VIX prediction midterm report

Hanyin Cao, Yiyang Wen, , Leqi Zhao
Cornell University
{hc936,yw892,lz469}@cornell.edu

Abstract

Uncertainty plays a crucial role in financial analysis, and it is recognized as a complex concept based on its correlations with macroeconomic environment, market expectation and investor's sentiment. Among various methods of describing uncertainty, volatility is the most practical one which gives us a simple way to quantify the uncertainty. Here, we introduce a ticker symbol for the Chicago Board Options Exchange(CBOE) Volatility Index—VIX, often referred to as the "fear index" as it is a measure of the stock market's expectation of volatility. It is constructed using the implied volatilities of a wide range of SP 500 index options.

In this project, we are going to develop an efficient model to predict the direction of VIX, so essentially it is a classification problem with three categories: up, stable and down. We utilized efficient models such as linear regression, ridge regression, lasso and random forests to achieve our goal.

Data Overview

The data we are using can be grouped into three categories:

a The main financial market indices which include SP 500, Nasdaq, Russell 2000 and Dow Jones Industrial Average. These indices can reflect the domestic market condition to a large degree as they include a wide variety of securities.

b Active options contracts indices including ESA index, SPA index and other 11 indices. Essentially, VIX reflects the implied volatility of the market, so the price of the active options contracts may have fundamental influence on the the VIX.

c Macroeconomic data can reflect general condition of macroeconomy, which influence the general expectation and further influence VIX index. The data we find include daily, weekly, monthly and quarterly indices including consumer price index for all urban consumers, crude oil prices, USD/EUR foreign exchange rate and other 24 data.

d Sentiment data indicate the market emotion. Such data can be divided into two parts. First parts are the 23 sentiment indices obtained from Bloomberg which could reflect the market emotion from various aspects such as ratios of different asset class, the performance of certain securities, social surveys of the investors' sentiment. The second parts are VIX related topics searching volume through Google trends.

More detailed data description including features full name, meaning, frequency, starting and ending dates, please refer the appendix file in our Github folder.

Data Gathering

The main source of our macroeconomic data is https://fred.stlouisfed.org/. The website includes a wide range of economic series. On the other hand, we obtain our financial and sentiment indices through the Bloomberg terminal on campus. Besides, we chose 50 most relevant keywords with respect to VIX and utilized a python package pytrend to acquire google trends data through API.

Data preprocessing

The data we are going to process is quite messy, with different time range, different frequency and different ratio of missing data. Before gathering all features, we decided to preprocess our three categories of data separately. Firstly, we convert the date to standard format. In particular, our macro data had different frequency: daily, weekly, monthly, quarterly. Since we want our final data is in daily basis, we fill in our lower frequency data value to a corresponding range of time.

Data Description

First, let's have a look at the distribution of each type of data. We draw the hist plot for each feature and violin plot for VIX index together with the distribution fitting function, which are as follows:

