Apple's Stock Price Prediction: A Study of Importance to Use External Data Assignment of Data Acquisition and Processing Systems 20/21

Anonymous Authors[1]

Abstract

Although many proposed stock price predicting models had great success in academia and industry, limitations can be experienced, i.e. they may fail to predict stock price accurately and in time. Some advanced data-driven methods have demonstrated stronger predictive capabilities, providing investors with the possibility of obtaining excess returns on investment. This project mainly investigates the importance of various features in predicting stock price with comprehensive data processing and analysis and conducts an empirical study to model the pattern of Apple Inc. (AAPL) stock price from April 2017 to May 2020 and predict its short-term trend. The experiment results show that features of AAPL stock trading volume, Nasdaq index and its trading volume, OCFPS and US retail sales can significantly contribute to improving the predictive capability of the model that uses historical stock price as the only predictor. The LSTM-based model achieves the lowest denormalized MSE at 0.06164 in the testing phase.

1. Introduction

Stock is a smart invention that enables individuals to share the future growth of a company, while, on the other hand, it also significantly contributes to company expansion by diversifying risks to every shareholder. Meanwhile, the stock price is also the most important information in the financial market, because most investment and financing decisions depend on it. Typically, investors who invest in a company's stock require not only returns on the time value of money but also compensation from taking a potential risk. A reasonable price of a stock is more likely to attract investors, but there is not a strictly objective price, which means investors have to evaluate if the stock price is overvalued. Therefore, stock pricing becomes crucial to be considered.

According to the Strong Form of Efficiency Market Hypothesis (Strong-Form EMH) proposed by Eugene Fama, in a completely efficient market, the stock price of a company should truly reflect company value and potential risk(Fama, 1960). To fairly price stock, various pricing models are proposed, e.g. Capital Assets Pricing Model (CAPM)(Sharpe, 1964), Fama Three-factor and Five-factor Asset Pricing Model(Fama & French, 2004)(Fama & French, 2015). In fact, the basic assumptions of these classic models are based on, or at least, contain EMH. Although these models not only provide a strong reference for academic research but also contribute to many witnessed successes in investment strategies, limitations can be experienced, i.e. they may fail to predict stock price accurately and timely. The reasons can be summarized as follow: *i) Stock market is still not Strong-form efficient, which implies information asymmetry.* Despite being monitored by a regulatory authority, insider trading, information leakage, stock price manipulation, etc. are suppressed, but still exist; *ii) The involvement of speculators and quantitative traders brings uncertainties to the market*, which is likely to lead to irrational behaviours, i.e. buy shares when the price is up, sell shares when the price goes down; *iii) These models only consider partial factors that can have an impact on stock price, as the riskiness is not comprehensively evaluated.* In fact, many studies(Adam et al., 2016)(Fama, 1995)(Malkiel, 1999) have shown that historical stock price is an approximate random-walk series, which is difficult to be accurately predicted by the traditional models. On the other hand, fundamental analysis lacks the support of systematic methodology, and thus it is less likely to reproduce and generalize.

With the advent of the big data era and the improvement of computer capabilities, some advanced data-driven methods have demonstrated stronger predictive capabilities, providing investors with the possibility of obtaining excess returns on investment, e.g. Facebook Prophet(Taylor & Letham, 2018), Random Forest(Loke, 2017), Hidden Markov Model (HMM)(Hassan et al., 2007), Recurrent Neural Network (RNN)(Selvin et al., 2017), Long-Short-Term Memory (LSTM)(Selvin et al., 2017), etc. In fact, a variety of data can be explored to improve the performance of these data-consumed models compared to using single historical price data. Firstly, open market transactions are all recorded and published to the public, which provides abundant financial market historical information that can be analysed(Ariyo et al., 2014)(Adebiyi et al., 2014)(Gokmenoglu & Fazlollahi, 2015)(Jain & Biswal, 2016). Secondly, required by regulatory rules, compulsory information disclosure of listed companies provides comprehensive financial data that is likely to enhance models' predictive capability(Heo & Yang, 2016)(Arkan et al., 2016)(Adebiyi et al., 2012). Thirdly, benefited from exhaustive statistics by the government, accessible macro-economic data reflecting the economic pros-

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

perity also has a certain impact on stock price(Borjigin et al., 2018)(Alamsyah & Zahir, 2018)(Nti et al., 2019). Fourthly, text-based data mining techniques provide additional and important information source for stock price prediction(Lee et al., 2014)(De Fortuny et al., 2014).

This project mainly investigates the importance of various features in predicting stock price and conducts an empirical study to model the pattern of Apple Inc.'s (AAPL or Apple) stock price from April 2017 to May 2020 and predict its short-term trend. In addition to using its historical price data, a variety of potentially useful data like financial market data, Apple financial data and macro-economic data are processed, explored and selected as predictive features. The project compares the performance of applying the LSTM model by using different combinations of features with using solely the stock price historical data. The experiment results show that the model using selected features has a better performance in predicting the AAPL stock price trend.

2. Data Description

2.1. Financial Market Data

Financial market data which shows historical trends and seasonality is commonly used for technical analysis by many investors. However, data in the financial market is diversifying and numerous, which means data selection is not only essential but also critical. Firstly, the historical stock price is able to provide an important reference for the future price trend. Researchers in (Ariyo et al., 2014) and (Adebiyi et al., 2014) indicate that advanced statistical models, e.g. Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN), can learn patterns well from the historical stock data like low, high, close prices and trading volume of the stock. Secondly, gold is a safe asset that investors always choose to avoid risk when people have a negative attitude toward the macro-economy or uncertainty in the financial market keeps high. Its price and trading volume are believed to be related to the prosperity of the stock market, and thus they might help models predict the trend of the stock price(Gokmenoglu & Fazlollahi, 2015)(Jain & Biswal, 2016). Thirdly, Apple is one of the constituent stocks of the Nasdaq Index. Their trends and trading volumes experience very similar changes, and thus Nasdaq Index and its daily trading volume are also considered as potential predictive features.

The collected financial market dataset contains 18 features, i.e. open, low, high, close, adjusted close prices/index and trading volume for AAPL stock, gold and Nasdaq Index, respectively. On the daily basis, the dataset covers the period from April 2017 to May 2020 with a total of 795 sample points. Before exploring and applying this dataset, features are selected and timestamps are aligned, while a detailed process is introduced in a later section. The adjusted price is adjusted for both dividends and splits and all features are in USD.

