VIX: The Use of Machine Learning Methods to Study & Predict Prices for a Volatility Benchmark

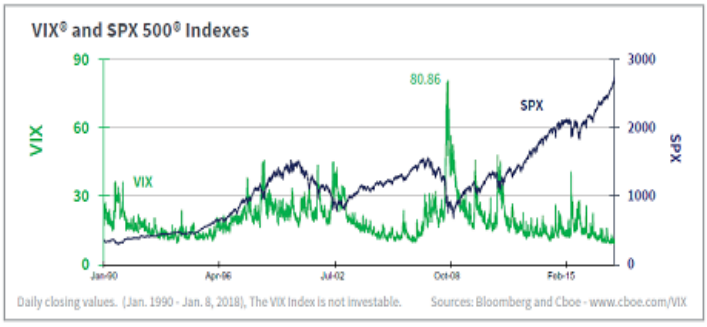
David Hamilton
NYU Center for Data Science
deh284@nyu.edu

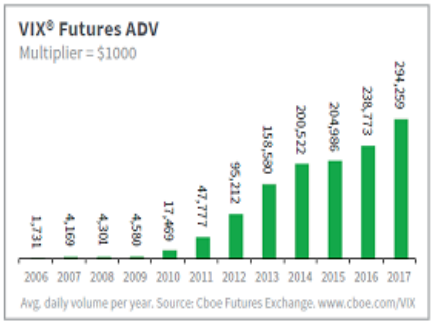
Abstract

This project is an attempt to gain new insights on a widely used, yet often misunderstood financial instrument. It looks to combine multiple machine learning techniques with time series analysis as a means of studying futures contracts on the VIX index, a barometer of stock market volatility. The end goal is to accurately forecast the price of said product or, at the very least, provide detailed information on how it interacts with related benchmark indices and derivatives.

1 Introduction

The VIX index (VIX) estimates expected short-term volatility on the S&P 500 stock index (SPX). It does so by averaging the weighted prices of SPX options (calls and puts) over a wide range of strike prices. Since 2004, there have been exchange-based futures contracts based on the price of VIX that, thanks to their rapidly expanding liquidity, have essentially established it as a stand-alone metric that can be traded and hedged.[1] Despite the product's widespread acceptance and use in recent years by institutional and retail investors alike, there have been multiple events that demonstrated the overall market might still not fully appreciate how to price and manage it properly. Focus on this relatively young, rapidly expanding area of derivatives could therefore unlock compelling information from research driven by data science.





[1] http://www.cboe.com/micro/vix/pdf/ vix-fut-and-options-cboe-vix-fact-sheet.pdf

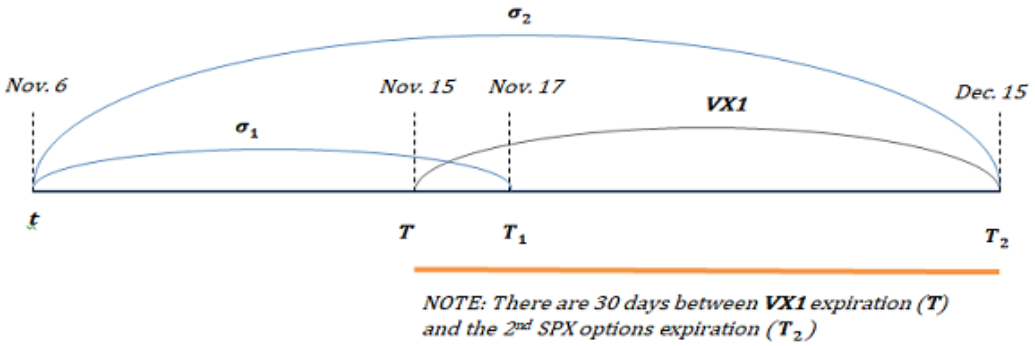
Of primary focus is the front-month futures contract on VIX (**VX1**), based on the forward 30-day implied volatilities of short-dated SPX options.[2] Fair value is derived by pricing the forward 30-day variance that underlies its own settlement values:[3]

$$VX1_t = \sqrt{\left(\frac{365}{30}\right) \cdot \left[P_t - \hat{\sigma}^2_{VIX_{t \to T}}\right]}$$

