Volatility Forecasting with Machine Learning Methods

Forecasting of future volatility is important for many economic and financial applications. Within financial markets, it has been repeatedly demonstrated that future volatility is considerably more predictable than future returns. My goal is to train machine learning-based forecasting methods for S&P 500 volatility, then compare the results of those methods to traditional forecasting approaches on holdout data.

My data set is daily S&P 500 returns for the 30 years ending April 14, 2023. In all cases, I seek to predict 21-day forward volatility (defined as standard deviation of daily returns). I use the first 80% of the time period for training, and the remaining 20% of the time period for testing. Within the training data, I use 10-fold cross-validation to tune hyperparameters. I also include a 126-day (6 month) gap between the training and test data to avoid any information leakage due to autocorrelation.

The table below lists the results of all methods attempted. Both the random forest and gradient boosted decision tree models outperform all of the classical methods of volatility forecasting, with boosted trees appearing especially promising. Creating a simple ensemble of the two machine learning models does not seem to outperform boosting.

	Tune	**R-Squared**	**RMSE**	**MAE**
Classical Methods:				
Unweighted Historical	History = 21 Days	22.5%	12.8%	7.3%
Exponentially Weighted Historical	Decay = 0.92	26.3%	12.3%	7.1%
GARCH	(1, 1)	30.8%	13.6%	8.1%
Machine Learning Methods:				
Random Forest	mtry = 30	32.0%	10.4%	5.8%
Gradient Boosted Trees	eta = 0.15, iter = 300	36.6%	10.0%	5.9%
Machine Learning Ensemble	50% RF/ 50% GB	35.6%	10.1%	5.8%