Investment and Trading Capstone Project

Build a Stock Price Indicator

Problem Description

Investment firms, hedge funds and even individuals have been using financial models to better understand market behavior and make profitable investments and trades. Machine learning engineers and fund managers are therefore working together to find financials model that can properly describe the behaviors of the stock markets.

However, statistical analysts once argued and (some of the researchers) proved empirically that stock returns (especially in U.S. Equity markets) are weak-form efficient. Therefore, historical prices bear no information on future stock price movements. They argued that stock prices are unpredictable martingale processes.

Recent rises in big data analysis, machine learning and deep learning have brought us to new horizons. With proper usage of these algorithms, we can probably be able to find reasonable models that can indicate the movement of stock prices.

I am therefore going to build a stock price indicator that can probably model the price movements of stocks. I will not simply model the prices of the stocks because it has been observed that stock prices are empirically not stationary process and linear regression on the prices against explanatory variables are likely spurious. The machine learning models associated with these problems are likely to be spurious as well. A more reasonable approach is to predict the stock returns instead of the prices. Stock returns are empirically proved to be stationary processes and therefore reasonable linear regression can be modeled. Complex machine learning and deep learning models can also be implemented. An extra benefit of stock return is stock returns are normalized to some number close to zero, while stock prices have different scales for each stock throughout the times.

Therefore, the problem becomes:

- Use machine learning / deep learning models to build the stock returns against a list of explanatory variables, each of them properly normalized.
- Analyze the stock price (return) predictability among the provided factors.
- Deploy the model to website or any platform through Amazon Sagemaker, such that users can input a stock symbol and get the prediction of movements tomorrow.

Data

I use the following three datasets to retrieve data and perform analysis:

- Yahoo Finance: Web API are used to get the stock prices.
- FRED: Web API are used to get the economic data. For example, interest rates, dollar index, etc.
- Professor Kenneth R. French's Website: Fama-French factor models used Fama-French factors. The up-to-date factors are freely available in the website.

The stock prices are what I want to forecast. For these datasets, we can get daily stock values such as Open, High, Low, Close, Volume and Adjusted Close. The returns of Adjusted Close are what I am trying to predict.

Tasks

Build and train stock predictor model

The training interface accepts a data range (start_date, end_date) and a list of ticker symbols (e.g. GOOG, AAPL), and builds a model of stock behaviors. Data will be directly downloaded from the web APIs through this training interface.

After the raw data are downloaded, the training interfaces will preprocess the data, clean the formats and feed it to the modeling processes. Since we are modeling returns instead of the direction of movements, I will use RMSE loss as the evaluation metrics. I will try to use simple neural networks or linear learner as benchmark models, and XGBoost or LSTM as desired models, figuring out whether these advanced models can beat the simple ones.

Predictability

If both the simple model and the complex model work perfectly, then it means stock returns predictability is significant through the markets. From historical evidence we believe it’s very unlikely. Moreover, the number of features and the observation times are limited. For example, our symbols are generally no more than 100 features, and daily data for 20 years contains only about 5000 observations. Advanced models like XGBoost or LSTM may not enhance the performance too much. Therefore, it’s likely that both the simple and advanced models are poor in forecasting. Stock returns unpredictability may be extended from statistical literatures to the modern machine learning frameworks.

Further question is, probably, whether a further inclusion of information are helpful in predicting stock returns? For example, prevailing data in options to inform insider's information, real-time data in news / twitters on the emotion and behavior of people to analyze the possible overreaction or underreaction of stock prices, real-time data reader of stock announcement and financial analysis reports… These techniques require NLP models and advanced methods. It may take a long time to finish the training so it's just an optional proposal to potentially improve my models.

Deployment

After training, I will implement the deployment method to build a simple user interface through the Amazon API Gateways. Users can input the stock of interest (assuming that we have trained them), and then I will show them my prediction of its predicted price movement tomorrow!