DATA ACQUISITION AND PROCESSING SYSTEMS ELEC0136 22/23 REPORTS

SN: 22103055

ABSTRACT

Forecasting stock price is essential for making investment choices and earning profit.However,randomness in stock prices makes accurate prediction challenging. We analyzed AAPL stock prices and auxiliary data using the standard Data Acquisition and Processing Systems (DAPS) process.Then,we trained relatively precise stock price forecast models for a one-month period using Facebook Prophet.The study found that Log-transformed AAPL stock prices resemble (but are not exactly) a random walk with a positive drift.Auxiliary data, especially T10Y2Y, can serve as a leading indicator for stock price prediction. Results show that models using multiple data sources outperform those using a single source.This project provides a basis for further stock price analysis, forecasting, and development of automated trading systems.

1. INTRODUCTION

Stock investment research is a complex field greatly influenced by market dynamics, investor psychology, and economic indicators. Traditional theories, such as the Efficient Market Hypothesis (EMH) proposed by Nobel laureate Eugene Fama in 1970, suggest that all publicly available information is already reflected in stock prices[1]. According to EMH, predicting stock price movements would be an exceedingly difficult task, as markets are assumed to be efficient, leaving little to no room for earning above-average returns.However, some research disagree with EMH[2].Basically, complex factors such as the market share distribution and crowd psychology cause stock prices to deviate from the results predicted by the EMH.

Contrary to the principles of EMH, Robert Shiller's "Irrational Exuberance" presents an alternative perspective, arguing that stock prices often deviate from their fundamental values due to investor irrationality. Shiller suggests that markets frequently undergo periods of undue optimism and pessimism, thereby creating profitable trading opportunities.[3]

Efforts to transcend the Efficient Market Hypothesis (EMH) and achieve gains in stock market prediction are generally categorized into two types: fundamental analysis and technical analysis.Some linear models, such as the autoregressive integrated moving average (ARIMA)[4], began to be employed for stock price forecasting.With the evolution of machine learning, algorithms like logistic regression and support vector machine have also seen success in this domain. Recently, the advent of deep learning has led to the extensive use of neural networks for stock price prediction, including graph neural networks, BiLSTM and AutoEncoders(AE)[5-8].

In stock prediction and analysis, selecting datasets and extracting features are crucial. On one hand, data that might impact stock prices exist in various forms: user comments, economic data, financial statements, and policy documents, which are often not directly recognizable by neural networks. On the other hand, the relationship between data and stock prices is usually complex. Thus, appropriate feature extraction is paramount to assist prediction models in better learning the characteristics within the data. HH Htun conducted a systematic study on feature extraction techniques relevant to stock prediction, identifying random forests, principal component analysis, and autoencoders as the most extensively applied and effective feature selection and extraction techniques in this field[9].

A noteworthy approach combines technical analysis with computer vision: Sezer et al. converted financial time series and technical indicators into two-dimensional images and trained convolutional neural networks, achieving favorable outcomes[10].

Since 2023, leading large language models such as Chat-GPT and GPT4 have pioneered new avenues for stock prediction. These LLMs are effective in analyzing textual content, including comments, news, and financial reports, as well as in evaluating stock prices. X Yu's research, employing language models based on Chat-GPT and Llama, analyzed stock price data and supplementary data, achieving outcomes surpassing those of the ARMA-GARCH model[11].

In this project, I followed the standard Data Acquisition and Processing Systems (DAPS) workflow to analyze Apple Inc.'s stock prices and other auxiliary data. The goal is to better understand the factors influencing stock price movements.Moreover, I employed Facebook Prophet models to predict future stock prices, seeking to comprehend the impact of various data sources on the predictive performance of the model, thereby offering useful insights for investment decisions.

The rest of this paper is organized as follows: Section 2 describes the datasets used in the study;Section 3 describes the methods for data collection;Section 4 describes the methods for data storage;Section 5 details the preprocessing

[1] The code is provided in github:https://github.com/UCL-ELEC0136/final-assignment-JacobChangUCL.git

steps taken to clean and prepare the data;Section 6 outlines the visualization process used to represent the data.;Section 7 employs EDA methods to analyze the datasets and conducts hypothesis testing to validate key assumptions; Section 8 details the training of Prophet models using single and multiple data sources for stock price prediction and compares the performance of models trained by single and multiple data sources ;Section 9 lists all the cited works.

2. DATA DESCRIPTION

The first dataset encompasses the historical stock prices of AAPL. I selected this dataset primarily because historical prices provide the most direct insight into a stock's performance. Stored in a CSV format, it captures daily stock prices for AAPL spanning from April 3rd, 2017 to April 1st, 2022, accumulating to a total of 1260 entries. The dataset is structured as follows.

●**Date**:Represents the record's date.Data type:string
●**Open**:Opening price of the day.Data type:float8
●**High**:the day's peak stock price.Data type:float8
●**Low**:the lowest stock price of the day.Data type:float8
●**Close**:the stock's closing price for the day.Data type:float8
●**Adj Close**:the stock's adjusted closing price, accounting for factors like dividends, stock splits, etc.Data type:float8
●**Volume**:the number of shares traded on that day.
Data type:int8

The second dataset is T10Y2Y, representing the yield spread between the 10-Year and 2-Year Treasury bonds. The 10-year yield reflects long-term economic expectations, while the 2-year indicates short-term views. A decreasing T10Y2Y suggests growing economic pessimism. Historically, when T10Y2Y nears zero or becomes negative, the U.S. is likely to face a recession within 12-24 months, which is highly likely to impact Apple Inc.'s stock value. Hence, T10Y2Y is instrumental for AAPL investment insights.The dataset is structured as follows.[12]

●**Date**:Represents the record's date..Data type:string
●**T10D2Y**:The spread between the 10-Year and 2-Year Treasury yields.Data type:float8

The third dataset is the Federal Funds Effective Rate, or DFF for short. The DFF signifies the prevailing risk-free rate in the market. A higher DFF typically reduces the attractiveness of investing in stocks, leading to a decline in overall market stock prices. Conversely, a lower DFF can boost the appeal of stock investments, pushing up stock prices. Therefore, I believe that the DFF data can provide insights into stock price predictions, especially for mega-corporations like Apple. By integrating the Capital Asset Pricing Model (CAPM) with the Dividend Discount Model (DDM), we can derive an expected stock price calculation formula:

$$E\left(Ri\right) = R_f + \beta_i\left(E\left(R_m\right) - R_f\right),$$

$$P_0 = \frac{D1}{E(R_i) - g},$$

It can be observed that the R_f term (DFF) is negatively correlated with stock price in the formula P0.(The meanings of specific economic symbols used in this formula are standard and can be referenced in [13].

●**Date**:Represents the record's date..Data type:string.
●**DFF**:Federal Funds Effective Rate of that day.data type:float8

3. DATA ACQUISITION

3.1. Data Acquisition.

I utilized the ”yfinance API“ to fetch stock price data, a process encapsulated in the download_data.py application. The Federal Funds Effective Rate (DFF) data was obtained from the Federal Reserve Economic Data (FRED) repository. This data is gathered using the “requests” library, with the specific operation encapsulated in the download_DFF() method of the download_data.py module.

