A comparative Study of Machine Learning forecasting Techniques in US Stock Market

Ali Abouyahia

September 2, 2024

School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
2023-24

Abstract

The purpose of this study is applying various machine learning algorithms to stock prices in attempt of predicting future prices.To achieve this goal, four major US companies were selected (JP Morgan Chase & Co, Apple Inc, Amazon.com Inc and BlackRock Inc) with their data collected at both daily and weekly level over the course of 23 years from Yahoo Finance.XGBoost,Long Short Term Memory Networks and Convolutional Neural Networks are all considered powerful algorithms to handle such problems and therefore will be use to conduct this research and investigate whether or not time granularity does affect US stock market forecasting in this particular cases.The Open,High,Low prices are considered as features of predictions.It is shown that weekly predictions are very challenging and not recommended at this time.On the other hand, neural networks shows promising results in daily data and conserve their dominance in the forecasting context.

Keywords: Long Short Term Memory · Convolutional Neural Networks · XGBoost · Forecasting · US Stock Market · Data Analysis

Acknowledgments

I would like to thank my supervisor Dr Qamar Natsheh that made this project possible with her assistance and guidance.

Abbreviations

- CNNs : Convolutional Neural Networks
- ANNs : Artificial Neural Networks
- RNNs : Recurrent Neural Networks
- LSTM : Long-Short Term Memory
- IPO : Initial Public Offering
- RoC : Rate of Change
- RSI : Relative Strength index
- SVR : Support Vector Regression
- ReLU :Rectified Linear Unit
- ARIMA :Auto Regressive Integrated Moving Average
- GRU: Gated recurrent units.
- GBDT :Gradient Boosting Decision Trees
- RMSE : Root Mean Squared Error
- MAE : Mean Absolute Error
- MAPE : Mean Absolute Percentage Error
- EST : Eastern Standard Time

List of Figures

List of Tables

Contents

1 Introduction

The stock market has always been a fascinating topic worldwide.As simple as it can be explained, there is nothing simple about investing in it.Each country has its own stock exchange platform with a list of companies however this report will focus on the US stock market which represents the double of its country's entire GDP.For decades, the question that has eluded everyone involve in stock market investment is simple : What will be the share price tomorrow ? With this, followed years of research and analysis which lead to two conclusions.On one hand, a fundamental analysis using the company's technique and fundamental information about their performances such as expenses, growth rates and revenue could help forecast the share price.On the other hand technical analysis which focuses on historical data can also be used to estimate future prices.In the past,such investigation could only be done by financial advisors.However, with the technology's progression, analysts have begun undertaking this ask with the help of machine learning.Using four stocks (Apple,Amazon,JP Morgan Chase and BlackRock) taken from Yahoo Finance from 2000-2023 inclusive this study will investigate the accuracy of neural Networks techniques (LSTM,CNNs) and ensemble learning techniques (XGBoost) using several metrics on daily and weekly data granularity.The choice behind these particular stocks is due to the fact that they are one of the most popular and powerful companies of the US stock exchange and therefore could potentially affect the performances of the S&P500 index or even the general trend of the market.According to Kroeller et Al [2] a majority of research in this field consist of daily data analysis and uses ANNs and lagged index features in their coding. This Survey helps us understand that the most optimal way to approach stock market predictions is to follow these criteria but it does not mean that further investigation into weekly data could not be useful especially to reduce day to day trading as forecasting an accurate weekly share price could lead to a more relaxed investment schedule.To approach this project, an extensive data analysis is to be performed to understand the stocks' behavior followed by a deep studies of machine learning techniques involved.Moreover, as this is a rich topic a thorough literature review was necessary to grasp the full concept of the field.Finally, the project will attempt to determine the best machine learning technique on the mentioned stocks and identify whether or not working with weekly data could lead to high accuracy in forecasting the close share price of each company.

1.1 Research Motivation

From the early 20th century, Charles H. Dow, the co-founder of Dow Jones & Company introduced what he called the Dow theory.This theory was among the first systematic approach in stock market analysis.Since then, the motivation to predict the stock market only grew bigger worldwide and hence the purpose of this project: Attempt to analyse which machine learning algorithm present today would be more suitable for the task.

1.2 Limitations

Due to its old history several factors are affecting the stock market,factors that until today remain unclear to the general public.Naturally, the sophistication of the stock market is evolving throughout time and therefore many new measures are implemented by corporations to encourage new investors such as using the share split.This split can occurs at anytime and is only announced a few weeks prior.

Economic situations (i.e crisis) are major factors on how the stock market behaves as usually the stock market mirrors how the country is doing economically.Sudden crisis or "bubble burst" could lead to the share price of certain companies to plummet or skyrocket depending on the factor(s) affecting said company.For instance, during the 2020 pandemic Amazon and Zoom have seen there stock skyrockets due to several lock downs happening worldwide with people relying a lot on remote work and deliveries while most of other companies have seen their share price go down.Due to these aspects, several stocks reach a high level of volatility due to their history and factors they encountered throughout time.Such stocks are labeled highly volatile and create an additional challenge in forecasting.

Finally,US companies only releases to the public their performance results at the end of the quarter that means that for 3 months it is unknown how the company is performing and therefore any effect that it could have on their stock is unknown.It can happen that some individuals have illegal access to this information and use it at their own benefit.This is known as *Insider Trading*.

1.3 Aim and Objectives

The purpose of this study is to identify at first on a daily level data which algorithm is more suitable for stock market closing share price predictions in the US exchange.Followed by this, the same stock at a weekly level data will be investigated using the same machine learning technique to identify whether or not weekly data would significantly affect the forecast's accuracy.Namely, how does time granularity (Daily/Weekly) affect the accuracy of Machine Learning's share closing price predictions in the US stock market on Yahoo Finance datasets spanning 2000-2023 for JP Morgan, Apple, Amazon and BlackRock?To answer this, the following objectives are put in place :

- Define financial background: Some key aspects of finance are required for this project following the history of the stock market and how does it function.(Completed in Section 2)
- Data Collection was made using Yahoo Finance while collecting data from four majors US companies.

- Data cleaning to be performed using SQL queries in order to get rid of unnecessary elements (if any) such as null values, duplicated rows or even mislabeled columns and/or rows.
- Exploratory Data Analysis (EDA) is a paramount process in any data science project. This will be conducted at a numerical level using SQL and a visual level using Python in order to detect patterns or irregularities in the data.
- Study of the required Machine Learning processes.Upon reviewing it seems that Convolutional Neural Networks (CNNs), XGBoost(Extreme Gradient Boosting) as well as Long Short-Term Memory (LSTM) Networks are the most adequate to undertake.
- Results and Interpretation of the above algorithms' outputs alongside an investigation of their accuracy using RMSE,MAPE and R-Squared.

1.4 Thesis Overview

Figure 1.1 Shows the entire project structure.The green represents the data section including the ticker symbols of each stock studies followed by the data analysis part with cleaning and pre-processing.In this particular case only duplicated values and null values were investigated as the data collected from Yahoo Finance is already "cleaned".The data analysis refers exactly to the Exploratory Data Analysis performed in sections 3.1.2 and 3.1.3.

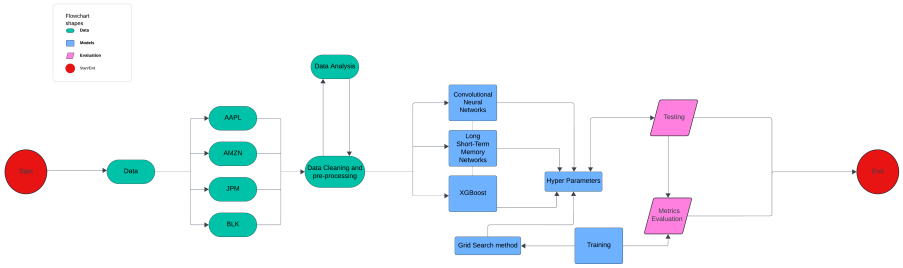

Figure 1.1: Fundamental concept of the proposed approach

1.5 Regulatory & Ethical Considerations

- *Legal Issues :* While using Yahoo Finance datasets for machine learning predictions, data licensing and compliance with financial regulations could lead to legal concerns.Making sure that the usage of the data adheres to copyrights laws and service terms is paramount as an unauthorised data use would most likely create legal disputes.Moreover, the use of machine for stock forecasting must always follow SEC regulations (as this project is about US stock market) especially when it comes to potential market

manipulation.At this scale, the data collected is only used for academic purposes and was available publicly.

