Predicting Volatility in the S&P 500 through Regression of Economic Indicators

Varun Kapoor kapoorvarun1999@gmail.com

Nishaad Khedkar npkhedkar@gmail.com

Joseph O'Keefe josephokeefe3@gmail.com

Irene Qiao ireneqiao9@gmail.com

Shravan Venkatesan shravmasterice842@gmail.com

Shantanu Laghate

shantanu.laghate@rutgers.edu

New Jersey Governor's School of Engineering and Technology

21 July 2017

***Abstract*—The unpredictability of domestic financial markets negatively affects millions of Americans on a daily basis. This paper proposes a predictive model for volatility in the S&P 500 stock index based on the Elastic Net regression algorithm. For the purposes of this analysis, market volatility is defined as the residual of any day's average S&P 500 value from the 200 day moving market trend. The data generated by this definition of volatility serves as a historical training set, on which the Elastic Net regression algorithm can train. This algorithm mathematically relates an array of economic indicators, which serve as a holistic and quantifiable representation of economic conditions in the United States, to generated values of historical volatility. After this training process, the Elastic Net algorithm generates a predictive model for S&P 500 volatility that has a standardized mean squared error of 0.2300 and predicts an increase in volatility over the next three months. Because the model is trained on a data set that is updated daily, the model can be utilized in the future to predict volatility as well.**

I. Introduction

Economic and financial data in conjunction can often offer insight into the economy. In the early 2000s, the United States economy experienced a financial housing bubble, during which economic and financial uncertainty increased substantially, resulting in a sharp drop in the stock market between late 2007 and early 2009. Financial instability rose during this period, resulting in millions of Americans losing their jobs.

Volatility, a measure of how sharply the S&P 500 index value changes over a period of time, can be indicative of such events and can show how uncertain investors are in the market. Market volatility may seem difficult to analyze and predict due to its dependence on seemingly random, yet interdependent, variables. However, the aggregation of multiple economic variables, or indicators, can create accurate predictions when large amounts of economic data, the likes of which are abundant and continually gathered, are analyzed.

Machine learning algorithms can be used in conjunction with these indicators, whose data is drawn from large economic databases, in order to analyze and predict the volatility of the S&P 500 stock index. Also, since the S&P 500 is a reflection of the economy as a whole, the prediction of its volatility can be extrapolated to approximate future economic instability in general.

The utility of this stock market analysis is apparent—when the margins of financial risk change, the decisions of individual consumers, corporations, and government entities are altered. With a means of analyzing the risk at a particular point in time, each of these groups will be motivated to make more conscious investment decisions. Allowing people to make more informed decisions can mitigate the impact of bubbles and crashes, which can improve the financial condition of both individuals and larger corporate entities. In addition to benefiting the investors in a market, an algorithmic model of volatility can also illustrate the general evolution of the market over different spans of time, which can prove useful in determining the magnitude of the effect of certain financial indicators on economic progress.

II. Elements of Economic Data Analysis

A. Data Analytics and Data Science

Data analytics is concerned with the aggregation, evaluation, and analysis of large data sets in order to draw conclusions and make observations about the trends and information they contain. Data analytics requires skills from a variety of different fields: machine learning, which is a branch of artificial intelligence in which a computer is trained and tested with data; computer science; information science, which is a field primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval, movement, dissemination, and protection of

information; and statistics, which is a study of the collection, analysis, interpretation, presentation, and organization of data. Data science, while involving the same mathematical and scientific concepts as data analytics, is instead concerned more with the strategy behind analytic decisions and how they will impact a system. [1]

With the evolution of computing capabilities, the translation of raw data into insightful results has become highly valued in the corporate world. In a subfield called business analytics, data analysts use large amounts of information to help companies reevaluate and structure many of their complex processes and systems. These data analysts also create predictive models of a company's progress, which drive the corporate decision-making process toward optimal financial policies. Other applications of data analytics and data science include genomics and healthcare analysis in the medical field, evaluation and documentation of particle accelerator data in experimental physics, and criminal activity documentation and prevention in the legal and military fields. [2]

B. Financial Markets

The Standard & Poor's 500 (S&P 500 or SP500) is a stock index of 500 US companies, ranging from the technological to the industrial. The degree of change in the index is calculated through a weighted average of changes in the stock prices of the constituent companies. In order to be listed on the index, these companies must either be listed on the New York Stock Exchange (NYSE) or the National Association of Securities Dealers Automated Quotations (NASDAQ). [3] The S&P 500 mirrors general stock prices in the US economy because it includes both growth stocks, which are shares that have values that are expected to grow, and value stocks, which are shares that trade below their expected value. The inclusion of these shares makes the S&P 500 the ideal representative index. It is often called the "bellwether" of the US economy. [4]

Economic instability in a financial market is marked by a series of extreme and unpredictable market fluctuations, but it is a broadly defined term, not mathematically constructed. Economic volatility, on the other hand, is the degree of deviation, or difference, from expected market returns, and so can be calculated with historical data from changes in the S&P 500.

C. Python

Python is a programming language typically used for general-purpose programming, but also has a number of libraries that makes it a useful tool for data analysis. Many of these libraries, such as numPy, pandas, and matplotlib, contain numerous built-in functions that facilitate data mining, manipulation, analysis, and visualization. Scikit-learn, a robust machine learning library in Python, also allows manipulated data to be extrapolated into powerful predictive models. It allows for the high-level and efficient implementation of an assortment of complex machine learning algorithms. Changing the algorithm implemented can be done with a few changes in the code, which makes it easy to compare many algorithms side by side. [5]

D. Machine Learning

There are three types of machine learning: supervised, unsupervised, and reinforcement learning. The model presented by this paper utilizes supervised learning, in which the algorithm is provided with the correct values of the target variable during training. This allows it to adjust its own parameters while training to reduce its own error so it can improve the generated model. Within the subset of supervised machine learning methods, there are classification and regression models. Classification deals with categorical variables and classifies a variable into one of many distinct groups. An example of this is the separation of spam emails from non-spam emails in the inboxes of users of an email service. Regression machine learning models, on the other hand, analyze continuous variables to determine a relationship between them and a target variable. Since volatility can be output on a range of values as opposed to discrete categories, using regression to model the relationship between variables over time was the most favorable for the purposes of this project. [6]

III. Choosing a Measure of Volatility in the S&P 500

A. Purpose

The goal of this project is to create a model that can accurately predict future volatility in the S&P 500. However, in order to achieve this, historical levels of volatility must first be defined. A set of volatility measurements is required to train any regression algorithm because without a data set of volatility values, there would be no target variable for the algorithm to train with. This aspect of gathering and using past data for the model is incredibly important because an error in this dataset would result in a flawed model and in turn, a flawed prediction.

B. Understanding the Ambiguity of Volatility

Theoretically, there is an infinite number of methods that can be used to measure volatility in data formatted as a time series, as volatility is simply a measure of change; therefore, it must be put in the context of a base value or pattern. The current volatility of the S&P 500 can be measured as today's change relative to its movement last week, last week's movement relative to this year, or even this year's movement relative to the last ten years. There is truly no consensus on what "volatility in the S&P 500" even means; thus, measuring volatility is almost entirely up to interpretation.

