Capstone Project | Matthew Aronowitz

Data Science | 11/25/2019

Definition

Project Overview:

Option prices are priced using the underlying asset's volatility. Options act as price insurance for an asset. The more volatile an asset, the more options cost because there is a greater chance these options will pay or end up 'in the money'. Volatility is an annualized standard deviation and a measure of how much an asset moves. Predicting the future volatility of asset is very difficult but essential for pricing options. Here is a link that shows the describes options in more detail and explains their relationship to volatility:

https://www.investopedia.com/ask/answers/062415/how-does-implied-volatility-impact-pricing-options.asp

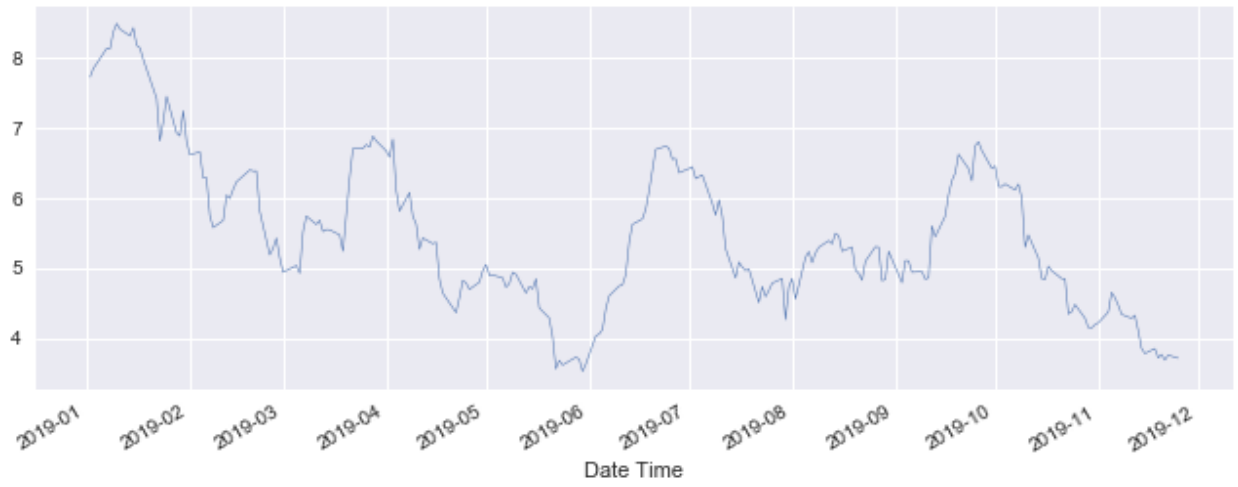
Problem Statement:

The goal is to find methods to predict the future volatility of an asset. Select the focus, the asset whose vol we are trying to predict. Predict the focus asset's vol by using it's own history and using the relationship between that asset and other similar assets within that class.