2.2. Apple's Financial Data

Another sort of data that is always used for fundamental analysis is the company's financial data. Many studies point out that some key financial indicators can reflect a company's operating status and affect the expectation of the stock value in a trading market. Research (Arkan et al., 2016) investigates the predictive capability of 12 financial ratios based on financial data from 15 companies listed in the Kuwaiti stock market and concludes that Return of Asset (ROA), Return of Equity (ROE), Price Earnings Ratio (PER) and Earning Per Share (EPS) are the most effective ratios on the stock price for service and investment sectors. Similarly, Heo et al. show that stock price can be better predicted by involving EPS and Book value Per Share (BPS) data(Heo & Yang, 2016). Additionally, a hybridized approach is proposed in (Adebiyi et al., 2012), where both company financial indicators and stock market data are applied for prediction. The hybridized model illustrates remarkable improvement compared to the model using only stock market historical data.

Data from Apple Inc.'s quarterly income statements and key financial ratios cover 38 features. On a quarterly basis, each feature in the dataset covers the period from 2017 Q2 to 2020 Q2 with a total of 13 sample points. Except for per-share features and ratio indicators, other features are in million and counted currency is USD. However, it is worth noting that there are many redundant features in the dataset, for example, the relationship between revenue, cost of goods sold and gross profit is completely deterministic. Therefore, feature selection is essential. By eliminating features with redundant information and selecting critical features according to the aforementioned studies, 7 features are considered as candidates for further exploration, i.e. revenue, EPS, ROE, research and development (RD) expenses, BPS, operating cash flow per share (OCFPS) and shares outstanding. Besides, to align the daily data of stock price, resampling is also required. Data processing details are presented in a later section.

2.3. Macroeconomic Data

It is also well-documented that macroeconomic variables can significantly affect the stock market returns. Many investigations focus on various macroeconomic variables, e.g. interest rates, exchange rate, gross domestic product (GDP), inflation rate, etc. Ramin et al. point out that a cointegrating relationship is found between Singapore's stock market and the short- and long-term interest rates, industrial production, price levels, etc.(Maysami et al., 2005) Similarly, Donatas et al. clarify that GDP and money supply have a positive impact on the stock market, but unemployment rate, exchange rate and interest rates have inverse effects(Pilinkus & Boguslauskas, 2009). Furthermore, researchers in (Khan, 2014) find that trade balance, exchange rate and interest rate negatively influence the KSE-100 index, while the consumer price index has a positive effect. However, research (Gay Jr, 2008) argues that exchange rate and oil price do not significantly affect the stock market of some developing countries, while the authors believe the possible reason is due to interference of other domestic and international factors.

In this project, 3 macroeconomic variables are considered as possible predictive features during the 3-year period, i.e. inflation rate, retail sales and interest rate in the US. The collected inflation and interest rates are recorded on daily basis with a total of 790 sample points, while retail sales data is sampled on monthly basis with a total of 38 sample points. While US inflation ratio and interest rate are counted in percentage, US retail sales are measured in million USD.

3. Data Acquisition

Data acquisition is the first step for data analysis once we have defined the goal. Generally, data can be acquired from sensors, webs, databases and statistical sampling. For this stock price prediction project, abundant data resources are available in various public webs and databases. In particular, we collect financial market and macroeconomic data from some open datasets, and scrape Apple's financial data by web scraper.

3.1. Open Datasets

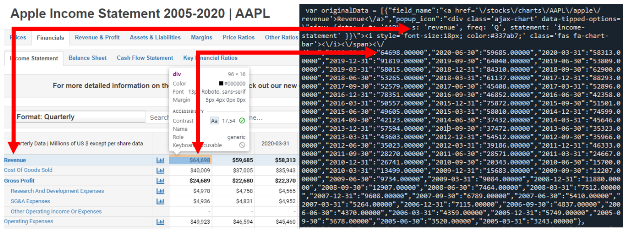
Acquiring structural data from open datasets such as prices, index and macroeconomic variables have certain advantages as follows: *It is the simplest and low-cost approach to acquire structural data with a reliable source; ii) The open datasets with structural data are usually formulated in some fixed formats (CSV, JSON, XLSX, TXT) which are easily readable in computer programs, e.g. modules in Python like Pandas and CSV serve well for Comma-Separated Values (CSV) file reading; iii) Open datasets are usually well maintained and stored, which eases further work of data pre-processing.* Therefore, the main method that this project applies to acquire financial market data and macroeconomic data is through reliable web sources such as Yahoo Finance, Nasdaq.com and macrotrends.net. However, acquiring data by downloading has some limitations, for example, it would be difficult to keep updated with the latest changes at the source.

For financial market data, 3 open datasets are acquired containing features of low, high, open and close prices/index and trading volumes of AAPL stock, gold and Nasdaq index, respectively. All the datasets in CSV format are publicly downloadable at Yahoo Finance and Nasdaq.com, where the AAPL stock and Nasdaq index datasets have tunable time granularity from daily to monthly and period from the start of AAPL stock going public and Nasdaq index opening. As for the historical data of gold, although only futures data is available at Yahoo Finance, while its spot price and volume daily data can be accessed at Nasdaq.com with a maximum period of 10 years. For macroeconomic data, 3 open datasets are collected from macrotrends.net with features of US retail sales, US inflation rate and Federal funds rate, respectively. In this case, the Federal funds rate is regarded as the fundamental interest rate as an approximation. Similar to the financial market data, the 3 datasets in CSV format are also available to be downloaded directly for further use. Indexed with daily/monthly timestamps, the period of them is optional ranging from 5 to 30 years.