- *Social Issues :* Having access to powerful tools (i.e. Machine learning algorithms) is not given to everyone.Therefore this already could constitute a social issue between the general public and engineers who would have such access on a daily basis. Moreover, said engineers could develop a complete dependence to such tools and therefore reduce human oversight. Investing in such manner could lead to problems within a company and potentially even increase market volatility if the trades are in significant amount.

- *Ethical Issues :* The use of machine learning in stock predictions raises some ethical questions regarding transparency, fairness and accountability. It is possible that these models lead to biases or even reach certain decisions without a clear explanation and therefore leading to some ethical dilemmas. It is critical to investigate how the deployment of these Machine Learning techniques could affect the US stock market, today due to their complexity it is very likely to make an impact and therefore some ethical checks are to be performed prior.

- *Professional Issues :* Analysts and data scientist should always take into account the reliability and accuracy of predictions based on time granularity when it comes to clients advisement or just making investment decisions for a corporation.It is their responsibility to make sure that predictions are consistent, transparent and scientists should be aware of the models' limitations.This would then ensure that forecasts are not misleading and used in a responsible manner during decision making as this the primary requirement in the financial sector.

1.6 Thesis Structure

Chapter 1 refers to the introduction section where the motivation ,limitations and objectives are defined followed by an extensive background and literature review in chapter 2.Chapter 3 and 4 corresponds to the data section and methodology respectively and to conclude chapters 5 and 6 shows results and conclusion as well as some potential future work related to this study.

2 Background and Literature Review

To understand the stock market and its analysis, several definitions and information along a comprehensive review of papers is necessary to grasp the full concept of this work.In this section all this will be investigate and detailed.

2.1 Definitions

- *A strike price* is the price in an option contract at which the underlying asset can be bought or sold.
- *Initial Public Offering* is a procedure where a private corporation first makes its shares available to the general public on a stock exchange, enabling investors to purchase and trade their stock.
- *The break even point* on a call option is the sum of the strike price and the premium. When you have a call option, you can calculate your profit or loss at any point by subtracting the current price from the break even point[3].
- *A share split* is a decision taken by a company's board to increase the number of outstanding shares in the company.New shares are issued to existing shareholders in a set proportion.
- *Market Capitalisation* represents the value of the company traded.
- *Insider Trading* refers to buying or selling a publicly traded company's stock by someone with non-public material information about that company.[4]
- *Shorting a stock :* is a investment strategy that refers to buying a stock in usually large amount and a low price and selling back for a profit quickly.Sometimes even within hours.

2.2 The Stock Market

As popular as this topic is a majority of people are not familiar with the stock market.Very correlated to the country's economy, the stock market is a place where people can own shares of companies at a certain value at the time of purchase.Later on if the value of the company increases so does the investment and vice versa.The purchase of stock is only possible if the company goes public on the market (i.e IPO definition 2.1).This is a very old concept, the first official stock exchange was the Amsterdam Stock exchange established in 1602 by the Dutch East India Company[5] being the first company to issue stocks and bonds to the public.However, throughout time, stock markets have developed all over the world and to each country their own.

The US stock market is considered to be the most popular today and the most complete with a lot of purchasing options for investors.People can buy shares but also invest in indexes which yield a lower but assured profit per year and also participate in the Initial Public Offering of a company.The IPO usually has a very low price that could spike in a few days which would make a lot of investor attracted to for shorting the stock.

Investing in the stock market became a full time job for a lot of people following the EST time zone.In fact, opening trading opens Monday to Friday from 9:30am to 4:30pm EST.In that time people and companies trade constantly monitoring anything that could make a stock go up or down.Which leads to the million dollar question : How will the stock perform today ? According to Warren Buffet[6], who is one of the greatest investor of all time "It is extremely challenging (if not impossible) to beat the stock market".In fact Buffet was so sure of himself that he made a million dollar bet and won against a big hedge fund company[7].

There exist two more types of purchases in the stock market, namely call and put options:

- A call option is a contract linked to a stock. it requires a fee payment for the contract called a premium.This then gives you the right to buy the stock at a fixed price, known as the strike price.This purchase can be made at any point until the contract's expires.It is not mandatory to execute the option. If the price of the stock increases well enough, then the buyer can execute it or sell the contract itself for a profit. If not it is wiser to let the contract expire and only lose the premium.

- A put option is a contract linked to a stock. it requires a fee payment for the contract called a premium.This then gives you the right to sell the stock at a fixed price, known as the strike price.This sell can be made at any point until the contract's expires. It is not mandatory to execute the option. If the price of the stock decreases well enough, then the buyer can sell the contract itself for a profit. If not it is wiser to let the contract expire.

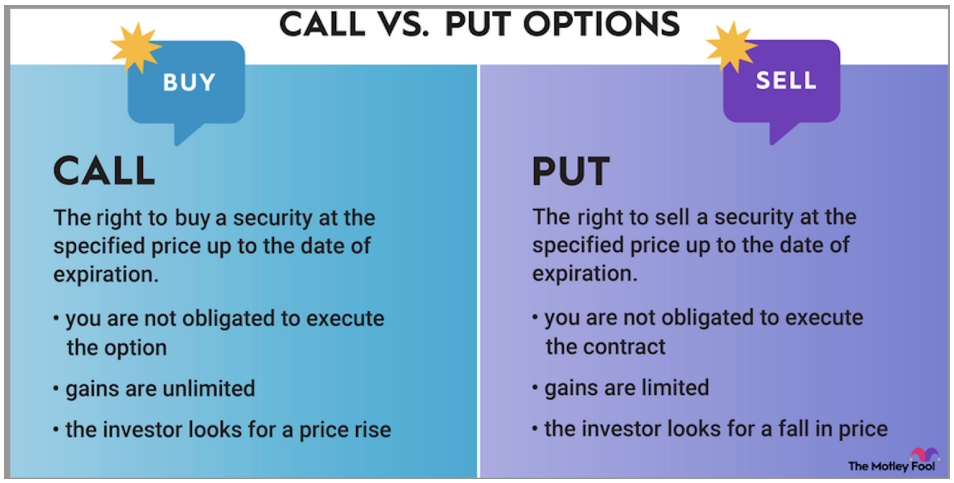

Figure 2.1: Summary of Call and Put Options

2.3 Related Work

The models studied in this paper by Kumar et Al.[8] are based on supervised learning techniques such as SVM,Random Forest, K-Nearest Neighbour Softmax and Naïve Bayes. Most importantly the key to this work is that technical indicators such as Moving Average for 10 and 50 days, Relative Strength index (RSI) and also RoC1 and RoC2 are taken into account while applying these algorithms.The data comes from different sources in this scenario , namely , Yahoo Finance, Quandl and NSE-India for the following stocks (Amazon, Cipla, Eicher, Beta and Bosch).The data has been split into a 70 : 30 ratio for training and testing (unlike the conventional 80:20 ratio).Kumar et Al. shows promising results for random forest algorithm when using large financial datasets and great results on small dataset using Naïve Bayes and then observed a decrease in accuracy when technical indicators are reduced highlights their importance.

Following the work of Chen[9] is essential to this study due to the similarity of the data. This paper concluded that SVR due to its comprehensive features has the most accurate outcomes and would make a perfect example to compare with ours.Moreover, it confirmed the fact that high volatility does affect predictions even when used with advanced models such as CNNs and that hybrid models improve accuracy.The reviewed material all consists on financial data taken from Yahoo Finance, which makes this very related to this project.In fact to conclude this, Chen attempted to train a model on a certain stock A and using that training to make the predictions on a stock B,showing that it is possible to do so while sacrificing some reasonable accuracy.This approach could be revolutionary for stocks with no data history.Moreover, combining two techniques to improve accuracy showed to be positive in this scenario however SVR still performed best even after combining two MLs due to a more comprehensive

set of features used in its implementation.

In Santosh A.S [10], the comparison of time series and deep learning models is investigated using ARIMA and GRU and studying the generalisability of the models in order to predict long term forecasts.After applying the models to a daily datasets of three major UK banks (Lloyds Bank Group , Barclays Bank Plc and Royal Bank of Scotland Plc) comprised of over 2800 data points along technical indicators that describes the movement of the stock (Similar to [8]) these are useful features for stock price predictions.It has been observed that the generalisability of deep learning models is greater than time series ones , especially in the banking sector. Santosh also states in his paper that further work is possible, for instance using the sliding window technique that trains deep learning models on time series models could be applied.

In the work of Yamada et Al., [11] a linear Kalman Filter is applied as a baseline for next day prediction on several stocks collected over a 10-year period from Yahoo Finance using several metric such as RMSE, R2 Scores and MAE.On the other hand, several layers of LSTM have been applied as well and while investigating, it has been concluded that volatility is a major factor in prediction.In fact while taking the Tesla stock (TSLA) as a volatile stock and Microsoft (MSFT) as a non volatile stock, the conclusion drawn was that a simple Kalman filter has no chance in forecasting when it comes to high volatility.The important part of this work is that stocks can be classified into different types (volatile and non volatile) and this could automate portfolio generation for a target return rate.

