Group Project

1. Use the selected stocks market data and sentiment data
2. Please build a machine learning model to predict **next day return for all stocks** in the basket (85%)
   - The standard model could be a tree model (e.g. Light Gradient Boosting Tree)
   - Use any techniques introduced to do data processing, labelling, etc. to improve the performance of the model
   - Please use the following evaluation model to evaluate your prediction and actual values
3. Build a trading strategy upon the prediction model and backtest the strategy in Backtrader (15%)
4. Project report and group presentation

Evaluation model:

Submissions will be evaluated on the R value between the predicted and actual values. The R value similar to the R squared value, also called the coefficient of determination. R squared can be calculated as:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}.$$

where y is the actual value, μ is the mean of the actual values, and $\hat{y}$ is the predicted value. Do not be discouraged by low R values; in finance, given the high ratio of signal-to-noise, even a small R can deliver meaningful value!

Note:

1. make sure you split the data into in-sample and out-of-sample data. NEVER leak future data into your training set
2. the data is not 100% clean. For example, you will see stocks come with 0 volume at a certain time (this may due to the stock didn't trade at that time for some reason, or pure data quality issue). Try your best to understand and clean the data before put it into the prediction model
3. any machine learning model besides a tree model is welcomed