Long-term Volatility Forecasting using Macro Economic Indicators

Xinyi Liu

Department of Mechanical and Aerospace Engineering, Princeton University

Motivation and Overview

Volatility is a measure of the variability of returns of assets and is an important measure of the market. Accurately forecasting the monthly volatility is critical for preparing a forthcoming financial crisis. This project aims to evaluate the effectiveness of various machine learning models on using 114 macroeconomic indicators to predict monthly volatility of S&P 500. An OLS, a random walk model, and an autoregressive model are used as benchmarks. Most of the models outperform the benchmark random walk model, and achieve a similar performance compared to an autoregressive model. In this application, linear models perform better than tree based non-linear models, where elastic net and linear gradient boosting have the highest performances. The selected features by LASSO and elastic net suggest that indicators of the labor market can provide predictive power to stock market volatility forecasting.

Data Processing

Data Source

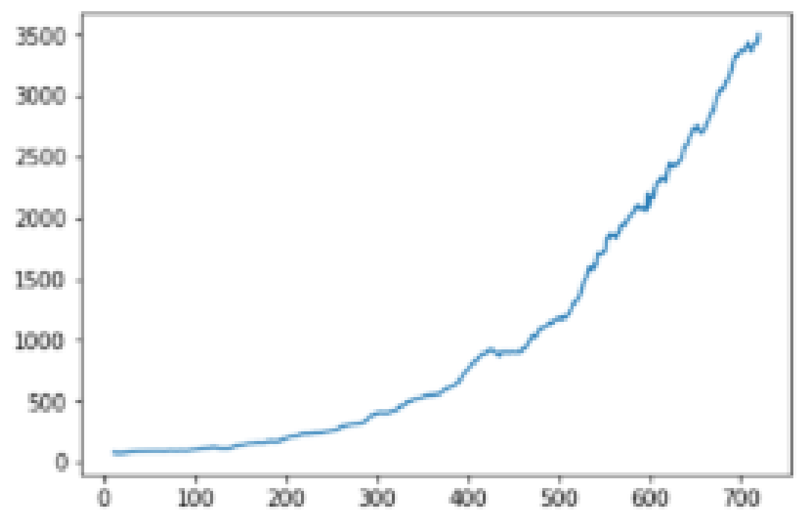
The financial data of S&P 500 was downloaded from Yahoo Finance from January 1959 to December 2018. The monthly realized variance is calculated using the daily closing price:

$$RV_t = \sum_{i=2}^{M} r_{i,t}^2 = \sum_{i=2}^{M} (log(P_{i,t}) - log(P_{i-1,t}))^2$$

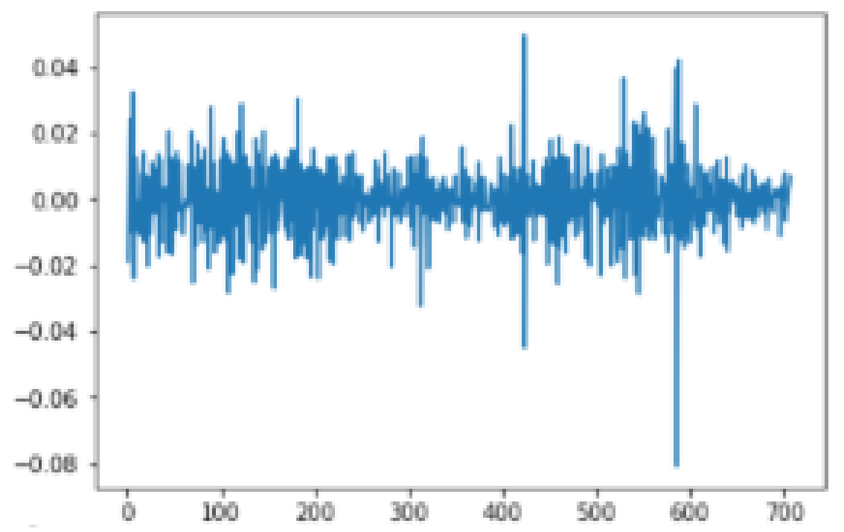
The macroeconomic data was downloaded from the FRED-MD database. The 114 macroeconomic indicators can be divided into eight subgroups: output and income; labor market; housing; consumption; money and credit; interest and exchange rate; prices; and stock market.

Data Imputation

Missing data were imputed using backward filling. Following recommendations in the FRED-MD database appendix, macroeconomic indicators are transformed to stationary.



