Analysis of the impact of sentiment on stock market volatility

Adam Kelbl, July 2025

1. Introduction

Stock markets are highly digitized today, which opens up various opportunities for data mining, machine learning, and other data related research. It is also well established in academic research that sentiment plays a significant role in financial markets, influencing stock prices, volatility, and investor behavior. In this project, I focus on analyzing the relationship between sentiment and market volatility, as well as its potential predictive power.

I used Python and its standard data science packages for the analysis and modeling. The full code including all visualizations and the data collection process is available here.

2. Data

Collecting the necessary data was a crucial part of the project, and also one of the most challenging.

I used 4 main sources of data for my project:

1. S&P500 (SPY) OHLC prices downloaded from Yahoo Finance
2. VIX index data, also from Yahoo Finance
3. Google Search Volume Index (SVI) for 10 keywords, downloaded using the LongTrends Python package
4. Financial news headlines downloaded from Kaggle.

The dataset covers the period from 2013 to 2024, includes only weekdays (Monday to Friday), and consists of 2437 observations in total. All the data is therefore daily. It includes 12 features and one dependent variable.

The variables are as follows:

1. S&P 500 daily volatility
   - Calculated using the Rogers-Satchell volatility estimator based on OHLC prices.
2. VIX index data
   - The VIX index measures the market's expectations of near-term volatility based on S&P 500 options prices, reflecting the level of investor uncertainty or negative sentiment.
3. Google SVI for each keyword
   - "stock market crash", "bear market", "bull market", "stock bubble", "market volatility", "S&P 500", "recession", "interest rates", "inflation", "financial crisis".
   - The data is normalized on a scale from 0 to 100, where 100 represents the peak popularity of the keyword during the selected time period. All other values are scaled relative to this maximum.
4. News headlines sentiment
   - Estimated using the FinBERT NLP model which outputs the probabilities of the headline being positive, neutral and negative.
   - Overall sentiment was computed by subtracting the negative probability from the positive one (*positive-negative*).

3. Exploratory data analysis

Before modeling relationships or making predictions, I explored the data to understand it better. Firstly, I plotted the evolution of S&P 500 volatility alongside each of the 12 independent variables to get an initial visual impression of their relationships (Figure 2). Some variables appear to follow similar patterns to volatility, notably the VIX index and certain keywords like "S&P 500" or "bull market", suggesting a possible relationship.

I also created X-Y plots, with volatility on the Y axis and each feature on the X axis (Figure 3). This helped visualize the relationship and direction of them. Again, some visible relationships emerge, for example volatility tends to increase with higher values of the VIX index (which is expected, as the index itself is derived from implied market volatility), as well as with certain keywords like "market volatility" or "S&P 500".

To assess the relationships more formally, I used a correlation matrix and conducted Granger causality tests. As we can see on the Correlation matrix, several variables (their lagged values) show moderate to strong correlations with S&P 500 volatility (Figure 4). The highest correlation appears with the VIX index (0.72), followed by Google SVI for "S&P 500" (0.59), "market volatility" (0.43), and "recession" (0.45). On the other hand, variables like news sentiment and "stock bubble" show only weak correlations, suggesting limited relationship. What is encouraging is that none of the Pearson correlation coefficients showed a p-value greater than 0.05, meaning that all observed relationships are statistically significant at the 5% level.

The same conclusion is supported by the Granger causality tests, where all tested variables showed p-values below 0.05 when predicting S&P 500 volatility. This means that lagged values (in this case by one day) of each variable provide statistically significant information about next day volatility.

4. Modeling the relationship

I modeled the relationship using Linear Regression via the OLS method (using Python library Statsmodels), with S&P 500 volatility as the dependent variable and the lagged values of sentiment indicators as features. Because of the presence of heteroskedasticity, I used robust standard errors to get more reliable results. The regression results are presented in the table below (Table 1).

