WorldQuant University

Machine Learning in Finance

Solutions Design

Version 1.0

Authors:

David Kwasi Nyonyo Mensah-Gbekor(mensah-gbekor@hotmail.com)

David Wonder Doe-Dekpey (wonderdoe85@yahoo.com)

Alexander Botica (alexbotica@yahoo.com)

Alexander Victor Okhuese (alexandervictor16@yahoo.com)

Project Instructions:

Submission One

Data:

- Start Date: 9/01/2013
- Raw Tick Data: Downloaded from

https://github.com/Jackal08/financial-data-structures/tree/master/raw_tick_data.

On a series of raw tick data:

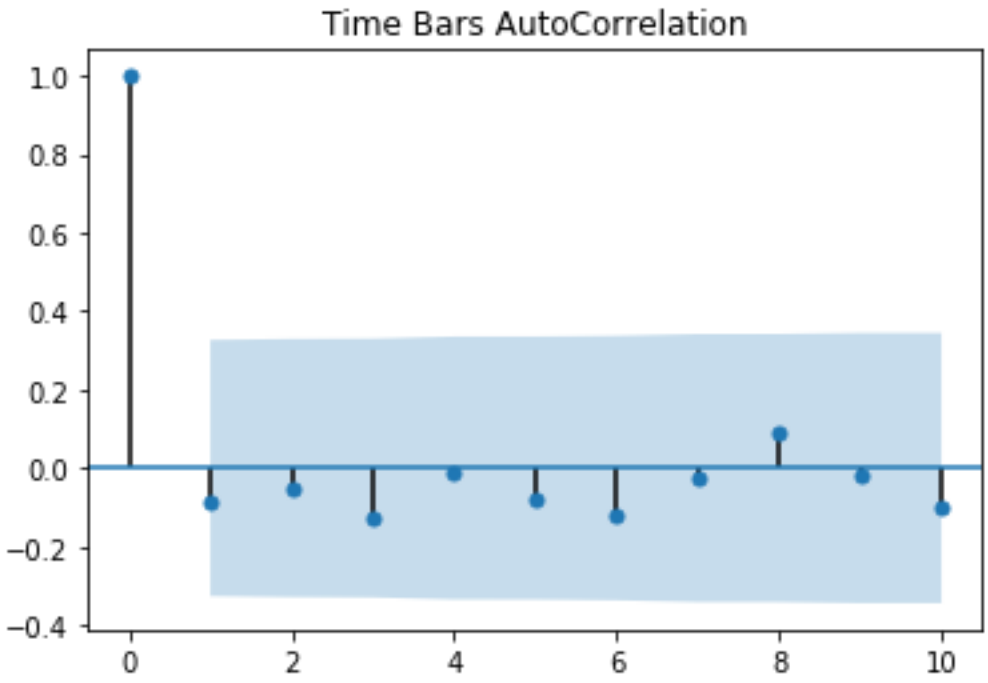
1. Create tick, volume, and dollar bars. (Bar must have open, high, low, and close values.) Students can do this from first principles or clone the following repo for an implementation.
2. Count the number of bars produced by tick, volume, and dollar bars on a weekly basis. Plot the time series of the bar count. What bar type produces the most stable weekly count? And why?
3. Compute the serial correlation of each bar type and report back on which method has the lowest serial correlation.