Before transformation

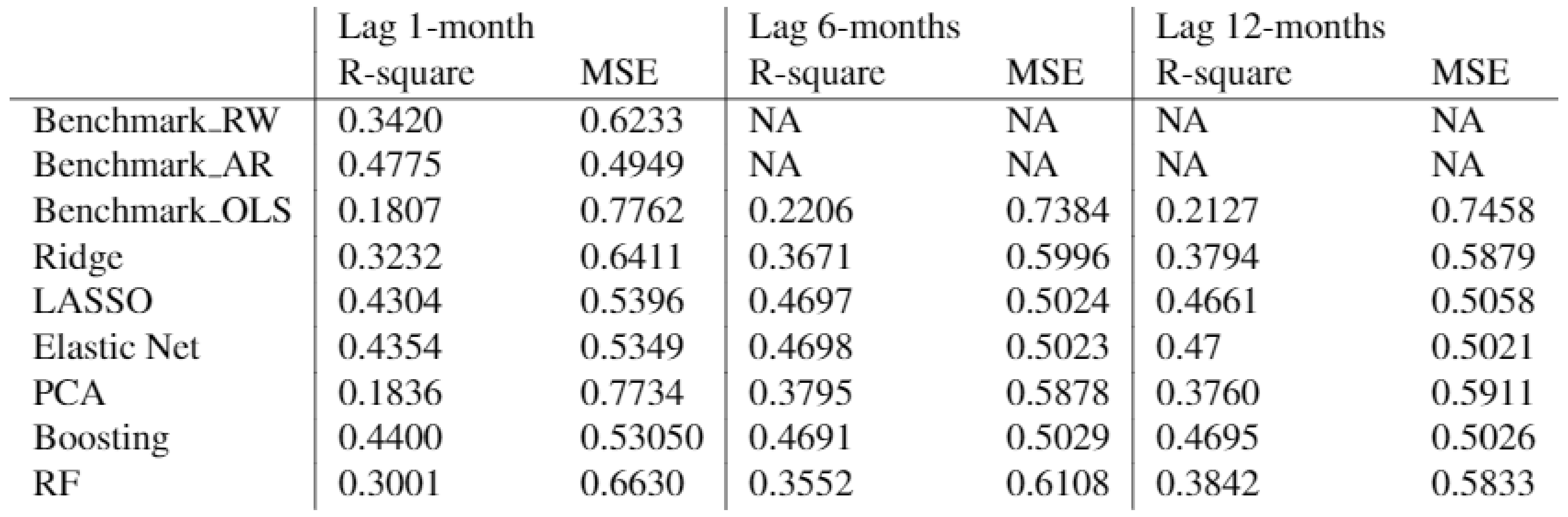


After transformation

Methods

- Benchmark models: a random walk model(simply uses past month result as the result for the current month), an autoregressive model(with 1, 6, 12 months lagged terms included), and OLS.
- Machine learning models: Ridge, Lasso, Elastic Net, Random Forest, linear boosting, and PCA.
- Evaluation methods: mean squared error and out-of-sample R^2
- A rolling window length of 30 years. The prediction starts fro January 1990 until Dec 2018.

Results



	Lag 1-month		Lag 6-months		Lag 12-months	
	R-square	MSE	R-square	MSE	R-square	MSE
Benchmark_RW	0.3420	0.6233	NA	NA	NA	NA
Benchmark_AR	0.4775	0.4949	NA	NA	NA	NA
Benchmark_OLS	0.1807	0.7762	0.2206	0.7384	0.2127	0.7458
Ridge	0.3232	0.6411	0.3671	0.5996	0.3794	0.5879
LASSO	0.4304	0.5396	0.4697	0.5024	0.4661	0.5058
Elastic Net	0.4354	0.5349	0.4698	0.5023	0.47	0.5021
PCA	0.1836	0.7734	0.3795	0.5878	0.3760	0.5911
Boosting	0.4400	0.53050	0.4691	0.5029	0.4695	0.5026
RF	0.3001	0.6630	0.3552	0.6108	0.3842	0.5833

Table 1: Out-of-sample R^2 and mean squared errors for different models, with 1-month lag, 6-month lags, and 12-months lags included.