C. The Different Measures of Volatility

The most explicit and accessible measure of current volatility is the absolute value of the daily percent change in the S&P 500. This measurement can be generated through a simple equation, as demonstrated in Equation (1), with V_n as the value of the S&P 500 on day n.

$$Percent\ Change\ (\Delta\%) = \frac{|V_n - V_{n-1}|}{V_{n-1}} \times 100 \tag{1}$$

A second and widely used measure of volatility is the Chicago Board Options Exchange (CBOE) Volatility Index (VIX). The VIX is a measurement of *implied* volatility, meaning it does not measure current volatility, but expectations of volatility for the next 30 days. The VIX also fluctuates with option expiration dates, as the index considers these dates in the determination of individual option volatility. This means that the index is practically an aggregate of the implied volatility of options available on the S&P 500. [7]

A final option for measuring volatility in the S&P 500 involves creating a custom definition that would measure uncharacteristic changes in the value of the S&P 500. This can be done in multiple ways, many of which require creating a base measure of what would be considered "characteristic" of the market. This would involve establishing a statistically defined range of "acceptable" changes, or by measuring the deviation of the indexes value from an established trend.

D. Selecting the Correct Method

Simply using daily percent change as a measure of volatility resulted in flawed and, in some cases, completely wrong measurements. This was because daily percent change itself was, ironically, very volatile. Though the data showed spikes for extremely volatile days, if value of the next day had little movement, it was characterized as non-volatile. Volatility should have been analyzed from a less nuanced point of view; it should have been characterized as existing or not existing in more extended periods of time, not just in daily spikes. A sample of daily percent change measurements shown in Fig. 1. shows that measuring volatility by daily percent change resulted in very frequent fluctuations that were difficult to fit a model to.

A second issue with daily percent change as a measure of volatility was that it failed to consider market trends. To illustrate, if the S&P 500 had two different large increases in value, the second large increase in value should not be considered as having high volatility, as it is consistent with market trends. Daily percent change recorded such an occurrence as an instance of volatility.

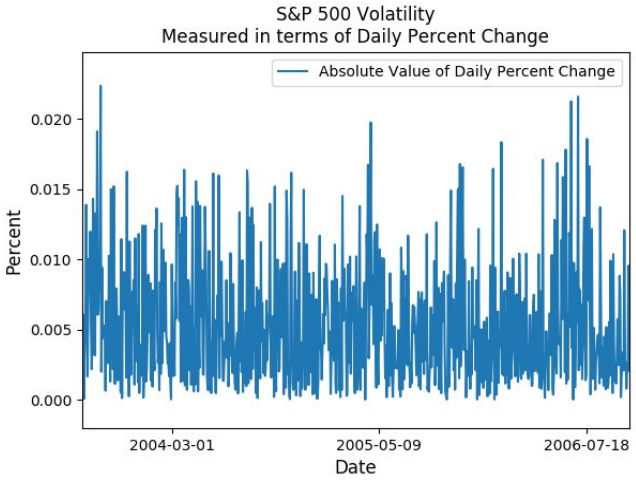

Fig. 1. Daily Percent Changes in the S&P 500.

With respect to existing volatility tools, the VIX is a relatively accurate predictor of volatility, but not one suitable for the purposes of this model. Since the VIX only attempts to predict volatility rather than measure it, it is an inappropriate tool on which to base an algorithmic regression model. Also, the VIX measures volatility using a process fundamentally different from that of this paper, involving high level economic measurements such as option price and implied volatility from current market realities. This analysis, which aims to predict volatility exclusively in the S&P 500, used economic indicators that measure realities at a lower level. [7]

The only measurement of volatility that provided the trend-sensitive, real time, and consistent data that this analysis required was a volatility measurement that was customized to match the indicator data.

E. Creating a Custom Measurement of Volatility

The first consideration taken into account to construct an instrument to measure volatility was market trends. The general pattern of the S&P 500 had to be understood in order to determine what makes a given day volatile. This was done by a simple linear regression—a line of best fit—of the past 200 S&P 500 index values for each day. The line generated from such a regression quantified the current "trend" of the market. Volatility can be understood simply as deviation from this implied market trend, or the difference between the real S&P 500 index values and the predicted S&P 500 index values, as shown in Fig. 2. The recorded volatility for a sample day in the S&P 500 is represented by the dotted residual.

Though using market trends through linear regression residuals had less variance than daily percent change, the data still contained daily spikes which were not suitable for regression analysis. Fig. 3. displays the volatile nature of the measured day by day residuals.

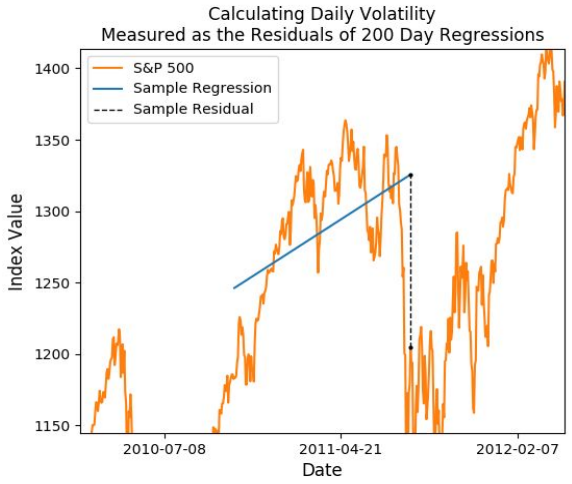

Fig. 2. A sample calculation of volatility based on the deviation of the daily value from the market trend.

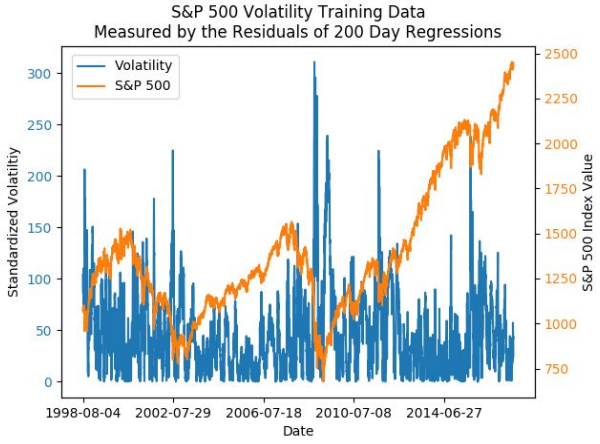

Fig. 3. Comparison of volatility measurements by daily S&P 500 values and the residuals of a 200 day regression.

To alleviate the effect of extreme spikes and dips in the data, the daily measurement of volatility, based on the residual from the regression, was "smoothed" by creating a continuous series of averages from different subsets of the full dataset. This *moving average* of the data was defined as the mean value of the 200 residuals preceding each day in the data set.

F. Choosing Parameters that Validate the Definition

Defining volatility as a moving average of daily residuals from a linear regression required the adjustment of two hyperparameters. The depth of both the linear regression and moving average smoothing function had to be adjusted to produce different shaped historical graphs of calculated volatility. Validating the calculation for volatility required adjusting the parameters so that they accurately reflected the common understanding of recent volatility in markets.

