UC SANTA BARBARA
Department of Statistics and Applied Probability

Forecasting SPY Volatility with Random Forests

Ben Wolfe
Professor: Mike Ludkovski | Graduate Mentor: Emre Duzoylum

CENTER FOR FINANCIAL MATHEMATICS & ACTUARIAL RESEARCH

Project Overview

- **Description:** In finance, the variability of asset prices are uncertain, and understanding their movements is important for risk management. We seek to improve **S&P 500** (SPY) forecasting with machine learning models.
- **SPY:** The largest and most traded Exchange Traded Fund in the US and a bellwether for forecasting stock market volatility.
- **Motivation:** The traditional Heterogeneous Autoregressive (**HAR**) model forecasts volatility well given its simplicity, but we suspect that more flexible (i.e. ML) models can do better. The Random Forest (RF) framework was chosen because of its ability to handle correlated features and capture complex relationships between inputs.
- **Objective #1:** Expand the feature set beyond the autoregressive lags found in HAR to improve one-day-ahead forecasts.
- **Objective #2:** Evaluate the RF model's predictive power by constructing an option trading strategy based on the next day's forecast; compare its performance to the HAR's predictions.
- **Neural-Network Comparison:** A peer used NNs to forecast volatility, results between models will be compared.

The HAR Model

- **Realized Variance**[1]**:** We cannot directly observe the Integrated (true) Variance, but Realized Variance is an observable and consistent estimator, defined as the sum of squared intraday log-returns: $RV = \sum_{i=1}^{M} r_{t,i}^2$
- **The HAR Model**[1]**:** The simple, linear model most often used to forecast future volatility:

$$\hat{RV}_t = \hat{\beta}_0 + \hat{\beta}_1 RV^d_{t-1} + \hat{\beta}_2 RV^w_{t-1} + \hat{\beta}_3 RV^m_{t-1}$$

where $\hat{RV}^p_{t-1} = \frac{1}{p}\sum_{i=1}^{p} RV_{t-i}$ p=1,5,22

- **Problems With HAR**: Poor performance when volatility is high; limited feature set; limiting mechanistic assumption.[1]

Random Forests

- **High Level:** Random Forest models combine the results of independent decision trees, each of which consider a random subset of features and data to ensure robustness and accuracy.

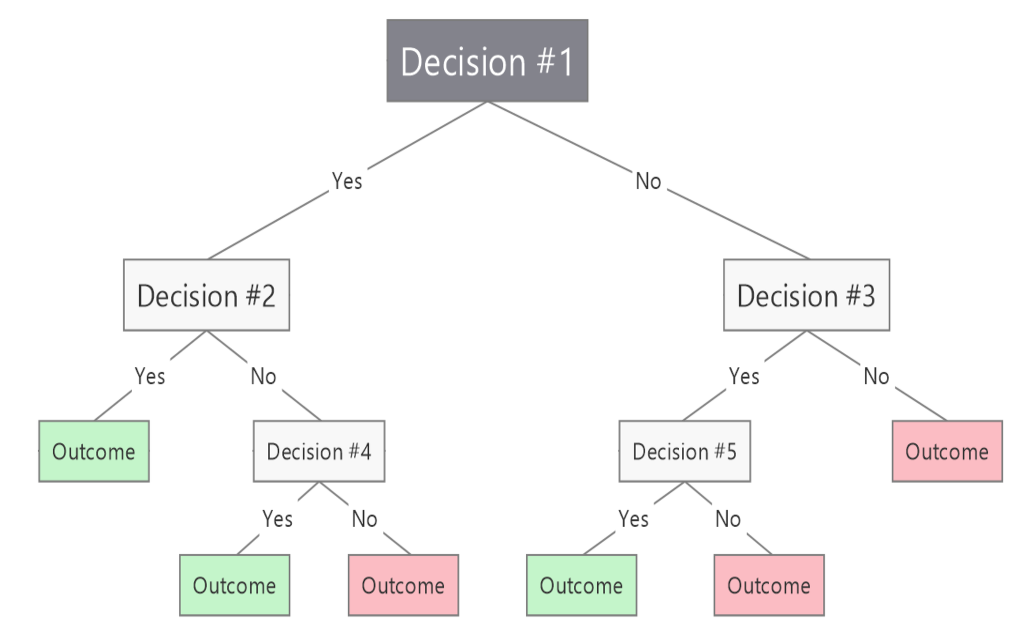

Figure 1: A single, generic decision tree.

- **Benefits:** Can handle correlated features (important for 'lags'), non-linear relationships between features and the outcome, and interaction effects between inputs. **Better predictions than a linear model**.
- **Limitations:** Computationally expensive; less interpretable than linear models; cannot extrapolate outside of training data.