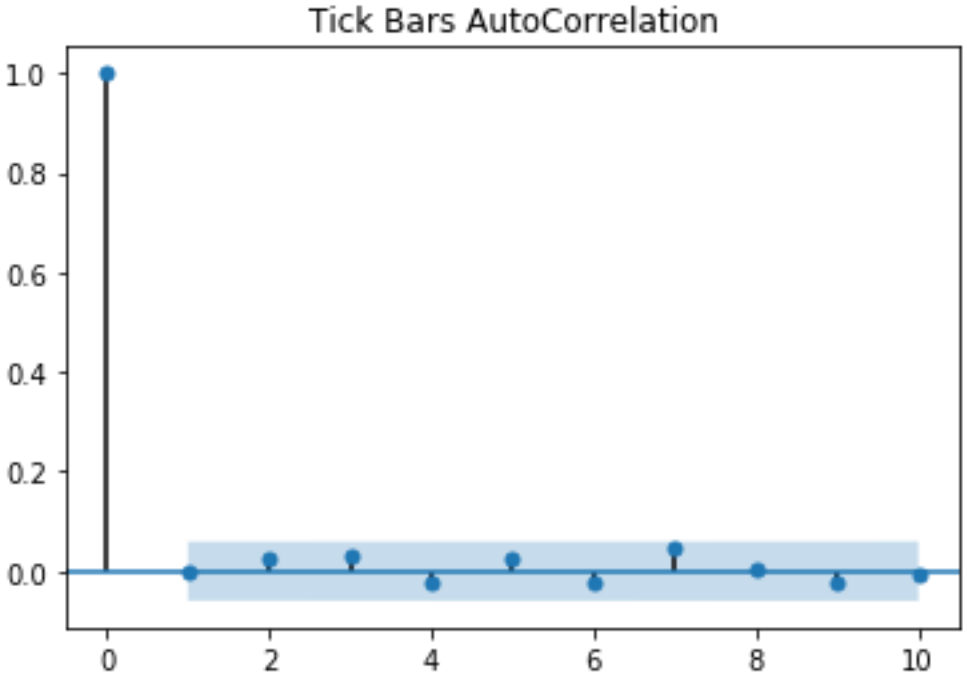
4. Apply the Jarque-Bera normality test on returns from the three bar types. What method achieves the lowest test statistic?

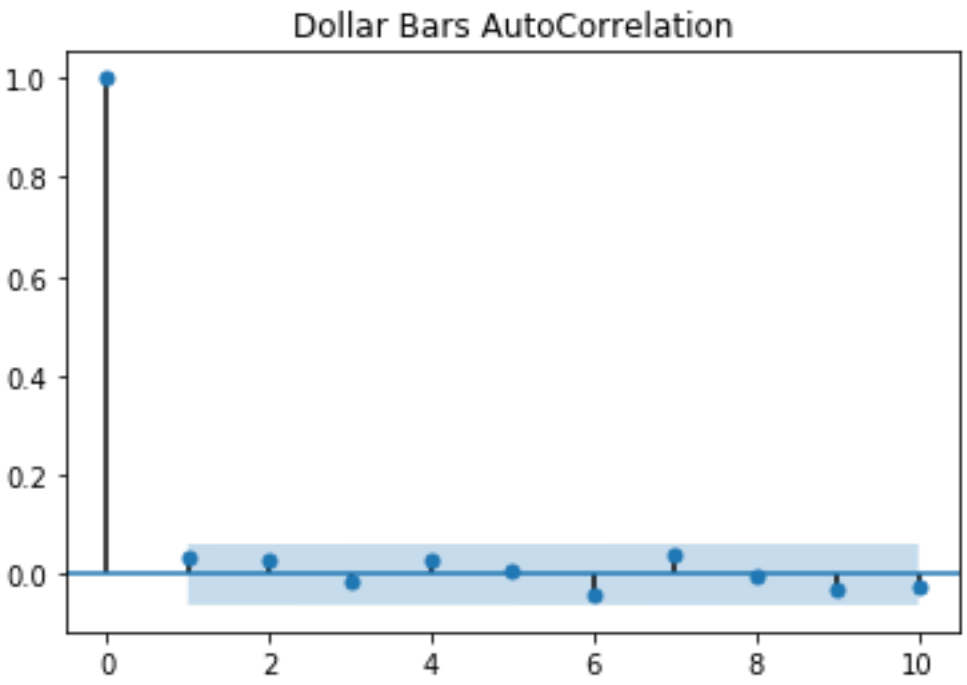
Write a 500-word report on your findings and research. Make sure to use the Harvard reference style.

