Udacity Machine Learning Engineer Capstone Project Proposal

Bo Ma

June 16, 2020

1 Introduction

Quantitative financial models have been used in investment firms and hedge funds to help understanding market behavior and developing profitable trading strategies. Stock market prediction tries to predict the future value of a stock or other financial instruments. However, the efficient-market hypothesis [1] claims that stock prices reflect all current information, and price changes are based on newly reveals information thus is unpredictable. Burton Malkiel, in his famous book A Random Walk Down Wall Street [2], states stock prices change is a "random" process and could not be accurately predicted by price history.

Given the unpredictability of the stock market, however, the goal of this project is to build a stock price predictor that estimates the stock price for any future dates based on past stock price patterns. Financial data is known to have a low signal to noise ratio, and it would be challenging to make accurate predictions just based on price history without considering any other sources such as fundamental, technical, sentiments, and factors data. Therefore, the focus is to get hands-on experience on the SageMaker pipeline and explore the performance of recurrent neural network (RNN) for making time-series predictions.

2 Problem Statement

This is a regression problem where the model takes a vector of stock prices of previous days and produces the expected stock prices for current day or any other future dates specified by users. The regression model will be trained for multiple tickers (specified by users) thus can make predictions for any of these tickers.

3 Datasets and Inputs

The SPX 500 Index is used as the stock universe. The tickers are obtained by scraping the wiki page "List of S&P 500 companies" [3], and the stock price data is downloaded using an open-source package yfinance [4]. Users can select any number of stocks in the universe as inputs to the model. The time horizon spans from January 1, 2006 to June 9, 2020. After the data acquisition, data cleaning is performed to fill out missing values with last valid

observation (to avoid look-ahead bias). Then, tickers with all NA values are dropped. As a result, we end up with 425 stocks in our universe as shown in Fig.1.

Date	A	AAL	AAP	AAPL	ABC	ABMD	ABT	ACN	ADBE	ADI	...	WYNN	XEL	XLNX	XOI
2006-01-03	20.609097	35.305672	41.280529	9.244411	16.633608	9.35	10.491420	22.193466	38.520000	25.182011	...	29.771486	10.574396	18.621290	37.85101
2006-01-04	20.664463	36.955460	41.574577	9.271619	16.523405	9.62	10.507332	22.314663	38.419998	25.415308	...	29.716314	10.625648	19.459511	37.91576
2006-01-05	21.205837	37.436264	41.773777	9.198654	16.318720	9.55	10.642622	22.481304	38.070000	26.224972	...	29.539719	10.619951	20.630112	37.72801
2006-01-06	21.316574	36.766911	41.726341	9.436100	16.137671	9.75	10.846883	23.594763	39.000000	26.327896	...	29.821154	10.659812	21.041996	38.47250
2006-01-09	21.255054	36.399246	42.267006	9.405183	16.137671	10.15	11.250092	23.526592	38.380001	26.698418	...	31.228340	10.631343	21.020315	38.45308

