VIX Prediction via Machine Learning

Jiqiang Zhang (jz3293@nyu.edu)

December 17, 2018

Abstract

VIX, namely the CBOE Volatility Index, is a popular measure of implied volatility of S&P 500 index options. VIX largely reflects future volatility and risk, so it's very valuable for us to predict VIX accurately.

In this project, I utilized some regression methods in machine learning to predict VIX. My features can be divided into three parts: google trends, macroeconomic factors and technical factors of VIX and S&P500. Then I did some exploration data analysis, selected some important features and fit data with machine learning methods like Regularized Linear Regression, Random Forest and LSTM. Finally, I evaluated the performance of each model by comparing MSE and calculating Information Ratio for simulated trading strategies.

Contents

1 Introduction

VIX, constructed with implied volatility of many S&P500 index options, is a very famous index for stock market volatility. VIX is often referred to as the fear index', with its value spikes during periods of extreme uncertainty. From Figure 1, we can see that spikes of VIX were often accompanied with financial crisis, such as 2008 Global Financial Crisis.

Figure 1: VIX in the last 15 years

It seems difficult to predict VIX due to its dependence on complex human behaviors. However, the aggregation of various economic indicators, both technical and fundamental, can create accurate predictions when large amounts of relevant data are used within scientific methods such as machine learning. Machine learning algorithms can be used to select proper indicators and build stable models for VIX prediction.

2 Data Collection

My features can be divided into three parts generally: macroeconomic factors, google trends and technical indicators of VIX and S&P500.

2.1 Macroeconomics

I retrieved about 40 macroeconomics indicators from St. Louis Fred and Bloomberg websites. Several indicators with high performance are explained below.

The **10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity** is the spread between ten-year and two-year Treasury Constant Maturity. When volatility increases, this indicator increases as a result of larger spread between long-term and short-term constant maturity.

The **Bank of America Merrill Lynch (BoAML) US High Yield Option-Adjusted Spread (OAS)** is the spreads between a computed OAS index of all bonds in a given rating category and a spot Treasury curve. The degree of spread increases sharply when volatility increases.

The **St. Louis Fed Financial Stress Index** measures the degree of financial stress in the markets and is constructed from 18 weekly data series: seven interest rate series, six yield spreads and five other indicators. It measures financial stress during volatile periods.

The **Real Gross Domestic Product** is a percent change of real gross domestic product from preceding period. It's the most famous indicator to compare economic conditions in different countries. This value drops sharply during periods of market volatility.

The **Russell 3000 Value Total Market Index** measures the performance of the largest 3,000 U.S. companies representing approximately 98% of the investable U.S. equity market. Similar to GDP, this index decreases significantly when volatility increases.

Meanwhile, I found several indicators with poor performance and I had to abandon them. For example:

The **Personal Saving Rate** measures personal saving as a percentage of disposable personal income (DPI). This rate has a very low volatility through past 10 years, so it may have low predictive power for future volatility.

The **CBOE Gold ETF Volatility Index** measures the market's expectation of 30-day volatility of gold prices by applying the VIX methodology to options on SPDR Gold Shares (Ticker - GLD). Although gold volatility is highly correlated with stock volatility, this index focuses more on ETF trading and has low predictive power for VIX.

2.2 Google Trends

Google Trends are the records of the frequency of certain keywords during a period of time. I selected about 20 keywords (stock, crisis, unemployment, etc.) and obtained the frequency series of each keyword during 2014-01-01 to 2018-11-20 on a daily basis.

2.3 Technical Indicators

I utilized about 15 technical indicators for both VIX and S&P500 to analyze and predict VIX series. All the indicators (simple moving average, relative strength index, momentum, etc.) are obtained from ta-lib library in Python.

3 Data Exploration

3.1 Data Preprocessing

Since many macroeconomics indicators are not daily data, I utilized forward filling to fill missing values for non-daily data, which means the value of certain indicators will be the same

for a week or a month. Since Google Trends start from 2004-01-01, I had to use data from all categories from 2004-01-01 to 2018-11-19. After removing null values of all features, I found many feature values were very volatile. Therefore, I transformed all feature values into scaled values using MinMaxScaler in Python.

3.2 Data Visualization

Now let's look at several distributions of macro factors and google trends.