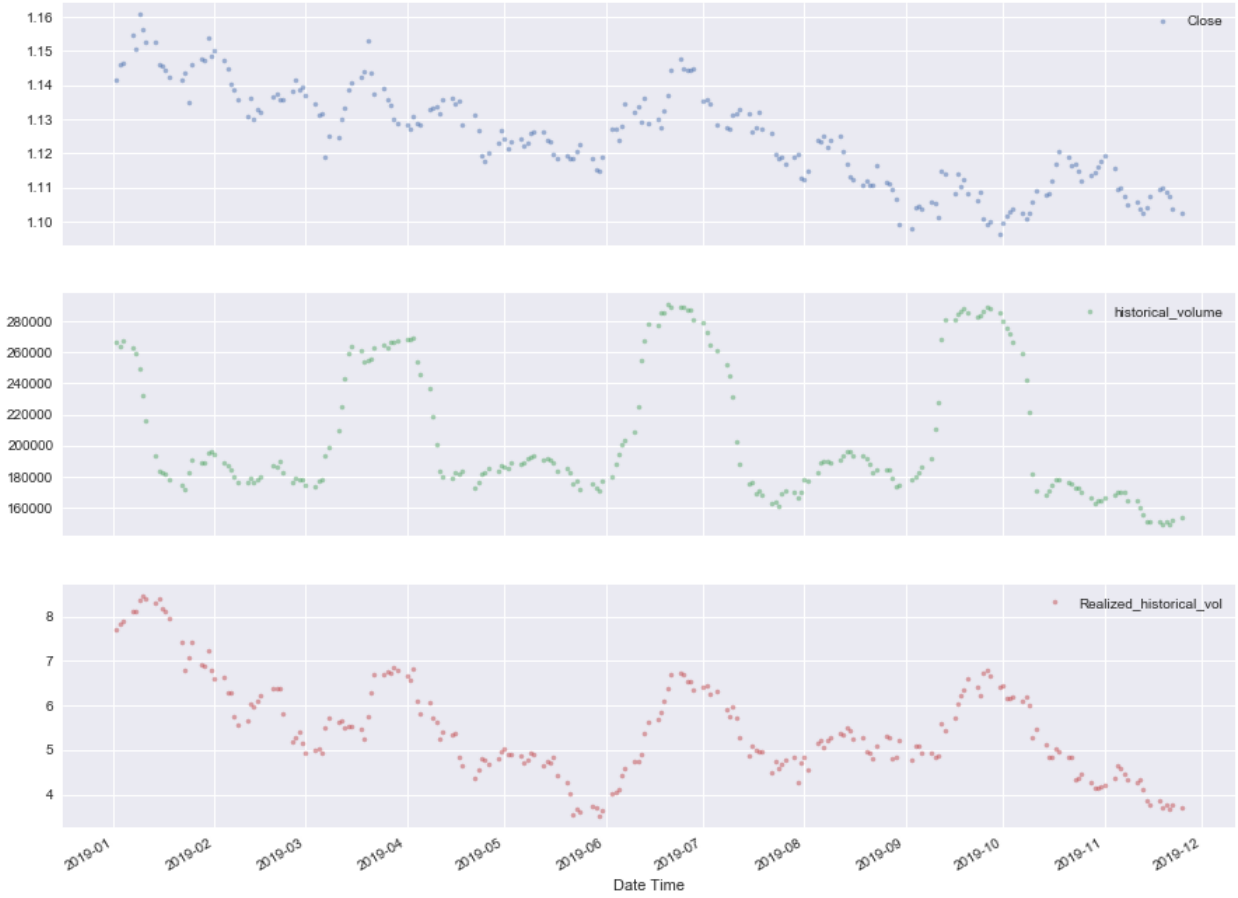
1. Download csv daily price history for the past 10 years from barcharts.com for interested assets. For this project I downloaded the price history a variety of currencies. Each currency is a separate csv file. I downloaded price history for the EUR/USD, JPY/USD, GBP/USD, CAD/USD, AUD/USD. Each download file includes the date, symbol, Open, High, Low, Close, Change, Volume and Open Interest.
2. Write functions for converting rates of change into vol and vol into rates of change.
3. Process files for time and exploratory data analysis.
4. Write functions for volatility prediction by sampling
5. Write processing functions for vol predictions by Machine Learning
6. Write Regression algorithms for vol predictions.
7. Write Decision Tree classification algorithm for vol predictions.
8. Set variables for predictions
9. Make predictions and score predictions, using sampling, regression and decision trees.
10. Compare and analyze results.

Data Exploration

For this set the focus asset is the eur/usd. Therefore, our models will focus on predicting the vol for the eur/usd. Looking at eur/usd chart of realized volatility below we can see that the volatility is erratic and does not trend. This time series chart is over the past year, so within a short period of time, volatility has had dramatic swings. This makes predicting an asset's volatility difficult without incorporating other techniques.



Next, we graphed the realized volatility against other possible predictors to see if any relationship exists. Looking at the graphs below, we start to see similar shapes forming between price and volume. Visually you start to see some correlation between these variables. Running some machine learning algorithms would allow us to see if these variables are actual predictors of realized volatility as opposed to mere correlation. Machine Learning algorithms would also allow us to define this relationship.