Methods

- **Time-Frame:** Option data access was limited, so the dataset spans 4/1/19 - 8/10/23.
- **Rolling-Window Fit:** Both models were fit to the previous W days, optimized at **300** for HAR and **150** for Random Forests. This allowed them to adapt to each market regime.
- **Implied Vol Features:** We used the average IV of the higher strike call and lower strike put, relative to SPY's close, to include market expectations in the model.
- **Feature Selection:** Forward selection was used: all were significant except the Realized Quarticity lags which are included in many extensions of the HAR.
- **Hyperparameter Selection:** The most flexible model was selected, with low values for min samples to split and min samples per node; the binding constraint was the depth of the tree (12 splits) to prevent overfitting.

Results

- **Feature Selection:** 11 features used; RQ not significant, exogenous inputs improved performance.

Model Features	Linear HAR	RF Only RV	+RQ	+Returns	+VIX	+ATM IV & ATR
Relative R^2	**0**	**.124**	**.125**	**.265**	**.245**	**.425**

- **Feature Importance:** SHAP values use game theory to allocate credit for the model's output among its inputs (Figure 2).[2] The Kernel SHAP method allows estimation of SHAP values without fitting the model to all possible feature sets.[2] Three exogenous features performed best.

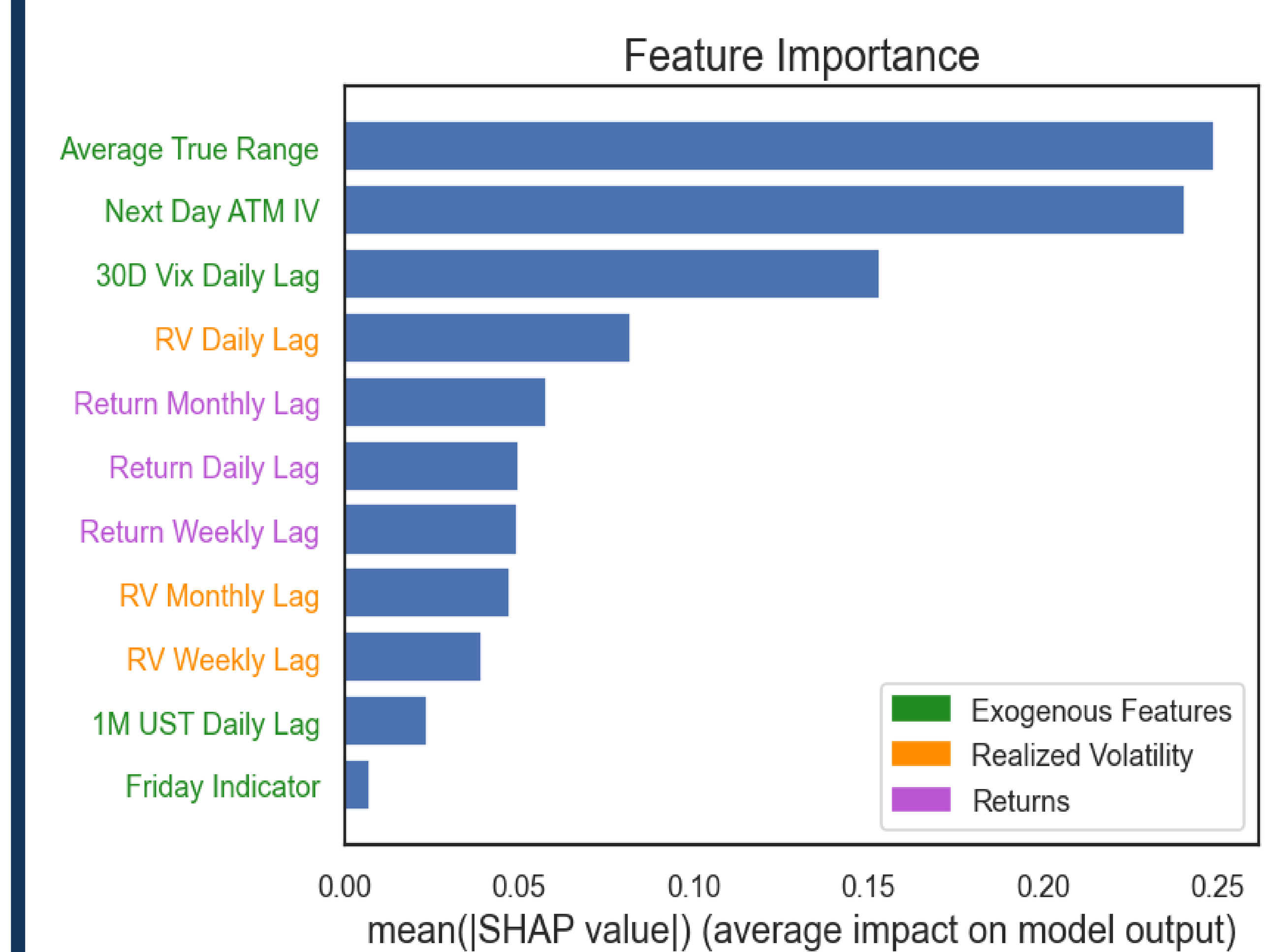

Figure 2: Feature Importance for Random Forest Model.

