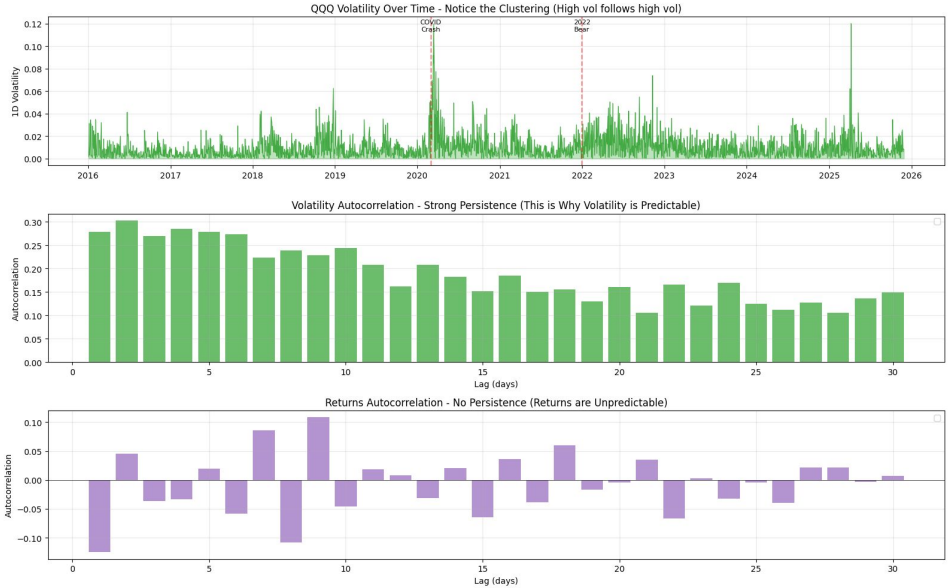






EDA – Volatility Clustering

- The key insight: volatility clusters, returns don't.
  - Volatility tends to persist and follow trends.
  - Returns have no persistence, knowing the previous days have little to no insight/trend.
- Why this matters for modeling:
  - Volatility has memory, returns don't, telling us up front to focus on predicting how much markets move, not which way.



Model Baselines

- Models Compared:
  - Ridge / Logistic Regression - linear baseline.
  - Random forest - captures non-linear patterns.
  - Naive baseline - “tomorrow = today” for vol, always guess “up” for direction.
- Key Findings:
  - Returns & Direction - all models fail equally - no better than naive, confirming minimal value in signal.
  - Volatility - RF clear performer, adding value beyond simple persistence.
- Based on these results, we target our focus to RF on returns and volatility.

RETURNS PREDICTION (Regression)

TARGET	MODEL	CORR	RMSE	MAE	R²
ret_1d	Ridge	0.013	0.03049	0.02177	-0.611
ret_1d	RF	0.021	0.02437	0.01617	-0.029
ret_5d	Ridge	0.004	0.07195	0.05353	-0.885
ret_5d	RF	0.072	0.05439	0.03792	-0.077

DIRECTION PREDICTION (Classification)

TARGET	MODEL	ACC	AUC	PREC	RECALL	F1
dir_1d	LogReg	50.8%	0.499	0.536	0.634	0.581
dir_1d	RF	52.1%	0.510	0.540	0.738	0.623
dir_5d	LogReg	50.5%	0.487	0.567	0.619	0.592
dir_5d	RF	54.9%	0.507	0.583	0.783	0.668

VOLATILITY PREDICTION (Regression)

TARGET	MODEL	CORR	RMSE	MAE	R²
vol_1d	Ridge	0.349	0.01714	0.01164	0.095
vol_1d	RF	0.417	0.01642	0.01045	0.169
vol_5d	Ridge	0.514	0.01272	0.00869	0.217
vol_5d	RF	0.565	0.01197	0.00749	0.306

Main Model Selection

- Model Selection: Quantile Random Forest
- Why predict distributions, not points?
  - Point estimates hide uncertainty - a 2% volatility prediction could mean [1%-3%] or [0.5%-5%].
  - Well-calibrated intervals tell you how confident the model is.
- Why Random Forest?
  - Handles nonlinearity.
  - Robust to outliers (good for fat tails).
- Why Quantile RF specifically?
  - Outputs full distribution (5th, 25th, 50th, 75th, 95th percentiles).
  - Intervals adapt to conditions (wider during volatile periods).