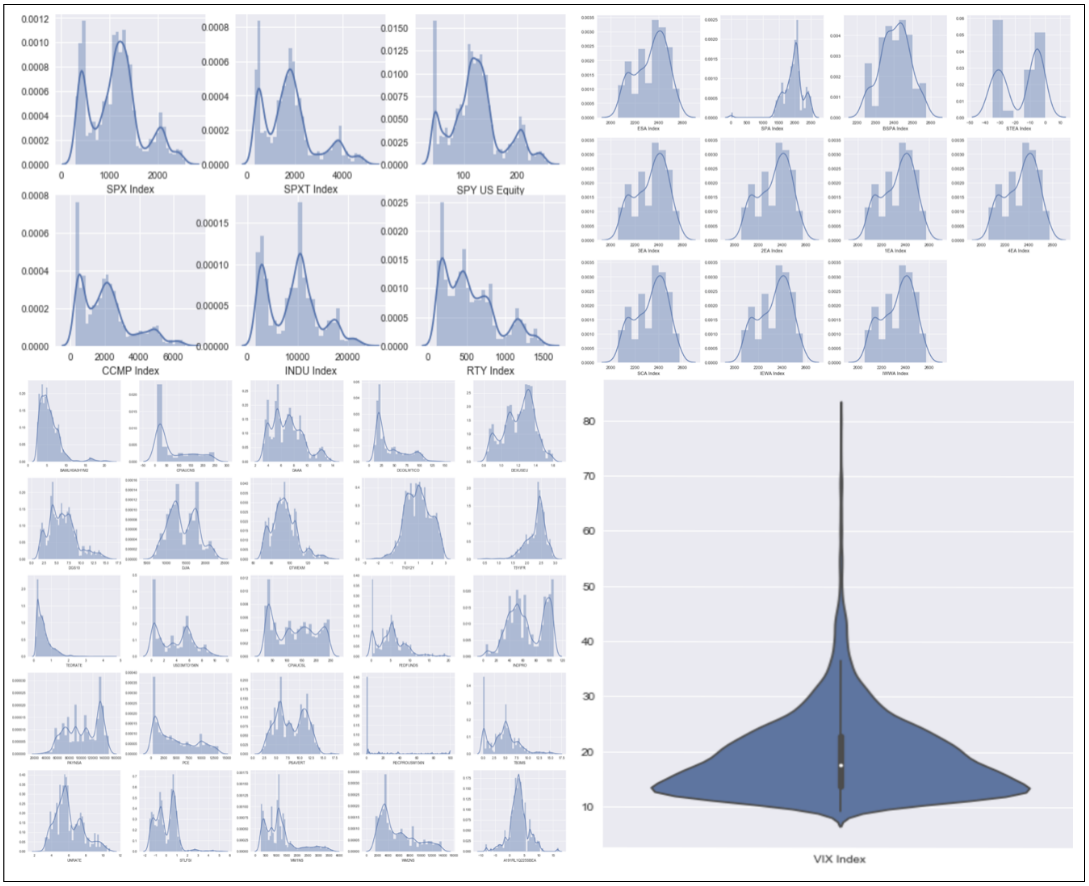


Figure 1: Distibution of Features and VIX index

Correlations

We found that for market index data and active contracts data, some distribution plots are so similar that we next conduct correlation analysis to figure out whether there exists strong correlation relationship between them.

As the figure 2 shows, we found that for market index data, SPX, SPXT, SPY have strong positive correlation and for active contracts data, all features expect STEA have correlation of 1. So we just eliminated the redundant features and only remained ESA index since it has the most data points.

Group the VIX data into three classes

In the real world, what we really care is the trend of the VIX index so that we can trade on them directly to gain profits. In this project we groups the VIX data into three classes based on the following formula:

$$Target = \begin{cases} 1 & \frac{VIX_t}{VIX_{t-1}} >= 0.2 \\ 0 & \frac{VIX_t}{VIX_{t-1}} \in (-0.2, 0.2) \\ -1 & \frac{VIX_t}{VIX_{t-1}} <= -0.2 \end{cases}$$

Features Aggregation

Since our VIX data is range from 1/2/1990 to 10/23/2017, we decide only remain the data for all features with the same start date.

Missing Values

As different features have different time ranges, they have various missing value ratios as the following figure shows:

In this project, we take several steps to handle the missing values.