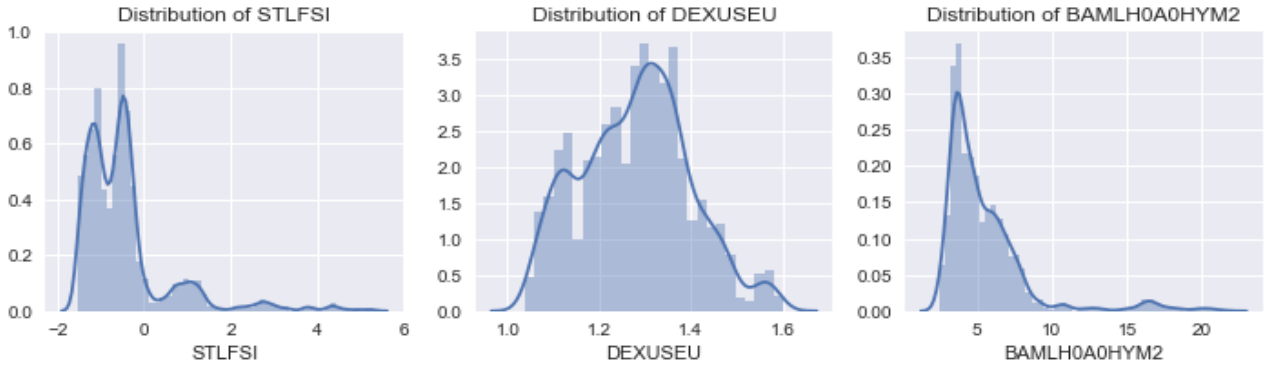

Figure 2: Distribution of three macro factors

From Figure 2, distributions of macro factors are different, but many of them have long right tails.

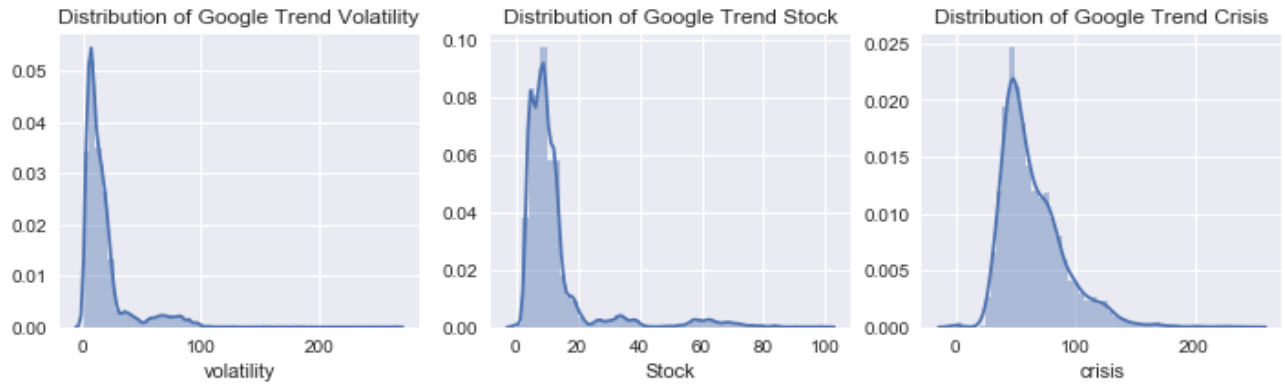

Figure 3: Distribution of three google trends

From Figure 3, many google trends have a long right-tail distribution.

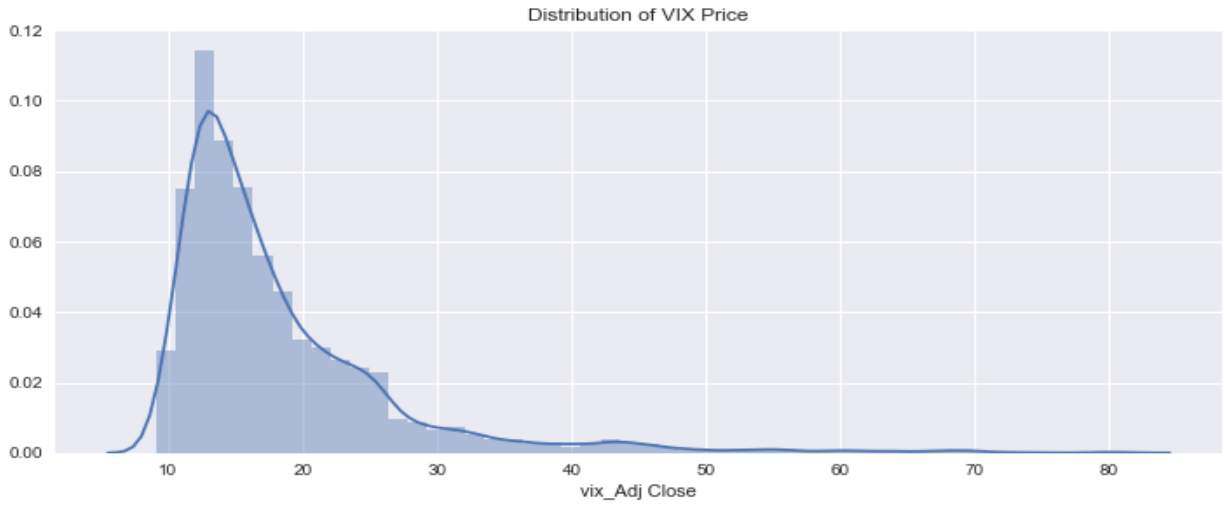

Figure 4: Distribution of VIX

From Figure 4, VIX distribution has a long right tail, which indicates relevance with many macro factors and google trends.

Next, I drew boxplots for several macro factors and google trends to know more details.