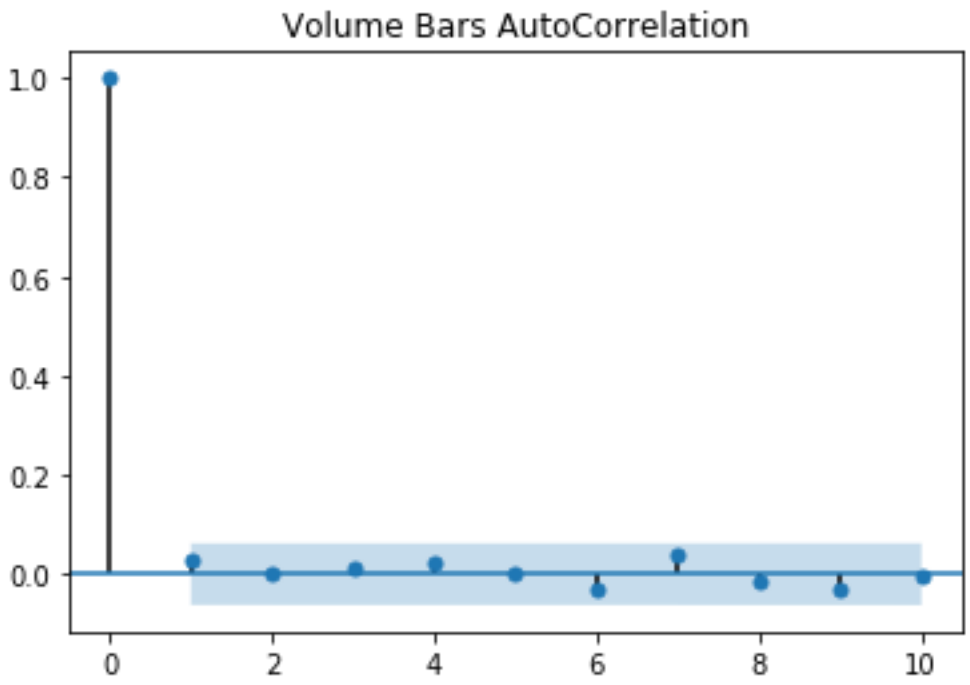
Steps towards Project Completion: Submission One

1. Created tick, volume and dollar bars from tick data downloaded from the repository https://github.com/Jackal08/financial-data-structures/tree/master/raw_tick_data. The implementation can be found in the accompanying notebook.
2. Counted each bar type for the standard deviation calculated we got 47.29% for tick, 47.77% for volume and 48.04% for dollar. The differences are not statistically significant; therefore, we cannot define the most stable one. The reason is because we do not have enough data to support a strongly based decision. Hence, more data would be necessary to answer this question.
3. Computed Serial Correlation of each bar type.


Time Bars AutoCorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
2
4
6
8
10


Tick Bars AutoCorrelation
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10

