Group Work

Solution Design Document

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

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.

Challenges	**Potential Solutions + Improvements**	**Current Solutions**
Create tick, volume and dollar bars from raw tick data		Time bars creation is hindered due to lack of adequate computing power

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).

Challenges	Potential Solutions + Improvements	Current Solutions
Investigate Feature Engineering Techniques	We could plot the serial correlation of the target variables since feature selection itself is also using domain knowledge.	Log transformation, scaling, filling null values with 0 and standardization
Investigate Cross Validation Techniques	We could find a better PC to aid running of the codes in the notebook and Select a technique	Implemented code but couldn't select which technique to be used during final submission