Where:

$\left(\frac{365}{30}\right)$ = Annualization factor for 30-day forward volatility
$VX1_t$ = current price of **VX1**, expiring on date T, between SPX option expirations T_1 and T_2
P_t = portfolio of SPX options with long, out-of-the-money positions expiring on date T_2 and short, out of-the-money positions expiring on date T_1
$\hat{\sigma}^2_{VIX_{t \to T}}$ = estimate of cumulative variance for **VX1** between t and T

The manner in which these relationships hold can be visualized on the following timeline using the November-December, 2017 VX1/SPX options expiration cycle:



Where:

t = current trade date
σ_1, σ_2 = implied volatility for 1st & 2nd term SPX options
T_1, T_2 = expiration dates for 1st & 2nd term SPX options

Also of interest is the behavior of VX1 with the instruments from which it is derived, as well as other closely linked indices and products. Chief among these are the SPX options underlying VIX and its futures contracts. Examining the formal

[2]CBOE Futures Exchange (CFE), http://cfe.cboe.com/cfe-education/cboe-volatility-index-vx-futures/vix-primer/the-basics

[3]CBOE Futures Exchange (CFE), http://cfe.cboe.com/cfe-education/ cboe-volatility-index-vx-futures/vix-primer/vix-features

index calculation process would require a weighted strip of calls and puts from every eligible strike price across two maturities and is too data and time-intensive for the scope of this project. A more reasonable approximation can be found by using a time-weighted combination of At-the-Money (**ATM**) SPX options whose maturities bracket VX1's expiration date. The formula for this proxy solution (hereby designated as **OPT**) is defined as:

$$OPT = \sqrt{\frac{\sigma_2^2 (T_2 - T_0) - \sigma_1^2 (T_1 - T_0)}{(T_2 - T_1)}}$$

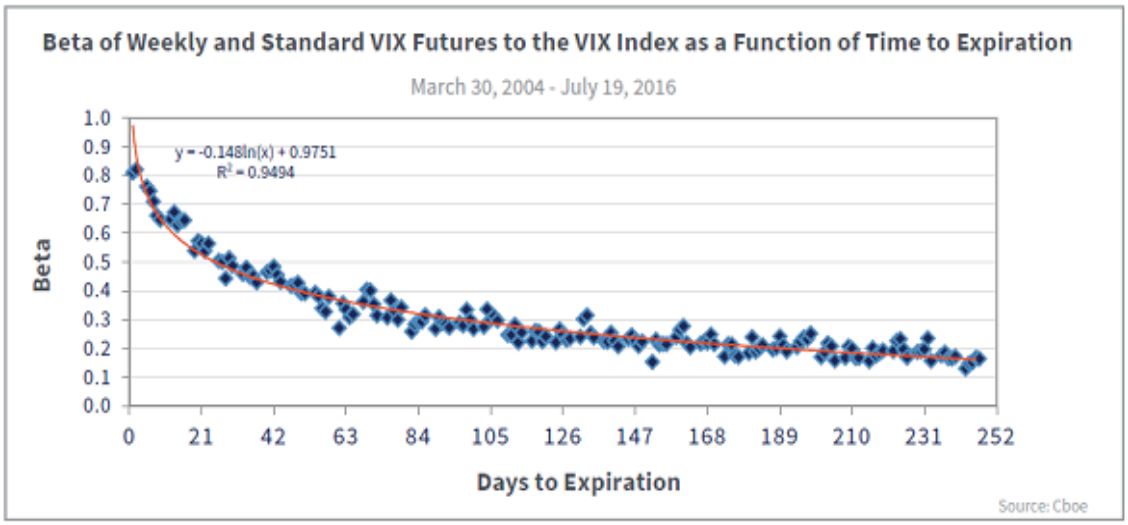
Where:

OPT = synthetic approximation of VX1 expiring on date T

σ_2^2, σ_1^2 = squared levels of implied volatility for SPX options expiring on T_2, T_1, respectively

The spread between VX1 and OPT (**VX1-OPT**) is studied as well, with the idea that a significant narrowing or widening of their price relationship could predict impending movement for the futures contract.