A very bright paper, written by Shahi et Al [12] is used as a guidance for the knowledge gap.The incorporation of financial news in forecasting shows prominent results in accuracy.Lags(in number of days) are also included in this work to improve accuracy using historical data of the stocks.Using LSTM and GRU, it has been concluded that except for training time, LSTM with news sentiment shows the best results.Unfortunately, the data comes from studying the Nepalese stock market and collected using a web crawler makes the comparison with other work non accurate.Moreover, another dataset type is used to incorporate financial news which makes the comparison even more challenging and it is important to note that the Nepalese stock exchange runs from Sunday to Thursday unlike the United States' Monday to Friday schedule.

To support the work of [12] J.Patel et Al. wrote a paper regarding the Indian stock market forecast.Investigating similar methods in different markets shows us that these machine learning techniques are powerful enough to adapt to the data given.In this work,J. Patel uses ANNs, SVM, Random Forest and Naïve Bayes with 10 years of data structured as open, high, low and close.After identifying that stock market data is considered to be non-stationary but rather made of trends, cycles and/or random walks at a particular time.The model will then first identify the pattern and follow it for that given year to achieve a

trend prediction.Finally, in this work the use of day lags and technical parameters is also taken into account to increase accuracy however the main focus was to first identify the up or down trend (also referred to as trend deterministic data preparation layer) before the stock share price.It appears that following their procedure Naïve Bayes using multivariate Bernoulli process gives the best outcome.This work could be very helpful in terms of contract trading and also helpful to this project as it is expected to identity trends only with weekly data.Identifying the close share price at weekly level is conjectured to be challenging.

Another interesting comparative study has been published by Ashfaq et Al.[13] This time focusing on machine learning regressors such as XGBoost which is used in this project and thus why it is relevant to investigate it.The data is collected from `https://www.nasdaq.com` using python's library (beautifulsoup4) for data scrapping.The target is the opening price of ten randomly selected stock within the NASDAQ index.This represents a portfolios of usually powerful companies, people can choose to invest in the entire index rather than just a single stock.This is to increase the chance of a positive return on investment as usually an index always grows.In fact, several big investors such as Warren Buffet always recommend the general public to do so as it is unlikely even impossible that all the companies of the same index crash.To perform such tasks, the forecast was done with several regression algorithms and metrics such as R2 Score and MSE were measure for the training and testing set.It has been concluded that the margin of increase or decrease in price is not found by determining the path of the share price and that historical data plays an important role in forecasting.This is the reason why no one has studied predictions on weekly data as there is always too much unknown between each data point.Finally, a flaw of this paper is that there is no visual representation whatsoever and the data results is focused on one table making it very overwhelming to read.

2.4 Knowledge Gap

Upon reviewing, there has not been a study on different time granularity as well as identifying economical events that occurred to understand the data's evolution. Potentially, working on two different time scales as well as including a research on economics could reinforce this project and the share price's predictions.A paper written by [12] comes close, however the investigation is done on the Nepalese Stock Market and not the US which makes the comparison unfeasible.Potentially being able to make a weekly accurate forecast could facilitate portfolio return.

3 Data

The exploratory data analysis is a paramount part of any data science project.It enhances the data for the reader and provides more details on the data and its source.This section will do just that.

3.1 Data Collection and Pre-processing

After conducting an extensive literature review, the data collection from Yahoo Finance seemed the more reliable and efficient.It will also be helpful as comparing results with other papers must have experiment conducted on the exact same data otherwise results will be inconclusive.

The data collected is from four major US corporations (Financial and Tech) these companies are paramount among investors on a daily basis and usually 'control' the performance of the market therefore it would be very beneficial to understand their share price behavior as it could help investors that also put their savings into indexes such as S&P 500 or NASDAQ.

3.1.1 Datasets

The Following datasets are all made of the following columns : Date, open, high, low, close ,Adj close and Volume.

- ***Apple Inc*** daily and weekly data stock price [14]
- ***JP Morgan Chase & Co*** daily and weekly data stock price [15]
- ***BlackRock Inc*** daily and weekly data stock price [16]
- ***Amazon.com Inc*** daily and weekly data stock price [17]

3.1.2 EDA (Python)

In this section the focus on EDA will be primarily visual.The main focus will be the daily closing price as well as the volume of shares traded in each company.The remaining metrics provided will also be mentioned however as the daily share price does not fluctuate it will be inefficient to also analyse them.As presented in section 3.1.1 the following study will follow the same order as the enumerated datasets, starting with Apple.To begin with a visual of the stock is necessary to identify any strange behaviour

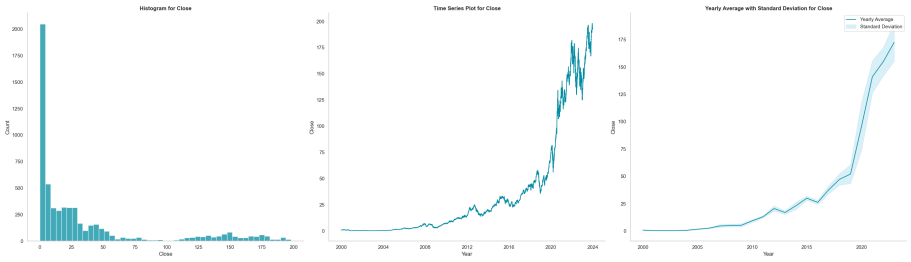

Figure 3.1: Plots for Apple's Daily Closing Share price

From left to right the histogram represents the number of times the stock opened at a certain price from 0 to 200 dollars, followed by the curve of the entire share price evolution from the moment the company went public, and finally a general trend.
From this we notice that from 100 dollars onwards there has not been many instances where the closing price was at that level.This is due to the splits of the company shares.This often occurs when the price of the share is to high to encourage small investors.Generally we can see that yearly apple almost always grows except in 2012 and 2016 due to stagnant sales at the time.Moreover, recently between 2022 and 2023 post COVID Apple share have taken a hit due to a lack of innovation in their products combined with high priced items.
As the data also contains the number of shares traded per day, it would be beneficial to identify if there is any correlation (visually) with the closing price.

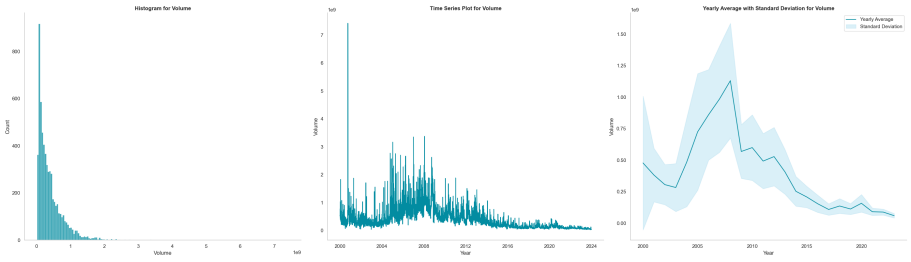

Figure 3.2: Apple's Daily share's Volume analysis

Similar to figure 3.1 the histogram represents the number of times a certain amount of shares has been traded followed by the time series plot which shows more details and changes followed by the yearly average curve that shows the yearly trend. The noticeable crash in this case that is shown to be around 2007 is due to the financial crisis.

According to figure 3.3,plots shows the original data with its corresponding trend, it can be observed that the trend is relatively constant at first and then increasing progressively.Usually, tech companies follow more or less the same

trend and are also affected by external factors and economical crisis.The full seasonal plots indicates that there is no seasonality in the data which could help us figure out a regular pattern and finally the residual plot seems steady matching the time series and actual residuals appear towards the end when there is a lot of fluctuation in the data.