1. We drop the features which missing value ratios larger than 0.5

2. We drop the features that has the low variance smaller than 0.1

3. We impute the missing values with the former nearest values because of the property of time series data.

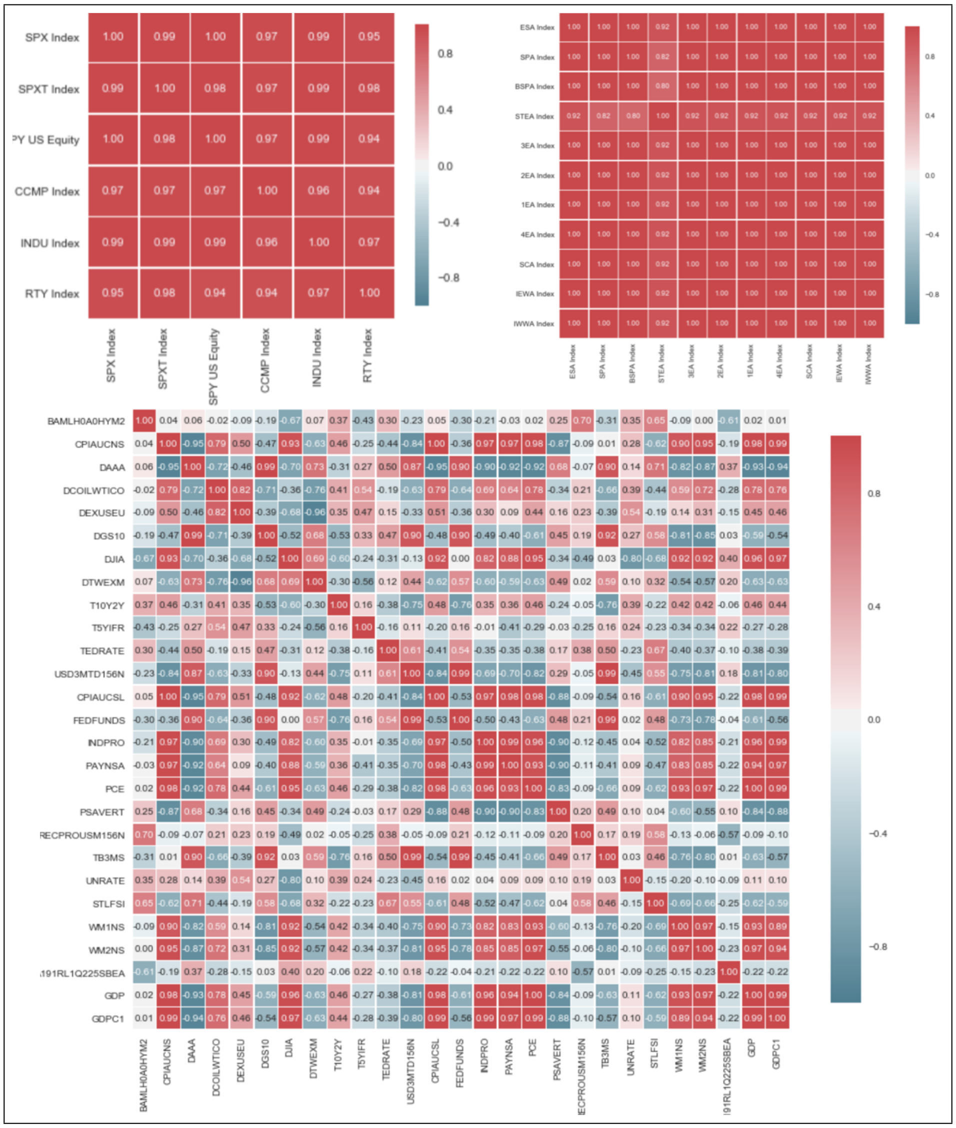
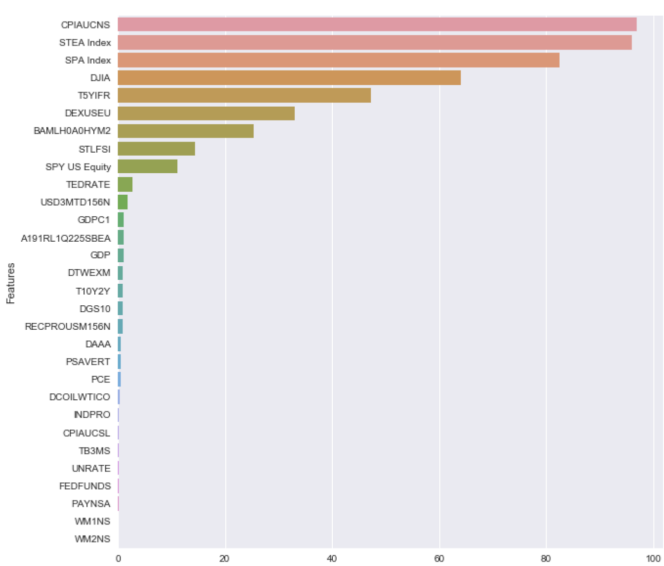
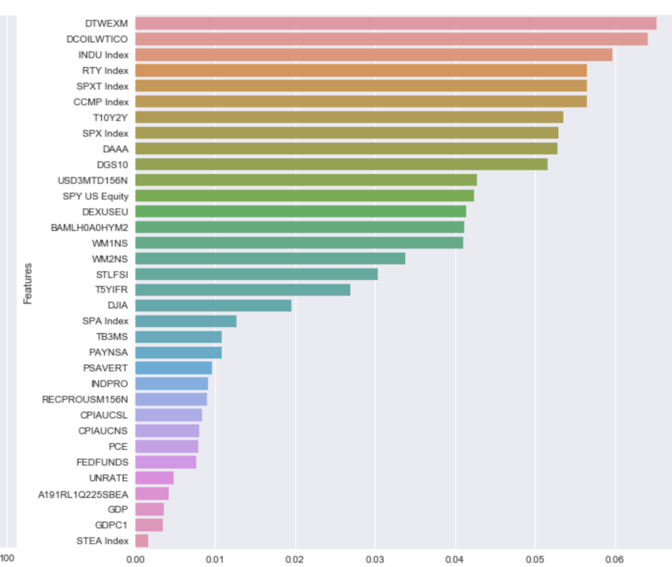


Figure 2: Correlation Analysis



(a) Missing Ratios



(b) Features Importance

Feature Importance and Feature Selection

Then based on the current features, we trained a random forest model and ranked the features according to their importance, which is a attribute of random forest model. The result is showed in figure 3 above.

Then we select the features with the feature importance larger than 0.008.

Baseline Model

Our preliminary analysis is to use extra trees model to train the data in the train dataset (from 1/2/1990 to 12/31/2012) and predict the directions of VIX based on the test dataset (from 1/1/2013 to 10/23/2017). The prediction accuracy of this baseline model is 0.56.

Next Step...

1. So far we have processed the macroeconomic data, market indices data and active contracts index and add them into the model. In the next step, we will furtherly absorb the sentiment indices and google trends data into the models, improving the prediction ability.
2. We have gathered some indices data which can be used to reflect the market sentiment, but we need feature engineering upon them. By researching, we will come up with proper way of feature engineering to enhance the quality of input features..
3. We have only used extra trees model to predict the trend of VIX. However, other valuable models can be utilized such as random forests, lasso and ridge regression. We will compare the predicting power of different models and analyze the reasons behind that.

Reference

1. http://www.investopedia.com/terms/s/sp500.asp
2. https://en.wikipedia.org/wiki/Macroeconomics
3. https://en.wikipedia.org/wiki/S