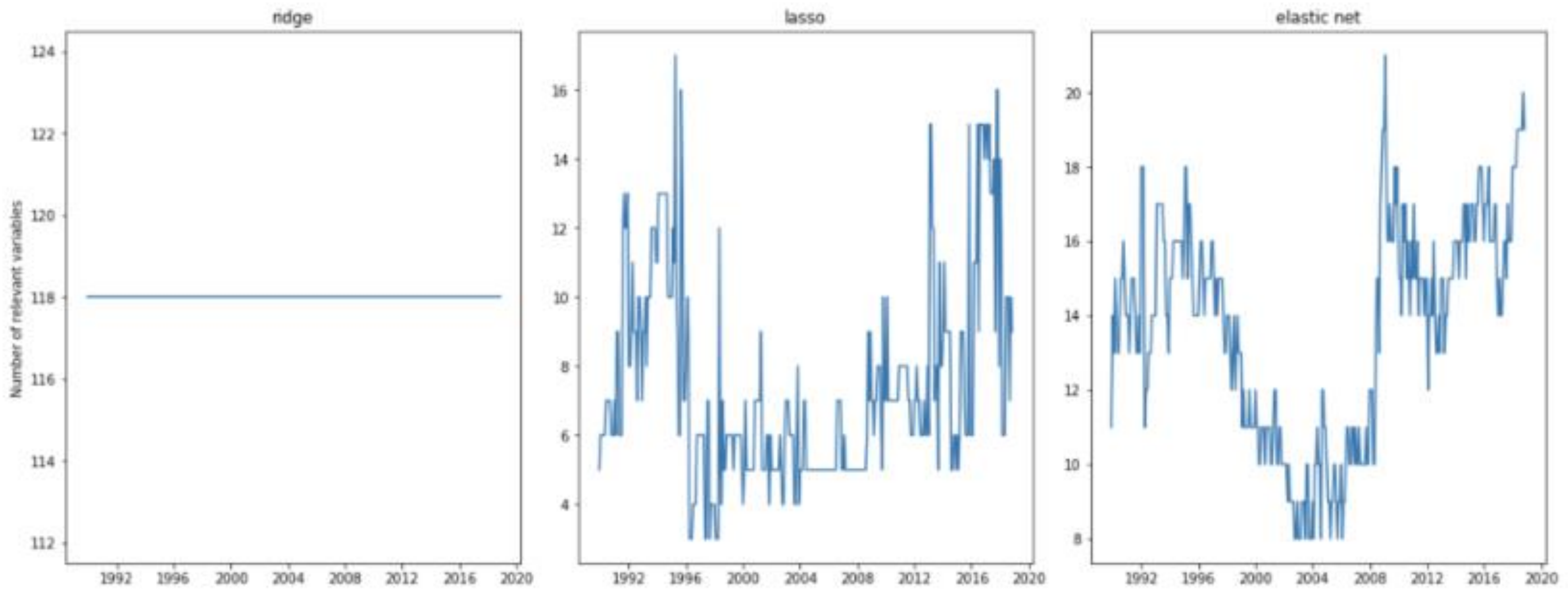
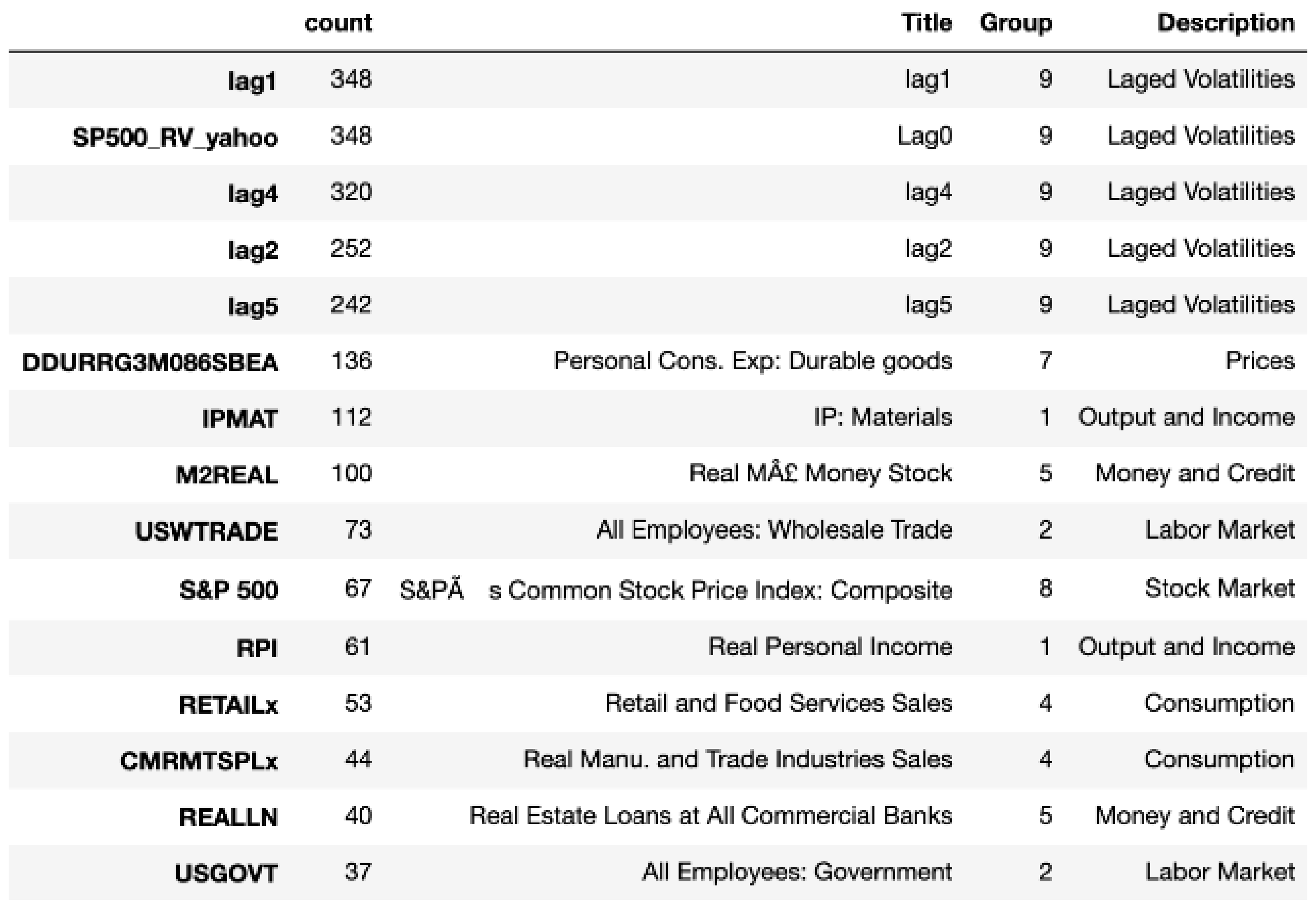