The first characteristic that any realistic graph of volatility needed to have is a sharp spike through the years 2008-2010. Economic uncertainty was a hallmark of the Great Recession, and any calculation of volatility should reflect the instability that permeated this period. [8]

A realistic calculation of volatility should also have increased values that reflect the uncertainty that surrounded the uncertain moments of European sovereign debt crisis (2011-2012), a period during which many eurozone members were unable to repay their loans. [9]

Finally, the most recent characteristic a calculation of volatility should have is heightened values during the 2015-2016 market crisis in China, which discouraged investors in the United States from investing and created more uncertainty in the market. [10]

The regression and moving average depth parameters were easily adjusted to both fit these historical characteristics and create a stable enough measurement for a regression algorithm that was interpreted accurately. Extensive fine tuning resulted in a depth of 200 days for both the regression and the moving average. These values created a numerical representation of historical volatility that was consistent with reality and was "learnable" for a machine learning regression algorithm. The blue line in Fig. 4. shows how the moving average made the volatility data more suitable for analysis while still remaining consistent with historical fact, with volatility increasing during periods when the S&P 500 experienced large fluctuations in value.

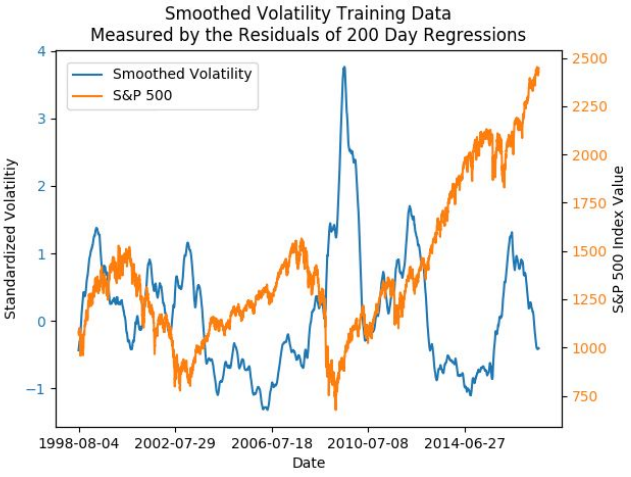

Fig. 4. Volatility as measured by a moving average of the residual of the 200 day regression of S&P 500 values.

IV. Determining Economic Indicators

A. Indicator Selection

Economic research and analysis were performed to determine which *indicators*—measures of different aspects of the economy—affected market volatility the most. Indicators with high data report rates and economic significance were selected, and any relationships or overlaps between the indicators were noted, as indicator relationships would be important for later regression analysis. [11][12][13][14][15][16] The indicators were then grouped according to a collection of overarching attributes.

B. Security Maturation

Securities are tools that possess monetary value that can be traded between institutions and individuals. Oftentimes, the rates at which securities mature—the dates specified, by which all interest on the securities must be paid—can vary. In this case, the Federal Reserve standardizes the different maturities to a constant maturity. The constant maturity can be indexed, and its rate of change measured.

The *10-Year Treasury Constant Maturity Rate* is one such index, which measures constant maturity over a ten-year time span. This rate decreases when there is an increase in volatility. [17]

The *10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity* is the difference between the ten and two-year measurements of constant maturity. As opposed to ten-year maturity alone, this rate increases sharply when there is an increase in volatility, as a result of the fluctuating differences between the two rates. [18]

The *Effective Federal Funds Rate (FFR)* is a measure of the rate at which banks and credit unions lend each other reserve balances, which are sums of money that institutions are required to hold in a central location. When considered by itself, the FFR decreases dramatically during periods of economic volatility. After the 2008 financial crisis, the FFR sank for several years before rising again in early 2016. [19]

The *10-Year Treasury Constant Maturity Minus Federal Funds Rate*, which is defined as the difference between the two rates previously discussed, is itself a distinct indicator. This difference increases when volatility increases. [20]

C. Credit and Loans

Standard loans and lines of credit are two different methods through which a borrowing party can obtain money temporarily from a bank or other corporate institution. From a financial perspective, credit is defined as a monetary relationship bounded by a predetermined borrowing limit, in which the borrowing party agrees to reimburse the lending party at a later date, oftentimes with interest charged on the sum. The interest rate is a specific proportion of the sum that the borrower must pay back at a certain intervals. When paid on an annual basis, the interest rate is also called the finance rate or annual percentage rate (APR).

A loan, on the other hand, is an instantaneously granted sum of money, usually for a specific purpose such as a mortgage, that is paid back gradually in installments. Both forms of borrowing include interest rates and installment plans that are determined by the lender.

The *Bank Credit for All Commercial Banks* is the total amount of money banks loan to institutions or individuals. In periods of volatility, this aggregate credit tends to decrease. [21]

The *Loans and Leases in Bank Credit* indicator for all commercial banks represents the fraction of the aggregate credit that is put toward loans and leases, which are essentially loans on property and other physical objects rather than money. The magnitude of this indicator tends to increase with time, except during times of economic volatility, when it decreases. [22]

Many of the indicators previously described have to do with credit agreements imposed by banks themselves; however, there exists credit imposed by corporations as well, in the Nonfinancial Corporate Business sector. The *Credit Market Debt as a Percentage of the Market Value of Corporate Equities* is an economic indicator in which equities are the values of assets after they have been readjusted to exclude liabilities. This measure also increases sharply during economically volatile periods. [23]

The *Commercial and Industrial Loans* indicator for all commercial banks represents the money borrowed from banks designated solely for loans. This sum of money sharply decreases in volatile circumstances. [24]

D. Interest Rates

While the indicators mentioned above require a conceptual understanding of interest rates to comprehend, those indicators are not measures of the rates themselves. There are several other indicators that deal specifically with the change in the interest rate itself for a financial instrument.

The *Finance Rate on Consumer Installment Loans at Commercial Banks* represents the interest rate that banks charge for automobile loans, specifically for loans spanning two years. A pattern of steady decrease, one which still continues today, began to show in this rate as a result of the 2008 crisis. [25]

The *Bank Prime Loan Rate* represents the best, or in other words lowest. interest rate that banks are willing to charge their borrowers. Most of the time, the ones who receive this privilege the bank's most valued customers, those who have had the most loyalty or experience with the bank. This rate drops sharply with volatility. [26]

The *Discount Interest Rates*, those that are charged to banking institutions by the Federal Reserve, can also serve as an accurate indicator of economic volatility. Like the Bank Prime Loan Rate, the discount rate decreases with an increase in volatility. [27]

E. Financial Returns and Profitability

A return, defined simply, is the profit made on a particular investment, measured in either hard dollar values or in percentage augmentation of the original investment sum.

The *Return on Average Equity for All US Banks* measures the ratio of the dollar profit generated by the bank to the dollar price of the share of the bank that the stockholder possesses. This is essentially the profitability of the bank relative to the investment in it. This ratio decreases sharply when volatility is present in a financial market. [28]

The *Return on Average Assets for All US Banks* is the ratio of the annual earnings of the bank to its total assets.

This is another measure of relative bank profitability, and decreases in a manner similar to that of Return on Average Equity during periods of high volatility. [29]

F. Yield Spreads

The yield spread, also known as the credit spread, for an investment type is the maximum range of difference on the quoted rates of return for bonds or other investments with similar maturities. It serves as an indicator of risk when choosing one investment over another.