3.2. Web Scraping

Web scraping is a popular technique to automatically extract wanted structural data from web pages which are a kind of unstructured data including text, pictures, descriptors, etc. Most web pages are coded by HyperText Markup Language (HTML), and general browsers download the page elements and source, and then demonstrate them. The wanted data is usually hidden in some pieces of the script mixing with other redundant information, and thus the key step of scraping the wanted data is to locate which element or source contains it and. Typically, fuzzy matching based on certain rules of regular expression contributes to segmenting the data from the redundant hypertext, which filters descriptors, useless text and images. Once the wanted data is cropped from the whole web page, it can be transformed into structural form and stored for further analysis. The advantages of acquiring data by applying web scraping are as follows: *i) Web scraping can be very fast and efficient. Once the scraping rules have been set, the web scraper can extract interested data from multiple web pages with similar coding structure; ii) What you can see is what you get. For those data without accessible Application Programming Interfaces (APIs) or scattered at many web pages, web scraper is able to collect them easily; iii) It makes acquired data updated. The web scraper can keep tracking updates of the data on the web, and thus it keeps the local data and application newest.* On the other hand, the limitations of using web scraping are also obvious. For instance, scrapers keep visiting websites very quickly, which may occupy many network resources and delay other users' viewing. One of the possible consequences is that the IP address would be blocked by the server.

To acquire Apple's financial data, manually collecting the 3-year data from its financial statement into tabular form is low-efficient. Furthermore, it is difficult to find an open dataset for Apple's financials. Therefore, a web scraper is set to automatically acquire Apple's financial data from macrotrends.com which inserts tables for Apple's quarterly income statement, balance sheet, key financial ratios, etc. Fig.1 shows an example of locating the wanted data in the browser and got source code's views, respectively. By the fuzzy matching regular expression "$var original Data = [(.*?)];$", feature name, data values and corresponding timestamps are located, where "$(.*?)$" represents lazy greedy matching. Further extractions can be realized by matching regular expression "$s :' (.*?)'$" for feature name and "$< \/div(.*?)\} >$" for data values and timestamps. The extracted data are subsequently transformed into a table with 38 financial features and 13 sample points for each feature.



Apple Income Statement 2005-2020 | AAPL

For more detailed information on th… …ck out our new …

Format: Quarterly

div 96 × 16
Color ■ #000000
Font 13p… Roboto, sans-serif
Margin 5px 4px 0px 0px
ACCESSIBILITY
Contrast Aa 17.54 ✓
Name
Role generic
Keyboard-focusable ⃠

…rterly Data | Millions of US $ except per share data				2020-03-31
Revenue	📊	$64,698	$59,685	$58,313
Cost Of Goods Sold	📊	$40,009	$37,005	$35,943
Gross Profit	📊	$24,689	$22,680	$22,370
Research And Development Expenses	📊	$4,978	$4,758	$4,565
SG&A Expenses	📊	$4,936	$4,831	$4,952
Other Operating Income Or Expenses		-	-	-
Operating Expenses	📊	$49,923	$46,594	$45,460

```
 var originalData = [{"field_name":"<a href='\/stocks\/charts\/AAPL\/apple\/
revenue'>Revenue<\/a>","popup_icon":"<div class='ajax-chart' data-tipped-options=
\"ajax: {data: { t: 'AAPL', s: 'revenue', freq: 'Q', statement: 'income-
statement' }}\"><i style='font-size:18px; color:#337ab7;' class='fas fa-chart-
bar'><\/i><\/span><\/
div>","2020-09-30":"64698.00000","2020-06-30":"59685.00000","2020-03-31":"58313.0
0000","2019-12-31":"91819.00000","2019-09-30":"64040.00000","2019-06-30":"53809.0
0000","2019-03-31":"58015.00000","2018-12-31":"84310.00000","2018-09-30":"62900.0
0000","2018-06-30":"53265.00000","2018-03-31":"61137.00000","2017-12-31":"88293.0
0000","2017-09-30":"52579.00000","2017-06-30":"45408.00000","2017-03-31":"52896.0
0000","2016-12-31":"78351.00000","2016-09-30":"46852.00000","2016-06-30":"42358.0
0000","2016-03-31":"50557.00000","2015-12-31":"75872.00000","2015-09-30":"51501.0
0000","2015-06-30":"49605.00000","2015-03-31":"58010.00000","2014-12-31":"74599.0
0000","2014-09-30":"42123.00000","2014-06-30":"37432.00000","2014-03-31":"45646.0
0000","2013-12-31":"57594.00000","2013-09-30":"37472.00000","2013-06-30":"35323.0
0000","2013-03-31":"43603.00000","2012-12-31":"54512.00000","2012-09-30":"35966.0
0000","2012-06-30":"35023.00000","2012-03-31":"39186.00000","2011-12-31":"46333.0
0000","2011-09-30":"28270.00000","2011-06-30":"28571.00000","2011-03-31":"24667.0
0000","2010-12-31":"26741.00000","2010-09-30":"20343.00000","2010-06-30":"15700.0
0000","2010-03-31":"13499.00000","2009-12-31":"15683.00000","2009-09-30":"12207.0
0000","2009-06-30":"9734.00000","2009-03-31":"9084.00000","2008-12-31":"11880.000
00","2008-09-30":"12907.00000","2008-06-30":"7464.00000","2008-03-31":"7512.00000
","2007-12-31":"9608.00000","2007-09-30":"6789.00000","2007-06-30":"5410.00000","
2007-03-31":"5264.00000","2006-12-31":"7115.00000","2006-09-30":"4837.00000","200
6-06-30":"4370.00000","2006-03-31":"4359.00000","2005-12-31":"5749.00000","2005-0
9-30":"3678.00000","2005-06-30":"3520.00000","2005-03-31":"3243.00000"},
```

Figure 1: Example of Locating the Wanted Data in the Views of Browser and Web Source

4. Data Storage

After the wanted data have been acquired, they should be properly stored that can be efficiently and easily accessed for further analysis. Storing locally or in the cloud database are two reasonable choices with certain benefits and drawbacks. This project applies and compares both approaches, storing the acquired data safely, efficiently and accessibly.

4.1. Local Storage

The collected data are stored locally in CSV format at first, where datasets are classified according to their type of features, i.e. financial market data, company's financial data and macroeconomic variables. CSV format stores tabular data in plain text, where each record is separated into the same sequence of fields by a specified delimiter, e.g. comma, tabs, etc. Because of this property, the CSV file has a very simple but standard structure and highly efficient storage usage, which can be read and written fast and consumes much less storage than the same data stored in other formats, e.g. JSON, XLSX, etc. Stored into the local disk, the data can be easily accessed and managed, while risks of data loss exist if the local machine is broken or the data is deleted by mistake. Furthermore, a local storage strategy prevents real-time data sharing. That means obstacles for others to obtain the datasets.