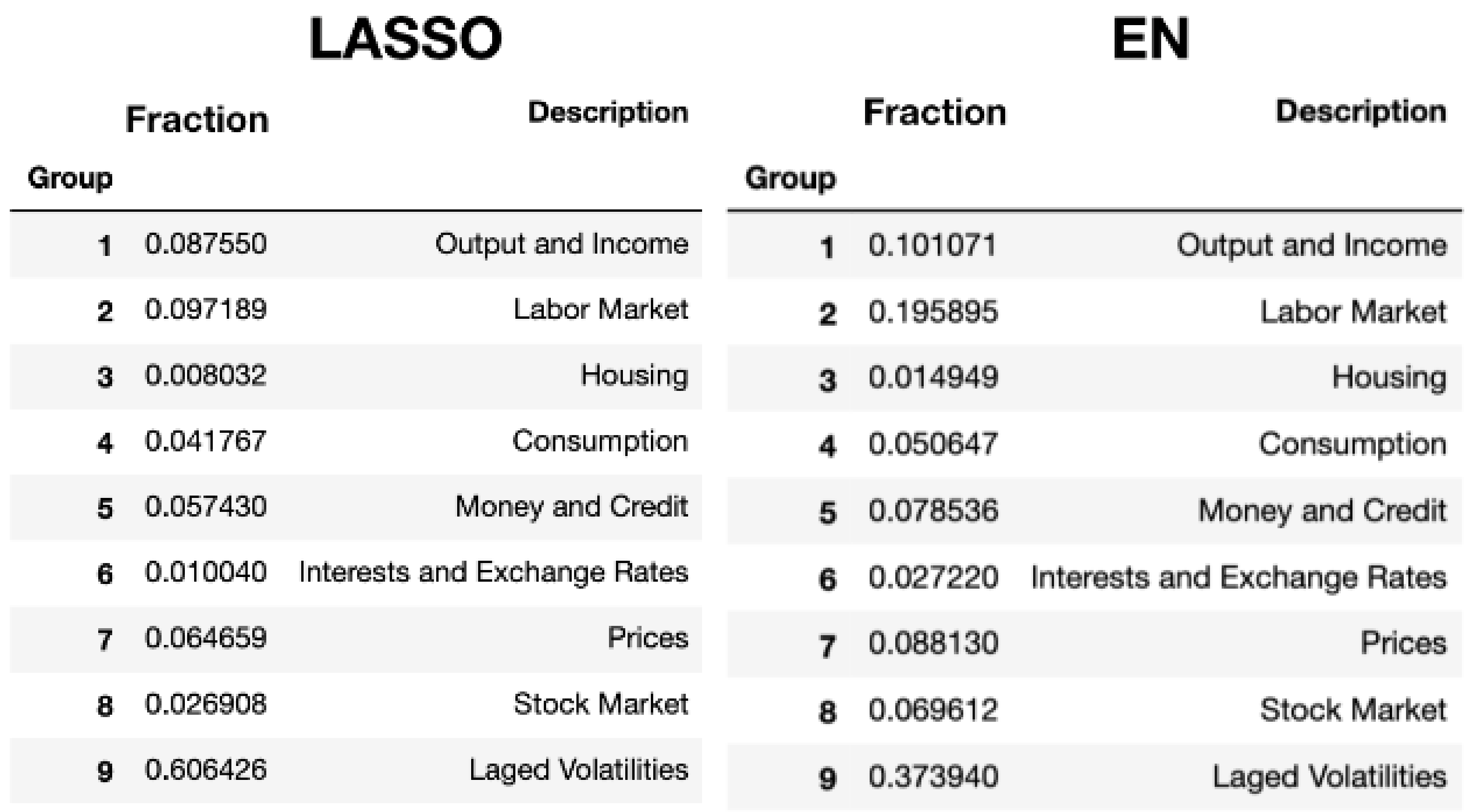