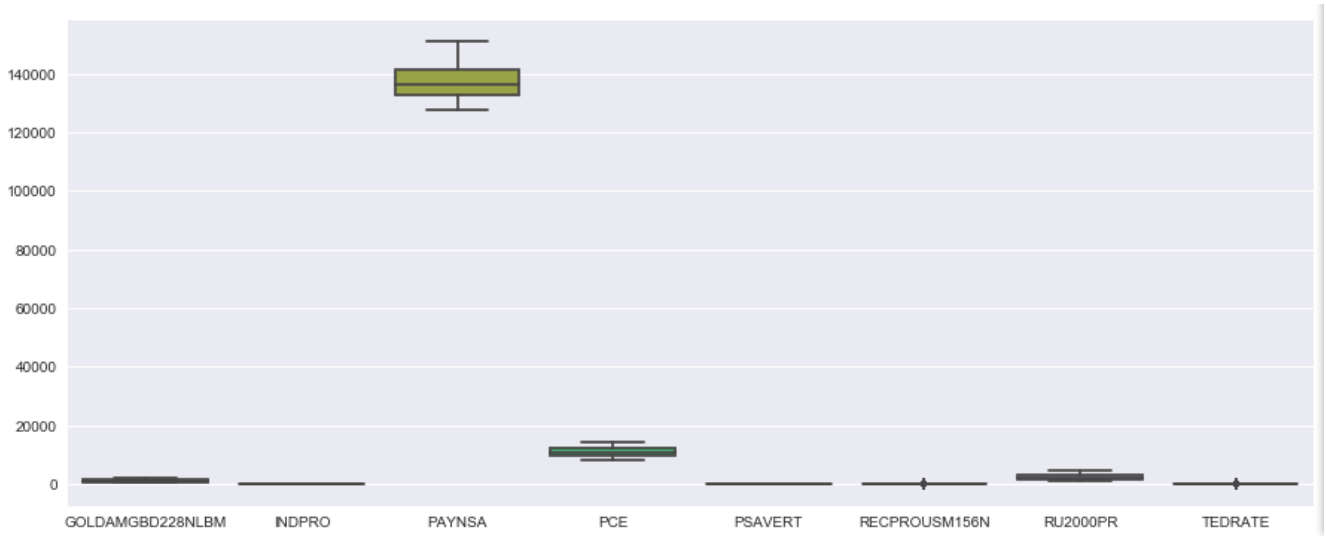

Figure 5: Boxplots for several macro factors

From Figure 5, some macroeconomics features, like 'PSAVERT', have very low volatility. Maybe they can't reflect the change of VIX.

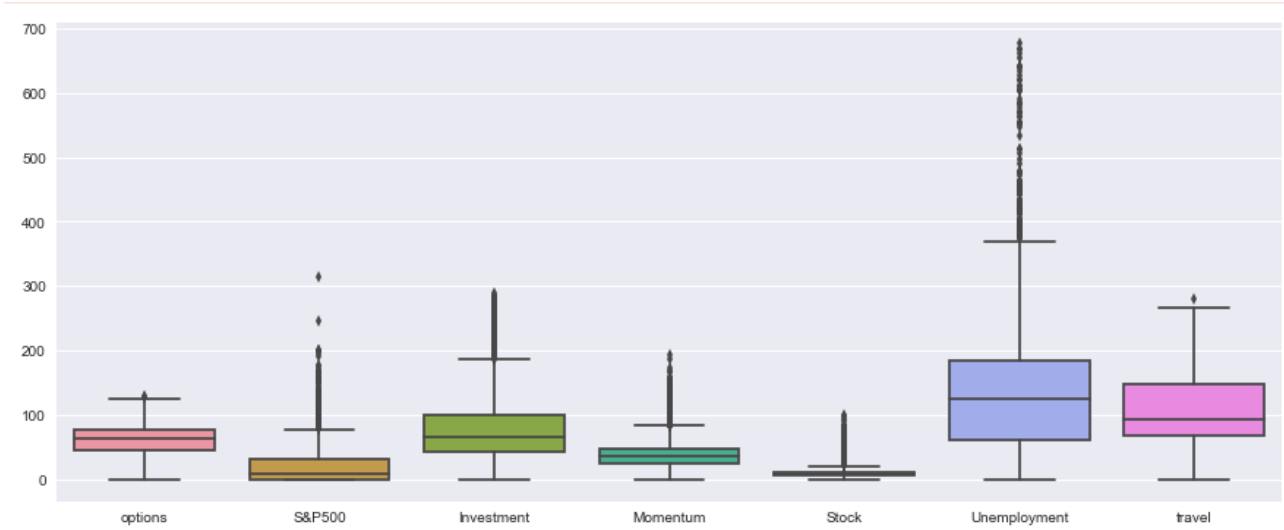

Figure 6: Boxplots for several google trends

From Figure 6, some google trends, like "Unemployment", have very high volatility. These features may be not stable for prediction.

Let's have a look at correlations among all the features.