4.2. Cloud Database

A good alternative to store data is to use a cloud database, e.g. Oracle Apex, Mango DB. On one hand, the abstractions of the database allow the application interface to be briefly defined and data management systems to be applied. On the other hand, cloud storage releases burdens of local storage through moving data into the cloud side where other permitted analysts can easily access the datasets at any time at any location. In fact, beyond the relational database model, many emerging database models are successfully applied to address horizontally scalable, unstructured, distributed data storage problems. For instance, a graph-oriented object database model stores data as a graph with nodes and edges, and thus data are manipulated by graph transformations(Gyssens et al., 1994). However, the used datasets of this project have limited size and structural form, and thus the SQL-based Oracle Database is applied for the data storage on the cloud. A data query API is built with a set of GET methods to obtain the specified dataset in JSON format. Table1 shows names, number of columns and number of items of datasets stored in the cloud database and their URLs.

Table 1: Datasets Query URLs on the Cloud

Dataset Name	Query URL
AAPL Stock Price	./AAPL_PRICE_DATA
Gold Price	./GOLD_PRICE_DATA
Nasdaq Index	./NASDAQ_INDEX_DATA
AAPL Financials	./AAPL_FINANCIALS_DATA
US Inflation Rate	./US_INFLATION_RATE_DATA
US Retail Sales	./US_RETAIL_SALES_DATA
US Federal Funds Rate	./US_INTEREST_RATE

5. Data Preprocessing

5.1. Missing Value Processing and Outlier Detection

In practice, the acquired data is always ditty due to incompleteness, noise and inconsistency, which introduces a bias in data inference. For example, missing values may lead to failure in some statistical models like support vector machine (SVM) and K-Nearest Neighbour (KNN); outliers have a severe impact on outlier-sensitive models like AdaBoost, decision tree. Therefore, detecting and cleaning missing values and outliers are of vital importance before further analysing the collected data. Table2 demonstrates the summarized information of missing values in the collected datasets for each feature, where merely few missing values can be found in the financial market and macroeconomics datasets. Considering these data are time series for trading days, time-series-based linear interpolation is applied to fill missing values. The advantage of using the time-series-based method is the interpolated data more determined by the data point closer in time, e.g. a missing value on Monday should be more like that on Tuesday rather than last Friday, even though it is closely adjacent to both in the dataset.

Table 2: Summary of Missing Values

Data Category	Feature	Number of Missing Values
Financial Market	AAPL Adjusted Close Price	1
	AAPL stock trading Volume	1
	Gold Close Price	1
	Gold Trading Volume	4
	Nasdaq Index	1
	Nasdaq Trading Volume	1
Company Financials (Quarterly)	Revenue	0
	EPS	0
	ROE	0
	R&D Expenses	0
	BPS	0
	OCFPS	0
	Shares Outstandings	0
Macroeconomics	US Inflation Rate	5
	US Retail Sales (Monthly)	0
	Federal Funds Rate	4

Furthermore, potential outliers are detected to have an insight into the noise of data. Boxplot is a good form to visualize and investigate the data location and spread (Fig.2), where the top and the bottom line is the measured upper and lower extremes, the box top and bottom are the upper and lower quartile of the dataset, line in the middle of the box is the average value, potential outliers are marked by red crosses. However, it is normal for the stock price and some economic data to fluctuate acutely due to randomness in the stock market and sudden impulse happens in the world. For example, disclosure of important change of company situation can lead to price rise and fall; the outbreak of COVID-19 may negatively influence the retail sales for months. Therefore, these detected ”outliers” should not be removed since they are supposed to contain special but important information that relates to the volatility of AAPL stock price.

5.2. Data Visualization and Comparison

After cleaning datasets, visualization and comparison help to perceive their properties, e.g. trends, cross-correlation, range, etc. Fig.3 demonstrates the aforementioned financial market data, where subplots are the close price/index

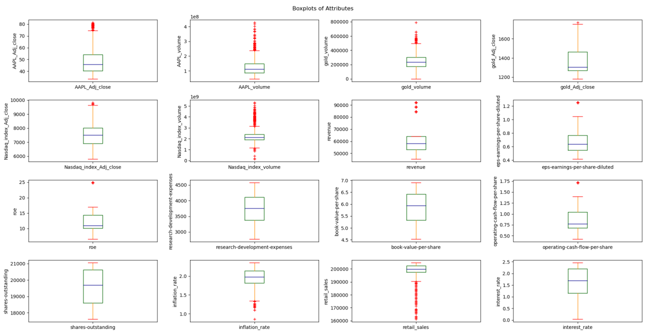


Figure 2: Boxplots of Features in Collected Datasets

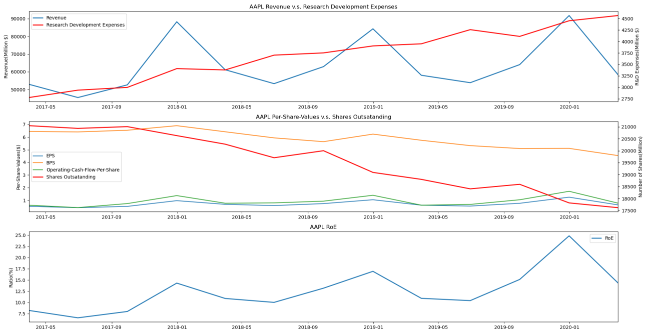


Figure 4: Selected Company Financial Data

(blue line) and corresponding trading volume (red line) of AAPL stock, gold and Nasdaq Index, respectively. Obviously, the trend of AAPL stock price is very similar to that of the Nasdaq index, but some opposite behaviours can be experienced in the trend of gold price in some periods. The inverse trends possibly because of the value-keeping function of gold when uncertainty grows. A sudden drop in March 2020 of the prices can be seen in every subplot, as it is deduced that the outbreak of COVID-19 significantly contributes to this change. It is more difficult to reveal a trend of trading volumes for all of them, as it fluctuates with a high frequency more randomly like a noise signal.