The *Real Gross Domestic Product* (GDP) is a measure of total economic output, which is the aggregate value of all goods and services produced per year of a nation. This is the indicator most commonly used to compare economic progress between nations. It is reported quarterly—every 3 months—which is not as frequent as more specific indicators. The GDP value drops sharply during periods of economic volatility. [30]

The *Bank of America Merrill Lynch (BoAML) US High Yield Option-Adjusted Spread (OAS)* takes into account the yield spread coupled with a benchmark yield curve that reveals embedded options, inherent facets of a bond that give the bond issuer certain rights related to how the bond can be converted and exchanged. The degree of spread, which is the magnitude of the difference in return rate value, increases sharply when volatility increases. [31]

The *BoAML US Corporate Master OAS* is a spread which applies specifically to corporate bonds rather than general financial investments, specifically indicating the risk factor and volatility associated with companies. Like the general High-Yield Spread, the Corporate Master OAS increases sharply during economically volatile periods, following an expectedly similar pattern. [32]

G. Individual-Consumer Measurements

Some economic indicators measure trends that concern individual people, rather than larger corporations and/or institutions. For example, the *Personal Saving Rate* is one of these indicators, and it describes the fraction of a person's net income, on average, that he/she designates for recreational activities. Since this trend is dictated by personal preferences, its patterns in volatile periods is not so obvious, but it fluctuated during the 2008 crisis, with a generally increasing value. [33]

H. General and Multivariable Indices

Many indicators are measurements of universal properties of the US economy, or simply indices that take many other distinct indicators into consideration for their construction. These indicators can serve as comparison tools when looking at different global economies, since many countries often measure the same types of data.

The *Trade Weighted US Dollar Index* measures the relative value of the dollar (USD) compared to the value of other world currencies. Depending on the global impact of economic volatility, this quantity can increase or decrease. Since the 2008 crisis deeply affected many other nations, the relative value of the dollar increased. [34]

The *St. Louis Fed Financial Stress Index* is a measure of stress on the market through a combination of 18 economic indicator variables, including multiple yield spreads and interest rates. This index measures high financial stress during volatile periods. [35]

Lastly, the *Russell 3000 Value Total Market Index* is another financial index constructed by the Russell Investment Group that serves as a standard for the entire US stock market. Like the St. Louis index, this index takes multiple indicators into account to measure the market capitalizations of 3000 stocks. Similar to the actual stock market, this index value drops significantly in periods of economic volatility. [36]

V. Determining the Model

A. Bulk Implementation of Regression Models

In order to gain a wide perspective on the types of machine learning algorithms that perform well on the selected economic and financial data, 10 regression models in the Scikit-learn Python library were analyzed, displayed in Fig. 5. These models were grouped by the type of analysis they perform. The categories of regression algorithms analyzed were Generalized Linear Models, Decision Trees, Ensemble Methods, and Supervised Neural Network models. Such a diverse group of algorithmic models allowed for a selective process that yields an accurate model that meets specified criteria.

B. Narrowing Down the Tested Models

Of the 10 total machine learning algorithms that were tested, the first subset was selected by removing models that were redundant; for each of the four categories of regression analysis models, a single representative model was chosen.

The narrowed down group of models, as shown in Figure 6, included the Gradient Boosting regression, Elastic Net regression, Randomized Forest regression, and Multi-layer Perceptron regression. Gradient Boosting uses an Ensemble Method, Elastic Net regression is a Generalized Linear Model, Randomized Forests utilizes Decision Trees, and the Multi-layer Perceptron is a type of Neural Network.

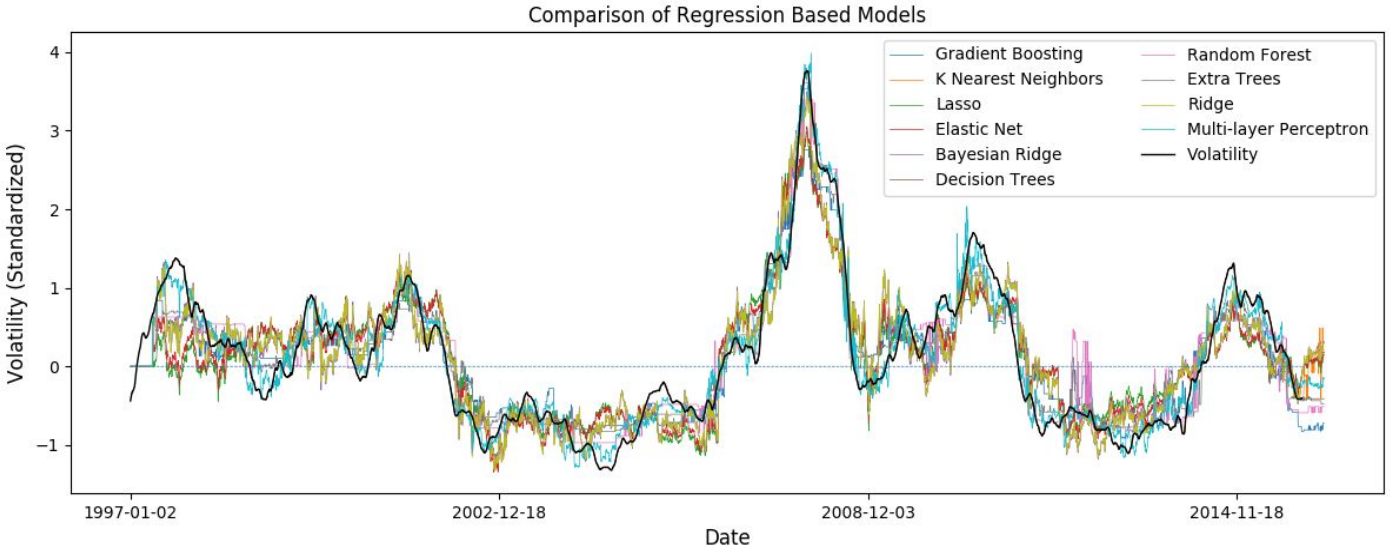

Fig. 5. Models generated by all tested regression algorithms.

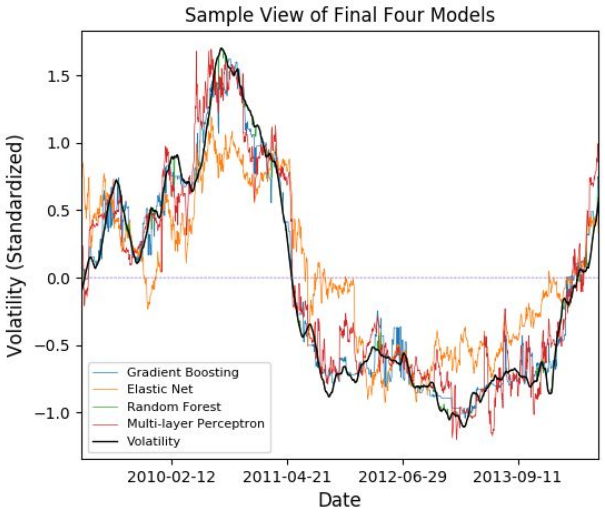

Fig. 6. The four most successful regression models of those tested.

C. *Picking a Final Model*

The Multi-layer Perceptron model was the first to be removed from the remaining group of models. This was because of the four models, as seen in Table 1, the Multi-layer Perceptron Model's mean squared error of 10.7553 was the worst error performance by a good margin. Such an extreme failure to predict was most likely due to the fact that the Multi-layer Perceptron is based on neural networks, which are notoriously poor at regression analysis. [6]

Model	Error
Gradient Boosting	0.3959
Elastic Net	0.2886
Multi-Layer Perceptron	10.7553
Randomized Forests	0.2053

Table 1. Mean squared error measurements for the most successful regressions of standardized data sets.