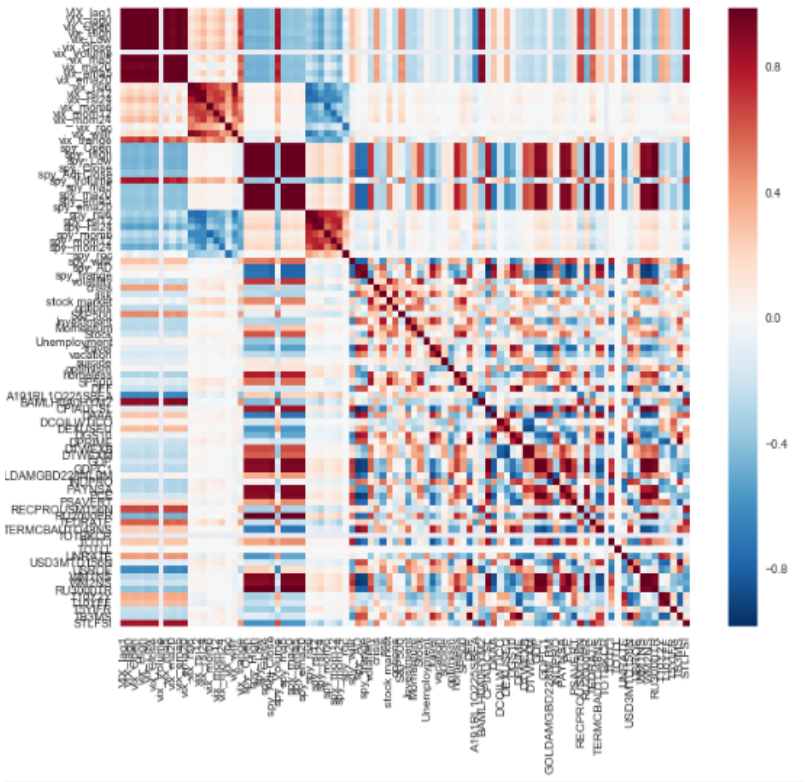

Figure 7: Correlation Analysis

As we can see from Figure 7, correlations among VIX technical factors and among some google trends are very high. Correlations between different types of factors tend to be low. So I eliminated some redundant features of technical factors and google trends.

To select features with an automatic method (not manually), I utilitzed random forest model to select all features with feature importance larger than 0.01.

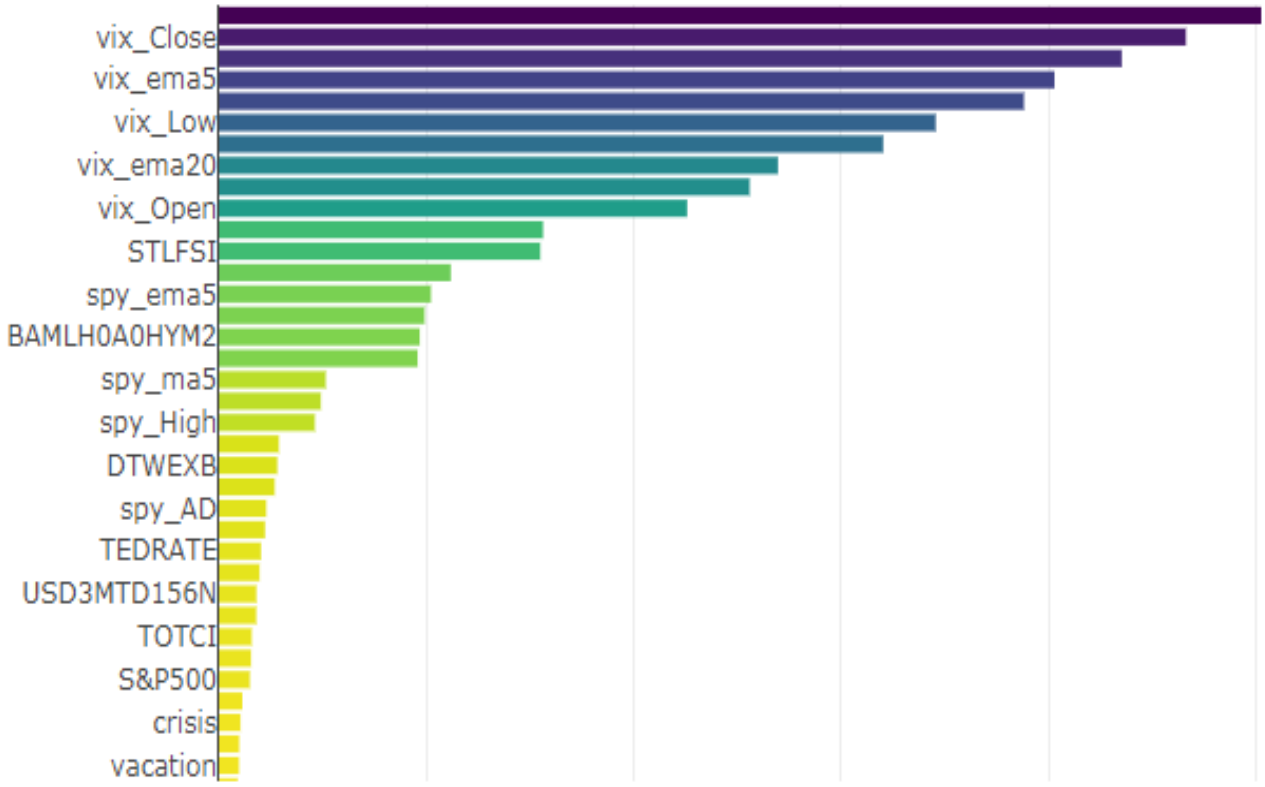

Figure 8: Feature Importance (part)

From Figure 8, I selected many technical indicators, several macro factors and one google trend with feature importance larger than 0.01. Totally, about 20 features were used in my final models.