All other data were also auto-downloaded from the Federal Reserve Economic Data (FRED) using scripts found in the download_data.py module.

4. DATA STORAGE

In this study, two data storage approaches are employed for the preservation and accessibility of data.All data are stored in the 'data' folder in CSV file format, as local storage offers the advantages of fast access, no need for internet connectivity and lower price.

Similarly, a MySQL database was established on Amazon Web Services (AWS) Relational Database Service (RDS), with a connection facilitated between the database and a Python program using the mysql.connector to store the data of this project. This approach was selected for several reasons: the free version offered by AWS, the stability of the service, the maturity of the MySQL database system, and the structured nature of the data involved in this project, which is particularly well-suited to a relational database model.

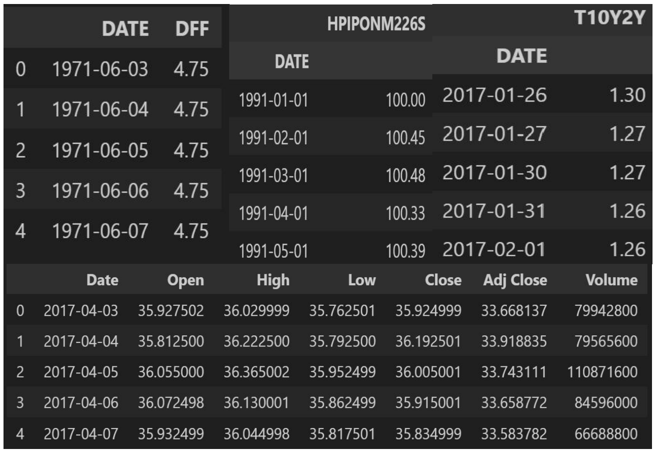
CRUD (Create, Read, Update, Delete) operations on the cloud database were executed using the cursor.execute() function in Python, allowing for efficient manipulation of SQL commands. Given the relatively modest volume of the dataset, both cloud and local storage options proved to be convenient for this task.

However, it is important to note that the volume of data that can be inserted in a single operation is limited by the maximum SQL command length that the cursor in mysql.connector can handle, typically several tens of MB. This limitation was not an issue for the current scope of the project. Nevertheless, for datasets significantly larger than this limit, errors might occur. In such cases, the SQLAlchemy package could offer a superior alternative to mysql.connector, especially in handling larger data sets.

5. DATA PREPROCESSING

5.1. Data Cleaning

Based on the 'DATA DESCRIPTION', it is evident that all datasets (AAPL, DFF, etc.) adhere to the principles of Tidy Data, where each column represents an individual variable, each row encapsulates a distinct observation, and every cell houses a singular value. Consequently, primary preprocessing endeavors were centered around the rectification of missing and anomalous values.



	DATE	DFF
0	1971-06-03	4.75
1	1971-06-04	4.75
2	1971-06-05	4.75
3	1971-06-06	4.75
4	1971-06-07	4.75

	HPIPONM226S
DATE	
1991-01-01	100.00
1991-02-01	100.45
1991-03-01	100.48
1991-04-01	100.33
1991-05-01	100.39

	T10Y2Y
DATE	
2017-01-26	1.30
2017-01-27	1.27
2017-01-30	1.27
2017-01-31	1.26
2017-02-01	1.26

	Date	Open	High	Low	Close	Adj Close	Volume
0	2017-04-03	35.927502	36.029999	35.762501	35.924999	33.668137	79942800
1	2017-04-04	35.812500	36.222500	35.792500	36.192501	33.918835	79565600
2	2017-04-05	36.055000	36.365002	35.952499	36.005001	33.743111	110871600
3	2017-04-06	36.072498	36.130001	35.862499	35.915001	33.658772	84596000
4	2017-04-07	35.932499	36.044998	35.817501	35.834999	33.583782	66688800

Fig. 1 head of four datasets

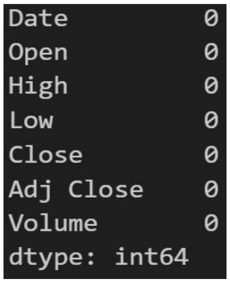
Missing value:

The stock market is closed on weekends and holidays, resulting in absent data for these days.

it's possible to ignore missing values because the absence of stock market data on holidays is not due to data collection issues but because trading does not occur on these days.Traders conduct their transactions before and after holidays, and the trading volume and prices already reflect the holiday factor.Several factors led me to choose linear interpolation for filling in the data gaps. First, not addressing these missing values would lead to numerous NaNs when calculating the rate of change in stock prices using a sliding window.Second, for consistency with other datasets that are recorded on a daily basis, filling in the gaps becomes necessary. Among various interpolation methods, I chose linear interpolation for two reasons: 1.theoretically, stock prices exhibit continuous trends with sustained directional momentum, which we refer to as the momentum factor. Therefore, linear interpolation and polynomial fitting offer higher accuracy, and 2) the lengths of the missing time intervals are not significant.Other methods like forward-filling lack the desired level of precision,while machine learning methods are more complex and require higher computational overhead. The advantage of linear interpolation is that it provides a good trade-off between quality and computational cost.We can also opt for machine learning packages like missingpy, which employ algorithms like random forests to address this issue, potentially yielding better results at the cost of increased computational resources.

I used different filling methods for the DFF dataset. On days where DFF data is absent, the market typically operates based on the most recent available DFF information. This suggests that data after forward-filling or interpolation would more closely represent the actual scenario. As a result, I employed a forward-filling technique to align it with the dates present in the AAPL dataset.

The data in other datasets is filled using interpolation methods, with reasons similar to the AAPL dataset.After filling, there are no missing data in the datasets, as shown in the following figure.

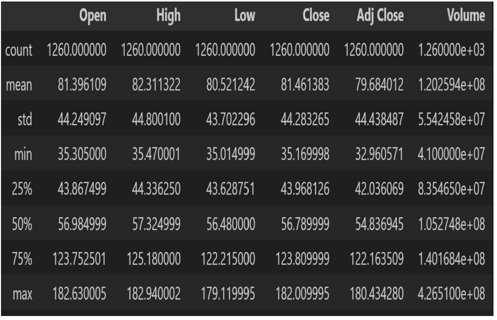


```
Date         0
Open         0
High         0
Low          0
Close        0
Adj Close    0
Volume       0
dtype: int64
```

Fig. 2 Missing data in the AAPL dataset after interpolation (other datasets are similar)

Noisy:

I use AAPL.describe() and DFF.describe() to see if there is strange maximum value or minimum value in the datasets.The same approach was also applied to the remaining two datasets.