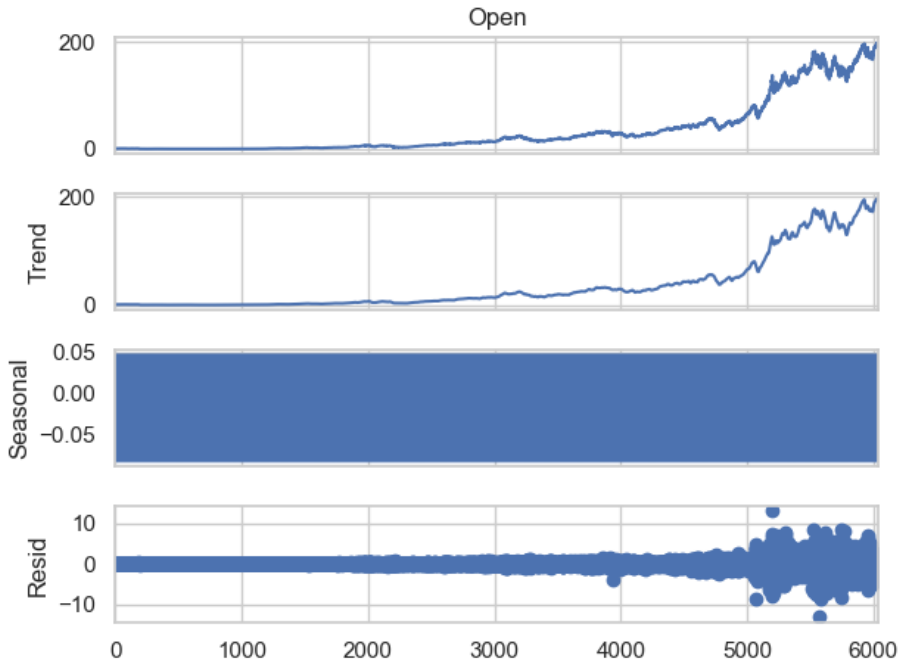

Figure 3.3: Analysis plots AAPL Daily

Regarding financial corporation, the investigation of BlackRock and JP Morgan has been conducted in this project.Financial institutions are usually more volatile than tech companies as seen in figure 3.4 the central graph represents the time series with several drops throughout the entire period, once again drops are matching several crisis that happened as well as other external factors.

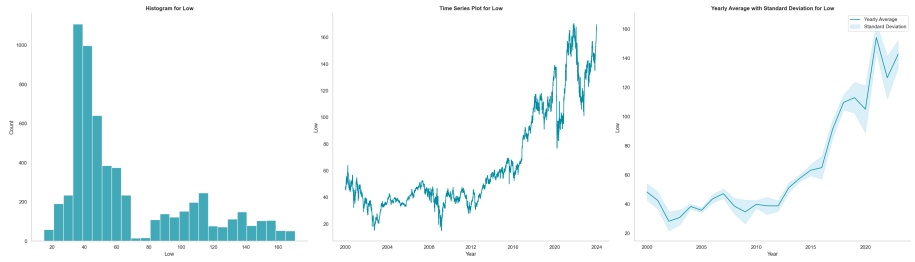

Figure 3.4: JP Morgan's Daily closing share price analysis

When it comes to amount of shares traded within JPM we can graphically identify a negative correlation suggesting that the more the stock increases the less shares are being traded.This could be due to the high price reached and/or external factor making people not willing to put their savings or just invest into such volatile stock.

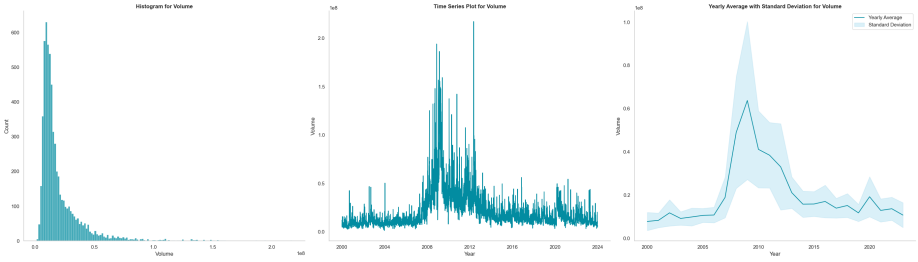

Figure 3.5: JP Morgan's Daily Volume analysis

Finally, using the plots shown in figure 3.6 it has been identified that as expected there is no seasonality in this stock and the trend showing to be relatively upwards with a bit more fluctuation.As supported by the residuals,the data is lot more volatile than AAPL or AMZN showing heavy disturbance in the data.This would help conjecture than it is more likely that the results in forecasting closing share price for financial institution would be more challenging than technological ones.

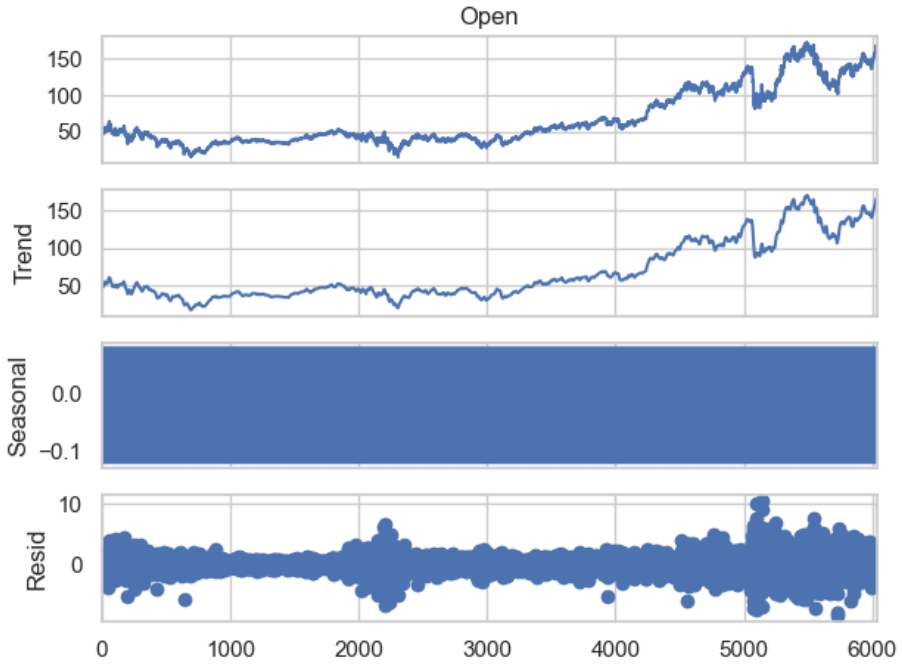

Figure 3.6: Analysis plots JPM Daily

To briefly conclude, when conducting a forecasting project a data analysis at a visual level is crucial to understand the trend taken by the data.

3.1.3 EDA (SQL)

Following the visual EDA section 3.1.2, a more accurate investigation is done using SQL to explore the numerical aspect of data analysis.At first a general overview gives us an idea on how the data behaves.Correlation between the closing price and the volume of share traded is also investigated.SQL helps us identify missing data, null values, duplicated values or any issue that could lead to corrupt data and therefore corrupted results in this work.From the below tables 3.1 and 3.2, a general overview is given showing that all datasets contain the same number of data points at each granularity.It can also be observed that the amount of data points dropped drastically between weekly and daily and thus could play an important role in the predictions accuracy.

Table 3.1: Summary Statistics for Daily Data

Ticker Code	# Data Points	Avg Closing Price	Std Closing Price	Max Closing Price	Min Closing Price	25th Percentile	Median	75th Percentile
AAPL	1252	35.86	51.09	197.57	0.23	2.19	14.46	41.07
AMZN	1252	37.87	51.47	185.97	0.30	2.26	10.27	59.84
BLK	1252	290.30	230.31	971.49	15.88	108.80	207.35	421.66
JPM	1252	66.84	38.56	171.78	15.93	39.15	48.57	95.60

Table 3.2: Summary Statistics Weekly Data

Ticker Code	# Data Points	Avg Closing Price	Std Closing Price	Max Closing Price	Min Closing Price	25th Percentile	Median	75th Percentile
AAPL	1252	35.86	51.09	197.57	0.23	2.19	14.46	41.07
AMZN	1252	37.87	51.47	185.97	0.30	2.26	10.27	59.84
BLK	1252	290.30	230.31	971.49	15.88	108.80	207.35	421.66
JPM	1252	66.84	38.56	171.78	15.93	39.15	48.57	95.60

Moving on to another aspect of this analysis part : Sudden Change detection. As mentioned previously, the stock market is very volatile and is affected by several outside factors.It is important to find out when these sudden changes occurred in the past to see if they correlate with real life events.For instance, most stocks have seen a drop in their share price during the COVID-19 pandemic and a recovery later around 2021.

A few other drops did happen during the big financial crisis of 2007-2008.These are mostly affecting BlackRock and J.P Morgan.The following table 3.3 shows the sudden change throughout the whole data period. At the exception of Amazon, all companies have been affected by the pandemic and show a drastic drop in their closing prices as well as a good recovery in 2022.Especially BlackRock due to the increase in their asset managements that of today are close to 10 trillions of dollars worth.Moreover, table 3.4 shows that all the peaks for all studied stocks occurred in 2021 at the exception of Apple which happened later in time.