3.3 Data Split

I split my dataset into 2 parts: 90% for training, 10% for testing. Since VIX is a time series and we cannot use future to predict history, traditional cross validation methods are not suitable. Instead, I utilized a rolling-window method for validation. This method built rolling windows to train data during a period of time and predict next several days so as to reduce overfitting.

4 Model Selection

4.1 Regularized Linear Regression

For all the models below, I utilized mean squared error(MSE) as my loss function. In the fitting of linear or logistic regression models, the Elastic Net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. After selecting the best parameters in the model, I drew two curves to compare predictions and real values for test set.

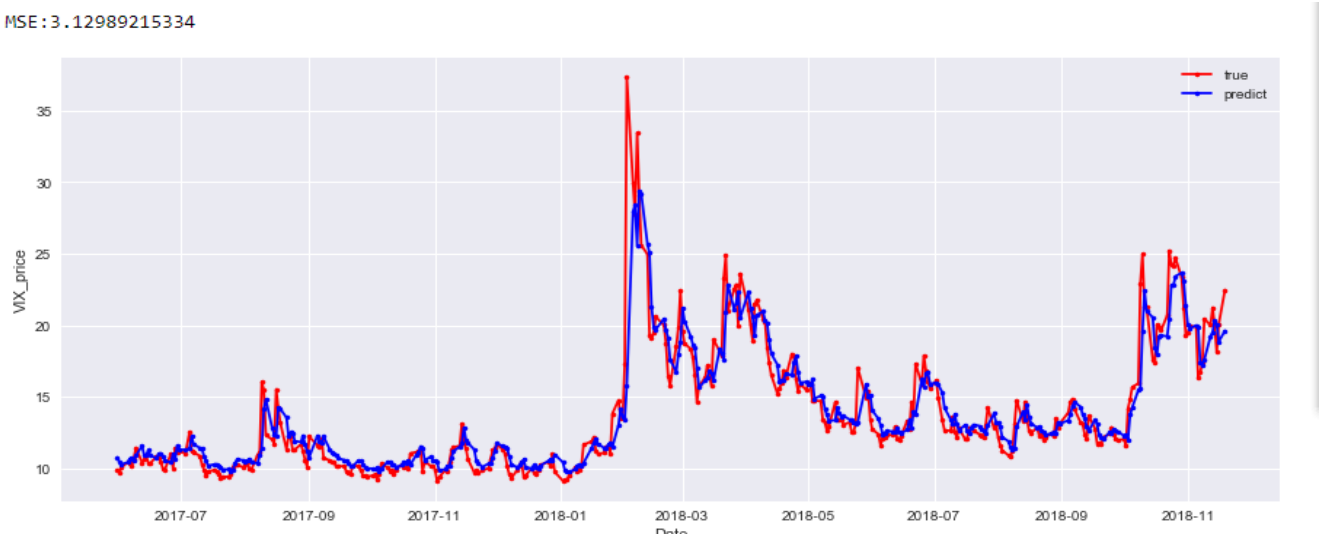

Figure 9: Regularized Linear Regression

4.2 KNN

The k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors. Model performance is shown in Figure 10.

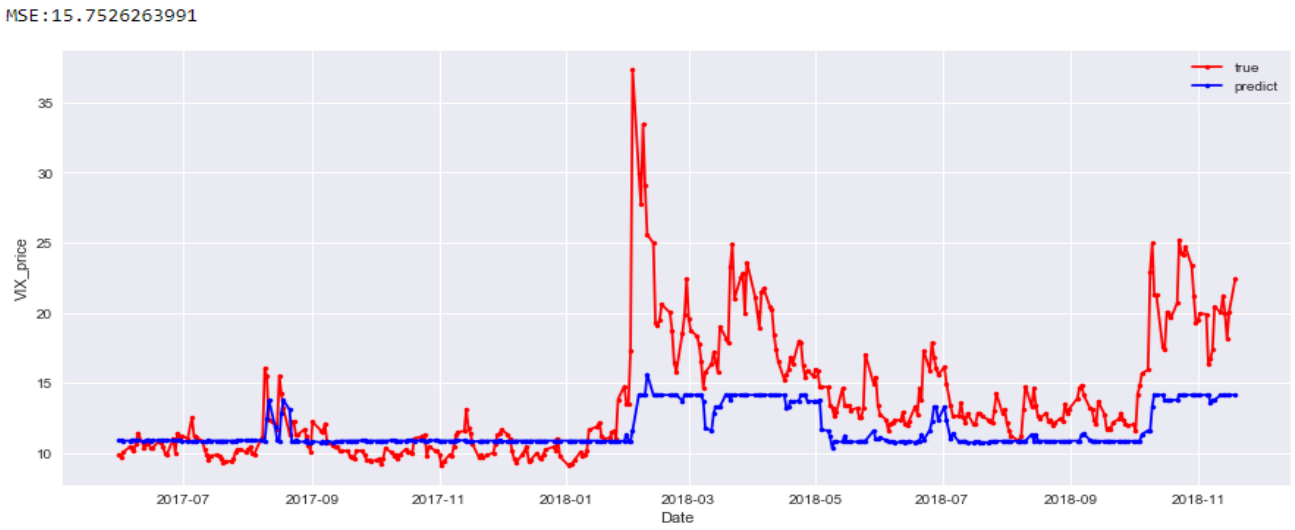

Figure 10: KNN

4.3 Support Vector Machine

A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Model performance is shown in Figure 11.