The Randomized Forest regression was also removed from consideration; even though this model had the best error performance, it also had a clear issue with *overfitting*, illustrated in Fig. 7. Overfitting occurs when an algorithm mirrors the data exactly as opposed to finding meaningful patterns within it. This means that the model may be very poor at making predictions about future volatility, even when its training data fit has very low error.

Of the last two models, the Gradient Boosting model was removed from consideration. The Gradient Boosting model did not appear to overfit the test data, but it still had relatively weak predictive power when compared to the Elastic Net regression.

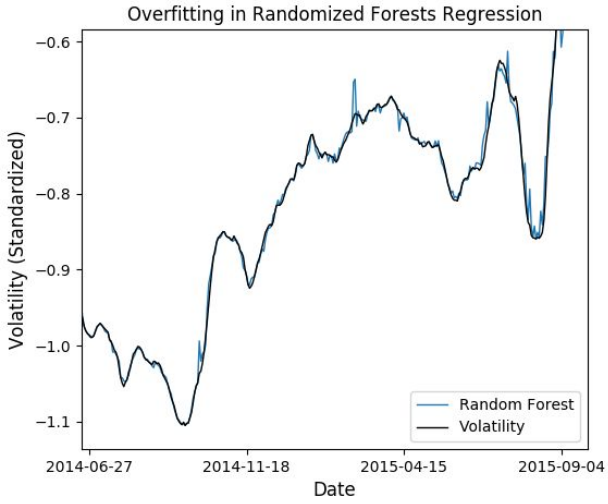

Fig. 7. Regression model produced by the Random Forest algorithm, indicating overfitting of the volatility function.

D. Benefits of the Elastic Net Regression

Multicollinearity is a phenomenon in which of two or more variables that are highly correlated are present and can linearly predict each other to a high degree of accuracy. Because the model deals specifically with many economic indicators that are highly correlated, multicollinearity is an important issue when selecting the right machine learning algorithm to implement. For this reason, algorithms such as the Ordinary Least Squares algorithm cannot be used. Although OLS has the lowest variance without setting restrictions on variable weighting—meaning that it does not attempt to correct for bias—it does not account for multicollinearity among variables, which causes the algorithm to artificially inflate the weighting of highly correlated variables, skewing the model's predictions. The characteristics of variance and bias are further explained in Section VI. Elastic Net regression accounts for multicollinearity and uses regularization values in its optimization function which inherently prevent overfitting and increase the predictive power of the model. [6]

Finally, in model testing the Elastic Net regression produced a low error, even with standardized hyperparameter values. The model's hyperparameters can also be adjusted and its predictive power increased further if it is deemed necessary through optimization. [37] A sample of an Elastic Net model fit to historical volatility can be seen in Fig. 8.

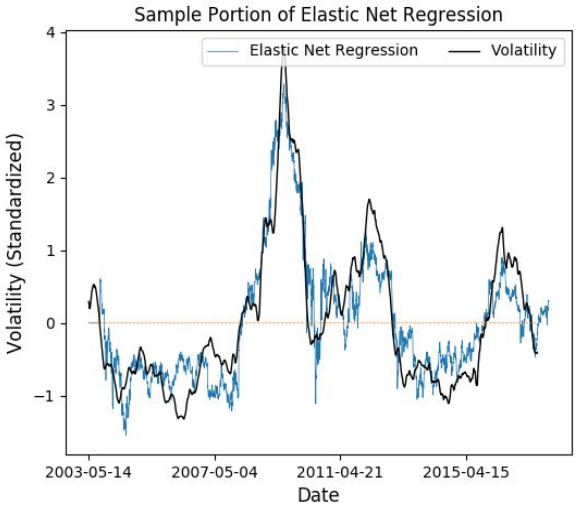

Fig. 8. Sample Elastic Net model, fit to historical volatility.

VI. The Volatility Model

Any linear model of a set of data aims to reduce the total error of the predicted values. This is most commonly done in two manners, the first of which is by minimizing the *variance* of the data from the generated model. Variance is defined in Equation (2).

$$\sigma^2 = \sum_{i=1}^{n} \|y_i - \hat{y}_i\|^2 \tag{2}$$

In the equation above, the variance σ^2 is the sum of the squared errors. The error is defined as the difference between the predicted value y_i and the actual value $\hat{y}_i$. This is a commonly used definition of error in statistics.

The second method to create a more accurate predictive model is to find and control for biases that may skew its predictions. Contrary to variance, which describes the range over which the model's predictions are spread, *bias* can negatively impact all of a model's predictions by shifting them in a certain direction. Fig. 9. shows a visual representation of the distinction between variance and bias. [38]

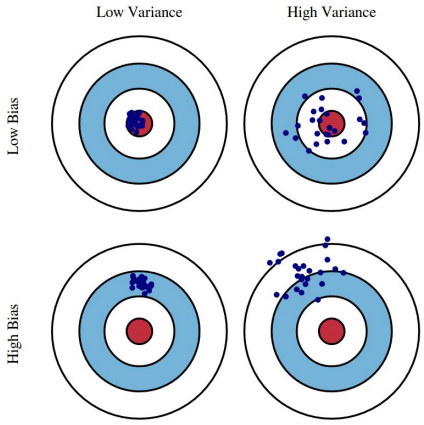

Fig. 9. Visual representation of variance and bias. [38]

Such bias is often present in regression models particularly because of the adverse effects of multicollinearity on variable coefficients, which cause all predictions to be strongly influenced by these interrelated variables. This bias can be controlled by decreasing the complexity of the regression model, which involves restricting the size of the coefficients of each independent variable and/or decreasing the number of variables used in the regression. This can prevent regression results from being drastically skewed.

However, it is important to note that the inevitable tradeoff of higher variance when controlling for bias may not always lead to predictions that better reflect the actual data or lower total error, as visualized in Fig. 10. Most often, a regression algorithm will find a balance between variance and bias by defining a *cost function*.

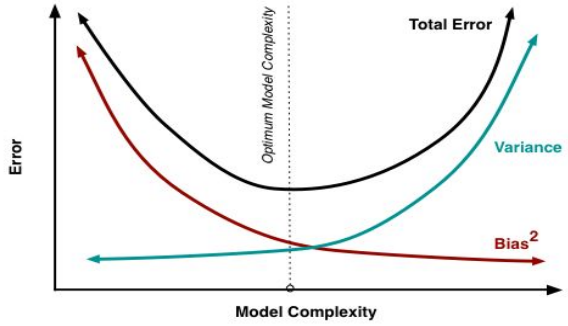

Fig. 10. The total error of a regression model as related to its variance and bias. [38]

A cost function is the function that defines error for a particular regression model, and it is the function that the regression algorithm attempts to minimize. When no bias is introduced into a data set, the cost function $J(\vec{\beta})$, in which $\vec{\beta}$ represents the *weight vector* (the combination of all of the variable coefficients in a multidimensional space), is simply defined by the variance of the model. The following section explores the how the cost functions of various multivariate linear regression models handle the tradeoff between variance and bias, which is necessary for a full understanding of how the Elastic Net model is designed.

A. Ordinary Least Squares (OLS) Regression

Equation (3) shows the cost function of the standard OLS regression method, which is equal to the variance, or the sum of the squared errors of each data point.

$$J_{OLS}(\vec{\beta}) = \sum_{i=1}^{n} ||y_i - \hat{y}_i||^2 \tag{3}$$

This model operates under the assumption that all of the independent variables lending to the regression are truly independent, and that there is no collinearity between them. However, if the data set being analyzed has multiple, intercorrelated variables, this assumption leads to these coefficients of these variables being magnified, resulting in an inaccurate regression. This is the primary reason why standard OLS is not used in the regression of large data sets.