5 rows × 425 columns

Figure 1: Sample stock data

Specifically, the tickers in the universe for users to choose from are 'A', 'AAL', 'AAP', 'AAPL', 'ABC', 'ABMD', 'ABT', 'ACN', 'ADBE', 'ADI', 'ADM', 'ADP', 'ADS', 'ADSK', 'AEE', 'AEP', 'AES', 'AFL', 'AIG', 'AIV', 'AIZ', 'AJG', 'AKAM', 'ALB', 'ALGN', 'ALK', 'ALL', 'ALXN', 'AMAT', 'AMD', 'AME', 'AMGN', 'AMP', 'AMT', 'AMZN', 'ANSS', 'ANTM', 'AON', 'AOS', 'APA', 'APD', 'APH', 'ARE', 'ATO', 'ATVI', 'AVB', 'AVY', 'AXP', 'AZO', 'BA', 'BAC', 'BAX', 'BBY', 'BDX', 'BEN', 'BIIB', 'BK', 'BKNG', 'BKR', 'BLK', 'BLL', 'BMY', 'BSX', 'BWA', 'BXP', 'C', 'CAG', 'CAH', 'CAT', 'CB', 'CBRE', 'CCI', 'CCL', 'CDNS', 'CE', 'CERN', 'CF', 'CHD', 'CHRW', 'CI', 'CINF', 'CL', 'CLX', 'CMA', 'CMCSA', 'CME', 'CMI', 'CMS', 'CNC', 'CNP', 'COF', 'COG', 'COO', 'COP', 'COST', 'CPB', 'CPRT', 'CRM', 'CSCO', 'CSX', 'CTAS', 'CTL', 'CTSH', 'CTXS', 'CVS', 'CVX', 'D', 'DD', 'DE', 'DGX', 'DHI', 'DHR', 'DIS', 'DISCA', 'DISH', 'DLR', 'DLTR', 'DOV', 'DPZ', 'DRE', 'DRI', 'DTE', 'DUK', 'DVA', 'DVN', 'DXC', 'DXCM', 'EA', 'EBAY', 'ECL', 'ED', 'EFX', 'EIX', 'EL', 'EMN', 'EMR', 'EOG', 'EQIX', 'EQR', 'ES', 'ESS', 'ETFC', 'ETN', 'ETR', 'EVRG', 'EW', 'EXC', 'EXPD', 'EXPE', 'EXR', 'F', 'FAST', 'FCX', 'FDX', 'FE', 'FFIV', 'FIS', 'FISV', 'FITB', 'FLIR', 'FLS', 'FMC', 'FRT', 'FTI', 'GD', 'GE', 'GILD', 'GIS', 'GL', 'GLW', 'GOOG', 'GOOGL', 'GPC', 'GPN', 'GPS', 'GRMN', 'GS', 'GWW', 'HAL', 'HAS', 'HBAN', 'HD', 'HES', 'HFC', 'HIG', 'HOG', 'HOLX', 'HON', 'HPQ', 'HRB', 'HRL', 'HSIC', 'HST', 'HSY', 'HUM', 'IBM', 'ICE', 'IDXX', 'IEX', 'IFF', 'ILMN', 'INCY', 'INTC', 'INTU', 'IP', 'IPG', 'IRM', 'ISRG', 'IT', 'ITW', 'IVZ', 'J', 'JBHT', 'JCI', 'JKHY', 'JNJ', 'JNPR', 'JPM', 'JWN', 'K', 'KEY', 'KIM', 'KLAC', 'KMB', 'KMX', 'KO', 'KR', 'KSS', 'KSU', 'L', 'LB', 'LEG', 'LEN', 'LH', 'LHX', 'LIN', 'LKQ', 'LLY', 'LMT', 'LNC', 'LNT', 'LOW', 'LRCX', 'LUV', 'LVS', 'LYV', 'MAA', 'MAR', 'MAS', 'MCD', 'MCHP', 'MCK', 'MCO', 'MDLZ', 'MDT', 'MET', 'MGM', 'MHK', 'MKC', 'MKTX', 'MLM', 'MMC', 'MMM', 'MNST', 'MO', 'MOS', 'MRK', 'MRO', 'MS', 'MSFT', 'MSI', 'MTB', 'MTD', 'MU', 'MXIM', 'MYL', 'NBL', 'NDAQ', 'NEE', 'NEM', 'NFLX', 'NI', 'NKE', 'NOC', 'NOV', 'NRG', 'NSC', 'NTAP', 'NTRS', 'NUE', 'NVDA', 'NVR', 'NWL', 'O', 'ODFL', 'OKE', 'OMC', 'ORCL', 'ORLY', 'OXY', 'PAYX', 'PBCT', 'PCAR', 'PEAK', 'PEG', 'PEP', 'PFE', 'PFG', 'PG', 'PGR', 'PH', 'PHM', 'PKG', 'PKI', 'PLD', 'PNC', 'PNR', 'PNW', 'PPG', 'PPL', 'PRGO', 'PRU', 'PSA', 'PVH', 'PWR', 'PXD', 'QCOM', 'RCL', 'RE', 'REG', 'REGN', 'RF', 'RHI', 'RJF', 'RL', 'RMD', 'ROK', 'ROL', 'ROP', 'ROST', 'RSG', 'RTX', 'SBAC', 'SBUX', 'SCHW', 'SEE', 'SHW',