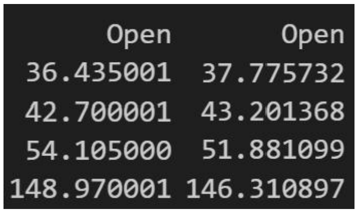
	Open	High	Low	Close	Adj Close	Volume
count	1260.000000	1260.000000	1260.000000	1260.000000	1260.000000	1.260000e+03
mean	81.396109	82.311322	80.521242	81.461383	79.684012	1.202594e+08
std	44.249097	44.800100	43.702296	44.283265	44.438487	5.542458e+07
min	35.305000	35.470001	35.014999	35.169998	32.960571	4.100000e+07
25%	43.867499	44.336250	43.628751	43.968126	42.036069	8.354650e+07
50%	56.984999	57.324999	56.480000	56.789999	54.836945	1.052748e+08
75%	123.752501	125.180000	122.215000	123.809999	122.163509	1.401684e+08
max	182.630005	182.940002	179.119995	182.009995	180.434280	4.265100e+08

Fig.3 The maximum value, minimum value, four quartiles, mean, and standard deviation of the AAPL dataset. (Other datasets are processed in the same manner)

Utilizing the Interquartile Range (IQR) as a metric for identifying outlier observations within the dataset is deemed unsuitable due to the upward trend manifested in the stock prices. In my analysis, two approaches were contemplated: 1) Transforming the time series into a stationary time series and then applying the IQR; 2) Abandoning the use of IQR. Ultimately, for the sake of convenience, the second option was chosen. A methodology that incorporates z-score computation based on a moving window is employed to detect anomalies within stock prices. Following the identification of these outliers, they are replaced with values sampled from a normal distribution characterized by the mean and standard deviation of the respective window. The formula used for the Shift-window Z-score is shown below:

$$Z_i=\frac{x_i-\bar{X}}{std(X)}, X=\{x_i,x_{i+1},...,x_{i+n-1}\}, n=window\_size$$

I set the window size to 6 and labeled values that exceed two standard deviations as anomalies. The relevant codeis encapsulated in the replace_anomalies_with_normal() function within the data_preprocessing.py file.



```
      Open        Open
 36.435001   37.775732
 42.700001   43.201368
 54.105000   51.881099
148.970001  146.310897
```

Fig. 4 left:raw anomaly value.

right:value replaced by normal distribution.

Data Integration: After data cleaning, I integrated all the data from various tables into a single master table using the "DATA" key as the primary key. This was done to save space and facilitate subsequent analysis.More specificly,I used outer join policy to join the tables,and filled the missing value(very little) with forward-fill method because there are only two lines of missing value,then I droped the repeated time columns.

5.2. Data Visualization

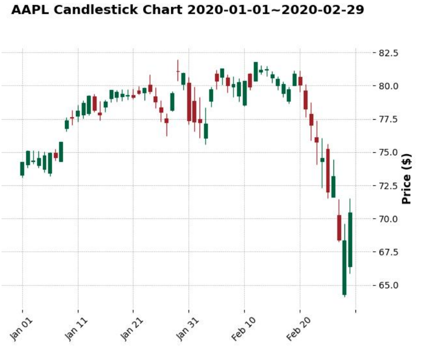


Fig. 5 candlestick chart of AAPL

As our study utilizes stock trading data that includes opening, closing, highest, and lowest prices, candlestick charts emerge as a particularly effective visualization method. For instance, within the sequences of stock prices, we can identify structures such as the Hammer Candlestick Pattern, Star Patterns, and Engulfing Candlestick Patterns. These patterns aid in forecasting future stock price movements and can act as precursors for uptrends or trend reversals[14].

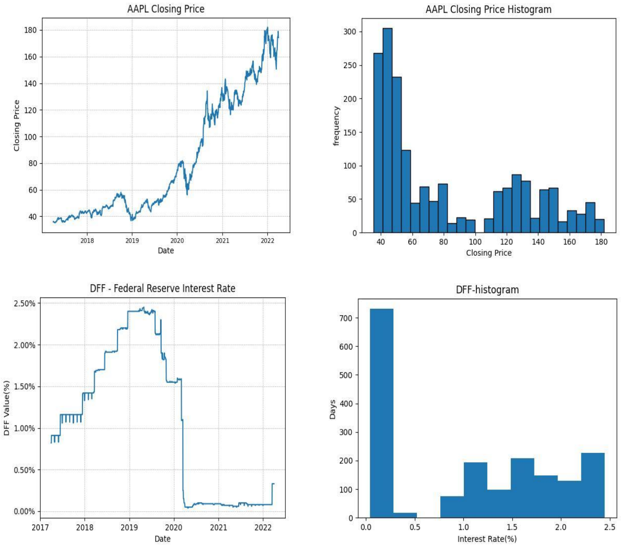


Fig. 6 Line charts and Histogram

Upper Panel: AAPL dataset; Lower Panel: DFF dataset.

No significant outliers were observed. The histogram and line graph indicate a concentration of AAPL's prices in the lower price range. This could be attributed to the choice of coordinates since investment returns are more accurately represented by rates of return rather than stock prices, which are better visualized on a logarithmic scale. Alternatively, it might indicate that AAPL's return rates remained low for an extended period. This hypothesis necessitates logarithmic transformation of AAPL stock prices, a topic I intend to revisit during the Exploratory Data Analysis (EDA) phase.

The histogram indicates a substantial concentration of DFF data near zero, corresponding to the line graph segment from March 2020 to April 2022. A comparative analysis with AAPL stock prices reveals a simultaneous sharp decline in DFF values and AAPL prices in March 2020. Following the cessation of DFF's decline, AAPL's stock price experienced a dramatic increase (post-April 2020). The end of DFF's downturn may potentially serve as an optimal buying point for AAPL stocks, aligning with economic rationale. Additionally, there appears to be an inverse relationship between DFF levels and AAPL's price gains - higher DFF regions correspond to smaller AAPL price increases and vice versa. This hypothesis warrants further validation. Our analysis suggests a complex, non-linear correlation between AAPL and DFF data.

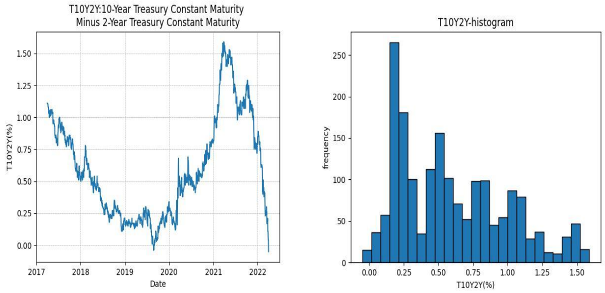


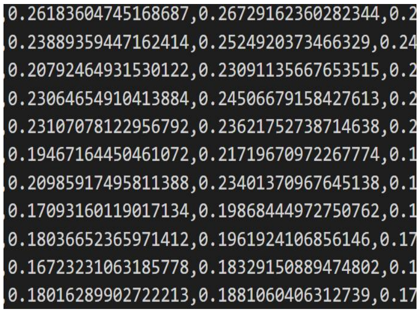
Fig.7 Line charts and Histogram for T10Y2Y

The relationship between T10Y2Y and AAPL stock prices is complex, with no clear linear correlation. T10Y2Y, an early indicator of economic risk, often precedes economic crises by 6-12 months. This connection, however, is hard to confirm due to the short time span of the dataset. Figure 7 shows T10Y2Y dropping below zero in July 2019, followed by a significant AAPL price drop eight months later. From April 2021, a T10Y2Y decline seemed to predict a future drop in AAPL stock, which indeed fell from April 2022 to January 2023, though this period is not displayed in the dataset.