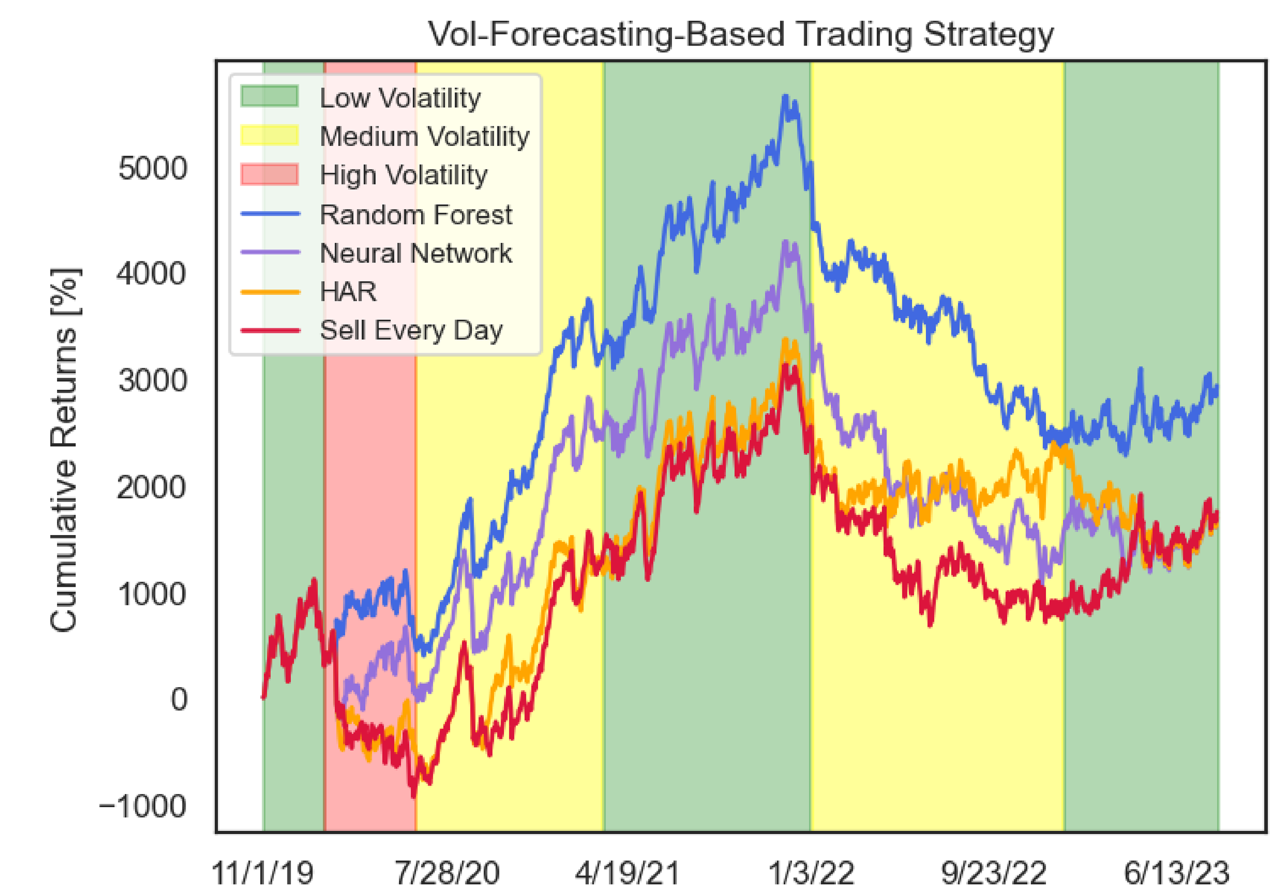

Figure 3: Results of executing the strategy based on model predictions.

- **Trading Strategy:** A model's prediction was compared to the "market's prediction" for next day volatility: if we predict more volatility than is priced in, buy an option strategy (ATM strangle) that profits from higher volatility, and vice versa. Selling the strategy is a control since the market generally overpredicts volatility. Our best RF model outperforms NN, HAR, and control.

Strategy Metrics	Daily Return	Return Std.	Sortino Ratio	Beta (95% CI)	Max Drawdown
	3.12%	**89%**	**.774**	**1.72**$\pm$4	**-80%**

Conclusions

- **HAR vs Random Forest:** Our results suggest that a RF model with enough useful features can outperform the HAR, translating to significant economic gains shown by the trading strategy.
- **Implications:** Future research should improve our short-term ATM IV feature. Recreating the 1D-VIX is suggested.
- **Neural Network vs Random Forest:** The RF model outperforms the NN, with Relative R^2 of **.425** and **.300**, respectively.

References

1. Clements, A., & Preve, D. P. (2021). "A practical guide to harnessing the HAR volatility model." *Journal of Banking & Finance, 133*, article 106285.
2. Lundberg, S. & and Lee, S. (2017). "A Unified Approach to Interpreting Model Predictions" *Advances in Neural Information Processing Systems 30*, pp 4768–4777.

Thank you to the FRAP grant program for funding this project!