# Capstone Proposal

Yaoru Peng

AWS Machine Learning Engineer Nanodegree

July 14, 2022

## Prediction on Stock Returns Using AutoML and Neural Network (The Winston Stock Market Challenge by Kaggle)

## Project Overview

To predict the future stock returns has been always a favorite topic for market investors. Market data is noisy, and some scholars even claim the short-term stock returns follow stochastic processes; in other words, returns are random and unpredictable. However, the existence of momentum in every stock market, no matter in NYSE or Europe stock market, demonstrates the delicate relationship between stock past prices(returns) and future performance.

In this project, I will create a stock return predictor using the AutoGluon framework and deep learning models. The predictor will take several stock features (25), daily returns of previous days (2), and intra-day one-minute return for two hours (120) as the input, and it will predict the ntra-day one-minute return for the next hour and also two future daily returns.

## Problem Statement

The goal is to create a stock return predictor on both future intra-day one-minute returns and daily returns for the next a few days. The tasks involved are listed below:

1. Fetch data from Kaggle website using Kaggle API and process the stock data
2. Use train_test_split to generate training sets and validation sets
3. Train the return predictor within the AutoGluon framework, which will automatically choose the machine learning model with the best performance.
4. Train another predictor using neural network, and compare the two predictors with chosen metrics.
5. After get the better predictor, generate test outputs and submit results to Kaggle.

## Evaluation Metrics

- Mean Absolute Error (MAE)

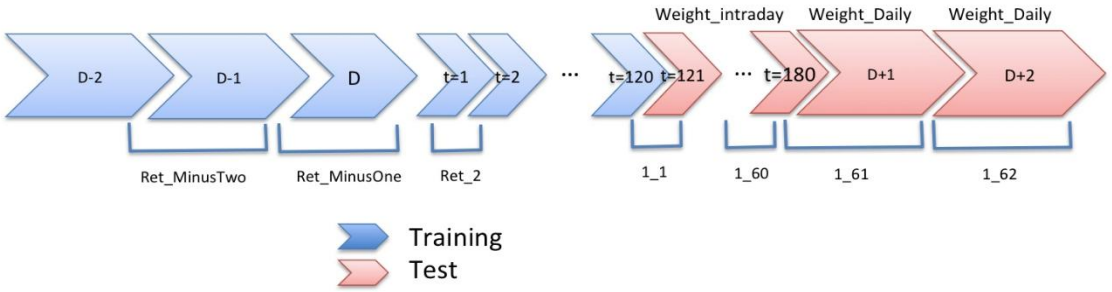
  This metric is widely used in machine learning projects, and it averages out the absolute error through the data. This metrics, in our project setting, would estimate the difference between expected return (model prediction) and real return, which has real financial meanings. Therefore, we apply MAE in this project for its financial sense.

- Root Mean square Error (RMSE)

  RMSE will penalize the large errors more than MAE. Since we does not want any large deviation between expected return and real return, RMSE would be a better metric for training our predictor to get rid of large errors.

## Datasets and Inputs

The datasets used in this project are downloaded from Kaggle, containing a training set and a test set. Training sets include provided 25 features (Feature_1 - Feature_25) of stocks, the daily returns in days D-2, D-1 (Ret_MinusTwo, Ret_MinusOne), and intra-day returns (Ret_2 - Ret_120) in day D. It will also contain our target variables, the intra-day returns in the rest of day D (Ret_121 - Ret_180) and the daily returns in day D+1, D+2. Then, the test sets contain the same columns except our target variables.




## Solution Statement

There are multiple data processing and modeling techniques involved in the solution. First, we will apply the method "train_test_split" to segment the training sets from Kaggle to training sets and validation sets. Since we would build two predictors to forecast daily returns for future few days, we need the validation sets to evaluate those two predictors, so that we could choose a better one. Then, to explore machine learning applications in stock return prediction, we will apply the AutoML, which could automatically choose the best machine learning models for the predictor. Last, we will apply neural network to make predictions, mainly through the PyTorch framework.

## Benchmark model

There exists multiple approaches to predict stock returns. One most well-known model is CAPM, which assumes a linear model between the stock premium returns and market returns. In this project, we will adopt the idea of linear modeling and build simple regression model between features and target variables. In fact, linear regression has the best interpretability, and it's widely used in financial institutions and researches.

$$Q_{j,t} = \sum_{i=1}^{k} P_{i,t} * F_{j,i,t-1} + u_{j,t} \qquad (1)$$

K = the number of independent variables
P = regression coefficient of independent variable I in month t
$F_{j, i, t-1}$ = independent variable I for stock j at the end of previous period (month t-1).
$Uj,_{t}$ = error terms for each regression
Qj, t = price of (dependent variable) stock j in month t

## Project Design

In this project, the whole process will be divided into two main parts. For the stock returns prediction, we will build both machine learning (AutoML) and deep learning (neural network)

model. Then, we will assess each predictor with our evaluation metrics using the validation sets. To predict the intra-day stock returns in a minute scale, we will only use the deep learning model to build the predictor.

# References

1. Mansouri, Ali, et al. “A Comparison of Artificial Neural Network Model and Logistics Regression in Prediction of Companiess Bankruptcy (a Case Study of Tehran Stock Exchange).” *SSRN Electronic Journal*, 2016, https://doi.org/10.2139/ssrn.2787308.