Two final features are the implied volatility of near-term VIX futures ($\mathbf{VX}_\sigma$) and the SPX index (**SPX**). VX_σ is used to estimate market expectation of how much VX1 might move. It is the time-weighted, composite level of implied vol for ATM options on the first two VIX futures, VX1 and VX2. The inclusion of this second maturity is twofold: it creates a smoothing effect on the levels being calculated, as options prices can fluctuate wildly close to expiration (especially those for VIX-linked products), and takes into account how contracts farther out on the futures term structure can show increasing independence from VIX while still influencing VX1.



SPX has been added because of the inverse correlation between its price and options contracts: when the stock index declines, implied volatility on its calls and puts (and therefore VIX and VIX futures) almost always increases.

2 Data

The primary dataset is an intraday time series, divided into 10-minute windows for every trade date over a two year sample period (2015-2016). This results in over 20,000 individual observations, or timestamps:

$$41 \textit{ observations per day} \cdot 252 \textit{ trade dates per year} \cdot 2 \textit{ years} = 20{,}664 \textit{ total observations}$$

Raw data components consist of historical intraday prices for multiple financial futures and options contracts, as well as one stock index. Two of these, VX1 and SPX, can be inserted directly into the master set, while the others are used to calculate remaining features OPT, VX1-OPT and VX_σ. Each of the independent variable classes is then standardized to have a zero mean and unit variance. This ensures all elements being studied will have common scaling, despite being expressed in different units of measurement.

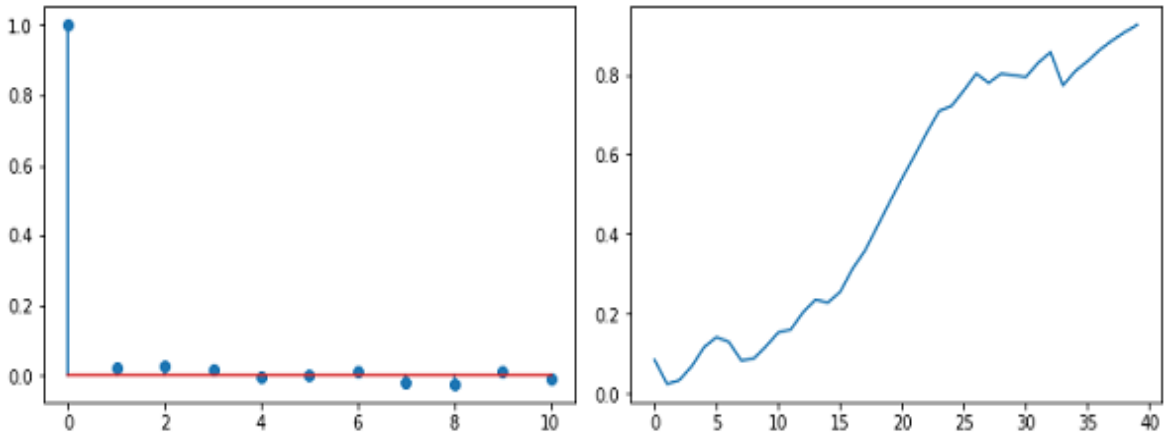
3 Modeling & Investigation

Once the master time series has been constructed, we look to build a bivariate Null model using the basic regression function $\widehat{\boldsymbol{y}} = \boldsymbol{\alpha} + \boldsymbol{\beta X}$ for each feature that attempts to predict the current, 10-minute log return of VX1, using a sample window of its own lagged log returns (n lags = 10). For the baseline null, this actually involves taking a previous snapshot of VX1 as a means of forecasting itself:

$$\widehat{ln\left(\frac{V_t}{V_{t-1}}\right)} = \alpha + \beta ln\left(\frac{V_{t-1}}{V_{t-2}}\right)$$

A lone exception is the case of VX1-OPT, where simple lagged prices are substituted in lieu of past log returns. We now run simple linear regression on the individual components to examine the quality of fit (R^2) and check for potential autocorrelation in its residuals as a means of model validation.