Table 3.3: Sudden Change Detection

Ticker Code	Max Change	Max Change Date	Min Change	Min Change Date
AAPL	12	10/11/2022	-10.5	03/09/2020
JPM	15.860001	13/03/2020	-15.6	16/03/2020
AMZN	18.793991	04/02/2022	-20.3	29/04/2022
BLK	90.190003	10/11/2022	-56.5	16/03/2020

Table 3.4: Max and Min Data

Ticker Code	Min Closing Price	Date Min Closing Price	Max Closing Price	Date Max Closing Price
AAPL	0.234286	17/04/2003	198.110001	14/12/2023
AMZN	0.2985	28/09/2001	186.570496	08/07/2021
BLK	15.75	05/01/2000	971.489999	12/11/2021
JPM	15.45	09/10/2002	171.779999	22/10/2021

The data collected also provides us with the volume of shares traded on a given day,using this a correlation between the closing price and said volume had to be investigated and yielded to mostly a negative correlation as shown in table 3.1.3

Table 3.5: Closing Share and Volume Correlation in Daily Data

Ticker Code	Correlation Closing Price-Volume	Kurtosis Closing Price
AAPL	-0.453	1.256
JPM	-0.266	1.417
AMZN	-0.300	1.609
BLK	0.251	1.085

This shows that as the closing price increases the volume of shares traded tends to decrease.This makes sense as most people either sell their shares or wait for a decrease in price before entering the market.Unfortunately, due to such a little correlation this information is not useful to help with stock market predictions.Using Volume as one of the features in forecasting does not yield a higher accuracy.

Table 3.6: Volatility Comparison Between Daily and Weekly Data

	Daily	Weekly
Ticker Symbol	Total Volatility	Total Volatility
AAPL	8.66	18.29
AMZN	10.72	22.64
BLK	7.51	15.74
JPM	7.85	16.77

4 Methodology

An overview of machine learning techniques used in this project are described in this section as well as a detailed explanation of the steps taken to answer the research question.

4.1 Experimental Setup

The following methodology will use several datasets regarding the US stock market spanning from 2000-2023 inclusive.The EDA has been performed on the visualisation aspect (i.e using Python) and on the numerical aspect (i.e using SQL). Moreover, as the dataset is quite extensive and consists of different stocks (i.e different companies) it will be split into tables. This is to ensure data quality and visualisation when coding.

JP Morgan Chase & Co, Apple Inc, Amazon.com Inc and BlackRock Inc will be the stocks studied here with the following ticker symbols JPM, AAPL, AMZN and BLK respectively.Convolutional Neural Networks (CNNs), XGBoost (Extreme Gradient Boosting) as well as Long Short-Term Memory (LSTM) Networks will be studied in this project and help us understand the behaviour of the US stock market as well as potentially forecast it at a weekly and daily level.Finally, comparing these results at different time scales will address the knowledge gap and potentially improve accuracy on shares price predictions at a daily level. Additionally, comparing this project's best results with those in other paper(s) used in the literature review would help identify any potential improvements to be made.The main objective is to see whether or not weekly data can give us any insights and/or lead to accurate forecast similar or close to daily data on stock market.A general rule states that : The more the data points the more the accuracy of the Machine learning techniques that are applied.Potentially a fair closeness could be achieve here using weekly granularity.

4.1.1 Resources and Libraries

- All resources used in this project are personal and no university resources are required.
- Apple MacBook Pro M1 Max Chip, 32GB RAM,10 Cores.
- OS : MacOs Sonoma 14.5
- IDE Visual Studio Code and DataGrip
- Python Libraries : Tensorflow,Numpy,Pandas,SciKitLearn

4.1.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks have gained an important popularity in data visualisation.Considered to be a class of deep neural networks, CNNs are used

for classification (numerical or imaging), object detection and segmentation.Moreover, CNNs can also be used for predictive problems using time series data using their unique ability to learn spatial hierarchies of features.This allows the adaptation onto various predictive modelling tasks.
For this project, the focus will be primary on the use of CNNs in time series predictions.In the later, the data is represented as sequences which consists of a defined number of time steps called *window size* from the original data.
Usually, in data analysis the use of filters is mandatory.For CNNs, the same process is applied throughout layers.This process captures temporal dependencies and a 1D convolutions are used in this project.The filters will be able to detect local patterns like trends or seasonality.
Considering :

- I the time series
- t as time
- K the kernel (or filter)

The following convolution operation is written as :

$$(I * K)(t) = \sum_i I(t+i) \cdot K(i) \tag{1}$$

It is important to also highlight the fact that CNNs uses an activation functions that allows the learning of more complex patterns by introducing non-linearity (ReLU is commonly used)

$$f(x) = max(0, x) \tag{2}$$

As seen in figure 4.1 the process of CNNs consists of the following supplementary steps :

- Max pooling operation : In order to make the model more efficient pooling layers in CNNs reduces the dimension of the feature maps.The model also becomes " immune " to small variations in the input and is computed following the formula where k is considered the pooling size.

$$P(t) = max\left\{I(t+i) : i \in [0, k]\right\} \tag{3}$$

- Finally, this process repeats several times which leads to flatting the output and passing it into fully connected layers defined bu the following where x is the input vector , f the activation function, b the bias vector and W the weight matrix

$$y = f(Wx + b) \tag{4}$$

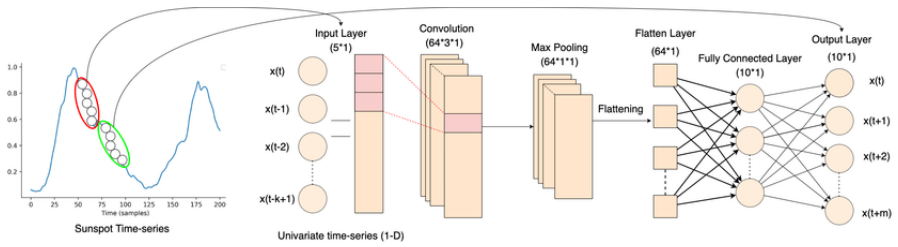

Figure 4.1: One-dimensional CNN for time series data.[1]

To summarise, it has been established that Convolutional Neural Networks can be utilised to perform time series predictions by using their ability to embrace patterns in sequential data.However, it is important to make sure that the data is prepared as sequence and that the model is properly trained using the correct optimiser and loss functions.CNNs can be very powerful in various predictive tasks in different fields provided that the steps described are properly followed.

4.1.3 Long Short-Term Memory Networks (LSTM)

LSTM networks acts as a solution to the vanishing gradient problem [18] which causes long term dependency in RNNs.Because of the feedback connections in LSTM they differentiate themselves with more traditional neural networks (feed forward) it allows them to process the entirety of time series data rather than considering each input in the sequence independently.What LSTM does is conserve information about previous data points in order to process the future data points.This leads to LSTM being very efficient when it comes to time series datasets.
Lots of models as mentioned before are affect by the long-term dependency problems.LSTM does not only rely on previous value to predict future ones but learns a pattern on a previously pre-defined period as well.This is particularly useful when there is a pattern that is separated by long time periods.
In a nutshell, Long short term memory networks' outputs depends on three elements[19]:

- The current long-term memory of the network called *Cell State*.
- The output at the previous point in time called *Hidden State*.
- The data input at the current time t.

4.1.4 Extreme Gradient Boosting (XGBoost)

Released first in 2014, XGBoost is a gradient boosting algorithm which builds model sequentially.It is considered to be a very powerful machine learning technique.XGBoost is particularly efficient when it comes to classification and regressions tasks because of its high performance, flexibility and ability to process large

datasets.It considered the leading machine learning library in those tasks.To form the strong predictive model XGBoost combines decision trees.[20].More accurately, it uses parallel tree boosting.There are a few things that one can do to improve machine learning techniques, such as using a lag.

By analysing a tree of if-then-else true/false feature questions and estimating the least amount of questions required to determine the likelihood of selecting the right choice, decision trees provide a model that predicts the label.Followed by a Gradient Boosting Decision Trees (GBDT) which is an ensemble learning algorithm close to random forest used for regression and classification.This results in a model consisting of multiple decision trees.Note that random forest uses bagging to build full decision trees in parallel from random bootstrap samples of the data set. The final forecast is an average of all of the decision tree predictions.