B. Ridge Regression

The problematic portion of the standard OLS regression is that coefficients can become arbitrarily large in the presence of multicollinearity. Ridge regression, while similar to OLS, controls for bias by constraining the weight vector of the model and preventing the inflation of coefficients. The model improves the cost function used in standard OLS by adding another term known as a *regularization term*. The cost function for Ridge regression is seen in Equation (4).

$$J_R(\vec{\beta}) = \sum_{i=1}^{n} ||y_i - \hat{y}_i||^2 + \lambda ||\vec{\beta}||_2^2 \tag{4}$$

The l_2 regularization term, which is also represented by $||\vec{\beta}||_2^2$, is defined in Equation (5). The λ preceding the regularization term is a hyperparameter that is specified when performing the regression. The larger the value of λ, the larger the impact of the regularization term on the cost function, and the more the weight vector is restricted.

$$||\vec{\beta}||_2^2 = \sqrt{\beta_1^2 + \beta_2^2 \ldots \beta_n^2} \tag{5}$$

The result is the model with the lowest total variance that remains within the bounds set by the weight vector. The point at which this occurs is where the Ridge algorithm's cost function is minimized. In Fig. 11, the variance is represented by the red shape, in which the point of the OLS estimate has the lowest variance, and variance increases as the lines move outward. The blue shape represents the area in which the weight vector stemming from the origin satisfies the bounds set by the regularization term. The

intersection point between the two shapes is where the estimate of the Ridge regression minimizes the cost function. [6]

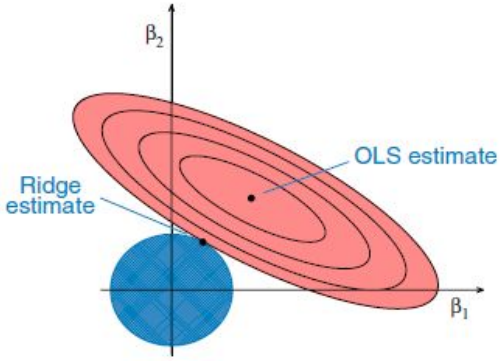

Fig. 11. The location of a Ridge regression weight vector as compared to a variance-minimizing OLS weight vector. [39]

While Ridge regression solves the issue of variable coefficients becoming arbitrarily large, it does not perform feature selection. This means that while it restricts the size of each variable's coefficient, it will never eliminate a variable altogether by making its coefficient 0 unless λ approaches infinity. This becomes problematic when examining a large number of variables. If multiple irrelevant sets of data are continuously involved in the prediction of the dependent variable, the error created by these irrelevant sets can compound even when their variance is minimized, resulting in overfitting the data.

C. Least Absolute Shrinkage and Selection Operator (LASSO) Regression

LASSO regression, in contrast to Ridge regression, is able to perform feature selection. LASSO regression has a cost function similar to both OLS and Ridge regression, which is shown by Equation (6).

$$J_L(\vec{\beta}) = \sum_{i=1}^{n} ||y_i - y_i||^2 + \lambda||\vec{\beta}||_1^1 \tag{6}$$

However, the regularization term here is the l_1 term, which is represented by $||\vec{\beta}||_1^1$, and defined in Equation (7). It is represented as the sum of the magnitudes of the weight vector's components rather than the magnitude of the weight vector itself.

$$||\vec{\beta}||_1^1 = |\beta_1| + |\beta_2| \ldots |\beta_n| \tag{7}$$

The resulting cost function serves a similar purpose to Ridge regression by constraining the absolute sum of the variable coefficients. However, because the regularization term is defined by the weight vector's components, it is often best for a component to be entirely eliminated for the cost function to be minimized. [6]

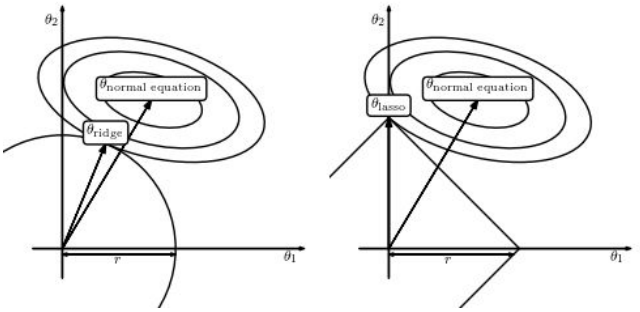

Fig. 12. The results of applying the l_2 (left) and l_1 (right) regularization terms on the resulting weight vector. [40]

As shown in Fig. 12, restricting variable coefficients using the l_1 term forms a different boundary for the weights than the l_2 method. The point along this boundary where variance is minimized, unlike with Ridge regression, is often at the point where one or more components of the weight vector are 0. When this constraint is generalized to a large number of dimensions—much more than is possible to show here—the LASSO regression method can perform large scale feature selection, setting the coefficients of many irrelevant variables to 0 and disregarding them.

With LASSO regression, instead of overfitting the data by utilizing too many variables, as Ridge regression does, there is the possibility for underfitting to occur, in which too many variables are eliminated for an accurate model to be generated.

D. Elastic Net Regression

The Elastic Net model functions as a compromise between the Ridge and LASSO regression methods. Its cost function is created by implementing both the l_1 and l_2 regularization terms, and is shown by Equation (8).

$$J_E(\vec{\beta}) = \sum_{i=1}^{n} ||y_i - y_i||^2 + \lambda_1||\vec{\beta}||_1^1 + \lambda_2||\vec{\beta}||_2^2 \tag{8}$$

Utilizing both of these terms, the total weight vector is still bounded close to the origin, but by a variable boundary that is between that shown by each regularization term separately. This is why the "net", or boundary, represented in Fig. 13. is said to be elastic; it can be adjusted using the hyperparameters λ_1 and λ_2 to restrict the weight vector more or less within the more rigid bounds that set by Ridge and LASSO regression. [6]

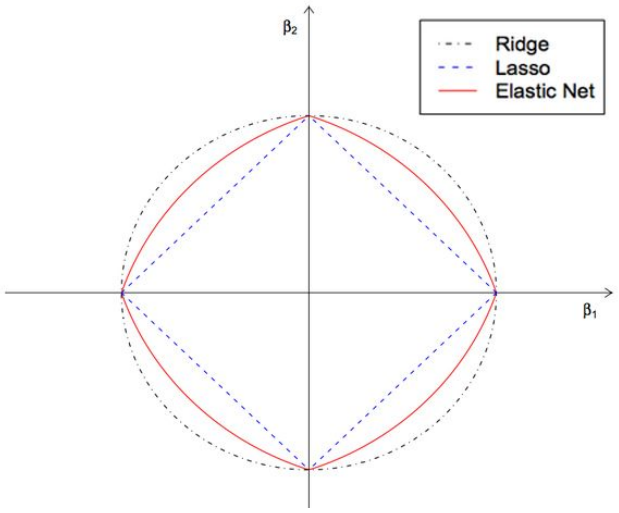

Fig. 13. The elastic net method's weight vector boundary in relation to that of Ridge and LASSO regression. [41]

The Elastic Net model ultimately presented the best prospect of creating an accurate predictive model. When working with economic data, in which every variable is somewhat correlated with each other, the set of possible indicators needs to be reduced enough so that this collinearity does not skew the model; however, it is still necessary to utilize a large amount of variables, many of which are, while related, not entirely dependent on each other. The Elastic Net model has the ability to restrict the weight vector's size, and to perform feature selection to an extent specified by the user, effectively creating a middle ground between the various other models discussed previously. Furthermore, the inherent feature selection that the algorithm performs eliminates the need to apply additional dimensionality reduction techniques before training the algorithm on the data set.