'SIVB', 'SJM', 'SLB', 'SLG', 'SNA', 'SNPS', 'SO', 'SPG', 'SPGI', 'SRE', 'STE', 'STT', 'STX', 'STZ', 'SWK', 'SWKS', 'SYK', 'SYY', 'T', 'TAP', 'TFC', 'TFX', 'TGT', 'TIF', 'TJX', 'TMO', 'TPR', 'TROW', 'TRV', 'TSCO', 'TSN', 'TT', 'TTWO', 'TXN', 'TXT', 'UAA', 'UDR', 'UHS', 'UNH', 'UNM', 'UNP', 'UPS', 'URI', 'USB', 'VAR', 'VFC', 'VLO', 'VMC', 'VNO', 'VRSN', 'VRTX', 'VTR', 'VZ', 'WAB', 'WAT', 'WBA', 'WDC', 'WEC', 'WELL', 'WFC', 'WHR', 'WLTW', 'WM', 'WMB', 'WMT', 'WRB', 'WST', 'WY', 'WYNN', 'XEL', 'XLNX', 'XOM', 'XRAY', 'XRX', 'YUM', 'ZBH', 'ZBRA', 'ZION'. Some sample statistics are shown in Fig. 2.

	A	AAL	AAP	AAPL	ABC	\
count	3638.000000	3638.000000	3638.000000	3638.000000	3638.000000	
mean	37.095672	25.725284	92.427215	84.788635	51.454688	
std	19.458048	16.203190	49.951199	72.144847	29.572422	
min	8.122367	1.659225	23.336365	6.266412	11.512946	
25%	22.460566	9.033596	40.568934	23.016733	20.564123	
50%	30.384507	28.426829	80.624096	67.367977	45.097704	
75%	45.125700	39.911220	144.144791	116.669054	81.516388	
max	91.139999	59.345577	198.368835	333.459991	105.652664	

	ABMD	ABT	ACN	ADBE	ADI	...	\
count	3638.000000	3638.000000	3638.000000	3638.000000	3638.000000	...	
mean	78.523560	34.272205	76.867654	92.661572	47.265778	...	
std	102.582496	21.052342	52.250997	89.347635	28.603947	...	
min	4.900000	10.491420	19.527224	15.980000	11.718954	...	
25%	13.160000	17.003025	31.114238	33.532499	24.299947	...	
50%	21.845000	29.605185	62.895391	44.424999	37.577780	...	
75%	116.584997	42.094382	110.049522	105.079998	60.299373	...	
max	449.750000	98.000000	214.950119	397.779999	124.589996	...	

Figure 2: Sample stock statistics

4 Proposed Solution

The idea is to use the Long Short Term Memory (LSTM) model that takes a sequence of closed prices in the past to predict the adjusted close price in the future. The LSTM is a RNN architecture that is particularly effective for time-series predictions since it could incorporate important events with long lookback into predictions.

5 Benchmark Model

Three simple models will be used to evaluate the LSTM model. The last value model makes predictions using the last observed model. The moving average model uses the average of past n values for predictions. The linear regression model fits a linear regression model to the previous n days and use the model for future predictions.

6 Evaluation Metrics

Users can specify any number of look-ahead days for making predictions. For example, 1, 7, and 20 look ahead days. As a result, the output of the model is a three-dimensional vector containing the respective predictions for the look ahead days. Therefore, the mean squared error (MSE) between ground-truth and predictions is a natural measurement for model evaluation.

7 Project Design

The datasets will be divided into train (85%), validation (15%), and test (10%) datasets in Chronological order. Then, data normalization will be conducted on each dataset (i.e., train, validation, and test) respectively. Data normalization is a crucial step as stock prices are generally rising over time. Without normalization, the model trained on small scales data would not be able to predict large scale data in the future. The LSTM model will use the lookback days (hyperparameters) to predict look ahead days specified by users. The hyperparameter tuner in SageMaker will be leveraged to select the best model with the smallest loss for validation data. In terms of output, we use the joint time horizon [5] i.e., the output contains multiple time horizons and they are determined simultaneously. The MSE will be used to compare the ground-truth and output vector for model evaluations with benchmark models.

References

[1] Malkiel, Burton G. "The efficient market hypothesis and its critics." Journal of economic perspectives 17.1 (2003): 59-82.

[2] Malkiel, Burton Gordon. A random walk down Wall Street: including a life-cycle guide to personal investing. WW Norton Company, 1999.

[3] *List of S&P 500 companies*, available at `https://en.wikipedia.org/wiki/List_of_S%26P_500_companies`.

[4] Yahoo! Finance market data downloader, available at `https://github.com/ranaroussi/yfinance`.

[5] *Stock market prediction*, available at `https://en.wikipedia.org/wiki/Stock_market_prediction`.