Project Based Learning: Machine Learning with Spark

Project Overview:

Predict the returns/prices of stock using Spark Machine Learning

Data Set:

a) *dailyPrices_AtClose.csv* **contains daily prices of a sample of 56 stocks (X1 to X56). Prices can be missing for some stocks in some periods due to many reasons such as the companies have not been listed on the exchange yet... It is safe not to consider these stocks during such periods.**
b) *X1Signals_AtClose.csv* **contains some daily signals that can help to predict future stock prices.**
c) *X2Signals_AtClose.csv* **contains other event signals that are only available for some stocks on some days. This type of signals is also expected to have some predictive power on future stock prices.**

Objectives:

- Learn Exploratory Data Analysis Using Koalas/PySpark/Pandas
- Learn to build Regression/Classification Model
- [Advanced] Hyper Parameter Tuning using Cross Validation and/or Train Test Splits

Sample Questions:

a) In the first step, we would like to see your approach to the exploratory analysis step which can help to uncover important characteristics of these temporal data sets**.**

b) In the second step, we would like to see how you build predictive models for future stock
returns using only the first data set ***dailyPrices_AtClose.csv***. In particular, our prediction outcomes of interest at the end of day ***t*** are stock returns in the next 3 days given by:

$$\boldsymbol{R_{t(10)} = \frac{P_{t+10} - P_t}{P_t}}$$

Where R_t is return of the stock, P_t is the price of the stock at time t