5.3. Data Transformation

I choose normalization for the data set:Because the range of data sets are significantly different,so I want to normalize them to same scale. The function is shown below:

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$



```
0.26183604745168687,0.26729162360282344,0.2
0.23889359447162414,0.2524920373466329,0.24
0.20792464931530122,0.23091135667653515,0.2
0.23064654910413884,0.24506679158427613,0.2
0.23107078122956792,0.23621752738714638,0.2
0.19467164450461072,0.21719670972267774,0.1
0.20985917495811388,0.23401370967645138,0.1
0.17093160119017134,0.19868444972750762,0.1
0.18036652365971412,0.1961924106856146,0.17
0.16723231063185778,0.18329150889474802,0.1
0.18016289902722213,0.1881060406312739,0.17
```

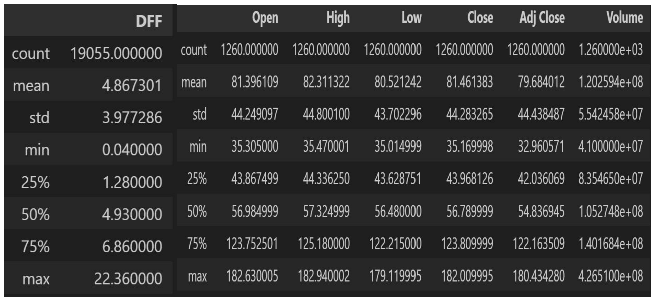
Fig. 8 normalized data

The result is shown in Fig 7. It can be observed that the data is perfectly normalized to between 0 and 1.

6.DATA EXPLORATION

6.1. EDA on Data

First,we check the basic information of these datasets using AAPL.describe() and DFF.describe() method.



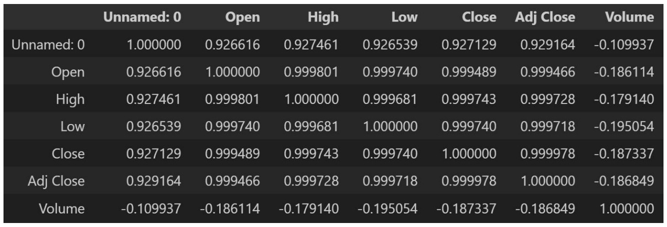
	DFF
count	19055.000000
mean	4.867301
std	3.977286
min	0.040000
25%	1.280000
50%	4.930000
75%	6.860000
max	22.360000

	Open	High	Low	Close	Adj Close	Volume
count	1260.000000	1260.000000	1260.000000	1260.000000	1260.000000	1.260000e+03
mean	81.396109	82.311322	80.521242	81.461383	79.684012	1.202594e+08
std	44.249097	44.800100	43.702296	44.283265	44.438487	5.542458e+07
min	35.305000	35.470001	35.014999	35.169998	32.960571	4.100000e+07
25%	43.867499	44.336250	43.628751	43.968126	42.036069	8.354650e+07
50%	56.984999	57.324999	56.480000	56.789999	54.836945	1.052748e+08
75%	123.752501	125.180000	122.215000	123.809999	122.163509	1.401684e+08
max	182.630005	182.940002	179.119995	182.009995	180.434280	4.265100e+08

Fig.9 The basic information of AAPL and DFF datasets

It can be observed that several attributes of AAPL, namely Open, High, Low, Close, and Adj Close, have very similar values. This suggests the presence of multicollinearity and redundancy among these attributes.We need to remove some of them.

We check the correlation matrix of these redundant data.



	Unnamed: 0	Open	High	Low	Close	Adj Close	Volume
Unnamed: 0	1.000000	0.926616	0.927461	0.926539	0.927129	0.929164	-0.109937
Open	0.926616	1.000000	0.999801	0.999740	0.999489	0.999466	-0.186114
High	0.927461	0.999801	1.000000	0.999681	0.999743	0.999728	-0.179140
Low	0.926539	0.999740	0.999681	1.000000	0.999740	0.999718	-0.195054
Close	0.927129	0.999489	0.999743	0.999740	1.000000	0.999978	-0.187337
Adj Close	0.929164	0.999466	0.999728	0.999718	0.999978	1.000000	-0.186849
Volume	-0.109937	-0.186114	-0.179140	-0.195054	-0.187337	-0.186849	1.000000

Fig.10 Line charts and Histogram for T10Y2Y

It can be observed that the correlation between attributes are close to 1, except for the volume attribute, which has a lower negative correlation without an apparent reason.

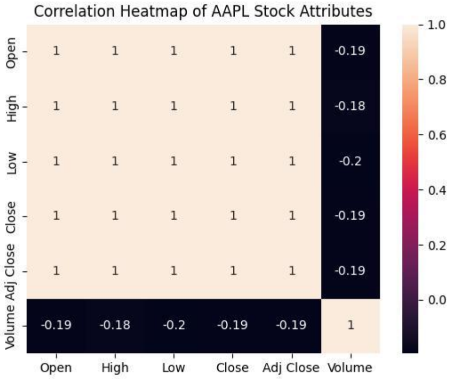


Fig.11 Line charts and Histogram for T10Y2Y

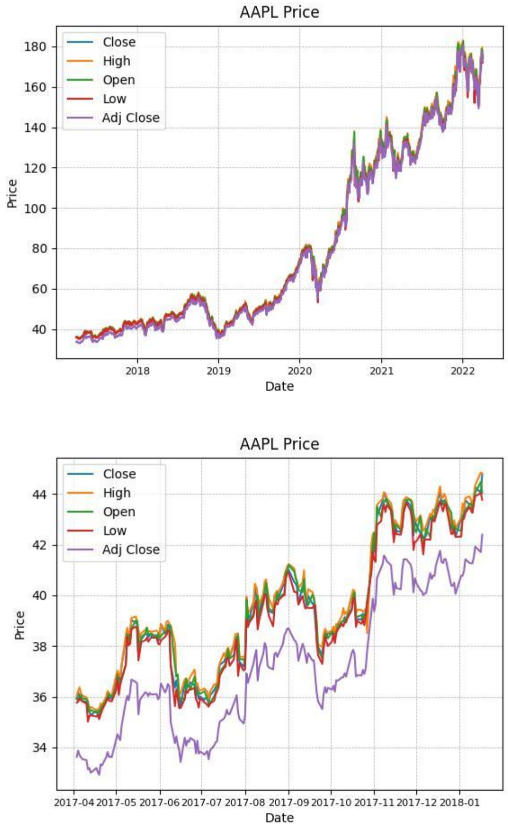


Fig.12 Line charts and Histogram for T10Y2Y

It is observed from Fig.12 that Adj Close is generally slightly lower than other values. This is because it includes earanings directly credited to the investor's account due to dividends, stock splits, or mergers[15].Therefore, using Adj Close instead of Close provides a more accurate prediction of actual returns. Given the task of predicting the "Close" price, I selected the Close attribute and removed the other columns (Open, High, Low, Adj Close) from the dataset.

