Q1)

A)

Data Analysis

- It is common practice to first analyse the data you are working with. It gives a solid overall understanding of what you are working with rather than blindly applying ML models. For example, if 2 of the data streams were nearly identical then it may only be necessary to include one of them. Simple insights like this are easily found with some basic analysis.

- Analysis of the data streams can be done with typical time series analysis. This includes exponential smoothing for data with no trends, double exponential smoothing for data with a trend and triple exponential smoothing for data that contains seasonality. More recent methods include Box-Jenkins which can also be applied to multivariate data.

Feature Extraction

-Many forms of feature generation are present with time series. Our goal is to extract as much useful information as possible. One simple feature that could be extracted for each entry is the time. Both the date and time can be turned into features for each entry of the series.

-Since this is a time series, the previous values of the series are of great significance when forecasting. We can simply include a desired number of entries for the previous hours. In addition, features can be extracted with moving averages and volatilities.

-For this context I would aim to extract as much information as possible and let the dimensionality reduction techniques take care of redundant variables. With the advent of Deep Learning and modern GPU’s, models are able to take high dimensional inputs so the more the merrier.

Dimensionality Reduction

-Dimensionality reduction not only eases the computational burden involved in training a model but also leads to less overfitting meaning it will perform better on unseen data. In addition, dimensionality reduction techniques remove features that are of no significance.

-Exhaustive feature selection is the optimal way that guarantees the best set of features for a given dataset. This, however, is not possible when working with high dimensional data. Simple iterative methods such as Forward Feature Selection and Backward Feature Elimination aim to estimate this.

-Another approach comes in the form of Principal Component Analysis (PCA). This method rewrites the dataset in a new orthonormal basis consisting of Principal Components. These Principal Components are estimates of the directions that most variance exists. To further explain, the 1st Principal Component is along the direction with the most variance, The second takes a subset of directions that are orthogonal to the 1st and selects a direction with the highest variance. This process is completed until the PCA space is of equal dimension to the initial data. Feature elimination is then done by simply selecting the first N Principal Components based on some variance metric given to them. It is important to note that the data is scaled before PCA is performed. One disadvantage of PCA is that you lose easy interpretability of the model as features are now linear combinations of columns in the initial data.

-Random Forest is the go to approach for feature selection. Using a bootstrapped aggregation of decision trees and by selecting a random subset of features at each step, the random forest algorithm is able to calculate ‘variable importance’ for prediction of the target variable. Using this, a set of features can be eliminated if they don’t exceed a desired importance. The advantage RF has over PCR is that you don’t lose interpretability. This method, however does not take into account the correlation between variables and separate analysis must be done before a model can be fit.

-For the context of this question I would use both PCR and RF and see which perform better during evaluation.

Performance evaluation & data splits

-When training and testing a model on a given data set it is common to split the data into training, validation and testing sets. Typically the data set is split into around 80% training and 20% testing. The training data is then further split into training and validation (again around an 80:20 split). The training set is used to fit a model and find suitable model parameters. The validation set is used to try different model hyper-parameters and the test set is used to evaluate the model on unseen data.

-When evaluating the performance of a model the MSE is typically used for regression problems. For Classification problems a confusion matrix gives a great understanding of where a model is working or not (and accuracy can be calculated from it). The AUC gives a good metric of how well a classification model has performed.

-Cross Validation is a staple of model evaluation. K-fold cross validation gives estimates of model accuracy as well as confidence intervals. Cross validation can also be used to get estimates for any other model metrics.

-For this question, since it is likely a classification task predicting Buy/Sell, I would cross validate the confusion matrix and accuracy in order to determine if a model is a good fit.