Variable	**Coefficient**	**Std. Error**	**P-value**
Constant	−0.0032	0.000	0.000
"Stock market crash" – Lag 1	0.0001	8.19e-05	0.146
"Bear market" – Lag 1	0.0002	6.84e-05	0.000
"Bull market" – Lag 1	−1.815e-05	1.54e-05	0.238
"Stock bubble" – Lag 1	−3.838e-05	1.01e-05	0.000
"Market volatility" – Lag 1	−4.554e-06	1.11e-05	0.683
"S&P 500" – Lag 1	2.011e-05	4.42e-05	0.649
"Recession" – Lag 1	−1.828e-05	2.77e-05	0.509
"Interest rates" – Lag 1	9.415e-05	2.85e-05	0.001
"Inflation" – Lag 1	−3.787e-05	1.32e-05	0.004
"Financial crisis" – Lag 1	5.964e-05	1.77e-05	0.001
VIX – Lag 1	0.0005	3.08e-05	0.000
News sentiment – Lag 1	1.376e-05	3.15e-05	0.663
R^2	0.574	**Adj. R^2**	0.572
F-statistic	82.51	**Prob (F-statistic)**	2.40e-170

Table 1: Linear Regression Results

Several variables show statistically significant effects on next day S&P 500 volatility at the 5% level. The strongest feature appears to be the VIX index (coefficient 0.0005, $p < 0.001$), which is expected given its direct link to market volatility. Other significant predictors include "Bear market", "Stock bubble", "Interest rates", "Inflation", and "Financial crisis", all with p-values below 0.01. In contrast, Google Trends keywords such as "Stock market crash", "Bull market", "Market volatility", "S&P 500", "Recession", and the news sentiment do not show significant p-values, suggesting insignificant relationship. The model achieves an R^2 of 0.574, meaning that approximately 57% of the variation in volatility is explained by the features.

5. Predictive ability

I then tested the actual predictive ability of sentiment by running Linear Regression and Elastic Net models (using Python library Scikit-learn) with a proper time-oriented train-test split and evaluation metrics. For the Linear Regression model, I included only those variables that had p-values below 0.05 in the previous OLS inference. The Elastic Net model was given all variables, allowing it to perform variable selection and regularization. We can see the out-of-sample results below (Table 2). The results show that the Elastic Net outperforms the Linear Regression model in all the metrics. It scores a lower MSE and MAE, and a higher R^2 score, having a better fit and predictive accuracy.

An important observation is that Elastic Net reduced most coefficients to zero, keeping only a three variables: VIX, “bear market” and “recession”. This shows that most variables had little predictive value and were excluded by the model. At least within the Elastic Net model.

Model	MSE	MAE	R^2
Linear Regression	0.000015	0.002721	0.2908
Elastic Net	0.000013	0.002585	0.4022

Table 2: LR vs EN - Predictive abilities

I am quite satisfied with the results, an R^2 of 0.4 is in fact a respectable outcome in this context. This means that our Elastic Net model is able to capture approximately 40 % of the variability in S&P 500 volatility. Below is a time series comparison showing the actual S&P 500 volatility alongside the predictions from the Linear Regression and Elastic Net models (Figure 1).

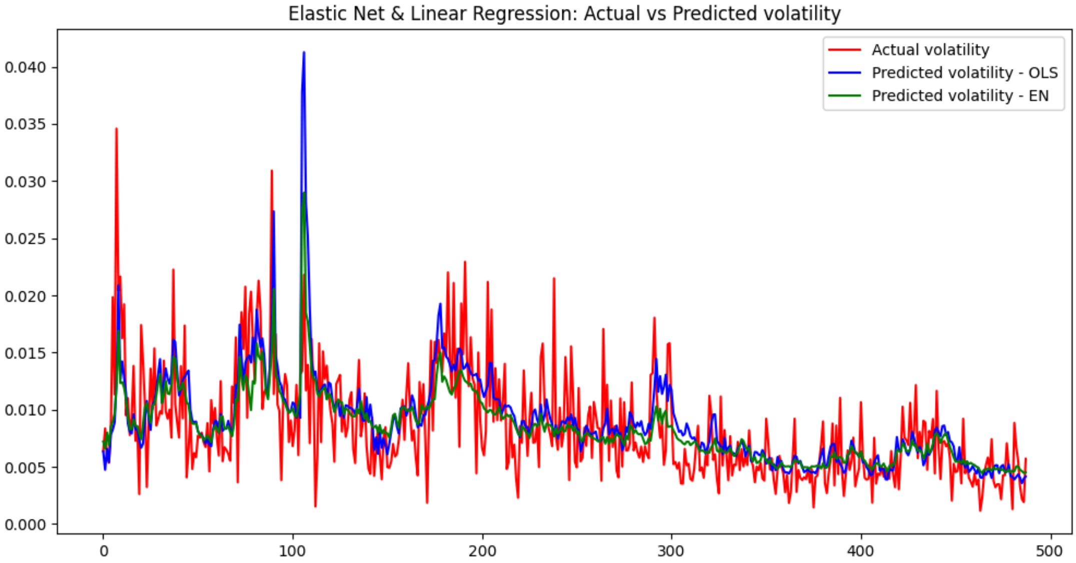


Figure 1: Time series comparison of actual S&P 500 volatility and model predictions

But do they beat the “benchmark”?

