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, we 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, we will use various machine learning methods to predict future stock price. More specifically, we 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, we 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.

2. Analysis

2.1 Data Exploration

The S&P 500 daily historical price data is downloaded from Yahoo Finance using Zipline (https://github.com/quantopian/zipline). We focus on the period from 1/1/2000 to 4/30/2016. The data includes different stock information, such as open, high, low, close, price, as well as volume. The explanation of each variable is as follow:

- open: the stock price when the market starts in a day
- high: the highest price in a trading day
- low: the lowest price in a trading day
- close: the stock price when the market closes in a day
- price: adjusted close price considering stock splits and dividends
- volume: the number of shares traded in a trading day

Here, the "price" is the adjusted closing price, which is a stock's closing price on any given day of trading that has been amended to include any distributions and corporate actions that occurred at any time prior to the next day's open. It is often used when doing historical price analysis.

Because S&P 500 index doesn't have splits and dividends like stocks, the "price" and "close" are actually the same for the S&P 500 index. For stocks, they may be different. The basic statistics of the dataset is as follow:

Table 1. Basic Statistics of S&P500 Index (2001-2016)

	open	high	low	close	volume	price
count	4107	4107	4107	4107	4107	4107
mean	1344.02	1352.55	1334.78	1344.16	2.99E+09	1344.16
std	332.03	332.04	331.94	332.13	1.59E+09	332.13
min	679.28	695.27	666.79	676.53	3.56E+08	676.53
25%	1121.96	1127.79	1114.83	1121.74	1.50E+09	1121.74
50%	1280.85	1287.61	1272.66	1280.70	2.96E+09	1280.70
75%	1471.35	1480.14	1460.48	1471.53	3.99E+09	1471.53
max	2130.36	2134.72	2126.06	2130.82	1.15E+10	2130.82

2.2 Exploratory Visualization

To explore the dataset, all variables (open, high, low, close, price, and volume) are plotted as follow.