Method 1

One method to analyze volatility is to take the price action over a period and randomly sample the daily moves for an average volatility over a time period. I wrote a function that processes and cleans the data. Then I wrote a sampling function 'hist_move_sampling' that takes the data and samples the daily volatility and returns the mean volatility over X iterations. This function also provides a standard deviation of those volatilities. In addition, it will provide a confidence level on a pre-set volatility level. So, if we are evaluating an implied volatility level for a time period, it would tell us what percentage of the time the volatility level was higher. This would enable us to have a confidence level in the trade. When I ran this function, I set the date range as the last 222 days. The option I am evaluating has 26 days to expiration, so I randomly sampled 26 days over the past 222 days, 10,000 times and found the average volatility over those 10,000 iterations to be 4.24. The implied volatility for that option is 3.90. Within these iterations we found that 66% of the samples had average volatilities over 3.90. Therefore,

our confidence in the sample volatility being higher than the implied volatility would be 66%. So if we were using historical volatility over the past 222 days as a proxy for future volatility we could be 66% confident that buying options with an implied volatility of 3.9 would be profitable.

Method 2

In a similar manner, method 2 involves sampling. Method 2 is trying to sample for an assets straddle price. Here is a link that describes a straddle price of an asset: https://www.investopedia.com/terms/s/straddle.asp.

In order to estimate the straddle price we need to determine the trading range. Within this function 'sim_option_value', the user selects the number of days to sample, iterations and starting price. It then calculates the total move from the starting price assuming those randomly selected days were consecutive. It then calculates the absolute move and multiplies it by two (put and call) to arrive at the straddle price. From the straddle price, using tools outside this project we can solve for the implied volatility using the black-scholes formula if we have a straddle price. Here is a description of the black-scholes formula:

https://www.investopedia.com/terms/b/blackscholes.asp.

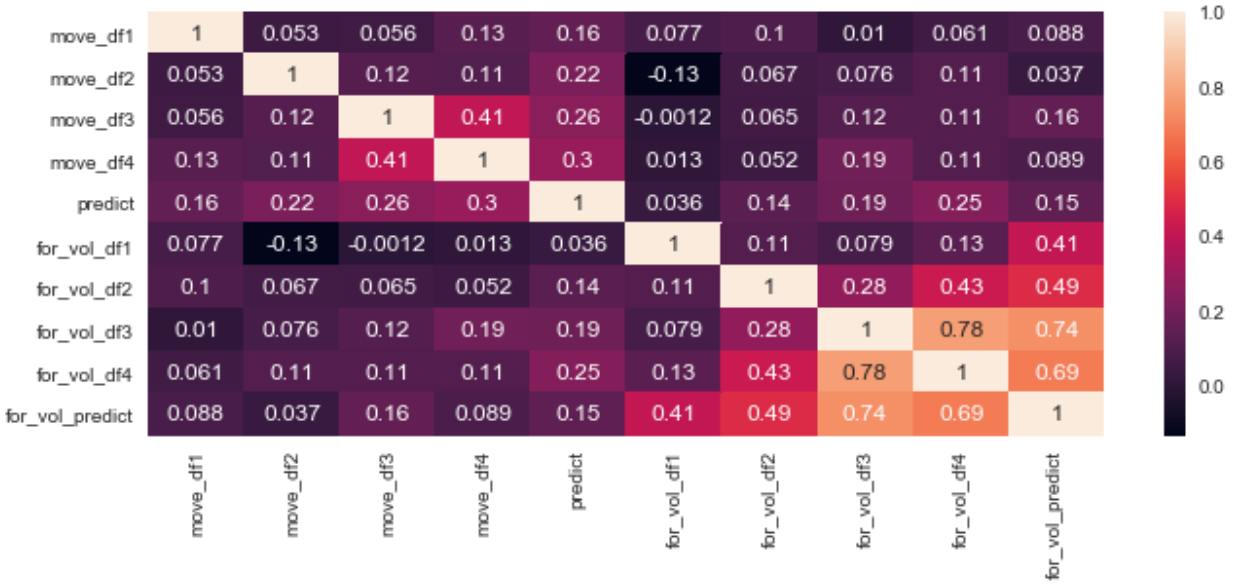
Using the same sample size and data range as in method 1, we found the straddle price to be .0115. Solving for the implied volatility we get a volatility of 4.25.