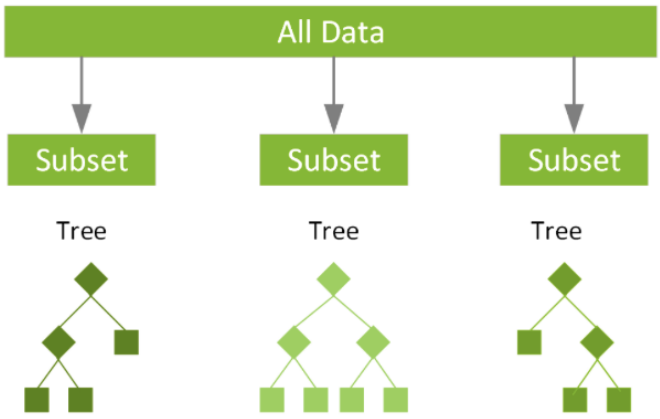

Figure 4.2: Decision trees built by GBDT and random forest

A lag refers to using previous observations as inputs to predict future values when it comes to time series modeling.For instance, a 10-day lag means that the model will use the values of a feature 10 days prior as input in order to predict the target variable. This is especially helpful in sequential data sets such as financial markets where past values have a big impact on the forecast.To incorporate such lag the data need to include current values plus the 10 previous days values.This can be defined mathematically as such :

$$Y_t = f(X_t + X_{t-1} + X_{t-2} + \ldots + X_{t-k}) + \epsilon \tag{5}$$

where

- X_t the feature value at time t

- X_{t-k} feature value k days prior.
- Y_t the target at time t
- k is the number of days.

As trees iterates, attempting to correct errors made by the previous tree iteration, XGBoost's main goal is to minimize a loss function $L(Y, \hat{Y})$ with Y being the actual value and $\hat{Y}$ the prediction.Considering the model to have m iteration.

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \alpha \cdot h_m(X_i) \tag{6}$$

where

- $\hat{y}_i^{(m)}$predicted value at the m-th iteration
- $h_m(X_i)$ decision tree at the m-th iteration
- α the learning rate

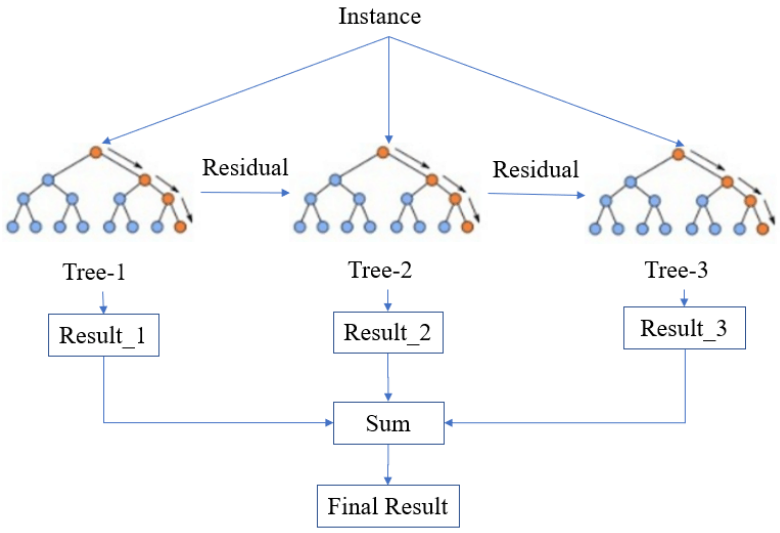

Figure 4.3: Simplified Structure of XGBoost process

4.1.5 Metrics

Upon outputting all the forecasts, metrics are needed in order to evaluate the performance of the models used.Thus in this paper the Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2) will help understand how well the models are performing.It is important to note that many factors are involved in this topic and therefore a highR^2 does not necessary mean that the model is fitted which is why evaluating a few metrics is always recommended.

- **Mean Absolute Percentage Error**

The Mean Absolute Percentage Error (MAPE) measures the accuracy's prediction in a predictive model by calculating the average absolute difference between predicted values and actual values $(\hat{y}_i)$,(y_i) respectively.This metric is given as a percentage by the following formula:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{7}$$

where:

- n is the number of observations.
- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.

Because it is a percentage, the use of MAPE is fairly intuitive.Say for instance the MAPE result is x %.Therefore it means that the predicted value are deviating by x % of the true ones.However, it is important to note that MAPE comes with disadvantages such as sensitivity to outliers.It can also be a misleading metrics when the data contains values close to or equal zero and will result in extremely high percentage of error.

- **Root Mean Squared Error (RMSE)**

Similar to MAPE, The Root Mean Squared Error (RMSE) is a used to calculate the differences between predicted $(\hat{y}_i)$ and actual values (y_i).RMSE is calculated as follow:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

RMSE gives an idea on how spread out residuals are and therefore indicating how concentrated is the data around the best fitting line.RMSE is basically a standard deviation of the errors in forecasting and unlike MAPE is sensitive to large errors because of the square of differences.On the other hand, it does not give a percentage measure and therefore a comparison with other MLs is more challenging.

- **Coefficient of Determination ($\mathbf{R^2}$)**

Consider an independent and non-independent variable R^2 measures the variance's proportion in the dependent variable that can be predicted from the independent variables as such :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

where:

- $\bar{y}$ is the mean of the actual values.
- y_i is the true value.
- $\hat{y}_i$ is the predicted value.

This metric is commonly used in a lot of field of computer science to indicated how good the predicted values actually fit the data with a perfect fit represent by a score of 1 while a negative score means that the model's prediction are worse than just calculating the average of the data.Finally,as all metrics there are pros and cons,in the R^2 case it is useful for a model comparison on the same dataset but does not account for complexity and can be misleading for non-linear data.

It is important to use several metrics in such project as the data's behavior can sometimes be considered chaotic and only one metrics would mean nothing on its own due to their individual weaknesses.

4.1.6 Project management

When undertaking a large project, having a Gantt Chart A.1 as project management tools is always a must.In this particular project, the chart was divided into several sections to ensure a logical schedule of the tasks throughout this three months.After conducting the proposal,the first steps were to look for dataset(s) and perform a extensive analysis.Followed by this was the applications of machine learning techniques found in the literature review and a first part of the report writing.This is to ensure to not leave all the writing for the end.The second part of the schedule was focused more on results and delivery method, a PowerPoint presentation was made in this case to summarise a part of the project and its results in a clear way.

5 Results and Discussion

Following the explanation of the methods this chapter will address the different outcomes of our data.Namely, how does the predictions behave in comparison to the actual data.All the predictions are done using the same features to ensure fairness and any lag parameter is the same among all predictions.

5.1 XGBoost Results

XGBoost's output follow the same parameter grid to ensure fairness between datasets.According to figures 5.1 and 5.2 it can be observed that there is less data points presents however the predictions trend seems to be similar (in red) while predicting the JP Morgan closed share price.With a volatility almost doubled in the weekly data 3.6 the results shown are promising for the weekly dataset as there is over five times more data points in the daily data.

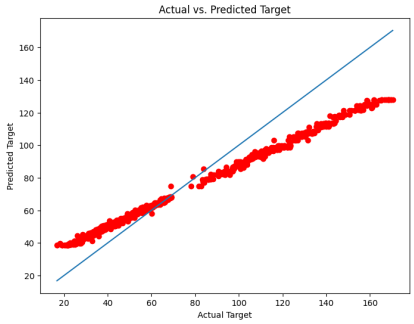

Figure 5.1: JPM Daily Data

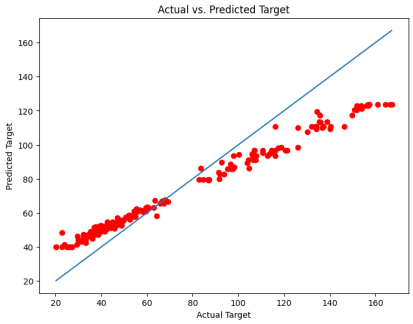

Figure 5.2: JPM Weekly Data

Unlike JPM, Apple stock seems to be slightly more volatile and therefore results in a less accurate graph as seen in figures 5.3 and 5.4.If compared between the weekly data on AAPL and JPM a bigger divergence is noticeable as predictions go on.Moreover, Apple has seen very sharp fluctuations towards the end of the data, this could be the reason why such divergence is visible.Similarly, in its daily data, the amount of data points near the true value seems to be slightly inferior to JPM.

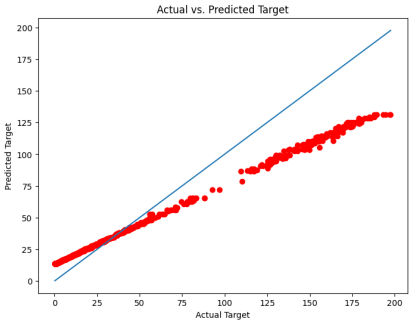

Figure 5.3: AAPL Daily Data

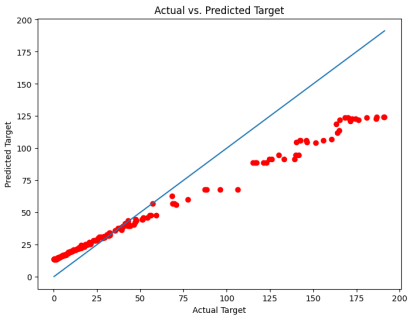

Figure 5.4: AAPL Weekly Data

Similarly, Amazon like Apple could both be considered as tech corporations both have high volatility and resulted in mediocre results especially at a weekly level Amazon has the highest volatility among all which results in lots of gaps as shown in 5.6.Potentially, due to its variety of services there could be more external factors that influenced the stock price of Amazon and created this high volatility.It is also important to recall the fact that unlike the other companies Amazon did thrive in the pandemic (recall figure 3.3 and 3.4.