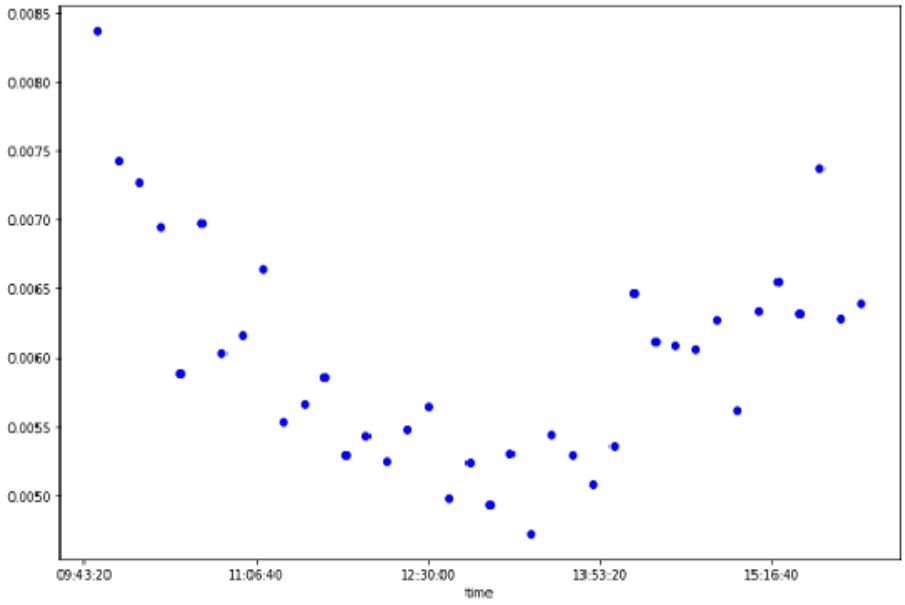
A look at the numbers shows this first attempt is unsuccessful in providing a deterministic solution for each of the features ($\boldsymbol{R^2 < 0.01}$ in every case) Our two-pronged approach to study the residuals however, produces a much more convincing result. First, we calculate the Autocorrelation (**ACF**) function for every set of observed errors and plot our findings. Second, we apply the Ljung-Box Test for additional verification:



VX1 residuals: ACF, nlags=10 (left); Ljung-Box p-values, nlags = 40 (right)

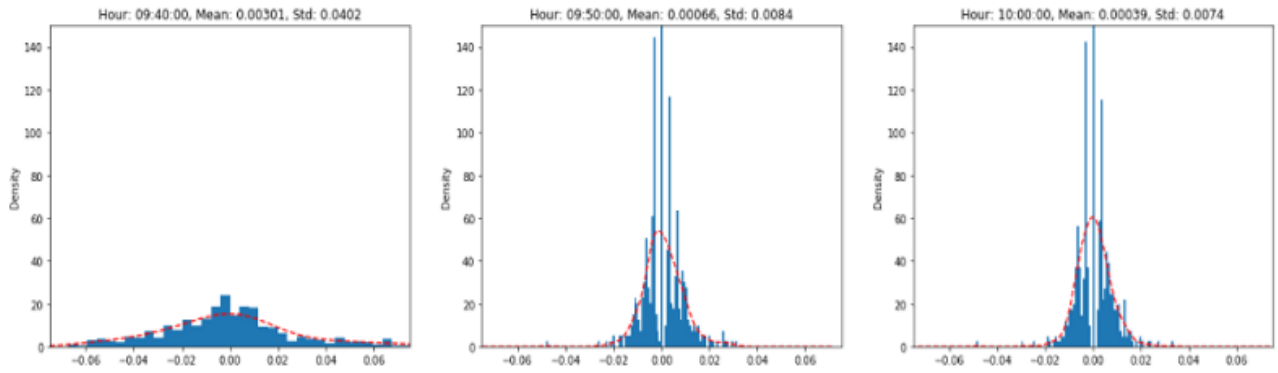
The example above (VX1 vs. its own lagged returns) displays negligible ACF values for each element of a specified window (n lags $= 10$), indicating that the residuals are inherently random. Further confirmation comes from a sample of Ljung-Box p-values (n lags $= 40$), nearly all of which exceed the 0.05 threshold that prevents rejection of the Test?s null hypothesis ($\boldsymbol{H_0}$) of no autocorrelation. Since this pattern is nearly identical for every one of the chosen predictors, we can therefore consider them all to be valid.