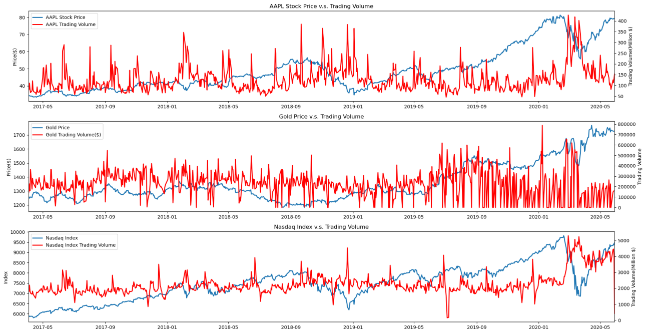


Figure 3: Financial Market Data

Fig.4 illustrates values over the 3 years of the selected features. Subplot 1 compares the trends of Apple's revenue and RD expenses, where the revenue fluctuates regularly with a peak at the beginning of each year due to new product announcements, while RD expenses keep increasing over the period. Subplot 2 shows the per-share features and corresponding shares outstanding, where despite a witnessed decrease of the shares outstanding, the other per-share features almost remain stable. Subplot 3 demonstrates the changes in Apple's ROE, where its fluctuation is similar to the revenue. Note that there are only 13 points for each feature in the figure, but the actual values between the two data points should be considered as a constant due to quarterly disclosure of company financials.

Fig.5 shows changes in the macroeconomic variables among the 3 years in the US, i.e. inflation ratio, federal funds rate and retail sales. Subplot 1 compares changes of two important indicators of monetary policy: US inflation ratio and Federal funds ratio. Sudden drops similar to financial market data is experienced for both variables due to the same reason at the beginning of 2020. Theoretically, a lower Federal funds rate tends to activate the market, as the low inflation rate may indicate a sluggish economy. In addition, the dramatical decrease in subplot 2 also reflects the depressing atmosphere in 2020 in US retails.

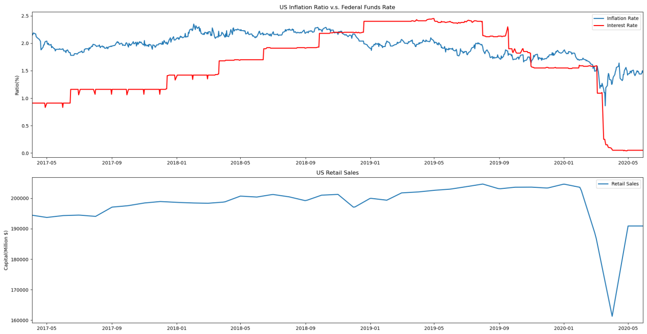


Figure 5: Macroeconomic Data

5.3. Data Transformation

Before further exploring statistical relations of the datasets, data transformation can be applied to improve the predictive capability of data and performance of model. Mainly two data transformation techniques are used in this project, i.e. series resampling, normalization and data dimensionality reduction.

Since the acquired company's financial data are obtained from quarterly reports and US retail sales are recorded on monthly basis, upsampling them in daily frequency to align daily AAPL stock close price is essential. For stock price prediction, I resample the quarterly and monthly data by padding the upsampling values with the closest former. That means the values of Apple's financials and US retail before the next report is considered to maintain constant. The rationale is that most investors in the stock market consider the latest disclosures without foreseeing the financials and macroeconomic circumstances in the next term, and thus they should only analyse the published data for prediction. Likewise, when the model predicts the future stock price, the accessible data is former disclosures.

Another applied transformation technique is normalization, which can avoid prediction models over-emphasizing some features simply because they have a bigger absolute value. For example, a feature with relatively bigger absolute values is more likely to lead the training process of neural network (NN), since its gradient can be bigger thus affecting the optima of using the gradient descent method. By applying normalization, the values of different features can be transformed into the same scale, i.e. ranging from 0 to 1, which are expected to have the same weights of importance for prediction initially. In this project, Max-Min normalization is applied for each feature, which can be expressed by Eq.1:

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

, where X_norm is the normalized feature vector of original feature X, X_min and X_max are the minima and maxima of the vector, respectively.

6. Data Exploration

6.1. Stationarity Test

The main purpose of classic time series analysis is to use the historical and current data to predict the possible future trends, that is, it is assumed that the basic characteristics of the time series must be maintained from the past to the predicting period. Otherwise, predictions tend to be unreliable. In fact, one of the main challenges to predict stock price is because it is normally not a stationary series. Indeed, previous observation of AAPL stock price (Fig.3) has shown an intuitive view that it should not be a stationary series, but a more quantitative analysis is needed. A good approach is analysing its autocorrelation and partial autocorrelation plots, shown in Fig.6. Particularly, autocorrelation coefficients describe the dependency of a variable to all of its past values, while partial autocorrelation coefficients intend to explore the dependency to the value of a single past time point, which can be expressed by Eq.2 and 3:

$$ACF_k = \frac{\sum_{i=1}^{n-k} (X_i - \bar{X})(X_{i+k} - \bar{X})}{\sum_{i=1}^{n-k} (X_i - \bar{X})^2} \tag{2}$$

$$PCAF_k = \frac{E[(X_t - \hat{E}[X_t])(X_{t-k} - \hat{E}[X_{t-k}])]}{E[(X_{t-k} - \hat{E}[X_{t-k}])^2]} \tag{3}$$

,where $\hat{E}[X_t]$ and $\hat{E}[X_{t-k}]$ are the short forms of $E[X_t|X_{t-1},...,X_{t-k+1}]$ and $E[X_{t-k}|X_{t-1},...,X_{t-k+1}]$, respectively. And ACF_k and $PACF_k$ represent the k-th order autocorrelation and partial autocorrelation coefficient, respectively.

There is a slow decay for AAPL stock price autocorrelation coefficients, while the partial autocorrelation coefficient is roughly 1 at first order and dramatically drops to about 0 for higher orders. This property proves the AAPL stock price series is not a stationary series, but it is likely a random-walk series.

ADF hypothesis test(Dickey & Fuller, 1981) is used to statistically test if a time series with first order auto-regression is stationary, which matches the property of first-order autocorrelated shown by the partial autocorrelation coefficient plot above. Setting null hypothesis *H0: there is a unit root*, alternative hypothesis *H1: there is no unit root* and $\alpha 0.05$, the test result shows that we should reject the null hypothesis, that is, it can be declared that the AAPL stock price series is not stationary under 0.95 confidence interval.