Method 3

Method 3 involves using simple linear regression to find the appropriate volatility for the focus asset based on the volatility of one predictor asset. We wrote a function that processes the data frame that combines both the focus asset data and the predictor asset. This regression function 'predict' trains on the data set to find the volatility relationship and then tests that relationship with R2 and RMSE scores. To make the prediction we use the implied volatility trading in the market. Therefore we get a relative relationship between the volatilities. The function uses Machine Learning to determine the two asset's historical relationship, but the market trading implied volatility on the predictor to solve for the volatility on the focus asset. In our case we selected the JPY/USD to be the independent variable to predict the EUR/USD volatility. The function optimizes the polynomials by searching for the model that returns the lowest RMSE. We also convert the RMSE into volatility so we get a sense of how many volatilities is the standard error. In our case, we ran a 3 degree polynomial model with a R2 of .21 and a RMSE in volatility terms of .99 vols. The predicted volatility on the EUR/JPY using an implied of 4.5% on the JPY/USD is 4.57%. That would imply a .07 predicted volatility difference between the JPY and EUR. In this case because the RMSE in vol is almost a full vol, I do not have much confidence in this number but it is interesting that this Machine Learning algorithm is predicting a slightly higher vol than the two sampling models.

Method 5

Method 5 uses a multiple regression model to see if we can improve on the RMSE by incorporating other assets. In a similar manner to method 4, our multiple regression models process the data frame but incorporates 5 assets. Running a correlation matrix on the features we can see some relationship between the variables against the focus asset's forward volatility represented by 'for_vol_predict'.



Similar to Method 4 we are using predictors to train and score a regression model and then use their implied volatilities to predict the volatility of the focus asset. Here we use the Yen, Pound, Canadian dollar and Australian dollar's forward volatility as a predictor of the Euro's forward volatility. The function 'predict_multiple' takes a data frame that has been processed with multiple variables and returns the scores and prediction. This function optimized at 2 polynomial degrees and returned a model with a RMSE in vol of .45, half that of the simple regression model used in Method 4. Here we get a predicted vol of 2.9 which is the lowest of all the methods we have run. This predicted vol is of 2.9 is relative to the implied vols we used to make the predictions. Therefore it is relative to the implied vols of the predictors which was JPY = 4.5%, GBP = 10.75, AU = 5.6 and CAD = 4. This model predicts a relative relationship as opposed to an absolute volatility level.

Method 6

Another method looks at other variables related to volatility and uses a machine learning classification algorithm, Decision Tree. Here I wrote the function 'find_vol' that tests a variety of volatility levels and returns a 1 if the forward volatility is predicted to be greater or a 0 if the forward volatility is not predicted to be greater than certain level. The user can set the various levels of volatility to test. The algorithm trains on a data set that includes various historical volatility levels, volumes and open interest levels. It then compares that to the asset's forward volatility level, denoting a 1 or 0 depending if the forward volatility level ended up being greater than the prescribed level. Here our focus is still the EUR/USD and when we run and test the various volatility levels using current historical volatilities, volumes and open interest we get 1 until the volatility tested reaches 5.75. Below 5.75 we got a 1 and above 5.75 we get a 0. This was scored with a F1 of 75% and a Recall of 82%. The test accuracy was 66%. We have more accuracy the lower the volatility level. We do not need to know if it will exceed 5.75 because the current implied volatility is 3.90%. Looking at the simple regression model which predicted a volatility of 4.57% on a relative basis, we might use this model to confirm that level on an absolute basis. At 4.55% we get a predicted 1 with a precision of .9464, test accuracy of .94 a recall of 1.

Conclusion

Looking at all these methods, the implied volatility of 3.90% seems very low. In Method 1 and 2 we are getting vols of 4.25%. Method 3 predicts a vol of 4.57 on a relative level. That level is confirmed with a high degree of confidence on an absolute level using the Decision Tree algorithm. Only the multiple regression model returns a vol lower than 3.90 but that is on a relative basis. The market could be pricing those relative assets too low which would then lower my multiple regression predicted volatility output. Looking at the vol levels as a trader, 3.90 seems low and my intuition is that it represents a buying opportunity. This seems to be confirmed with most of my models predicting a vol level between 4.25% and 4.60%. Given these models, I think there is a low probability of loss buying EUR/USD options at the 3.90 volatility level.