Before constructing models on the combination of available features, we run one last check on the data for seasonality by looking at the log returns of VX1. There appears to be a noticeable intraday pattern that emerges with increased price volatility around the market open and close:



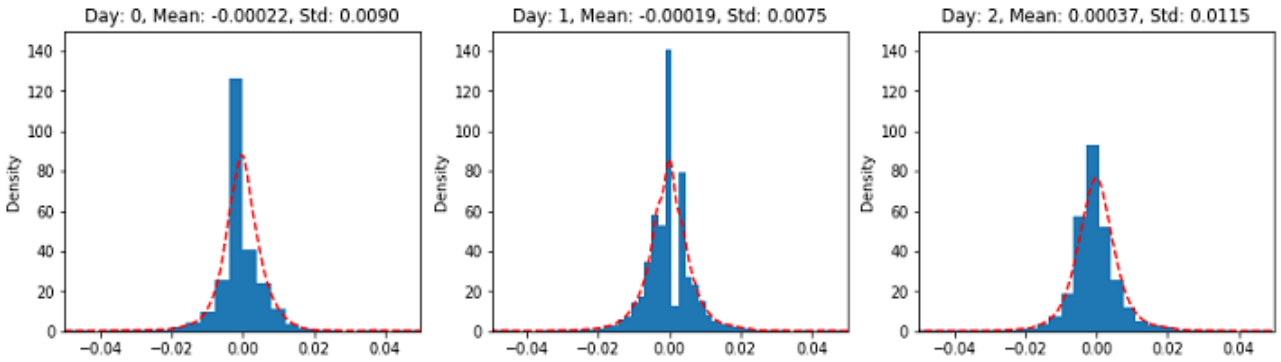
Average std. dev. of VX1 log returns for each individual 10-minute timestamp

This is further illustrated when all observed log returns are divided and viewed according to the 10-minute timestamp in which they occur during the trading day:



Histograms of VX1 log returns for (in order): 9:40am, 9:50am, 10:00am

With regards to seasonality on a daily basis, it appears that Wednesdays stand out with a more positively skewed mean and larger standard deviation. One possible explanation is that this is the day for the monthly (and weekly) expiration of VIX futures and options. While this remains relatively small in absolute terms, it nonetheless differs from other trading sessions during the week:



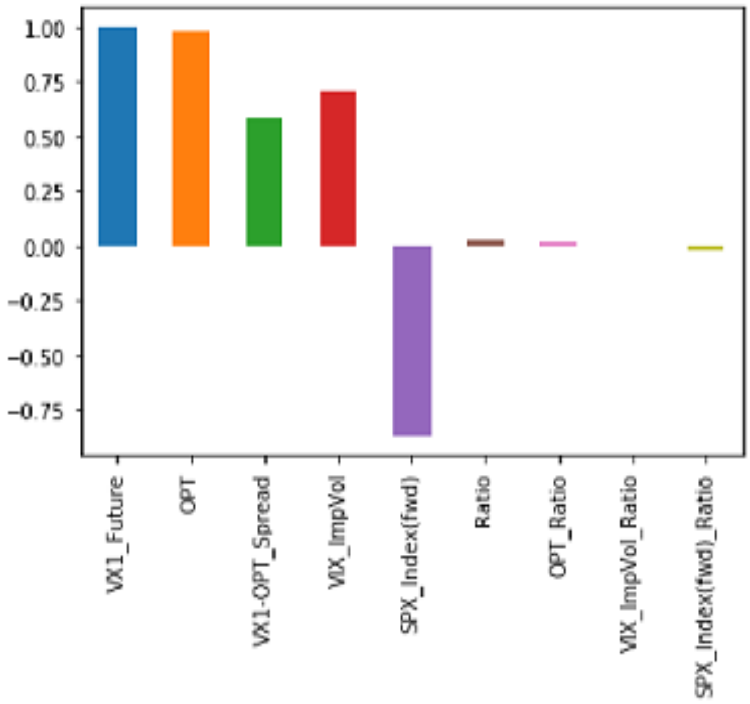
Histograms of VX1 log returns for (in order): Monday, Tuesday, Wednesday

