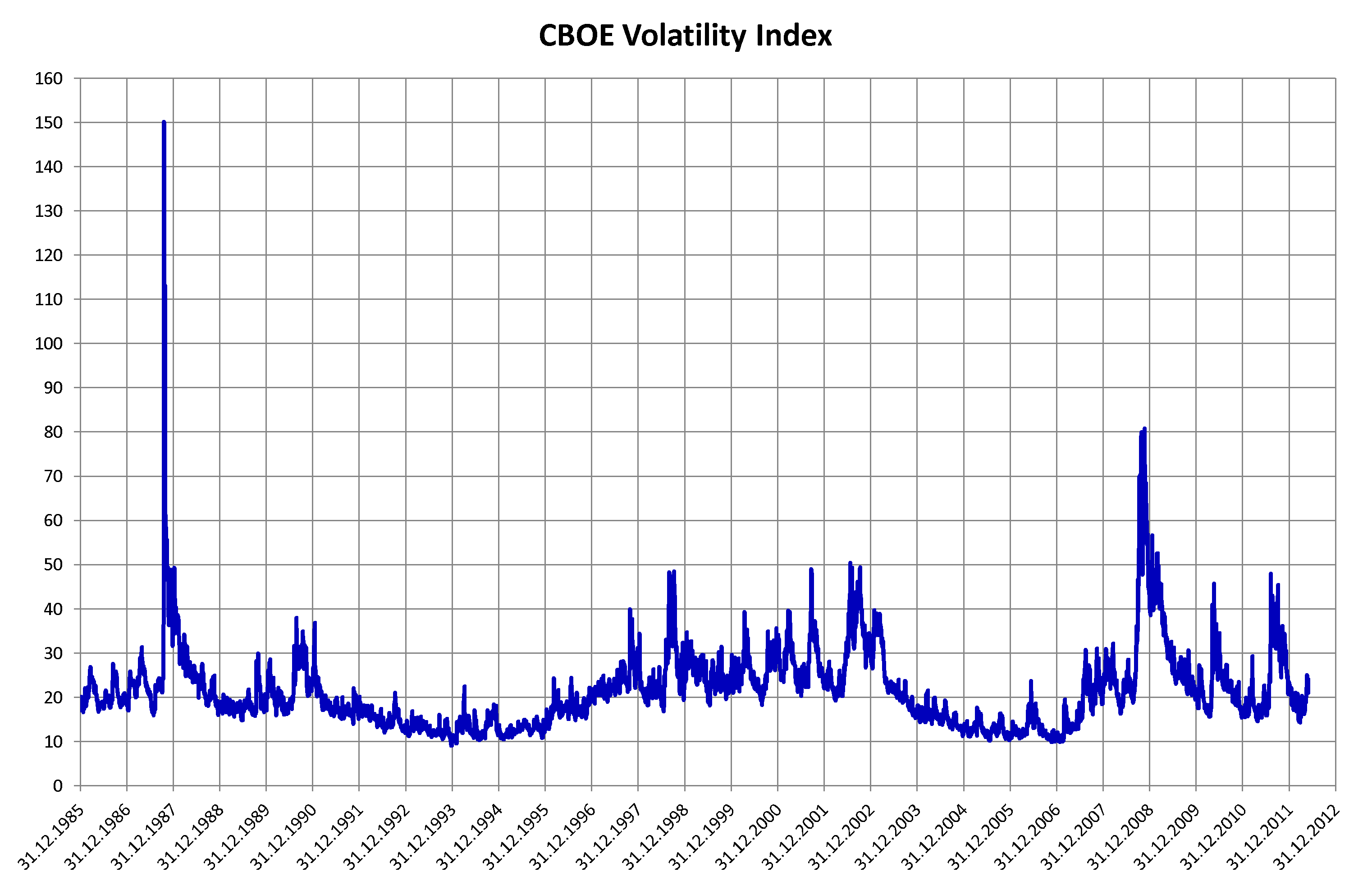
1. Time scale of the train file is 2002 – 2009. See below the comparison between realized cross-sectional stdev of asset returns vs historical implied volatility curve (CBOE VIX index)