Model Results

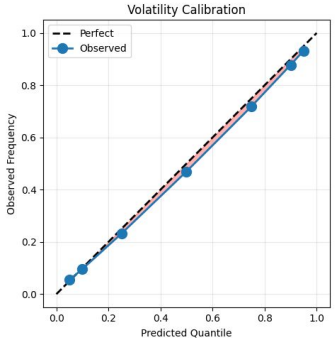
- Key Takeaways:
  - Volatility correlation ~0.4-0.6 confirms predictability established in baselines.
  - Well-calibrated intervals (90% CI contains ~88% of actuals).
  - Return correlation low, but intervals still well-calibrated, model “knows what it doesn’t know.”
  - 2020 & 2023 graphs show high and low volatility periods respectively, where CI bands are adapting well between different conditions for both returns and volatility.

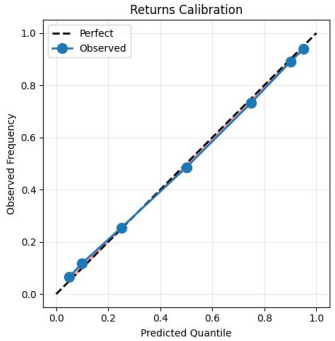
Distribution Prediction Results (Quantile RF):

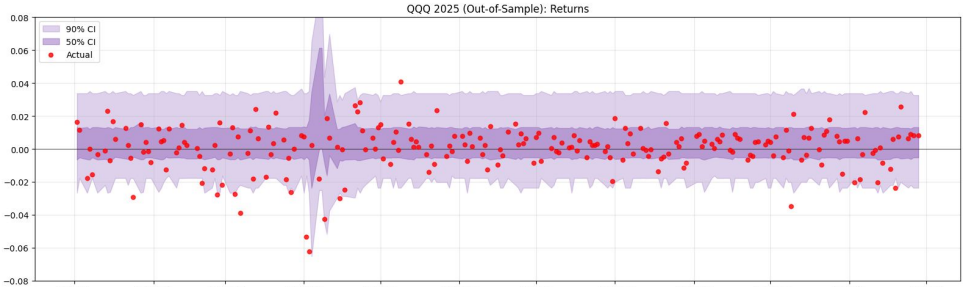
Target	Corr	90% Cov	50% Cov
ret_1d	0.045	87.3%	47.9%
ret_5d	0.061	84.5%	44.1%
vol_1d	0.405	87.8%	48.8%
vol_5d	0.553	88.2%	48.4%

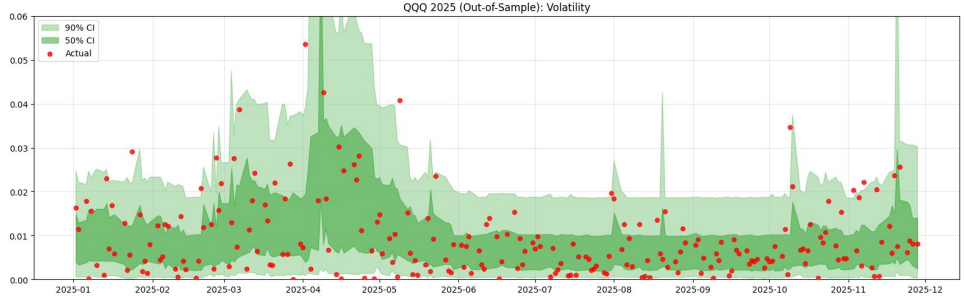
Model Results Continued...

- After retraining on full data and testing on the out-of-sample 2025 data, the model still performs and adapts during changing conditions, while retaining the calibration for confidence intervals.
    - Note: returns calibration looks near-perfect because model is being conservative with predictions.









So What can This Model do?

- What works:
  - Volatility prediction - correlation ~0.4-0.5 with confidence intervals, generalizes well and beats naive baseline and classical methods for volatility prediction.
- What doesn't:
  - Return/direction prediction - very low results consistent with baseline, no better than naive approach.
- Key insights:
  - Model "knows what it doesn't know"
  - Returns: wide intervals - don't bet on direction.
  - Volatility: tight, adaptive intervals - actionable forecasts.
- Possible applications/strategies:
  - Risk management (position sizing based on predicted vol).
  - Options pricing (better implied volatility estimates with confidence bounds).
  - Alert systems (flag when actual vol exceeds intervals).
  - Vol-timing strategies (reduce exposure when vol is high).

What Could Be Added Next

- Try more advanced models like LSTMs or Transformers
- Add option market volatility signals
- Enhance data, such as adding more predictive features and adding other sectors for wider breadth.
- Test with strategies that deal with volatility forecasts to interpret reliability and application impact.

Conclusion

- Predicting stock returns was not helpful since it was inherently random.
- Volatility yielded measurable predictive value with calibrated uncertainty - driving the potential applications for risk management.