However, I have also run linear regression using only one feature, the VIX index. The results are interesting. This regression beats both the previous models with R^2 of 0.4318. This means that the VIX index alone predicts volatility better than both the full Linear Regression and Elastic Net models. So even though the complex models use many variables, they don’t outperform a simple model using just VIX. Closer look at the actual and predicted volatility time series can be seen below (Figure 2).

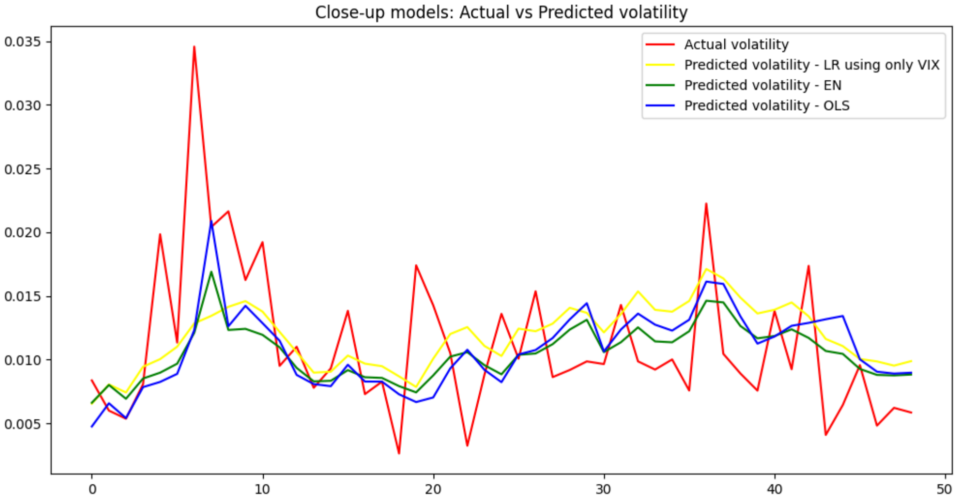


Figure 2: Close-up time series comparison of actual S&P 500 volatility and model predictions

The VIX index, being a measure of expected market volatility based on option prices, probably does have the strongest predictive power in this context, which makes sense, given that it directly reflects investor sentiment, “mood” and fear levels.

6. Summary

In this project, I analyzed the relationship between S&P 500 volatility and various sentiment indicators. I started with exploratory analysis using plots, a correlation matrix and Granger causality tests. Then I modeled the relationships using multiple linear regression (OLS), followed by predictive modeling with a train-test split using both OLS and Elastic Net. While both models performed quite well ($R^2 = 0.29$ and 0.40 - out-of-sample results), neither outperformed a simple linear regression using only the VIX index ($R^2 = 0.43$), which turned out to be the most powerful predictor.

However, the correlation matrix, Granger causality tests and partially the model results suggest that sentiment variables do carry some predictive potential, or at least show a visible relationship with volatility. This indicates that with more complex models or features, they could contribute to the volatility prediction.

Personally, I am satisfied with my work. This was my first larger project, and it was a valuable challenge. Even though the results were not exactly what I hoped for, that’s not the main point of data science. The key takeaway is that I learned a lot during the process.

Possible additional work

- Using more complex ML models, possibly XGBoost, LSTM networks, or even some econometrics models like ARIMA etc.
- Gathering a larger dataset covering a longer time period
- Expanding the set of sentiment variables, possibly including social media sentiment and more credible source of the news sentiment data
- Improving feature engineering in some way