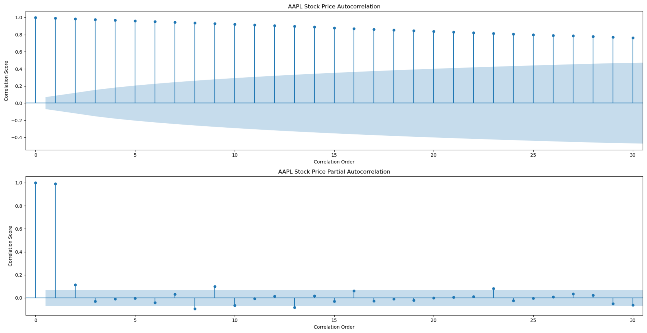


Figure 6: Autocorrelation and Partial Autocorrelation Plots of AAPL Stock Price

6.2. Data Distribution

Data distribution can affect the analysis method and model performance significantly, where standard normal distribution is the most preferable one in most cases. Typically, data in such a distribution have better statistical properties and predictive capability. An intuitive approach to check if the acquired data satisfy normal distribution is a Q-Q plot. For each feature, their Q-Q plots are demonstrated in Fig.7. It can be deduced that none of the features nicely matches a normal distribution accordingly.

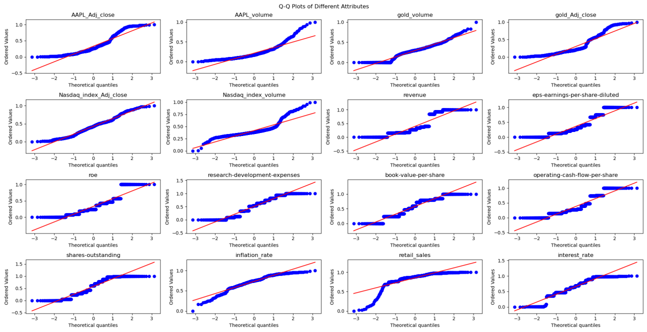


Figure 7: Q-Q Plots of Features

I further verify the above results based on D'Agostino and Pearson's test(d'Agostino, 1971)(D'AGOSTINO & Pearson, 1973) that combines skew and kurtosis to produce an omnibus test of normality. Setting the null hypothesis *H0: samples come from a normal distribution*, alternative hypothesis *H1: samples do not come from a normal distribution* and α 0.05, the test results for all features reject the null hypothesis, that is, the data of all features do not satisfy normal distribution under a 0.95 confidence interval.

6.3. Feature Correlation Analysis

In the EDA process, the feature correlations are investigated using the Spearman correlation coefficient, as it has been proved that none of the features is in normal distribution and assumed that any feature has nonlinear relationships with others. Heatmap (Fig.8) below illustrates the spearman correlation coefficient matrix, where some highlights correlations can be found. For example, Apple's revenue is highly positively related to its EPS, RoE and OCFPS; its share outstanding and BPS are negatively related to the stock price, gold price, Nasdaq index, RD expenses, etc. Although it does not indicate any causality between features, data redundancy can be deduced, and thus feature selection before using them in the model is crucial.

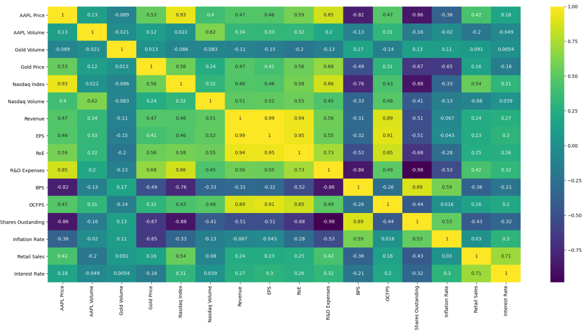


Figure 8: Spearman Correlation Coefficient Matrix

6.4. Seasonality Analysis

Seasonality has been proved to exist in many economic and financial variables. For instance, the prosperity of the stock market affects the stock price(Jegadeesh, 1991), so it is important to analyse the dependency of AAPL stock price on different time granularities before building the inference model. Two heatmaps are drawn in Fig.9 to show how features in the datasets vary as a function of the day of week and month. The aggregated value for each month/day is represented by the average within this period.

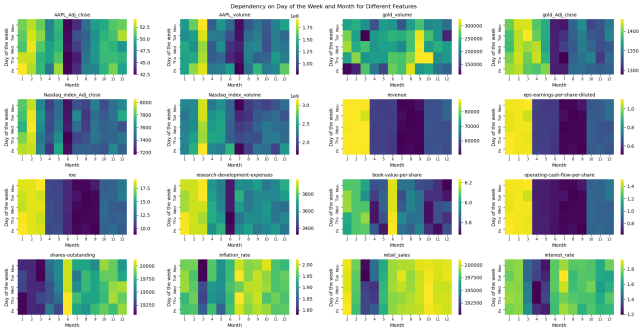


Figure 9: Dependencies of AAPL Stock Price on Month and Year

It can be found in Fig.9 that the lowest number of almost all the financial market data is always observed in June and July except for gold trading volume which has no significant trend. However, there is no seasonality on a workaday basis. As for Apple's financial data, the first quarter always enjoys higher revenue, EPS, RoE, RD expenses and OCFPS, but the opposite trend is experienced by shares outstanding. Noticeably, BPS usually reaches its peak in June. Interestingly, the macroeconomic variables also show a significant seasonality, where retail sales usually approach to its peak in the last months of a year and the number of others reaches their maxima in the middle of each year. Similarly, their daily trends are so insignificant to observe.

6.5. Feature Selection

Due to the high dimensionality of data with high correlations and redundancy, feature selection is needed to extract features that have good independence and predictive capability. Using Facebook Prophet(Taylor & Letham, 2018) which implements a generalized additive model and allows taking external features as additional regressors, a set of Prophet models with an external feature adding to the regressors are built and compared with the baseline model using only historical stock price data. By observing how better the model performs with the additional regressor, the predictive capability of features can be investigated. Note that the training set is the data starting from April of 2017 to March of 2020, and the testing set is the data in the April of 2020, which does not use the testing set of data in May of 2020 for inference later. Fig.10 a) shows the baseline model regressed on historical AAPL stock price only, where the poor predictions have an inversed trend to the ground truth. Fig.10 b) demonstrates the predictions by adding different features as an additional regressor, and some of the features significantly contribute to the improvement of model performance, e.g. AAPL stock trading volume, Nasdaq index and its trading volume, OCFPS and US retail sales.