I think I don’t need to elaborate how similar they are. Also below is the full curve for CBOE VIX index,



There are two more ranges after 2009 with significant heteroscadesticity. So pay attention to that as the test set is very likely to be after 2009, why? There are 8.8 years in the train file, and there are 6 years between 2009 and 2015 (2016 is ruled out because it’s not completed yet), hence the train file takes about 60 % of total data point with 40 % goes to test set, which is what I would do if I am the guy who designed this challenge from two sigma. But of course, this is only my deduction.

If you are interested in digging more into the time scale, try to draw the curve of average missing rate of a specific factor (just randomly select one) vs time stamp, you will see strong seasonality at two periods that confirms my above conclusion.

1. Those features are called “factors” in practice, i.e. Fundamental factors, technical factors, etc. And for a specific stock they tend to be slow-moving, for example, you can’t expect book value or market cap of a company to swing a lot. Hence a reliable way to impute missing value is look at the historical factor value for the stock you are predicting.
2. Don’t try to classify assets based on their asset-return performance, and I can guarantee you that you will find two types of stocks, stocks within each type has positive correlation with each other, and stocks between types has negative correlation with each other. If you do see this, I am afraid it’s pointless, because you are most likely gonna find out the first type is called “good performing stocks”, the second type is called “bad performing stocks”, which is completely useless. In fact, no matter what your time scale is, there are always good and bad performing stocks, this doesn’t mean there is fundamental difference between them, so it’s dangerous to classify them.
3. There is indeed one fact in this train file that has strong sign to classfy assets into two types: the missing pattern. I won’t go into details here, but if you are interested, you should check for each feature / factor, which of those stocks are not applicable / missing. You will find there is a group of stocks for which a set of factors are always missing, yet those factors are filled for other stocks. This implies there is a possibility that the first group of stocks are either small-companies that don’t publish enough announcement or they are simply from emerging markets.
4. I am sure all of you already noticed factor values are quite irregular, most of them have a peak and very narrow dispersion, one way to handle this is transform factor values, which is very popular among portfolio management. One way is to do:

X’ = 1 / (1 + exp(-X))

1. I wouldn’t recommend try any exotic machine learning technique like CNN, RNN. Based on my experience, they are very effective in highly structured features, for example, human eye can recognize whether a guy is drunk or not by looking at his move, listen to his voice. But mathematically, this is very hard to describe. I call this kind of data “structured”, which means in order to find useful feature, you need to use highly non-linear model to capture them, hence deep learning models. But financial data is not exotic at all, in fact, they are highly noisy, and there is a big difference between being super non-linear and being super noisy. So I would recommend to use solid basic model to capture the super weak signal in financial data. Only use exotic models if you have strong belief and statistical significance to do so.
2. One typical way to solve this portfolio return prediction problem is “fundamental factor model”, which simply do a cross-section linear regression at each time stamp, and try to study the time series of regressed coefficients (mostly, the average is taken for future prediction because financial data is too noisy to assume an exotic model, remember?). If you try out this method (you have to use y values of previous step you just predicted, which is not allowed in this competition unfortunately). You will get high R-value, I got R = 0.8 in my backtest. Hence if you want to stick with this model, I suggest you focus on how to mitigate the problem of “not allowed to use y values after prediction” issue, which, in my personal opinion, is quite weird.) One way I thought about but haven’t tried out is: regress “the coefficients / slopes” at each time stamp mentioned above onto the “reward” for previous time stamp, see if there is any prediction power, if you can somehow get a rough estimation of the regression coefficients for next time stamp based on the reward you got at current time stamp, then I think you will have a sweet R value. Btw, the cross-section regression coefficient is called “factor premium” in “fundamental factor model” in case someone want to do more background research.