We can observe several features from Figure 12:

1.The investment return from 2017 to 2019 was slightly less than that after 2019.

2.Recommending AAPL stock to clients in September 2018 or January 2020 would have resulted in a loss for at least six months.

3.Every significant drawdown (-20% to -30%) presented a good buying opportunity.

To analyze the cyclical components of stock prices, we conducted time series decomposition on AAPL's stock price.

TimeSeries Decompsition：

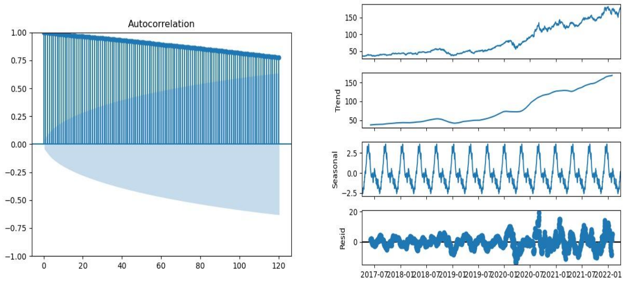


Fig.13 left: autocorrelation plot of AAPL
right: seasonal_decompose plot of AAPL

Before proceeding with time series decomposition, it is essential to determine the appropriate period for the decomposition. To explore potential cyclical components within the time series, I utilized the acf() function from the statsmodels library to plot the autocorrelation of the time series at various lags. These autocorrelation plots aim to uncover hidden periodicities in the series through their local peaks. However, the autocorrelation plots suggest that this time series does not exhibit significant cyclical components.

I used the seasonal_decompose() function from the statsmodels library to perform time series decomposition with a period of 30 days. It was observed that the time series exhibited a clear trend component and an increasing noise component, while the seasonality was very weak.

This is a logical conclusion. If AAPL's stock prices exhibited a clear seasonal pattern, hedge funds could exploit this by buying at relative highs and selling at relative lows, thereby achieving significant excess returns. Consequently, such arbitrage actions would effectively neutralize the seasonal component, leading to the absence of a pronounced cyclical pattern in AAPL's stock prices.

HYPOTESIS TESTING 1:
Test for the Random Walking Hypothesis

Through prior exploratory data analysis, I have raised a question: aside from trends akin to exponential functions, it is difficult to find regular features within AAPL's stock price. Consequently, I hypothesize that AAPL's stock price may essentially represent a random walk with a positive drift component. Hypothesis testing will be used to validate (or refute) this hypothesis,

We can use the Ljung-Box test to determine if the time series is white noise, but the precondition for the Ljung-Box test to be effective is that the time series must be stationary. Therefore, I first conduct a unit root test on the series; if it is non-stationary, I stabilize it through first-order differencing. Afterward, I verify the effect of stabilization with another unit root test.

1. Logarithmization

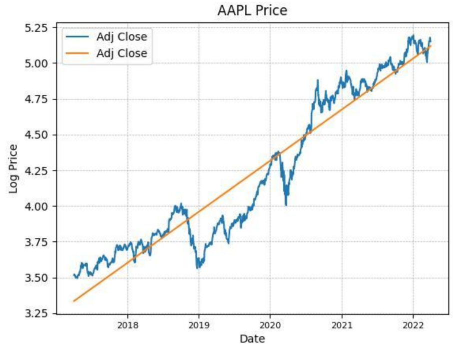


Fig.14 Log-transformed stock prices and their linear approximation.

Given that stock prices do not accurately reflect true investment returns, my initial step involves logarithmically transforming them. Logarithmic transformation helps stabilize the volatility and rate of change of stock prices across different levels. Additionally, the percentage changes in stock price increases and decreases have the same absolute value (albeit with opposite signs) after a logarithmic transformation. This allows us to model the step sizes of a random walk in a more equitable manner.

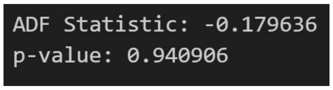
HYPOTESIS TESTING 1.1 Unit Root Test:

The Augmented Dickey-Fuller (ADF) **[sdasdasd]**test is employed to determine the presence of a unit root, necessitating the stabilization of the series if one is found.

Null Hypothesis(H0): The time series has a unit root, indicating it is non-stationary.

Alternative Hypothesis(H1): The time series does not have a unit root, indicating it is stationary.

critical value:p=0.05



```
ADF Statistic: -0.179636
p-value: 0.940906
```

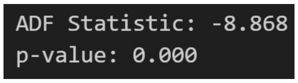
As illustrated in the figure above, consistent with observational results, the original time series almost certainly contains a unit root.

1. First Difference

The diff() method is used to remove the unit root through first differencing.

2. The ADF test is then applied to ascertain the removal of the unit root and to verify the stationarity of the time series.

HYPOTESIS TESTING 1.2 Unit Root Test:

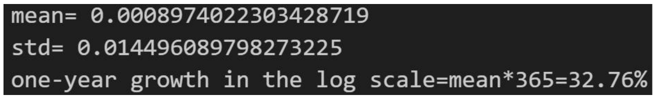


```
ADF Statistic: -8.868
p-value: 0.000
```

critical value:p=0.05

As shown in the figure above, the p-value is below the critical value, indicating that the time series becomes stationary after first-order differencing.

EDA of the First-Differenced Series



```
mean= 0.0008974022303428719
std= 0.014496089798273225
one-year growth in the log scale=mean*365=32.76%
```

Fig.15 Basic information of the First-Differenced Series

Check the average value, variance, and other information. After first-order differencing, the time series has an average value of 0.0009. Analyzing this value on a logarithmic scale indicates that the underlying time series exhibits an annual growth rate of approximately 33%.It is crucial to note that this growth rate does not correspond to the annualized growth rate of AAPL, as the data has undergone logarithmic transformation.

Assume that a function can be represented as log(y) in a logarithmic coordinate system.

$$\log(y) = k \cdot x + b$$

therefore,

$$y = e^{kx+b} = C(e^{k})^{x}$$

$$u = e^{k} - 1 = e^{0.33} - 1 = 0.39$$

In this context, u represents the annual growth rate. This suggests that Apple Inc. (AAPL) is capable of generating an approximate annual return of 39% for investors.

We now proceed to visualize the first-order difference to identify patterns of value.

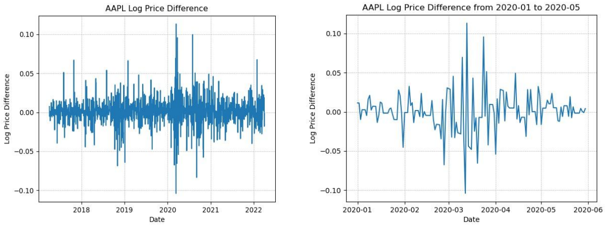


Fig.16 Left:the first-order difference of AAPL Right:Image of the locations of increased volatility

The graphical analysis indicates that while the data look like noise, there are discernible periods of amplified volatility within this noise, particularly before 2019 and after 2020.