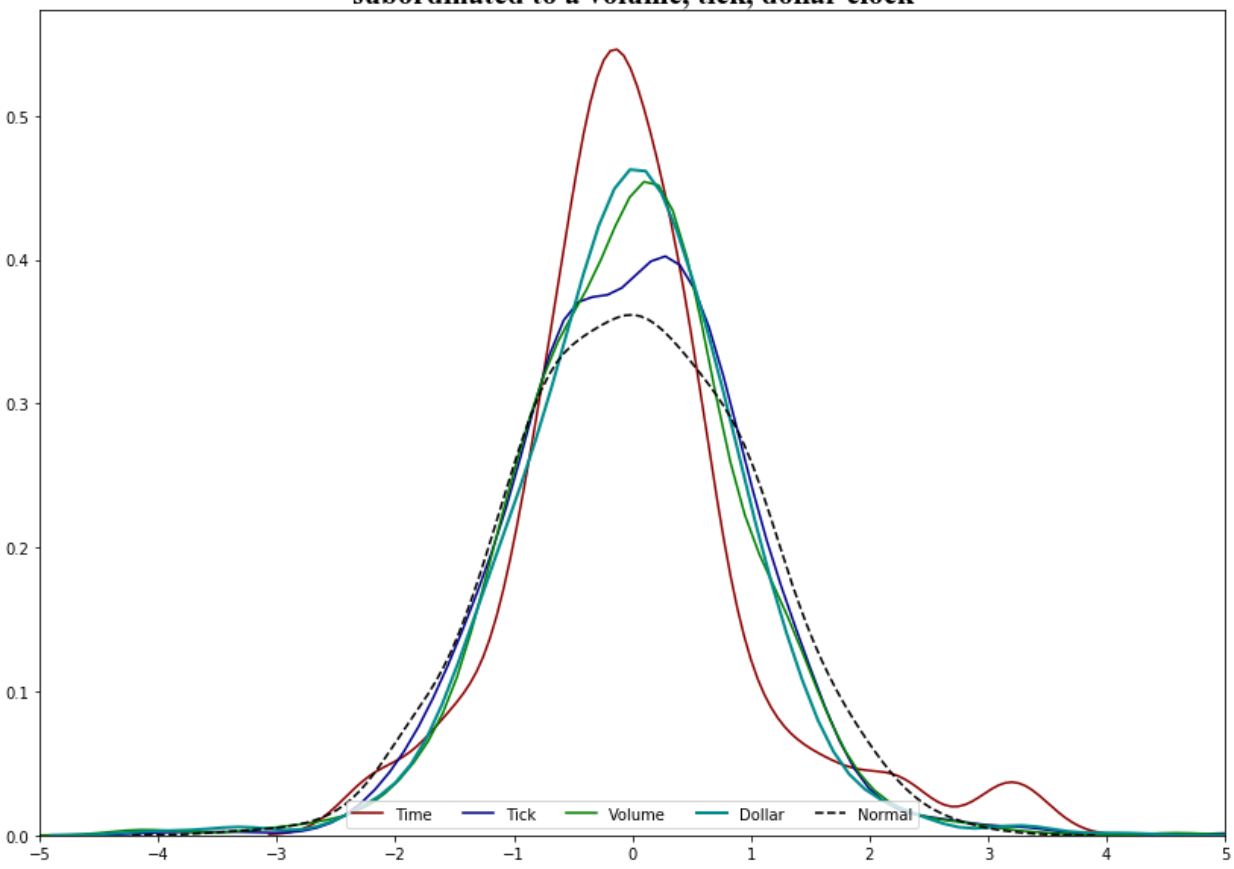

Dollar Bars AutoCorrelation
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10



The relationship between a variable and a lagged version of itself over various time intervals with the lowest correlation was presented by tick bars, with a coefficient of -0.002.

4. For this task, besides the Jarque-Bera normality test on returns, we also implemented another normality test known as Shapiro-Wilk Test. This strategy was used because the Jarque-Bera tests statistic asymptotically and has a Chi-squared distribution with 2 degrees of freedom; consequently, we should have at least 2000 data samples in order to this test to work (Jarque and Bera, 1980). As we only had a few samples over 1000, we implemented the Shapiro-Wilk Test. The result for both, Jarque-Bera and Shapiro-Wilk were consistent and indicated ticks bars with the lowest statistic test (more normal).

Exhibit 1 - Partial recovery of Normality through a price sampling process subordinated to a volume, tick, dollar clock



Submission Two:

1. Select at least four explanatory variables and perform the necessary transformations so that they are useful in the model phase. You are encouraged to use more than four variables. Investigate feature engineering techniques such as PCA and encoding target variables using one-hot encoding.

2. Write a short paragraph about each technique investigated and show an implementation of it in a Jupyter Notebook. Make sure to include references that indicate where the ideas were sourced.

3. At this stage groups should take the opportunity to familiarize themselves with the cross-validation techniques for forecasting financial time series – for example, traditional k-fold cross-validation versus walk forward analysis, and Purged K-Fold CV. Write a short paragraph explaining each technique researched. Research at least three (they don't have to be the 3 mentioned here).

Steps Taken towards Project completion

1. Explanatory Variables

Close Price of Peer S&P500

We chose this variable because they are correlated. They are correlated in a sense that if the price of one is in a particular trend we can easily tell the trend in which the other will be in. It's implementation can be found in the accompanying notebook.

Close Price of Peer Dow Jones Industrial(US30)

We decided to add this variable to the input variables due to the fact that it is positively correlated with the target variable. An implementation can be found in the accompanying notebook.

Moving Average

We also decided to use the 200 Exponential Moving Average also used to identify long term momentum by many traders. An implementation is also found in the accompanying notebook.

Momentum

Momentum is a measure of overall general sentiment that can support buying and selling with and against the trend (Investopedia).

$M = V - Vx$

where: V=the latest price,

Vx=the closing price x number of days ago

We understand that a positive momentum can indicate a bullish trend while a negative momentum can indicate a bearish trend. This is why we chose this. An implementation can be found in the accompanying notebook.

Average True Range

The Average True Range is a technical analysis tool that measures the market volatility. We decided to choose this as an input variable because we noted that when the ATR value is increasing the underlying trend becomes stronger. It has been implemented in the accompanying notebook.

2. Techniques can be found in the accompanying report and the implementation is also in the accompanying notebook.

3. Cross Validation techniques investigated can also be found in the accompanying report