Figure 1. The number of relevant features selected by Ridge, LASSO, and elastic net using Lag 6 model for each rolling window



	count	Title	Group	Description
lag1	348	lag1	9	Laged Volatilities
SP500_RV_yahoo	348	Lag0	9	Laged Volatilities
lag4	320	lag4	9	Laged Volatilities
lag2	252	lag2	9	Laged Volatilities
lag5	242	lag5	9	Laged Volatilities
DDURRG3M086SBEA	136	Personal Cons. Exp: Durable goods	7	Prices
IPMAT	112	IP: Materials	1	Output and Income
M2REAL	100	Real MÂ£ Money Stock	5	Money and Credit
USWTRADE	73	All Employees: Wholesale Trade	2	Labor Market
S&P 500	67	S&PÄ s Common Stock Price Index: Composite	8	Stock Market
RPI	61	Real Personal Income	1	Output and Income
RETAILx	53	Retail and Food Services Sales	4	Consumption
CMRMTSPLx	44	Real Manu. and Trade Industries Sales	4	Consumption
REALLN	40	Real Estate Loans at All Commercial Banks	5	Money and Credit
USGOVT	37	All Employees: Government	2	Labor Market

Table 2. The top 15 most relevant features selected by Lasso.



LASSO

	Fraction	Description
Group		
1	0.087550	Output and Income
2	0.097189	Labor Market
3	0.008032	Housing
4	0.041767	Consumption
5	0.057430	Money and Credit
6	0.010040	Interests and Exchange Rates
7	0.064659	Prices
8	0.026908	Stock Market
9	0.606426	Laged Volatilities

EN

	Fraction	Description
Group		
1	0.101071	Output and Income
2	0.195895	Labor Market
3	0.014949	Housing
4	0.050647	Consumption
5	0.078536	Money and Credit
6	0.027220	Interests and Exchange Rates
7	0.088130	Prices
8	0.069612	Stock Market
9	0.373940	Laged Volatilities

Table 3. The relative importance of macroeconomic indicators from the 9 groups, for LASSO and elastic net. The fraction is calculated using the number of times a group indicator is selected.

Lagged volatilities are most relevant, followed by labor market and output/income.

Conclusion

- Performances of multiple machine learning models in using macroeconomic indicators to forecast monthly S&P volatilities are evaluated.
- For financial data, it is challenging to compete with the autoregressive model, as a lot of financial information is self-contained in the lagged terms.
- Elastic net and linear gradient boosting demonstrate strong predictive power. In this application, linear models tend to perform better than non-linear tree based models.
- Macroeconomic indicators associate with labor market can provide additional predictive power to the model.