VII. Implementation and Optimization

A. Data Importation and Formatting

The data for each chosen economic indicator was first downloaded in comma separated values (*.csv*) file format, which is one of the simplest formats to sort and search for data. Each set of data imported from FRED was stored in a separate Python pandas DataFrame object, each with the values stored in one column and the dates on which the values were collected as the index. All of the indicators were then conglomerated into one DataFrame object by date, and any dates with 'null' (missing) values were assigned the value collected on the date immediately preceding it. Missing values were not very common in any of the data sets, and so filling these values with close approximations was more favorable than eliminating any dates entirely. Any dates without reported values for a large amount of variables, however, were deleted from the frame. The resulting DataFrame object contained well-reported data for each indicator variable from January 1997 to July 2017.

B. Selecting the Cross-Validation Method

A common problem in creating predictive models is finding the balance between improving the model to fit the data more closely and overfitting. In the case of overfitting, the predictive model appears to match the training data very closely, but fails to create accurate predictions when tested on new data. Because the amount of available data is limited, a portion of the data must be set aside to test for overfitting, known as the test set. The rest of the data is then used for training the model.

There are many methods of cross-validation. Cross validation attempts to split the data into different combinations of training sets and testing sets so that the best estimate of predictive error can be made. However, most cross-validation methods cannot be used to test the predictive model in this project because the methods involve shuffling the order of the data when splitting into training and test sets. The economic indicators and S&P 500 volatility data are formatted as time series, and thus, the order of the data must be preserved when training and testing the machine learning algorithm.

Walk Forward Validation was chosen for estimating error in the model's predictions because this method preserves the order of the data when splitting it into training and testing sets. [18]

C. Implementing Walk Forward Validation

In using Walk Forward Validation, the first 300 samples, which includes data from the earliest dates, were set as the training data, and the first 90 samples after the last training data sample were set as the set of test data. Then, using these training and test sets, the mean sum of squared errors of the prediction was calculated. These steps were performed in a loop in which the starting date of the 300 training data samples was increased in each iteration, as well as the earliest date in the test data set, which comes chronologically after the dates in the training data set. This process is broadly illustrated in Fig. 14. After exiting the loop, the average of the mean sum of squared errors from all the prior evaluations was calculated to estimate the error of the Elastic Net regression model's error. [42]

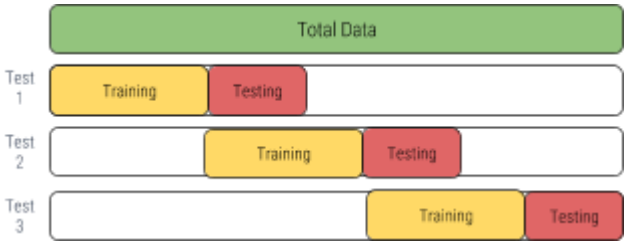

Fig. 14. Visualization of Walk Forward Validation process.

D. Algorithm Optimization

The selection of the Elastic Net regression model was based on its ability to perform limited feature selection as well as continuously retain low errors, as described in Section VI and Section III. F, respectively. However, once the model was selected, it was still desirable to lower the error of the model even further so that any predictions made would be as accurate as possible. The initial mode used standardized hyperparameter values of $\lambda = 1.0$ and $l_1\ ratio = 0.1$, which meant that the l_2 term was much more impactful on the cost function. This resulted in an average mean squared error of 0.2887 through the Walk Forward Validation method specified in Section VII. F. Ideally, the mean squared error would be brought as close to 0 as possible, meaning the model would predict volatility values for the specified date range with 100% accuracy.

To lower this error value, the regression model was repeatedly generated, using all of the indicator data as well as a wide range of l_1 ratios and λ values, testing a multitude of possible combinations. The combination of hyperparameter values that generated the lowest error was used in the final model. The hyperparameters for this final model were $\lambda = 0.9$ and $l_1\ ratio = 0.1$.

VIII. Results

A. Results of Optimization and Interpretation of Error

The λ value determined in the optimization process signifies that while the model's coefficients did need to be restricted to some extent, the cost function was ultimately most effective when the regularization terms, which attempted to control bias, were secondary to the variance of the model. Additionally, the low l_1 ratio of the model signifies that the algorithm performed best when the l_2 regularization term was weighted much more than the l_1 term, making the model much closer to a Ridge regression than a LASSO regression.

These two parameter values are likely indicators that while weight vector restrictions and feature selection were sometimes needed to regulate collinearity within the data set, the independent variables were different enough that each of them contributed a sufficient amount to the accuracy of the final model, which produced a mean squared error of 0.2300.

B. Extrapolation and Future Expectations

The results from running the final optimized model are shown in Fig. 15. The graph shows the standardized values for volatility. As is apparent, the Elastic Net regression model fits the actual data closely, and shows that according to this model, the volatility level after the last date used for training the model will increase.

The Walk Forward Validation method discussed in Section VII. E was used to find the estimated error of the model's predictions, using a training data set size of 300 samples, an alpha value of 0.9, determined by the results of the optimization process, a step forward size of 300 samples, and a test data set size of 90 samples, or 90 days. By performing Walk Forward Validation, the estimated error of the model's predicted values, which is the mean sum of squared error, was returned as 0.2300.

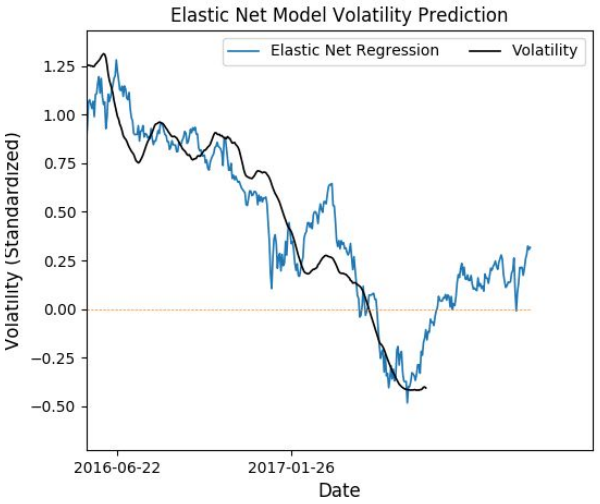

Fig 15. Graph of Elastic Net prediction in relation to actual volatility in the S&P 500.

IX. Conclusion

The custom volatility index based on the S&P 500 was created and trained on an Elastic Net regression model and financial data from the past 20 years. Based on the results of Walk Forward Validation, which is used to generate a predictive error for the volatility model, this model showed potential to accurately predict the volatility levels in the S&P 500 up to 90 days in the future.