Appendix

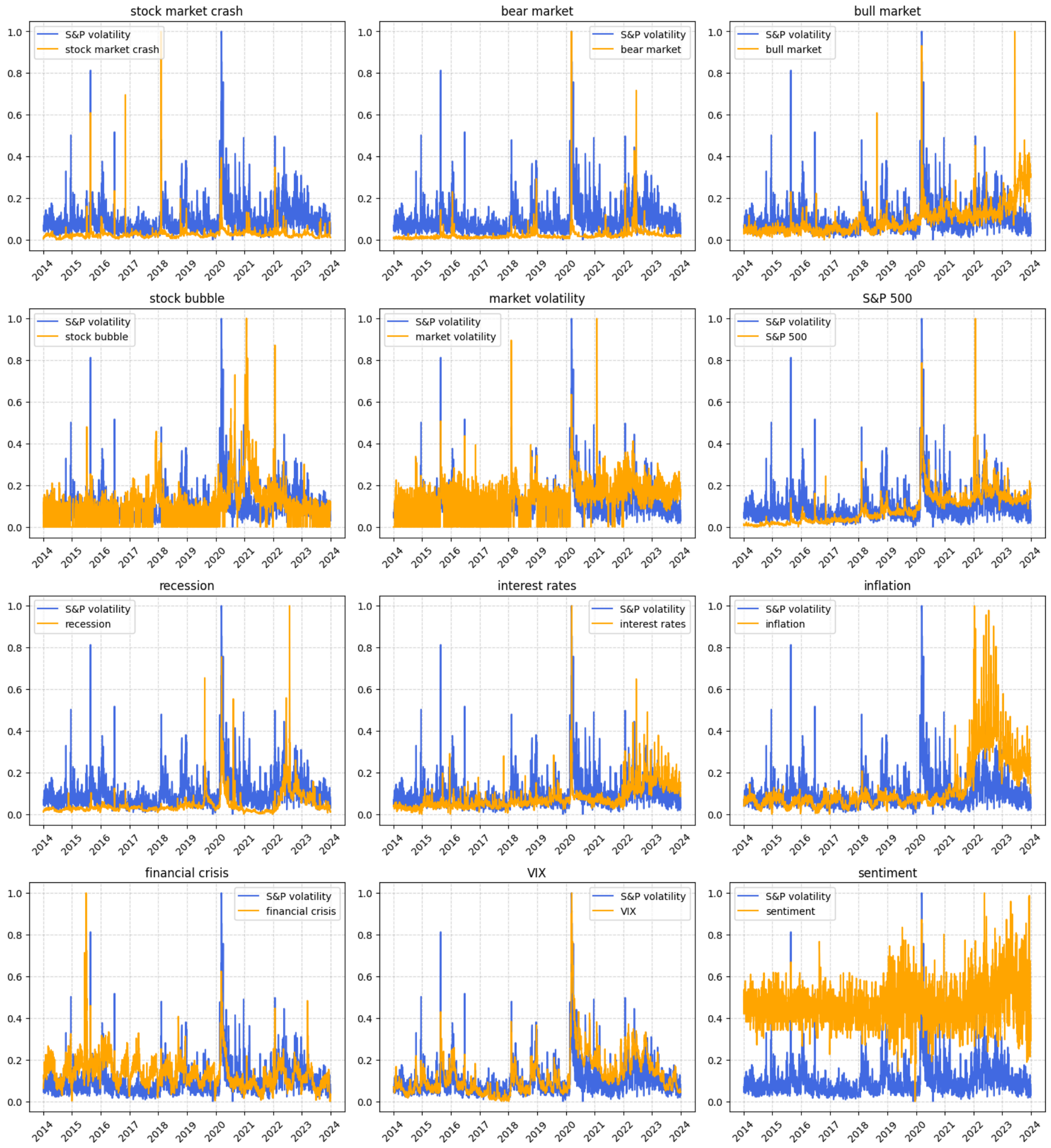


Figure 3: Overlayed time series plots showing the evolution of S&P 500 volatility alongside each of the 12 independent variables