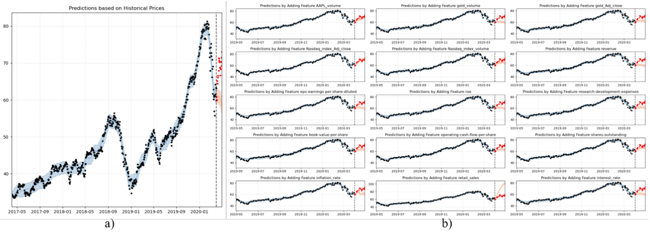


Figure 10: Baseline Model and Models with Different Regressors

To further statistically test if the predicted sequence is close to the ground truth, a hypothesis test is applied. As analysed in Section 6.2, none of the features satisfies normal distribution, and thus the parametric method cannot be used in this case. Therefore, the Wilcoxon signed-rank test(Wilcoxon, 1992) which is a non-parametric version of paired T-test is used to test if two related paired samples (sequences) come from the same distribution. Setting the null hypothesis *H0: The median of differences between two sequence is zero* and alternative hypothesis *H1: The median of the differences is not zero* and α 0.05, the test results for features of AAPL stock trading volume, Nasdaq index and its trading volume, OCFPS and US retail sales cannot reject the null hypothesis, while others should reject the H0. That means the predictions of the aforementioned features can have very close predictions to the ground true value sequence under a 0.95 confidence interval. Therefore, these features are selected for the following data inference.

Finally, the correlation coefficients are calculated again for the selected features. It is found that there is no strong correlation between them and thus can be further applied for data inference.

7. Data Inference

7.1. Model Description

Although traditional time series models like HMM, ARIMA have shown their success in solving time series prediction problem, LSTM provides a more robust and accurate solution that can fully consider the current input and previous state in the long and short term. Typically, an LSTM cell is composed of three gates, i.e. input, forget and output gate, where the forget gate calculates the percentage of the previous information (state) that needs to be forgotten, the input gate computes how much new information should be learned and updated to the current state, and the output gate generates a non-linear output according to its cell state.

Determining the structure of the LSTM model is critical as well as its training approach. A fine-tuned LSTM model is implemented with 64 LSTM cells followed by a dropout layer and an output neuron activated by a sigmoid function. The predictions of next-day AAPL stock close price are made based on the most recent two days' data. To fairly compare the predictive capability of adding selected external source data, the LSTM structure is kept in the models driven by solely price data, price data combined with selected features, dimension-reduced data using PCA and selected data only, respectively. To train the LSTM modes, Mean Square Error (MSE) that is given by Eq.4 is chosen as the loss function to be minimized using *Adam*(Kingma & Ba, 2014) optimizer.

$$Loss = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n} \tag{4}$$

, where y is the true value of AAPL stock price, $/haty$ is the predicted price and n is the number of predictions.

7.2. Experiment Results and Discussion

To evaluate the predictive capacity of the model, the whole dataset is split at the time of the start of May 2020, which means the data after this time point is used as the testing set, whereas the data before this time point is used as the training set. The model performance is shown in Fig.11, where the transparent blue region is the predicting region. Obviously, the predictions made by the model using both price and selected features (orange) well fit the ground true price (blue) and slightly better than predictions by others, although the other models using other combinations of features also have a satisfactory performance in predicting.

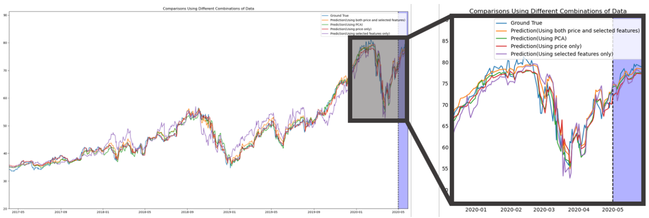


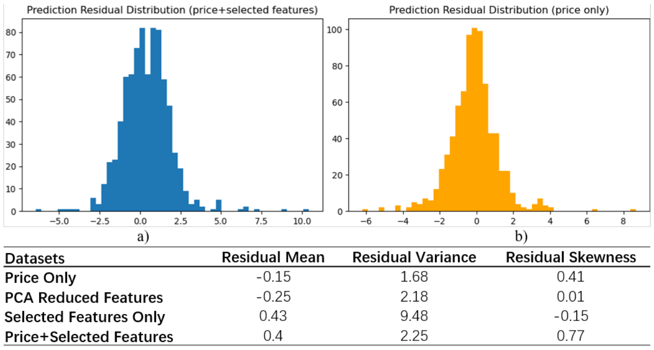
Figure 11: Predictions Made by Models Using Different Combinations of Feature

For further quantitative evaluation, the denormalized MSEs are calculated to compare their predictive capability shown in Table3. Although the models using price only and PCA-reduced features have a better training MSE, they perform worse in the testing set than the one using both price data and selected features. While PCA is believed to perform well to extract features automatically, the carefully selected features show a stronger predictive capability. The improvement indicates the significance of using external data in stock price prediction.

Table 3: Denormalized MSEs using Different Features

Datasets	Training MSE	Testing MSE
Price Only	0.00215	0.16274
PCA Reduced Features	0.00289	0.12040
Selected Features Only	0.01268	0.20979
Price+Selected Features	0.00315	0.06164

Fig.12 a) and b) demonstrate the distributions of the prediction residual of models using concatenated data and only price data, where both of them are approximately in a normal distribution with zero means. The inserted table demonstrates some statistics of residuals for all the models tested in the experiment. It can be seen that most means and skewness of residual are similarly close to zero, while the residual variance of the third model is much bigger than that of others.