The ability to predict the volatility of the S&P 500 will have widespread influence. If the volatility of the S&P 500, which is a standard measure of overall economic instability, can be predicted with a high degree of accuracy, this information could help all participants of the economy make sound financial decisions. By knowing future volatility levels, participants in the economy will be able to review their options in the face of potential economic fluctuations, and therefore make wiser decisions. On a small scale, individuals will be able to adjust their decisions on stock investments, while on a large scale, corporations will be able to evaluate their financial risk exposure in the face of a changing market.

Because of the possibility of such a widespread influence, the predictive model must be continuously improved. First, the accuracy of the model must be verified after the 90 day prediction interval has ended. Next, different machine learning algorithms will need to be researched, and new data will need to be collected from the 90-day period for further data testing and training. Additionally, if the model is to have public influence, people must be made aware of its

existence. From then on, there will be a standby period for another 90 days, after which this cycle will repeat.

In terms of future research, there is the possibility of modifying the model to not only predict the next 90 days, but significantly longer time spans of years or even decades. With the extension of the predictive time span, the predictive error will likely increase, so adjustments will have to be made to keep this error constant, or possibly even lessen it. These steps, if taken, will create predictions that will be more accurate for longer periods of time.

Acknowledgements

The authors would like to sincerely thank Shantanu Laghate, Project Mentor for the Data Analytics Research Group; Dean Ilene Rosen, the Director for New Jersey Governor's School of Engineering & Technology; Dean Jean Patrick Antoine, the Associate Director for the New Jersey Governor's School of Engineering & Technology; the Rutgers School of Engineering, for hosting the research group and allowing it access to its resources and advisory staff for the duration of the project; The State of New Jersey, for their support for the Governor's School of Engineering and Technology program; and Lockheed Martin, Silverline Windows, and Alumni of the New Jersey Governor's School of Engineering & Technology for their generous funding and continued support.

References

[1] J. A. Smith, "Data Analytics vs Data Science: Two Separate, but Interconnected Disciplines," Data Scientist Insights, 10-Sep-2013. [Online].
[2] B. Marr, "The Awesome Ways Big Data Is Used Today To Change Our World," LinkedIn. [Online].
[3] I. Staff, "Standard & Poor's 500 Index - S&P 500," Investopedia, 27-Oct-2014. [Online].
[4] R. Berger, "Total U.S. Stock Market Vs. The S&P 500 Index - An Investor's Guide," Forbes, 29-Jun-2017. [Online]
[5] I. Bobriakov, "Top 15 Python Libraries for Data Science in 2017," Medium, 09-May-2017. [Online].
[6] S. Raschka, *Python Machine Learning*. Birmingham , UK : Packt Publishing, 2015.
[7] "CBOE Volatility Index® (VIX®)," VIX Index. [Online].
[8] "The Financial Crisis of 2008: Year In Review 2008," Encyclopædia Britannica. [Online].
[9] "The Eurozone in Crisis," Council on Foreign Relations. [Online].
[10] "The causes and consequences of China's market crash," The Economist, 24-Aug-2015. [Online].
[11] "Financial Soundness Indicators and the IMF," International Monetary Fund. [Online].
[12] B. Gadanecz and K. Jayaram, "Measures of financial stability - a review," Bank for International Settlements. [Online].
[13] A. Gersi and J. Hermanek, "FINANCIAL STABILITY INDICATORS: ADVANTAGES AND DISADVANTAGES OF THEIR USE IN THE ASSESSMENT OF FINANCIAL SYSTEM STABILITY," Czech National Bank. [Online].
[14] J. Haltiwanger, "Globalization and economic volatility," World Trade Organization. [Online].
[15] A. Petreski and E. M. Mihajlovska, "Aggregate indices for financial stability as early warning indicators for monetary measures in the Republic of Macedonia," National Bank of the Republic of Macedonia. [Online].
[16] E. M. Petrovska, "Measures of Financial Stability in Macedonia ," National Bank of the Republic of Macedonia. [Online].
[17] "10-Year Treasury Constant Maturity Rate," FRED, 19-Jul-2017. [Online].
[18] "10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity," FRED, 19-Jul-2017. [Online]
[19] "Effective Federal Funds Rate," FRED, 03-Jul-2017. [Online].
[20] "10-Year Treasury Constant Maturity Minus Federal Funds Rate," FRED, 19-Jul-2017. [Online].
[21] "Bank Credit of All Commercial Banks," FRED, 14-Jul-2017. [Online].
[22] "Loans and Leases in Bank Credit, All Commercial Banks," FRED, 14-Jul-2017. [Online].
[23] "Nonfinancial Corporate Business; Credit Market Debt as a Percentage of the Market Value of Corporate Equities," FRED, 12-Jun-2017. [Online].
[24] "Commercial and Industrial Loans, All Commercial Banks," FRED, 14-Jul-2017. [Online].
[25] "Finance Rate on Consumer Installment Loans at Commercial Banks, New Autos 48 Month Loan," FRED, 10-Jul-2017. [Online].
[26] "Bank Prime Loan Rate," FRED, 03-Jul-2017. [Online]. Available: https://fred.stlouisfed.org/series/MPRIME. [Accessed: 19-Jul-2017]. [Online]
[27] "Interest Rates, Discount Rate for United States©," FRED, 01-Jun-2017. [Online].
[28] "Return on Average Equity for all U.S. Banks," FRED, 11-May-2017. [Online].
[29] "Return on Average Assets for all U.S. Banks," FRED, 11-May-2017. [Online].
[30] "Real Gross Domestic Product," FRED, 29-Jun-2017. [Online].
[31] "BofA Merrill Lynch US High Yield Option-Adjusted Spread©," FRED, 19-Jul-2017. [Online].
[32] "BofA Merrill Lynch US Corporate Master Option-Adjusted Spread©," FRED, 19-Jul-2017. [Online].
[33] "Personal Saving Rate," FRED, 30-Jun-2017. [Online].
[34] "Trade Weighted U.S. Dollar Index: Broad," FRED, 17-Jul-2017. [Online].
[35] "St. Louis Fed Financial Stress Index©," FRED, 13-Jul-2017. [Online].
[36] "Russell 3000® Value Total Market Index©," FRED, 19-Jul-2017. [Online].
[37] "Forward and Reverse based Gradient filled Hyperparameter optimisation", Researchgate [Online].
[38] Fortmann-Roe, S. (2017). *Bias and variance contributing to total error*. [image] Available at: http://scott.fortmann-roe.com/docs/BiasVariance.html [Accessed 15 Jul. 2017].
[39] "5.1 - Ridge Regression," 5.1 - Ridge Regression | STAT 897D. [Online]. Available: https://onlinecourses.science.psu.edu/stat857/node/155. [Accessed: 20-Jul-2017].
[40] Vanderplas, J. (2017). *Ridge and Lasso Regression Geometric Interpretation*. [image] Available at: http://www.astroml.org/book_figures/chapter8/fig_lasso_ridge.html [Accessed 16 Jul. 2017].
[41] Zou, H. (2017). *2-dimensional illustration of Ridge, Lasso, and Elastic Net regression*. [image] Available at: http://web.stanford.edu/~hastie/TALKS/enet_talk.pdf [Accessed 16 Jul. 2017].
[42] Gupton, G. (2005). *Advancing Loss Given Default Prediction Models: How the Quiet Have Quickened*. Economic Notes by Banca Monte dei Paschi di Siena SpA. [online] Oxford: Blackwell Publishing Ltd, pp.227-228. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.164.917&rep=rep1&type=pdf [Accessed 16 Jul. 2017].