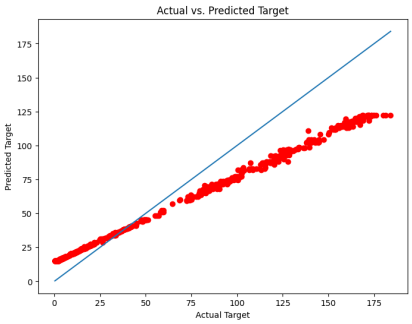

Figure 5.5: AMZN Daily Data

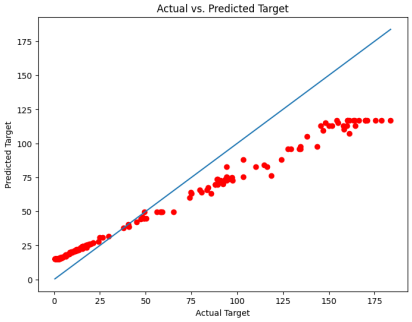

Figure 5.6: AMZN Weekly Data

Finally, the share price forecasting of BlackRock using XGBoost appears to be the most promising, potentially due to the power of the company always being in control of so many assets.Surprisingly, BlackRock has the lowest volatility on a daily granularity and shows a reasonable output in both figures 5.7 and 5.8.

From a first experiment, it seems that financial institution are "easier" to study compared to tech cor

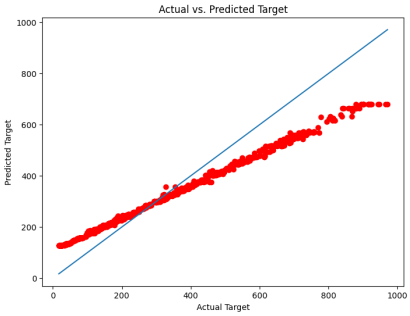

Figure 5.7: BLK Daily Data

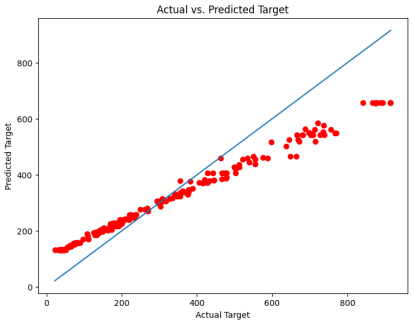

Figure 5.8: BLK Weekly Data

5.2 CNNs Results

Taking on a graphical approach here between weekly and daily data in the AAPL.While training the model on the following parameters(see Table A.It appears that the model fits the daily data quite well and follows the stock price pattern with a slight over-fitting.In most of the training on the daily data set the AAPL stock does not encounter many fluctuations.On the other hand due to the lack of data points, the weekly share price only manages to follow the trend of the true values and not the values themselves.This is due to the fact that the volatility in a weekly granularity is very high and therefore challenging to predict.

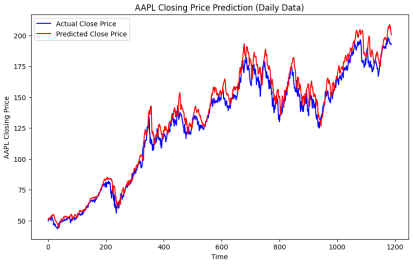

Figure 5.9: AAPL Daily Data

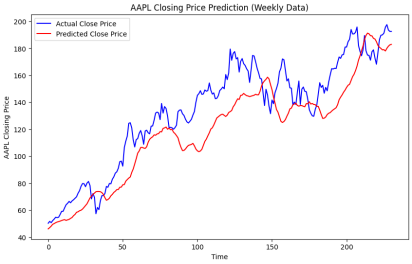

Figure 5.10: AAPL Weekly Data

A similar scenario is occurring for the case of JPM however in this case on a daily level it seems that the model is slightly under-fitting but still follows very well the trend and price forecasting.The weekly data however here is concerning.It is true that the weekly volatility of JPM is high as seen in 3.6 however after passing half of the prediction the model goes in opposite direction of the true trend.This could be tackled by an investigation on hyper parameters of CNNs or a different training way for this specific case.

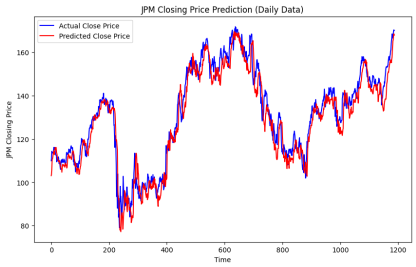

Figure 5.11: JPM Daily Data

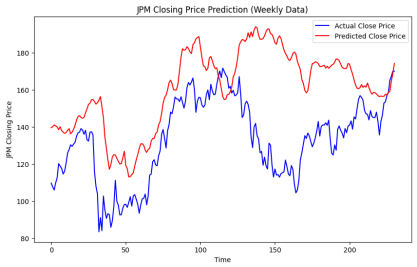

Figure 5.12: JPM Weekly Data

The model also fits the BLK daily data well and gives quite promising metrics as shown in 5.1.It also has the lowest volatility which could be a reason why at a weekly level the model's performance hold with reasonable metrics 5.2.The pattern is well followed even at the end where there appears to be lots of fluctuations.

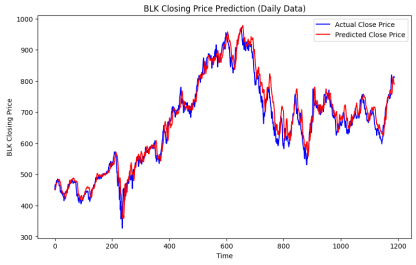

Figure 5.13: BLK Daily Data

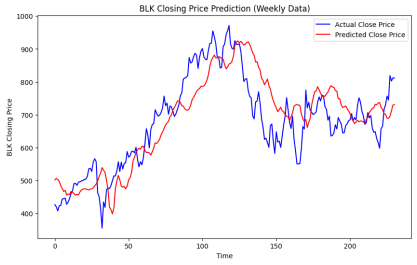

Figure 5.14: BLK Weekly Data

Similar to AAPL, in fact as both stocks have a similar history throughout the data it was expected to see a similar output in the predictions whether is at a daily or weekly level.On a daily level AMZN seems to be slightly under-fitting and on a weekly level the trends are well preserved

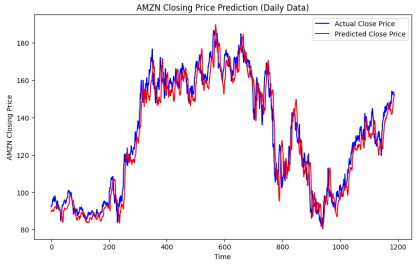

Figure 5.15: AMZN Daily Data

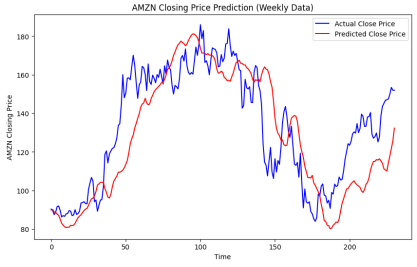

Figure 5.16: AMZN Weekly Data

To summarise, CNNs performs quite well in this case and aligns with the

work of [9].There always are improvement to be made in this field especially at the weekly level.Using CNNs would quite accurately give us the trend of the data in terms of predictions and this can be used if an individual wanted to invest in call or put options.However, the results are far off and it seems that a traditional weekly trading is not possible yet.In the following section, an investigation around LSTM will be conducted to investigate potential improvements.

5.3 LSTM Results

According to figures 5.17 and 5.18 there is clearly improvement on both sides (Daily and Weekly) but most importantly the weekly data behaves much more accurately in the case of AAPL compared to CNNs.