In the original data, these two periods correspond to significant decreases and increases in stock prices, respectively, and during these times, excellent buying opportunities were present. The concentration of volatility may have an economic basis, as suggested by the concept of 'Irrational Exuberance' [3]. According to behavioral finance, market overreactions and herd behavior can lead to short-term increases in market volatility, thereby offering rational investors (like us) the opportunity to achieve above-average returns.

We now employ hypothesis testing to ascertain the degree to which the first-order difference exhibits characteristics of white noise.

3. **HYPOTESIS TESTING 1.3 white noise test:**

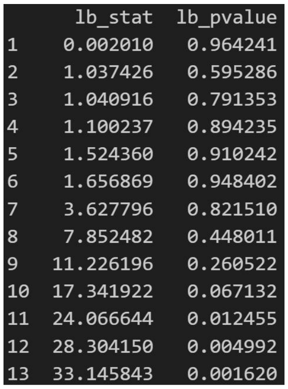
The Ljung-Box test is used to examine whether the log-transformed stock price series, after first-order differencing, is white noise.

Null Hypothesis(H0): The series is white noise

Alternative Hypothesis(H1): The time is not white noise

critical value:p=0.05

The results are as follows:



```
      lb_stat  lb_pvalue
1    0.002010   0.964241
2    1.037426   0.595286
3    1.040916   0.791353
4    1.100237   0.894235
5    1.524360   0.910242
6    1.656869   0.948402
7    3.627796   0.821510
8    7.852482   0.448011
9   11.226196   0.260522
10  17.341922   0.067132
11  24.066644   0.012455
12  28.304150   0.004992
13  33.145843   0.001620
```

Fig.17 Result of the Ljung-Box test. The leftmost column represents the number of days considered.

The Ljung-Box test results indicate that the p-value is less than 0.05 within 11 days, and greater than 0.05 for periods exceeding 11 days. This suggests that the time series closely approximates white noise, yet it does not fully conform to it. Consequently, our initial hypothesis (the null hypothesis) is incorrect, indicating that despite It's hard to identify useful characteristics, the time series exhibits autocorrelation and is not a white noise sequence.

The above hypothesis testing underscores the value of further EDA.(Because conducting EDA on a random sequence is futile.) Therefore, we will proceed with the EDA on the data.

EDA for DFF and T10Y2Y

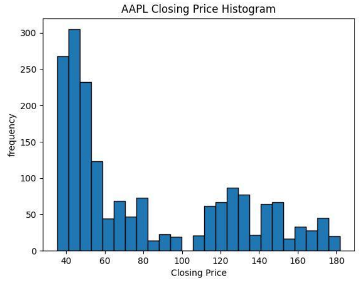


Fig.18 Histogram of AAPL Closing Price

By plotting a frequency distribution histogram for AAPL stock prices, it was observed that the prices lingered within the $40 to $60 range for an extended period. Given the high rate of the Discount Federal Funds Rate (DFF) during the same period, a potential correlation between these two variables is suggested. The elevated risk-free rate may diminish the investment appeal of risk assets such as stocks.

I tried to convert the AAPL stock price into a future return rate, assessing the relationship between this return rate and economic indicators like DFF and T10Y2Y (the 10-year Treasury yield minus the 2-year Treasury yield). The transformation function used is:

$$g(x) = \frac{f(x+t) - f(t)}{f(x)} \bullet \frac{365}{t}$$

In the equation, t represents the investment duration, denoting the time interval for calculating the rate of return, measured in days. Here, f(x) signifies the stock price, while g(x) denotes the function for the annualized rate of return.

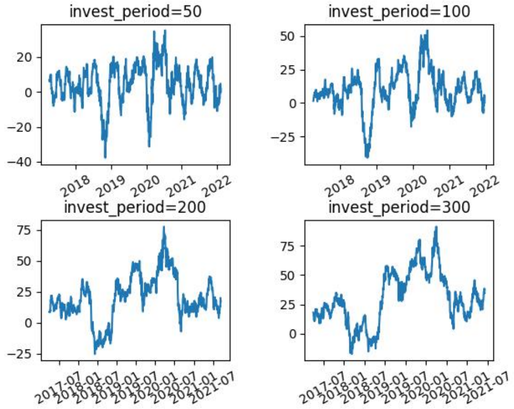


Fig.19 Investment returns of AAPL stock across different investment periods.

We analyze the relationship between this function and T10Y2Y, DFF. Taking an investment period (Invest_Period) of 100 days as an example, we examine the correlation between AAPL's return rate and DFF, T10Y2Y. A scatter plot is created to illustrate the relationship between AAPL's return rate and DFF.

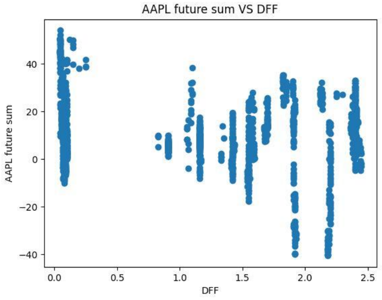


Fig.20 The scatter plot of AAPL's return rate and DFF

It has been observed that in regions with a low DFF, AAPL's return rate is higher; conversely, in areas with a high DFF, AAPL's return rate tends to be lower.

HYPOTHESIS TESTING 2 CORRELATION TEST

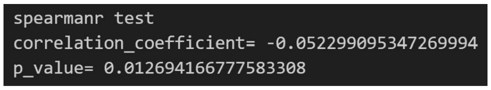
We conducted a correlation test between AAPL and DFF using the Spearman rank correlation coefficient.

I chose Spearman rank correlation because it does well in measuring non-linear correlation

Null Hypothesis (H0): There is no correlation between the two variables.

Alternative Hypothesis (H1): There is a correlation between the two variables.

Significance Level: 0.05



```
spearmanr test
correlation_coefficient= -0.052290095347269994
p_value= 0.012694166777583308
```

The p-value is 0.0126, which is less than 0.05. So we reject the null hypothesis. This indicates a significant correlation, though the strength is pretty low.

The yield spread between 10-year and 2-year Treasury bonds (T10Y2Y) is a critical predictive indicator. We examined the relationship between the inverted T10Y2Y and the 300-day return curve of AAPL investments.

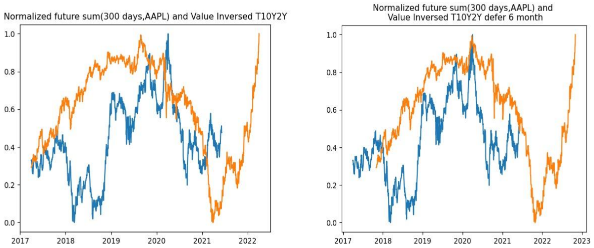


Fig.21 The relationship between T10Y2Y and AAPL's return rate
Left: No delay.
Right: After delaying T10Y2Y by 7 months.

An really interesting phenomenon was observed: the two appeared to change synchronously. When I postponed the leading indicator (T10Y2Y) by seven months (the lead time often suggested in past papers), they exhibited a very high correlation. This is particularly noteworthy considering that T10Y2Y is a leading indicator, making it invaluable in our analysis.