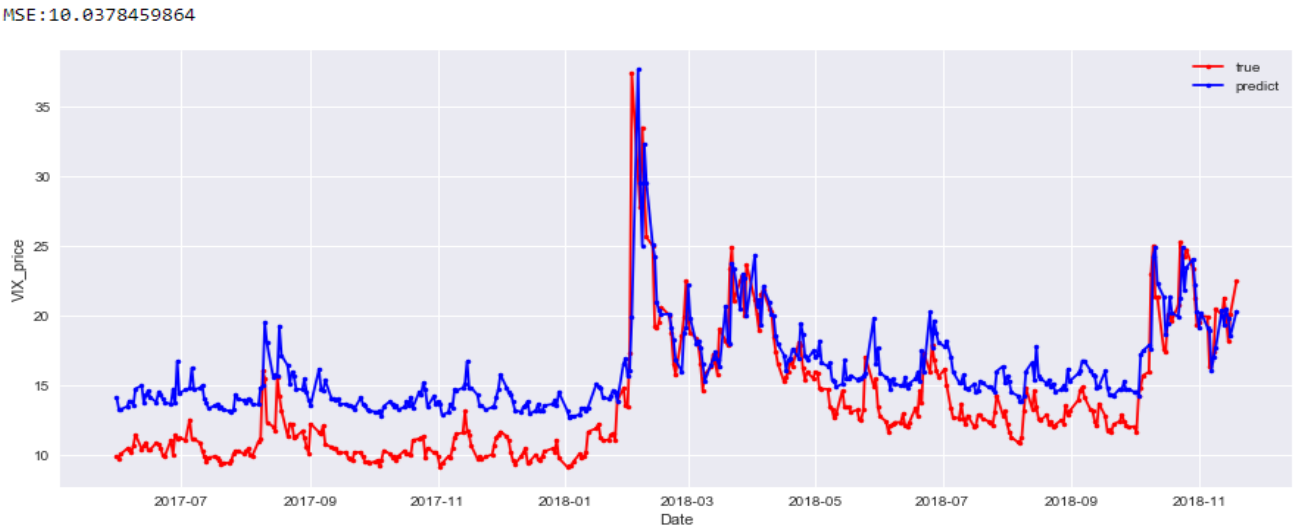

Figure 11: Support Vector Machine

4.4 Decision Tree

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Model performance is shown in Figure 12.

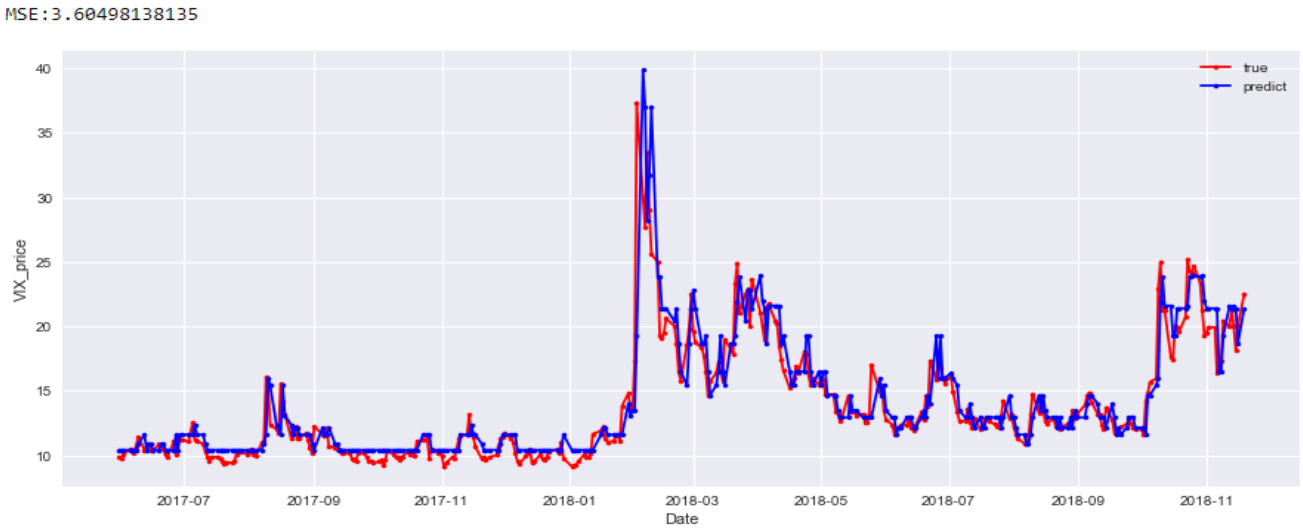

Figure 12: Decision Tree

4.5 Random Forest

Random forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Model performance is shown in Figure 13.