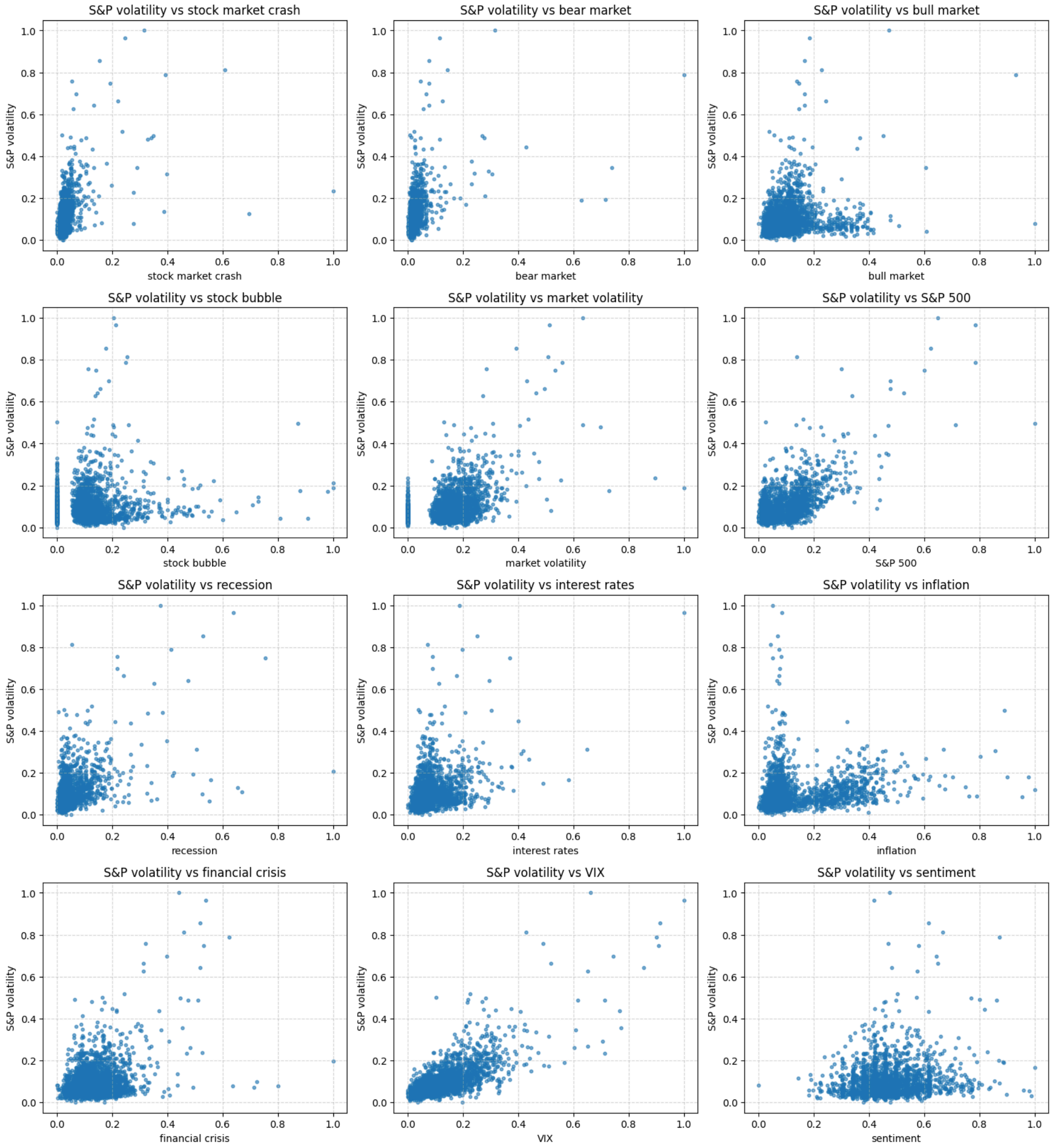
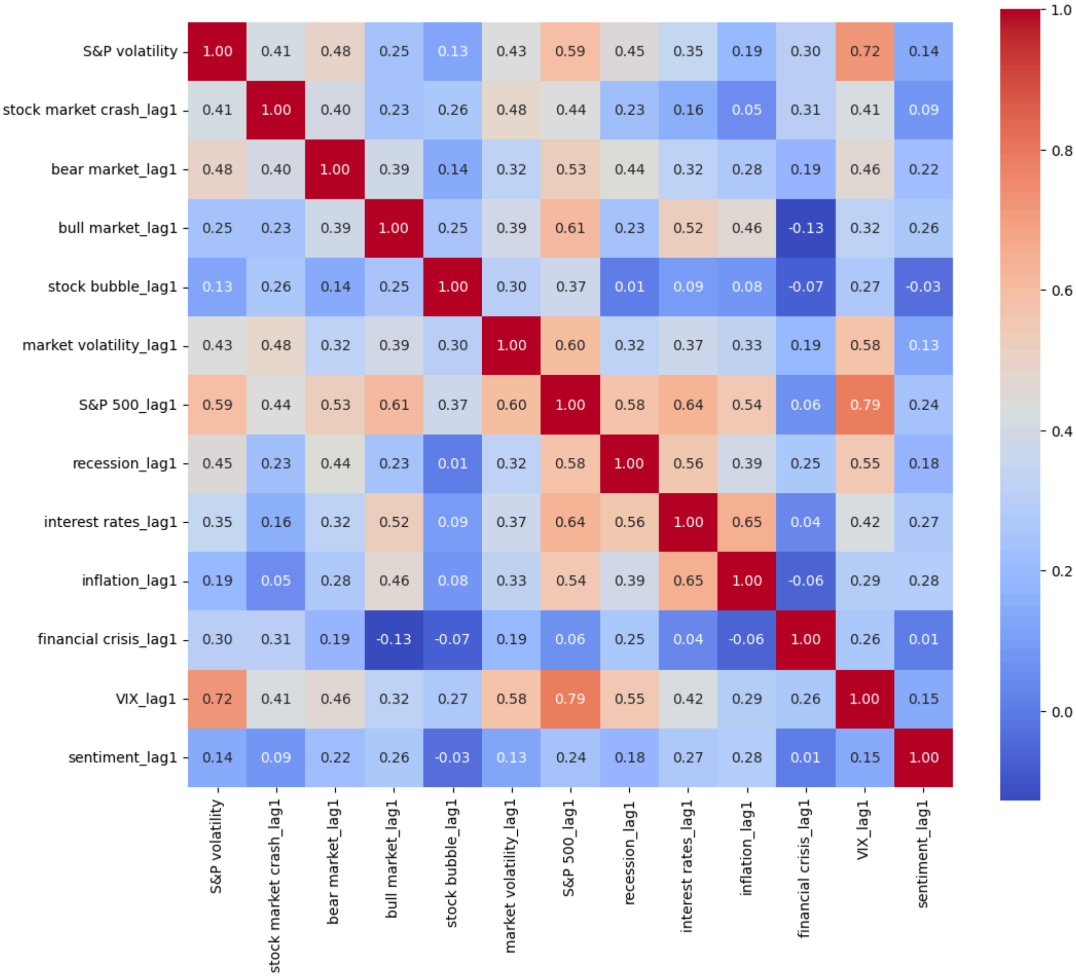


