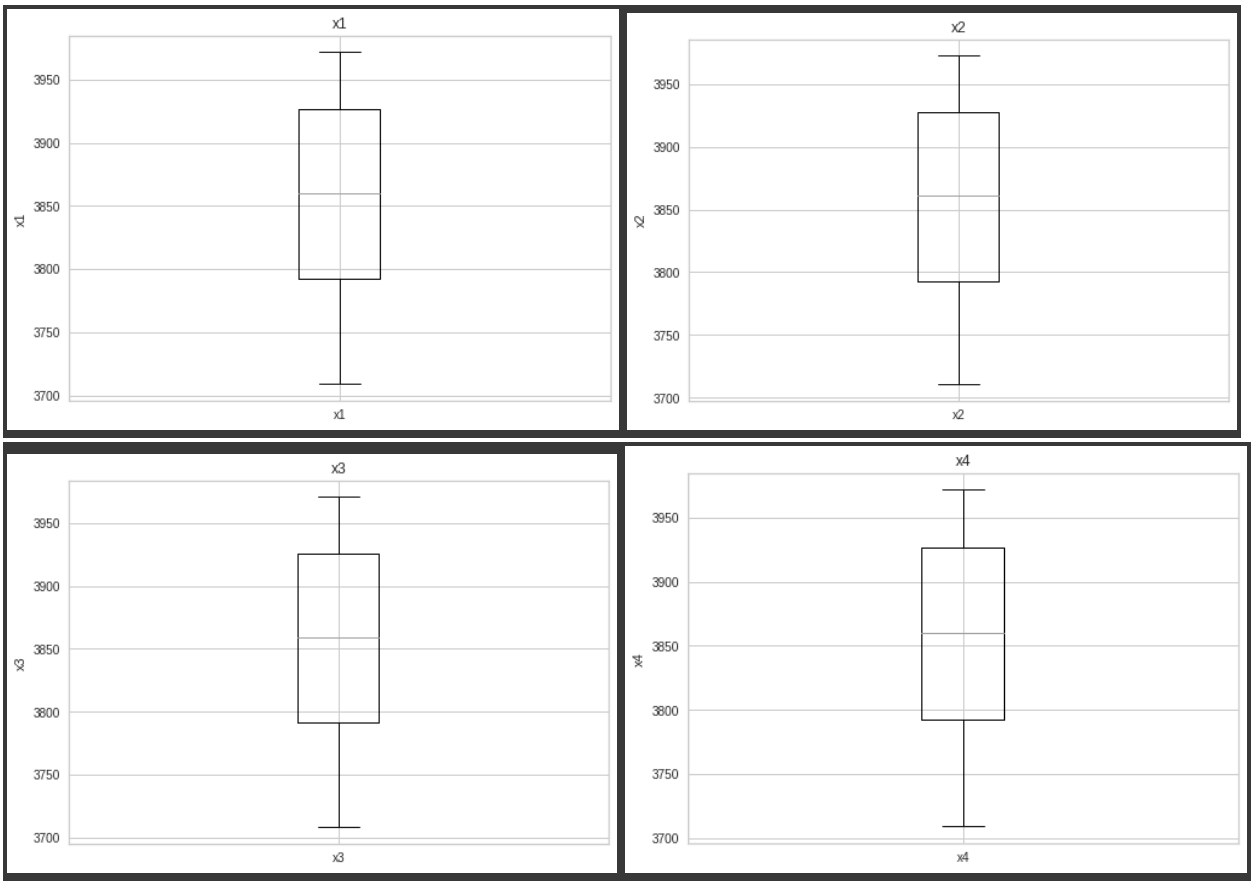
Analytics GC

Bhuvan Aggarwal (solo), 190040026, H9

Data provided:
Training data : OHLC data 10000x4 and y labels
Test data : OHLC data 1000x4