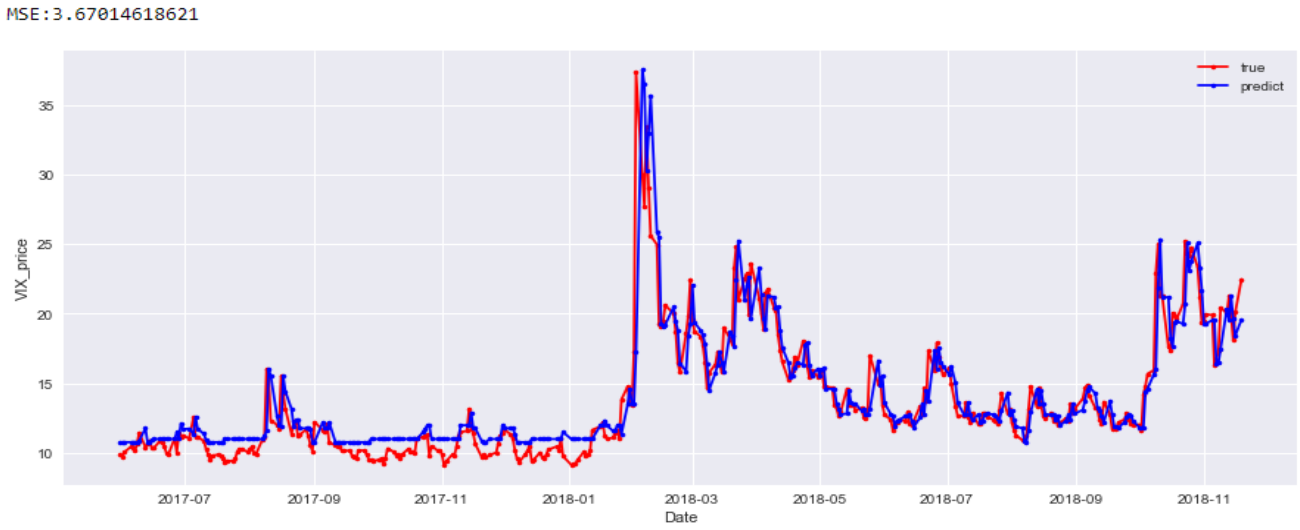

Figure 13: Random Forest

4.6 Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods

do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Model performance is shown in Figure 14.

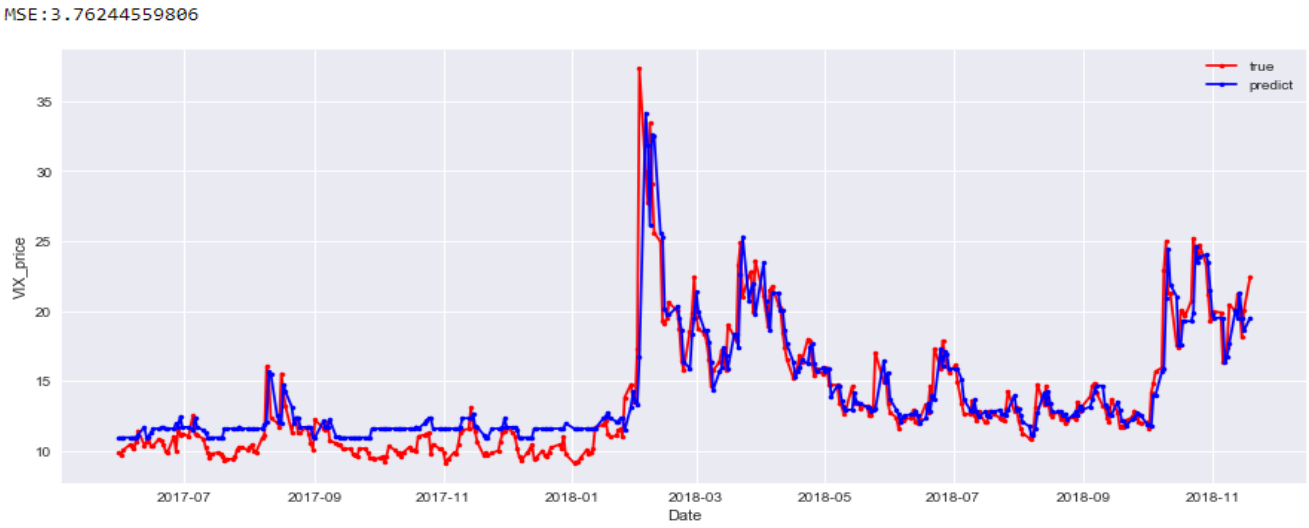

Figure 14: Gradient Boosting

4.7 LSTM

Long short-term memory (LSTM) units are units of a recurrent neural network (RNN). An RNN composed of LSTM units is often called an LSTM network. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. Model performance is shown in Figure 15.

