Capstone Project: Stock Price Prediction

1. Definition

1.1 Project Overview

Investment firms, hedge funds and even individuals have been using financial models to better understand market behavior and make profitable investments and trades. A wealth of information is available in the form of historical stock prices and company performance data, suitable for machine learning algorithms to process.

The price variation of stock market is a very dynamic system. Predicting stock price and its movement has been considered as one of the most challenging applications of time series prediction. Three common analytical approaches are fundamental analysis, technical analysis, and quantitative analysis.

Fundamental analysis relies on the statistics of the macroeconomics data such as interest rates, money supply, inflationary rates, and foreign exchange rates as well as the basic financial status of the company. After taking all these factors into account, a decision of selling or buying a stock will be made.

Technical analysis is based on the historical financial time series data to explore different patterns and indicators, such as trends, abrupt changes, and volatility patterns. However, because the stock price data is nonstationary and even chaotic, technical indicators, generated by technical analysis of historical data, sometimes are not able to reveal the real variation of the stock market.

Quantitative analysis applies scientific methods, such as statistics and machine learning, while takes advantage of all kinds of data including datasets used by both fundamental and technical analysis, to make better and complex evaluation of future stock market, and help investor to make better decision.

In this project, I will build stock price predictors that takes daily trading data over a certain date range as input, and outputs projected estimates for given query dates with the help of different machine learning methods.

1.2 Problem Statement

In this project, I will use various machine learning methods to predict future stock price. More specifically, I will focus on the S&P 500 index, which is widely regarded as the best single gauge of large-cap U.S. equities. There is over USD 7.8 trillion benchmarked to the index, with index assets comprising approximately USD 2.2 trillion of this total. The index includes 500 leading companies and captures approximately 80% coverage of available market capitalization.

Because stock price is more like a random walk, models with original price information as independent variables will not give us much accurate prediction. Therefore, following previous studies, I use various technical indicators generated from historical stock price information as our model input. Because of stock splits and dividend, not all stock price changes are due to the market variation. Therefore, adjusted price (no influence from stock splits and dividends) is taken as model output.

The basic procedure is as follow:

1) Download daily historical stock data (Open, High, Low, Close, Adjusted Close, Volume)
2) Generate various technical indicators
3) Define prediction window (e.g., use previous 10 days' data to predict next day's price)
4) Split data into training and testing
5) Build different models and evaluate prediction performance based on defined metrics
6) Choose the best model based on defined metrics

1.3 Metrics

The model prediction performance is defined as the mean absolute error (MAE) of testing dataset. That is

$$-\sum_{1^{5}=_{167}} |_{1}-_{1}|$$

MAE measures the mean absolute difference between the prediction and real stock price, which is a typical choice for regression problem. It can directly tell the price difference between prediction and real data, and has been using for regression model performance for a long time. Therefore, MAE is taken as model performance metrics. The MAE metrics are also used for evaluating the training process of our models.