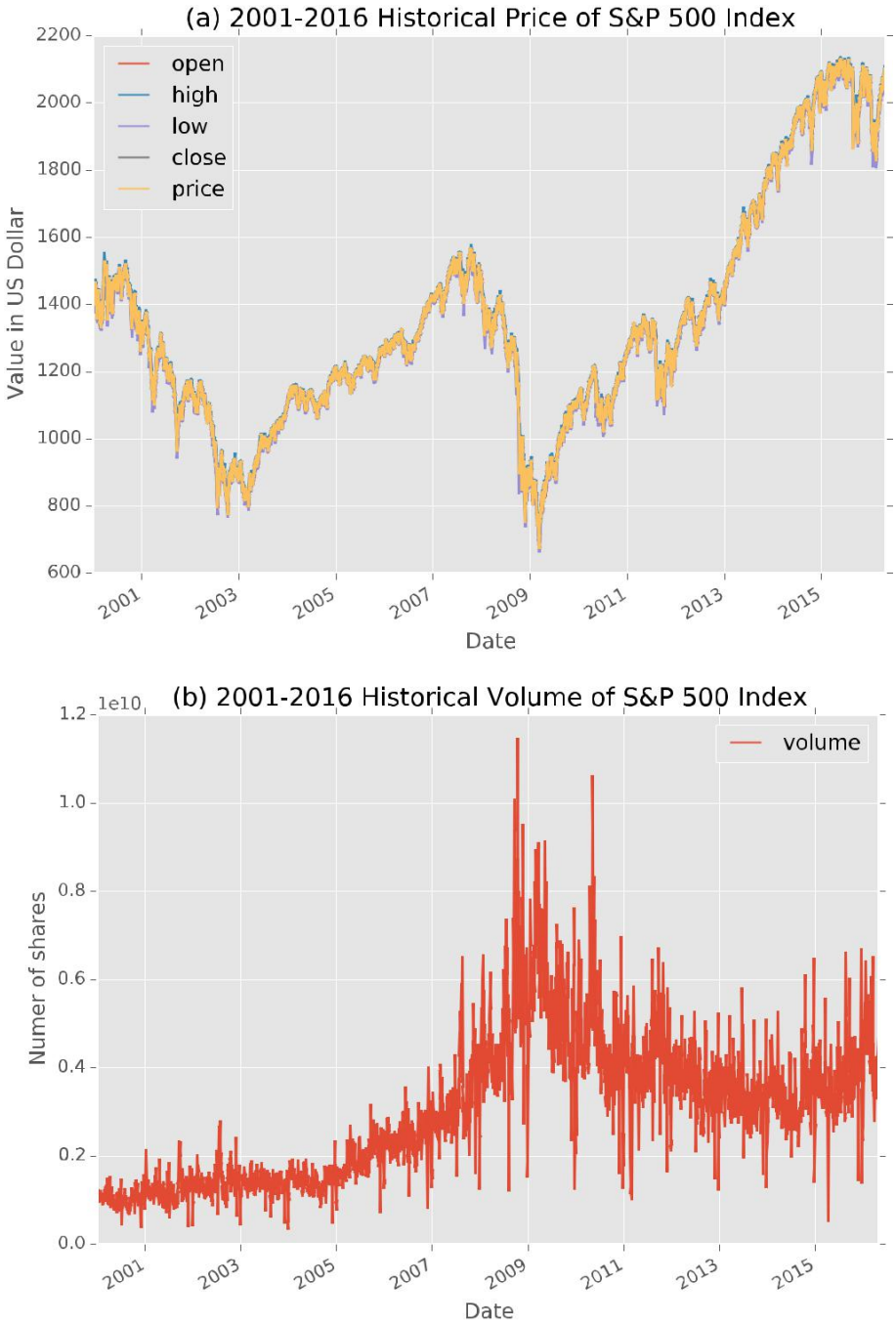

Figure 1. S&P500 Index historical price (a) and volume (b) from 2001 to 2016.

From the above figures, we can see the stock price has a lot of variation over the fifteen years (2001-2016). Generally, two low peaks exist, one in 2003 and other one in 2008, which are corresponding to the two financial crises. Because the "open", "high", "low", and "close" do not change a lot over a trading day, they tend to overlap with each other over such a long time period. Because we choose the S&P 500 index as our example, the adjusted close price ("price") is equal to the close price ("close"). Over the study period, the standard deviation of the daily "price" variation ($price_{n+1} - price_n$) is 15.18.

From the "volume" plot, we see more variability compared to that of the price over the last fifteen years. The largest variation over the study period occurred in 2008, which again corresponds to the 2008 financial crisis.

2.3 Algorithms and Techniques

Based on previous historical stock price, the future price will be predicted. Apparently, this is a regression problem. For the inputs, various technical indicators will be constructed based on the historical price information. This will be done in the Data Processing section. After the technical indicators generated, different models will be built with various technical indicators as inputs. Because this is a regression problem, the models chosen are as follows:

- **Linear Regression:**

This is the basic model for regression problem. The output ("price" variable) will be taken as a linear combination of all technical indicators. We will take the linear regression model results as the benchmark.

- **Lasso Regression:**

Lasso is the linear regression with L1 regularization. It is a shrinkage and selection method for linear regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. Because the input variables themselves may be correlated or overlapped, and some input variables may not be related the output variable, Lasso can help us get rid of some variables.

- **Support Vector Machine (SVM):**

SVM is a set of supervised learning methods used for classification, regression and outliers detection. It is effective in high dimensional spaces. It uses a subset of training points in the

decision function, which is more efficient. For our cases, the various technical indicators can be better analyzed using the SVM in a high dimensional space.

- **Gradient Boosting Regression:**

It produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

- **Artificial Neural Network (ANN):**

ANN is a family of models inspired by biological neural networks which are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.

Different models will be tested in the following sections. The performance of the models will be analyzed using the testing dataset.

2.4 Benchmark

Over the study period, the standard deviation of the S&P500 index daily price variation is 15.18. The long-term S&P500 Volatility Index (VIX) from Yahoo Finance shows that the volatility (standard deviation) of the daily S&P500 index ranges from 10 to 60. Therefore, we set the MAE benchmark of all the models to be 15, which is equal to the standard deviation of the S&P500 index daily price variation during the study period. That is, if the MAE of model prediction on the testing dataset is less than 15, we consider the model can beat the benchmark. Otherwise, the model cannot beat the benchmark. This is a reasonable choice for the benchmark given the historical variation of the VIX from Yahoo Finance.

Using the previous defined metrics (MAE), we can demonstrate the performance of different models and compare them with the benchmark in later sessions.