Datasets	Residual Mean	Residual Variance	Residual Skewness
Price Only	-0.15	1.68	0.41
PCA Reduced Features	-0.25	2.18	0.01
Selected Features Only	0.43	9.48	-0.15
Price+Selected Features	0.4	2.25	0.77

Figure 12: Prediction Residual Distributions and Statistics

8. Conclusion

This report mainly analyses the importance of using diversified features from multiple sources, including the stock market, the company's financial report and macroeconomic statistics, to predict stock price. After the raw data are acquired from open datasets and scraped from web pages, they are properly stored both locally and in the cloud for efficient access to future applications. The collected data is then cleaned, visualized and transformed before EDA is applied to understand the statistical properties of these data features. In the EDA phase, it has been proved that the AAPL stock price is non-stationary time series with seasonality. High correlation relationships and non-normal distributions are experienced across features, and thus an exploratory feature selection process is applied. Based on LSTM trained by training data (data before April 2020), the inference model achieves a promising prediction of AAPL stock price in May 2020. Meanwhile, the experiment results show that the features of stock's trading volume, Nasdaq index and its trading volume, OCFPS and US retail sales can contribute to the model predictive capability compared to that using historical price data only. In particular, the model trained by data concatenated stock price with selected features has a similar denormalized MSE in training set like MSEs of others but the best (0.06164) in the testing set.

References

Adam, K., Marcet, A., and Nicolini, J. P. Stock market volatility and learning. *The Journal of Finance*, 71(1): 33–82, 2016.

Adebiyi, A. A., Ayo, C. K., Adebiyi, M. O., and Otokiti, S. O. Stock price prediction using neural network with hybridized market indicators. *Journal of Emerging Trends in Computing and Information Sciences*, 3(1):1–9, 2012.

Adebiyi, A. A., Adewumi, A. O., and Ayo, C. K. Comparison of arima and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*, 2014, 2014.

Alamsyah, A. and Zahir, A. N. Artificial neural network for predicting indonesia stock exchange composite using macroeconomic variables. In *2018 6th International Conference on Information and Communication Technology (ICoICT)*, pp. 44–48. IEEE, 2018.

Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pp. 106–112. IEEE, 2014.

Arkan, T. et al. The importance of financial ratios in predicting stock price trends: A case study in emerging markets. *Finanse, Rynki Finansowe, Ubezpieczenia*, (79):13–26, 2016.

Borjigin, S., Yang, Y., Yang, X., and Sun, L. Econometric testing on linear and nonlinear dynamic relation between stock prices and macroeconomy in china. *Physica A: Statistical Mechanics and Its Applications*, 493:107–115, 2018.

D'AGOSTINO, R. and Pearson, E. S. Tests for departure from normality. empirical results for the distributions of b 2 and b. *Biometrika*, 60(3):613–622, 1973.

d'Agostino, R. B. An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2):341–348, 1971.

De Fortuny, E. J., De Smedt, T., Martens, D., and Daelemans, W. Evaluating and understanding text-based stock price prediction models. *Information Processing & Management*, 50(2):426–441, 2014.

Dickey, D. A. and Fuller, W. A. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: journal of the Econometric Society*, pp. 1057–1072, 1981.

Fama, E. F. *Efficient market hypothesis*. PhD thesis, Ph. D. dissertation, University of Chicago, Graduate School of Business, 1960.

Fama, E. F. Random walks in stock market prices. *Financial analysts journal*, 51(1):75–80, 1995.

Fama, E. F. and French, K. R. The capital asset pricing model: Theory and evidence. *Journal of economic perspectives*, 18(3):25–46, 2004.

Fama, E. F. and French, K. R. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22, 2015.

Gay Jr, R. D. Effect of macroeconomic variables on stock market returns for four emerging economies: Brazil, russia, india, and china. *International Business & Economics Research Journal (IBER)*, 7(3), 2008.

Gokmenoglu, K. K. and Fazlollahi, N. The interactions among gold, oil, and stock market: Evidence from s&p500. *Procedia Economics and Finance*, 25(Supp. C):478–488, 2015.

Gyssens, M., Paredaens, J., Van den Bussche, J., and Van Gucht, D. A graph-oriented object database model. *IEEE Transactions on knowledge and Data Engineering*, 6(4):572–586, 1994.

Hassan, M. R., Nath, B., and Kirley, M. A fusion model of hmm, ann and ga for stock market forecasting. *Expert systems with Applications*, 33(1):171–180, 2007.

Heo, J. and Yang, J. Y. Stock price prediction based on financial statements using svm. *International Journal of Hybrid Information Technology*, 9(2):57–66, 2016.

Jain, A. and Biswal, P. Dynamic linkages among oil price, gold price, exchange rate, and stock market in india. *Resources Policy*, 49:179–185, 2016.

Jegadeesh, N. Seasonality in stock price mean reversion: Evidence from the us and the uk. *The Journal of Finance*, 46(4):1427–1444, 1991.

Khan, M. S. Macroeconomic variables & its impact on kse-100 index. *Universal Journal of Accounting and Finance*, 2(2):33–39, 2014.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lee, H., Surdeanu, M., MacCartney, B., and Jurafsky, D. On the importance of text analysis for stock price prediction. In *LREC*, volume 2014, pp. 1170–1175, 2014.

Loke, K. Impact of financial ratios and technical analysis on stock price prediction using random forests. In *2017 International Conference on Computer and Drone Applications (IConDA)*, pp. 38–42. IEEE, 2017.

Malkiel, B. G. *A random walk down Wall Street: including a life-cycle guide to personal investing*. WW Norton & Company, 1999.

Maysami, R. C., Howe, L. C., and Rahmat, M. A. Relationship between macroeconomic variables and stock market indices: Cointegration evidence from stock exchange of singapore's all-s sector indices. *Jurnal Pengurusan (UKM Journal of Management)*, 24, 2005.

Nti, K. O., Adekoya, A., and Weyori, B. Random forest based feature selection of macroeconomic variables for stock market prediction. *American Journal of Applied Sciences*, 16(7):200–212, 2019.

Pilinkus, D. and Boguslauskas, V. The short-run relationship between stock market prices and macroeconomic variables in lithuania: an application of the impulse response function. *Inžinerinė ekonomika*, (5):26–34, 2009.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K., and Soman, K. Stock price prediction using lstm, rnn and cnn-sliding window model. In *2017 international conference on advances in computing, communications and informatics (icacci)*, pp. 1643–1647. IEEE, 2017.

Sharpe, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

Taylor, S. J. and Letham, B. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

Wilcoxon, F. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pp. 196–202. Springer, 1992.