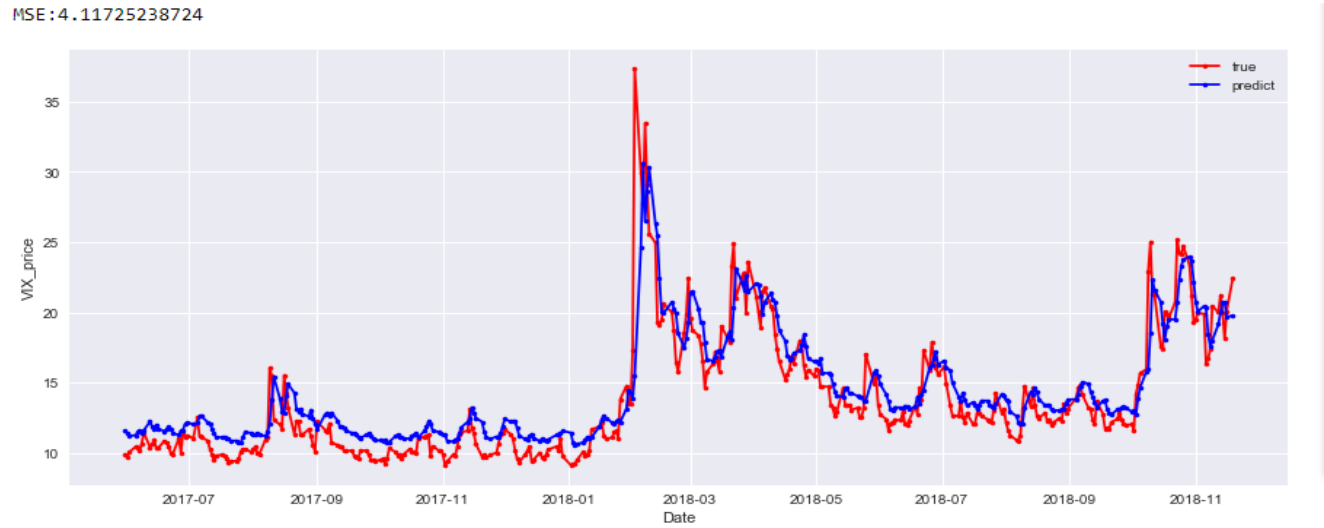

Figure 15: LSTM

To sum up, the Elastic Net model gives the lowest MSE 3.13. Random Forest, Decision Tree, Gradient Boosting and LSTM are also accurate models with MSE about 4. However, KNN and Support Vector Machine have poor performance for prediction.

Model	MSE
Elastic Net	3.13
KNN	15.75
SVM	10.04
Decision Tree	3.60
Random Forest	3.67
Gradient Boosting	3.76
LSTM	4.11

Table 1: MSE for different models

5 Trading Strategy

Although we cannot trade the VIX Index directly, we could trade some derivatives of VIX Index following the index, such as iPath S&P 500 VIX Short-Term Futures ETN VXX. Considering transaction costs, I designed a simple trading strategy with three parts: '1' means long, '0' means empty, '-1' means short. Then I simulated a trading process of VIX index.

Assuming that transaction cost is 0.5%, if predicted daily return is larger than 0.5%, long VIX index, if it's less than -0.5%, short VIX index, otherwise, do not invest at all. The PnL curves for all the machine learning models with the best parameters above were shown below. I also calculated Information Ratio as a simple measure for trading performance.

Figure 16: PnL curves for different machine learning models

Model	IR
Elastic Net	1.33
KNN	-0.02
SVM	2.42
Decision Tree	1.78
Random Forest	1.39
Gradient Boosting	0.49

Table 2: Information Ratio for different models

To sum up, SVM gives the best Information Ratio 2.42. Elastic Net, Decision Tree and Random Forest are also valuable models with IR larger than 1.3. But KNN and Gradient Boosting have poor performance for trading.

6 Conclusion

In this project, I predicted VIX index through different machine learning methods. Some models presented excellent performance, which means VIX is predictable with large amount of

data and scientific methods. More importantly, the prediction results can be used for trading and earning profits.

From the perspective of VIX prediction, Elastic Net(alpha = 0.001, l1_ratio=0.6) gives the lowest MSE 3.11. But from the perspective of trading, Support Vector Machine Regression gives the best Information Ratio 2.42, which means trading profit is a different measure of performance other than mean squared error.

Since this project is focused on predicting VIX accurately, regularized linear regression (Elastic Net) is the best choice for predicting VIX. However, I think LSTM would present the best result if better parameters and hyper-parameters were set. In the future, I will learn more about deep learning and try to improve my LSTM model.

Also, feature engineering is very important. Most of my features came from many books and papers. The features are claimed to be highly correlated with stock market volatility. I analyzed some characteristics of those features and utilized boxplots, histograms and random forest model to select the best features and avoid noise. In the future, I need to find better features and transform all the features into more proper forms for better machine learning results.

References

[1] Kiran Manda, Stock market volatility during the 2008 financial crisis, 2010.

[2] Wes McKinney, Python for Data Analysis, 2012.

[3] Jake VanderPlas, Python Data Science Handbook, 2016.

[4] Varun Kapoor, etc. Predicting Volatility in the S&P 500 through Regression of Economic Indicators, 2017.