While there appears to be proof of both intraday and daily seasonality, we judge that its presence is ultimately not significant enough to warrant immediate action. Instead, it is identified as an element of the time series that can be adjusted or removed in the future as a means of improving modeling performance.

We now turn our focus to the forecasting ability of all independent variables used together. As a primary step, this entails taking a combined set of lagged log returns (n lags $= 10$) for the previously outlined dataset components (except for VX1-OPT, which is simply lagged prices) and applying multiple regression algorithms:

Linear Regression	Random Forest	Gradient Boosting
-0.009473	0.01958	0.048059

Unfortunately, the results show little improvement over the initial null models. One takeaway, however, is the high correlation (or inverse correlation) between current (non-lagged) feature values:



Current features on the LEFT, Lagged features on the RIGHT

When contrasted against the lagged returns and the negligible relationships they demonstrate, it begs the question of whether or not a much smaller interval than 10 minutes should be used as our timestamp frequency.

In order to find some type of predictive metric, we divide the 20,000+ log returns of VX1 from the time series into percentage quartiles:

(-0.1369, -0.0033)	(-0.0033, 0.0)	(0.0, 0.0032)	(0.0032, 0.2587)
5065	8140	1986	5058

# of VX1 log returns by quartile (1st quartile = -13.69 % : -0.33 %)

These splits are then used as the new y-input for a number of classification methods:

KNeighbors (KNN)	Random Forest	Gradient Boosting	Stacking Classifier
0.404941	0.44664	0.457115	0.459091

NOTE: Stacking Classifier includes Random Forest (gini & entropy), AdaBoost & Gradient Boost

This technique displays a significant score increase from previous models. It is important to note, however, that the baseline in this case starts at 0.25 as opposed to 0.0 with the regressors (a random guess produces a 1 in 4, or 25% chance of correctly classifying a return). Still, while none of the algos produce results superior to a coin flip (0.50), they come reasonably close, forecasting 0.40-0.46.

4 Conclusion

In our study we used various machine learning methods as a means of time series analysis to gain a better understanding of VIX futures prices and attempt to predict their upcoming returns. While initial forecasting results were less than ideal, the investigation was successful in providing detail on how individual features impacted the overall process. Two areas for potential improvement are the modeling of observed seasonality in the data and the use of higher frequency data (timestamps at intervals of 1-minute, 1-second etc). This last point is of particular importance and raises the possibility that observations taken in 10-minute increments are too far apart given the high volatility at which VX1 moves.

Acknowledgments

Thanks to Gerald Hanweck and Clarence Corbett of Hanweck, the leading provider of real-time risk analytics for global derivatives markets. It is difficult to express how grateful I am for Jerry's generous donation of all intraday data used and keen interest in the subject matter, as well as Clarence's invaluable assistance with all my query needs. This project would not have been possible without them.

Thanks to Marco Avellaneda, PhD, Professor of Mathematics at NYU Courant Institute of Mathematical Sciences. Marco is one of the foremost experts on the rapidly expanding field of volatility-linked derivatives and was kind enough to lend his considerable knowledge and time to shape and guide the early stages of this study.

Thanks to Michael Gill, PhD, Moore-Sloan Data Science Fellow at NYU Center for Data Science. I was extremely lucky to have Michael as a primary faculty advisor and time series guru- someone who could provide both general project guidance and a true practitioner's knowledge for specific areas of data analysis.

Code

All supporting code can be found at the dedicated project repository on GitHub: https://github.com/NYU-CDS-Capstone-Project/VIX