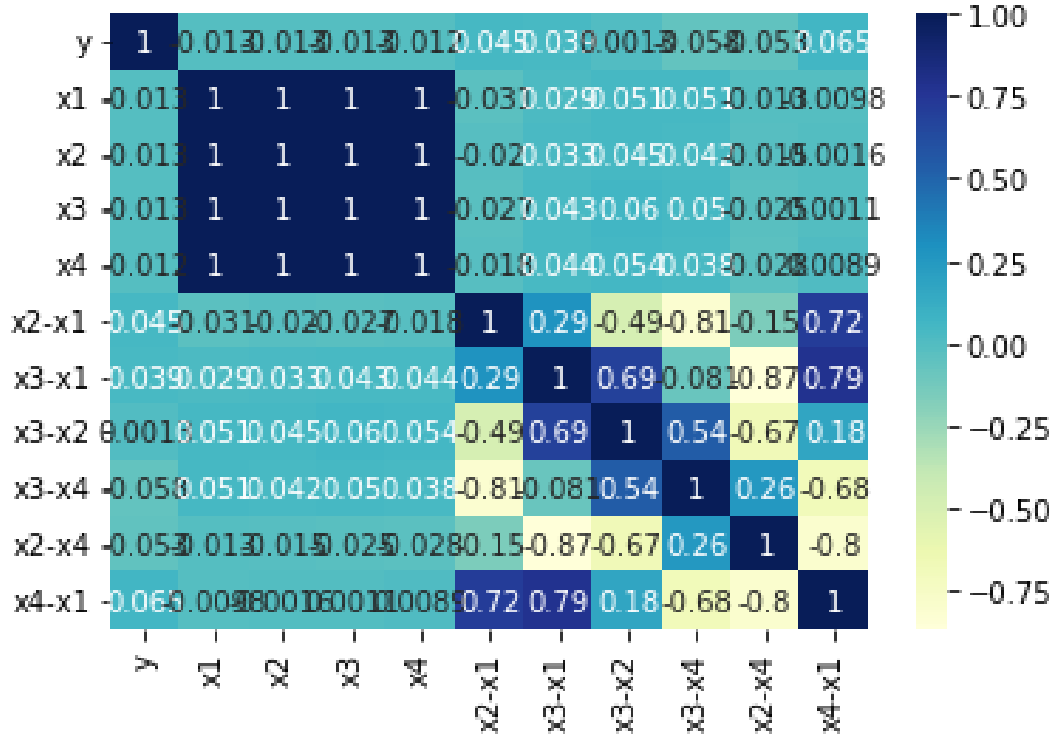
Data processing:

```
#   Column  Non-Null Count  Dtype
 0   x1      10000 non-null  float64
 1   x2      10000 non-null  float64
 2   x3      10000 non-null  float64
 3   x4      10000 non-null  float64
```



Above box plots show that there are no outliers.

OHLC datapoints were highly correlated so added new features x1-x2, x2-x3 and so on and removed the original ones.



	x2-x1	x3-x1	x3-x2	x3-x4	x2-x4	x4-x1
0	0.18	-0.57	-0.75	-0.07	0.68	-0.50
1	0.56	-0.09	-0.65	-0.40	0.25	0.31
2	0.56	-0.22	-0.78	0.00	0.78	-0.22
3	0.10	-0.78	-0.88	-0.53	0.35	-0.25
4	0.84	-0.10	-0.94	-0.91	0.03	0.81

Models:
Tried out various models but the major problem seen is that very less amount of features is present and therefore the complex models are performing bad in comparison to the simpler models.
Below are few of the results

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
RidgeClassifierCV	0.56	0.56	0.56	0.56	0.04
LinearDiscriminantAnalysis	0.56	0.56	0.56	0.55	0.04
CalibratedClassifierCV	0.56	0.56	0.56	0.55	0.57
LinearSVC	0.55	0.56	0.56	0.54	0.17
RidgeClassifier	0.54	0.54	0.54	0.52	0.03
LogisticRegression	0.52	0.53	0.53	0.48	0.06
KNeighborsClassifier	0.52	0.52	0.52	0.52	0.08
ExtraTreeClassifier	0.51	0.51	0.51	0.51	0.03
Perceptron	0.52	0.51	0.51	0.45	0.02
QuadraticDiscriminantAnalysis	0.51	0.51	0.51	0.46	0.02
DecisionTreeClassifier	0.51	0.51	0.51	0.51	0.13
NuSVC	0.51	0.51	0.51	0.50	1.97
XGBClassifier	0.51	0.50	0.50	0.51	0.58
PassiveAggressiveClassifier	0.50	0.50	0.50	0.50	0.03
DummyClassifier	0.50	0.50	0.50	0.50	0.02
SGDClassifier	0.49	0.50	0.50	0.34	0.06
RandomForestClassifier	0.50	0.50	0.50	0.50	3.89
GaussianNB	0.50	0.50	0.50	0.49	0.02
AdaBoostClassifier	0.49	0.50	0.50	0.49	0.46
BaggingClassifier	0.50	0.49	0.49	0.49	0.90

ExtraTreesClassifier	0.49	0.49	0.49	0.49	1.13
LGBMClassifier	0.49	0.49	0.49	0.49	0.18
NearestCentroid	0.49	0.49	0.49	0.49	0.03
BernoulliNB	0.49	0.49	0.49	0.49	0.03
SVC	0.48	0.49	0.49	0.45	3.51
LabelSpreading	0.48	0.49	0.49	0.43	4.13
LabelPropagation	0.48	0.48	0.48	0.43	2.79

A neural network -

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense (Dense)                (None, 16)                176
 batch_normalization (BatchN  (None, 16)                64
 ormalization)
 dense_1 (Dense)              (None, 8)                 136
 batch_normalization_1 (Batc  (None, 8)                 32
 hNormalization)
 dense_2 (Dense)              (None, 1)                 9
=================================================================
Total params: 417
Trainable params: 369
```

Also tried a similar LSTM model but both failed miserably, therefore I finally settled for ridge classifier. Even tried a time series window classification but nothing else seemed to be working.
Next, I also went to Algorithmic trading strategies to see if the y is a buy/sell prediction but could only try a few due to the time constraint.

Final Accuracy: Ridge Classifier gave a final cross validation score of 54.45% with peaks around 57-58%.

Answer for label guess:
I think the label y is a buy/sell prediction for the financial institution based on a technical indicator.

References:
Websites like stackoverflow for doubts
Sklearn and keras documentation