HYPOTHESIS TESTING 3：CORRELATION TEST

We aim to accomplish two things:

1. Determine the lag period of the 300-day AAPL return rate relative to T10Y2Y by iterating over the cross-correlation function (CCF) values at different lags to find the maximum value.

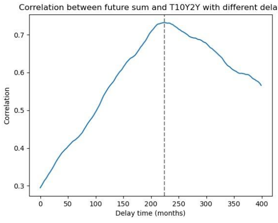


Fig.22 Diagram of the relationship between lag period and correlation.

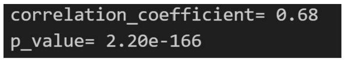
Clearly, there is only one maximum value within the domain, corresponding to 7.47 months, or 224 days.

2. Conducting a hypothesis test on their correlation using the Pearson correlation coefficient test.

Null Hypothesis (H0): There is a correlation between the two variables.

Alternative Hypothesis (H1): There is no correlation between the two variables.

Significance level: 0.05



```
correlation_coefficient= 0.68
p_value= 2.20e-166
```

The P_value is 2.2e-166, less than the critical value, indicating a strong correlation between the two variables. With a correlation strength of 0.68, it serves as an important leading indicator.

CONCLUSION OF EDA:

EDA of Apple Inc.'s (AAPL) stock prices and economic indicators reveals a complex nonlinear relationship between AAPL's stock prices, the risk-free rate (DFF), and the yield spread between 10-year and 2-year Treasury bonds (T10Y2Y). High correlation among stock price attributes within the AAPL dataset indicated data redundancy, so the superfluous data was removed. Time series analysis of AAPL showed no significant periodicity or seasonal effects, suggesting that Market traders' preemptive trading has smoothed out these fluctuations., Therefore, the long-term holding strategy(Buy and hold Strategy) offers a high cost-effectiveness. "Buy and hold" strategy is the Nash equilibrium strategy in a strong-form efficient market[13].In this case,it offers a satisfactory annual return rate of 39%.

After removing the trend component, the stock price resembles a random walk (though it is not, as confirmed by the Ljung-Box test). This indicates that, post-trend removal, the stock price data contains little information amidst lots of noise, requiring a machine learning model with high noise robustness.

Volatility clustering observed in the first-order difference of stock prices, coupled with potential buy points in these clusters, may be attributed to investors' irrational behavior and herding effects, offering opportunities for generating excess returns. EDA findings suggest that T10Y2Y could serve as a leading indicator of significant stock price declines; meanwhile, higher DFF values subtly decrease investment returns, with DFF inflection points potentially signaling buying opportunities, necessitating a joint analysis with prevailing U.S. economic policies.

Considering the intricate nonlinear relationship between auxiliary data and stock prices, inadequate feature transformation, or the model's inability to capture these complexities may not enhance predictive performance and could even detract from it. Thus, model selection and feature engineering should be carefully considered to harness the predictive power of auxiliary data effectively.

7.DATA INFERENCE

In this section, we need to develop time-series forecasting models using both single and multiple data sources to predict AAPL's stock price.

I have decided against adopting any linear model for predicting the stock price movements of AAPL. This decision stems from the no-arbitrage assumption and the strong-form efficient market hypothesis[], according to which any simple, exploitable attributes for excess returns would be neutralized by market transactions. This is particularly evident in the case of Apple Inc., whose stock is constantly monitored by numerous traders, thereby enhancing this characteristic. Therefore, I think that attempts to secure excess returns by identifying linear patterns in the market through linear models are often unsuccessful.

In light of this, I posit that only models capable of integrating multiple data sources and capturing their intrinsic nonlinear relationships stand a chance of outperforming the benchmark returns. Consequently, I have chosen Facebook Prophet as the forecasting tool. Facebook Prophet, an advanced time series forecasting library, implements a time series prediction framework based on generalized additive models. Its notable advantage lies in its ability to automatically handle seasonal variations, trends, and holiday effects, which are common challenges in traditional time series forecasting methods.

Most importantly, it supports external variables and provides interpretability of the predictions, meaning that we can easily analyze the impact of different data sources on the forecasting results.

Table 1. Model Performance

	MSE	RMSE	MAE	MAPE	MDAPE	SMAPE
Single Data Sources	0.010	0.100	0.082	0.018	0.015	0.018
Multiple Data Sources	0.0087	0.093	0.076	0.016	0.013	0.016

It can be observed that after using multiple data sources, every model metric experienced a slight improvement.(Note that every model metric has improved, which indeed exceeded my expectations. I had thought it would be difficult for the model to learn useful features from the additional datasets.)

7.1: Development of model using stocks

I trained my model using a single data source(AAPL Stock Price).The dataset was divided into training (80%) and testing(20%) sets.

The training results are as follow:

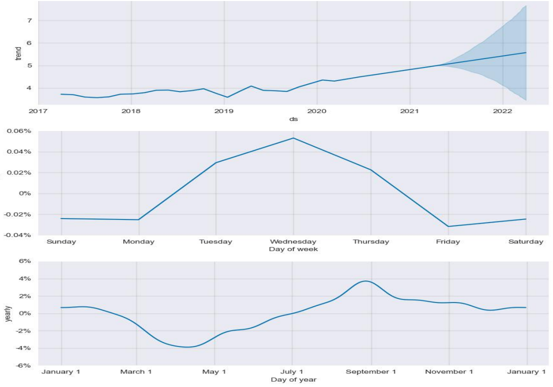


Fig.23 The trend and seasonality of AAPL

It can be observed that the trend component predominates, with both weekly and monthly seasonal effects being considerably weak and lacking in exploitative value. The confidence interval of the model's predictions rapidly widens as time progresses.

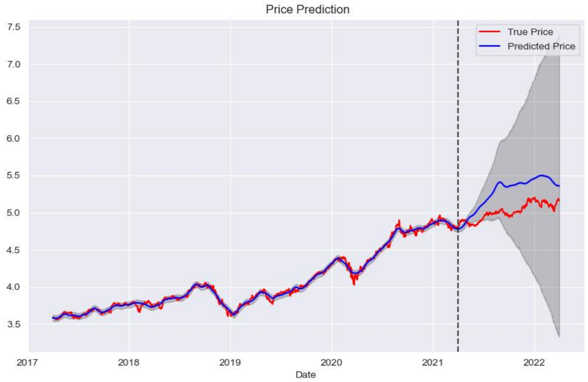


Fig.24 Actual and predicted prices in logarithmic scale, with a forecast period of one year.

The model fits the training set perfectly but shows larger errors on the test set. Considering that it predicts the stock price changes for the coming year (rather than just one month), this level of error is acceptable.

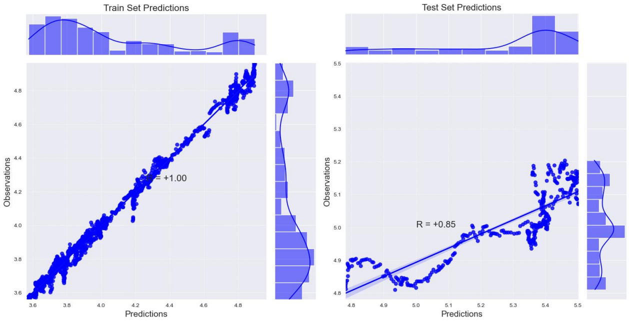


Fig.25 joint plots showing marginal distributions.Left:training set. Right:testing set.