Figure 4: Scatter plots showing the relationship between S&P 500 volatility and each independent variable



	S&P volatility	stock market crash_lag1	bear market_lag1	bull market_lag1	stock bubble_lag1	market volatility_lag1	S&P 500_lag1	recession_lag1	interest rates_lag1	inflation_lag1	financial crisis_lag1	VIX_lag1	sentiment_lag1
S&P volatility	1.00	0.41	0.48	0.25	0.13	0.43	0.59	0.45	0.35	0.19	0.30	0.72	0.14
stock market crash_lag1	0.41	1.00	0.40	0.23	0.26	0.48	0.44	0.23	0.16	0.05	0.31	0.41	0.09
bear market_lag1	0.48	0.40	1.00	0.39	0.14	0.32	0.53	0.44	0.32	0.28	0.19	0.46	0.22
bull market_lag1	0.25	0.23	0.39	1.00	0.25	0.39	0.61	0.23	0.52	0.46	-0.13	0.32	0.26
stock bubble_lag1	0.13	0.26	0.14	0.25	1.00	0.30	0.37	0.01	0.09	0.08	-0.07	0.27	-0.03
market volatility_lag1	0.43	0.48	0.32	0.39	0.30	1.00	0.60	0.32	0.37	0.33	0.19	0.58	0.13
S&P 500_lag1	0.59	0.44	0.53	0.61	0.37	0.60	1.00	0.58	0.64	0.54	0.06	0.79	0.24
recession_lag1	0.45	0.23	0.44	0.23	0.01	0.32	0.58	1.00	0.56	0.39	0.25	0.55	0.18
interest rates_lag1	0.35	0.16	0.32	0.52	0.09	0.37	0.64	0.56	1.00	0.65	0.04	0.42	0.27
inflation_lag1	0.19	0.05	0.28	0.46	0.08	0.33	0.54	0.39	0.65	1.00	-0.06	0.29	0.28
financial crisis_lag1	0.30	0.31	0.19	-0.13	-0.07	0.19	0.06	0.25	0.04	-0.06	1.00	0.26	0.01
VIX_lag1	0.72	0.41	0.46	0.32	0.27	0.58	0.79	0.55	0.42	0.29	0.26	1.00	0.15
sentiment_lag1	0.14	0.09	0.22	0.26	-0.03	0.13	0.24	0.18	0.27	0.28	0.01	0.15	1.00

Figure 5: Correlation matrix with all variables