Quantitative indicators of model performance:
RMSE in this month: 0.071
Correlation: +0.85

7.2: Development of model using stocks and other data sources

I trained my model using multiple data sources(AAPL Stock Price,DFF,T10Y2Y).The dataset was divided into training (80%) and testing(20%) sets.

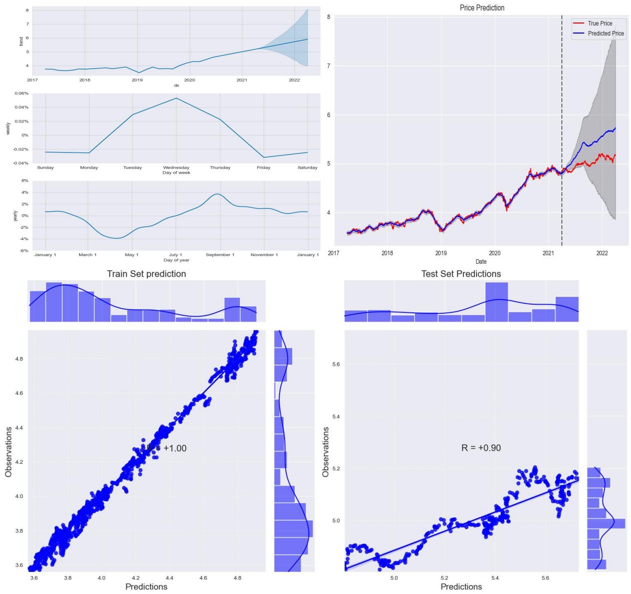


Fig.26 Results after adding additional data sources.
Top left: The trend and seasonality of AAPL
Top right: Actual and predicted prices in logarithmic scale
Bottom left:joint plots for Training set
bottom right:joint plots for Testing set

Quantitative indicators of model performance:
RMSE for one month: 0.044<0.071
Correlation: +0.90>0.85
better than first model

We can observe that the predictive performance with multiple data sources is slightly better than with a single data source. However, over a longer period (one-year forecasting period), there is a pretty greater bias than with a single data source. I need to use cross-validation to determine whether this is a coincidence or a real issue.

7.3: Evaluation metrics implementation

I used the cross_validation tool built into Facebook Prophet to perform cross-validation on two models. Based on the recommendations from the paper[17], I chose the parameters “initial='1095 days', period='15 days', horizon = '30 days'”, and the results are shown in **Table1**.

I used a total of six evaluation metrics to assess the results after cross-validation.:MSE(MSE (Mean Squared Error),
RMSE(Root Mean Squared Error),
MAE(Mean Absolute Error),
MAPE(Mean Absolute Percentage Error),
MDAPE(Median Absolute Percentage Error),
SMAPE(Symmetric Mean Absolute Percentage Error).The introduction of these evaluation metrics can be find in [17].It can be seen from **Table 1** that all the evaluation metrics have improved when we use multiple data sources.Simply put, over a 30-day forecasting period, the average difference between the forecasted values and the actual values is about 1.6% (MAPE*100%).

8. CONCLUSION

In this project, I acquired Apple's stock price data (AAPL) along with auxiliary data, performing storage, preprocessing, visualization, and exploratory data analysis (EDA). These datasets were used to train time series forecasting models, which were then evaluated in terms of data sources and model performance. The research identified a relatively complex, nonlinear relationship between AAPL stock prices and economic indicators like DFF and T10Y2Y, particularly noting that T10Y2Y could serve as an indicator for predicting downturns in AAPL stock prices. The findings indicated that models incorporating multiple data sources slightly outperformed those with a single source, achieving better forecasting accuracy. This has practical significance for guiding investors towards more informed stock investment decisions and lays the groundwork for automated trading software development.

However, the study has limitations. Firstly, as a smartphone manufacturer, Apple's perpetual exponential growth appears unrealistic due to constraints like Gartner's growth curve and market capacity. To enhance forecasting accuracy, incorporating data on Apple's sales volumes, profits, market share, market growth, and the stock growth patterns of similar companies at comparable stages could offer valuable insights for predicting Apple's stock price. Due to space constraints, these aspects remain to be analyzed further.

Secondly, the study did not compare the performance of different time series models. In semi-strong efficient markets, identifying and leveraging factors that improve stock price prediction is challenging, demanding more from the models' learning capabilities and feature engineering.

Lastly, while using the Prophet model, I assumed a linear growth (growth = linear), implying that the actual stock prices() should be close to an exponential function(A linear function becomes an exponential function after undergoing an exponential transformation.). This assumption is problematic because sales data growth has limits, suggesting that Apple's stock price should exhibit behavior a bit close to a logistics function. Consequently, models without this consideration might overestimate Apple's stock price in the long run.

8. REFERENCES

[1] Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, 38, 34–105.

[2] Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17 , 59 – 82.

[3] Shiller, Robert J. "Irrational exuberance: Revised and expanded third edition." *Irrational exuberance*. Princeton university press, 2015.

[4] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

[5] Alpaydin, E. (2014). Introduction to machine learning. MIT press.

[6] Matsunaga, D., Suzumura, T., & Takahashi, T. (2019). Exploring graph neural networks for stock market predictions with rolling window analysis. arXiv preprint arXiv:1909.10660 , .

[7] Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLOS ONE, 12 , 1 – 24.

[8] Lu, Wenjie, et al. "A CNN-BiLSTM-AM method for stock price prediction." *Neural Computing and Applications* 33.10 (2021): 4741-4753.

[9] Htun, Htet Htet, Michael Biehl, and Nicolai Petkov. "Survey of feature selection and extraction techniques for stock market prediction." Financial Innovation 9.1 (2023): 26.

[10] Sezer, Omer Berat, and Ahmet Murat Ozbayoglu. "Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach." Applied Soft Computing 70 (2018): 525-538.

[11] Yu, Xinli, et al. "Temporal Data Meets LLM--Explainable Financial Time Series Forecasting." arXiv preprint arXiv:2306.11025 (2023).

[12] Lee, Unro. "Another Look at the Predictive Power of the Yield Spread: New Evidence." *Journal of Accounting and Finance* 21.2 (2021).

[13] Bodie, Zvi, and Alex Kane. "Investments." (2020).

[14] Nison, Steve. *Japanese candlestick charting techniques: a contemporary guide to the ancient investment techniques of the Far East.* Penguin, 2001.

[15] Murphy, John J. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications.* Penguin, 1999.

[16] Taylor, S.J. and Letham, B., 2018. Forecasting at scale. *The American Statistician*, 72(1), pp.37-45.

[17] De Gooijer, J. G. & Hyndman, R. J. (2006), ‘25 years of time series forecasting ’ , *International Journal of Forecasting* 22(3), 443 – 473.