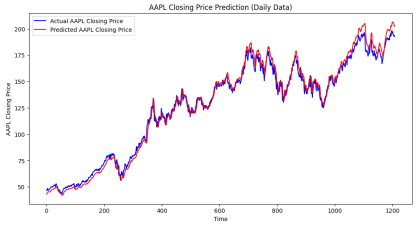

Figure 5.17: AAPL Daily Data

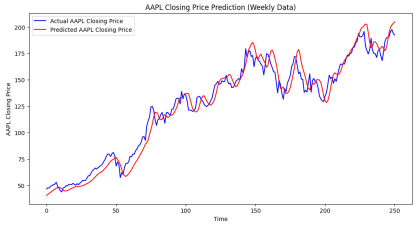

Figure 5.18: AAPL Weekly Data

Another case of significant improvement here in figure 5.19.The model seems to still be under-fitting however seems to give accurate close price forecast at a daily level.On the other hand, at a weekly level in figure 5.20 the model captures the trend with close to no error and even follows most of the fluctuations in the correct directions.

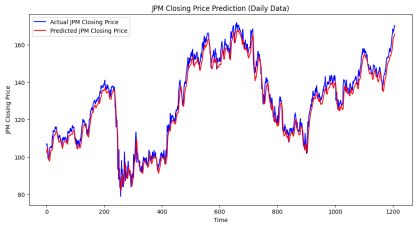

Figure 5.19: JPM Daily Data

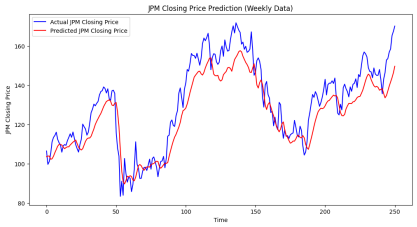

Figure 5.20: JPM Weekly Data

It appears clear now that the BLK stock shows the most promising results in this study.In this case the model seems to fit very well the data and at a weekly level a very high portion of the trend is predicted correctly.

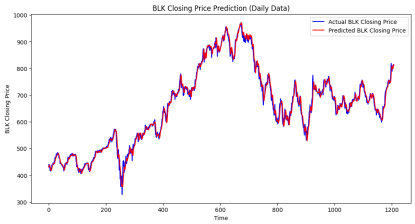

Figure 5.21: BLK Daily Data

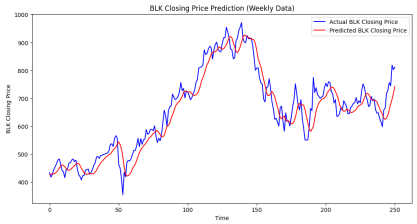

Figure 5.22: BLK Weekly Data

The AMZN dataset has the highest volatility.Although the model seems to be quite under-fitting in its predictions it still manages to follow the pattern at both granularity levels especially in the middle section where quite a lot of zig-zags patters are occurring.

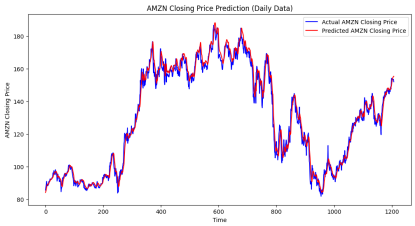

Figure 5.23: AMZN Daily Data

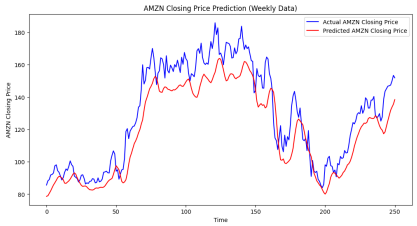

Figure 5.24: AMZN Weekly Data

Long short term memory networks are known to be dominant in forecasting.Several papers reviewed in this study focus on performing LSTM for accurate predictions and in this case shows promising pattern detection's in weekly data, much more accurate than CNNs.It also align with the research that LSTM are the most accurate model however their time computation is quite intensive.To run at the selected parameters (See Table A takes much longer than XGBoost and CNNs.

5.4 Comparative Analysis

Due to its complexity.A study on the stock market forecasting can not only rely on graphs or metrics but on both combined.From table 5.1, at a daily level the most accurate prediction are done on the BlackRock stock price with a MAPE of 1.82% for the daily dataset and 5.74% on the weekly dataset.This follows the path of several papers stating such accuracy level for LSTM such as Chen [9] while finding a MAPE as low as 0.67%.

It is also important to highlight the power of CNNs,which provided relatively fair results on a daily level but under performed at a weekly level until showing a complete incompatibility with a negative R^2 score for the weekly JPM data.

Finally, in this case XGBoost performances are not promising.This can be mostly due to high volatility and hyper parameters tuning in the model.Moreover, the MAPE metrics is prone to 0.In the case of XGBoost as denoted by the * the MAPE value was over 100% and therefore we used a capped MAPE to 100% to make sense of the results and be able to give a cross model/datasets comparison

Table 5.1: Performance summary on daily data

Model	**Stock**	**RMSE**	**MAPE**	**R^2**
LSTM	AAPL	4.82	3.35%	0.99
	AMZN	4.16	2.41%	0.98
	BLK	15.83	1.82%	0.99
	JPM	3.97	2.50%	0.97
CNN	AAPL	8.19	4.91%	0.96
	AMZN	7.57	4.54%	0.94
	BLK	28.71	3.29%	0.96
	JPM	5.92	3.55%	0.93
XGBoost	AAPL	19.04	60.37*%	0.86
	AMZN	18.81	66.87*%	0.86
	BLK	84.85	71.83%	0.86
	JPM	14.00	21.18%	0.86

Table 5.2: Performance summary on weekly data

Model	**Stock**	**RMSE**	**MAPE**	**R^2**
LSTM	AAPL	8.74	6.35 %	0.96
	AMZN	13.75	8.60%	0.79
	BLK	45.65	5.74%	0.90
	JPM	9.71	5.95%	0.80
CNN	AAPL	20.71	13.05%	0.73
	AMZN	19.40	12.11%	0.57
	BLK	81.12	10.35%	0.66
	JPM	34.45	23.23%	-1.50
XGBoost	AAPL	30.17	59.07*%	0.85
	AMZN	20.28	67.34*%	0.85
	BLK	90.31	61.39%	0.85
	JPM	14.73	21.72%	0.86

6 Conclusion and future work

Years have been spent on creating,modifying and tuning machine learning models to tackle this forecasting task and most the research tends towards neural networks as the most prominent forecasting technique in this field.This research confirms that, LSTM and CNNs are the best performing model at this stage whether the analysis is regarding trends predictions or price prediction.Moreover,it seems that in some cases it is possible to predict fairly accurately the **trend** of the close share price and thus allowing investors to trade weekly using neural networks algorithms and purchasing **contracts**.

However, due to several external factors and the higher MAPE it will not be advised to trade on the close share price of a stock and especially not in large sums.In this field a seven days gap between the data is considered significant and therefore until technology evolves it would not be wise to move to a full weekly trading schedule.

6.1 Contributions

This paper, with improvements can potentially joins the numerous research available out there emphasising at first that neural networks in general are now more suited to conduct forecasting on stock market data.It is also highlighting the fact that weekly level predictions are too inaccurate and volatile to currently move to a weekly trading schedule.

6.2 Challenges and limitations

The project did run quite smoothly for the majority of its duration thanks to the extensive papers available in this challenging field.The available material gave an extensive knowledge on the stock market and what scientist are trying to achieve.However, it was quite overwhelming as there was many papers to read with lots of different results and different data sets.In complete fairness except the complexity of the stock market patterns there was no other limitation but there were two main challenges.

Time,eight weeks to do such project including the EDA and this report is not enough due to its complexity and richness in solutions that could be computed to achieve higher accuracy.Also, in this particular case the coding was challenging,having close to no experience in machine learning was thought and created so much doubt during the project.

6.3 Future Work

It is evident that neural networks are the key to forecasting in time series data, potentially in any datatype and frequency.Due to time constraint this was a relatively basic study as the field of stock market share prediction is immense.To

begin with, hyper parameter tuning could be investigated to increase the model compatibility with the data and including external factors would help increase the accuracy's prediction.Moreover, when it comes to CNNs, there exist several layers that could be applied.

It could also be possible to investigate if there is a correlation between the behavior of cryptocurrencies and the stock market especially when it comes to big indexes such as S&P500.Finally, the stocks chosen have a significant amount of historical data however would this process work for new stock ? Would the model have enough data to train on ?

On another aspect , the data split of 80:20 could be modified, after investigation the first 80 percent of the data ends when COVID begins and therefore where the data is the most volatile.This could also lead to poor results and therefore changing the way the model is trained from the conventional 80:20 rules could be better.

References

[1] Rohitash Chandra, Shaurya Goyal, and Rishabh Gupta. Evaluation of deep learning models for multi-step ahead time series prediction. *IEEE Access*, page 1–1, 2021. ResearchGate, Available from : https://www.researchgate.net/publication/350457058_Evaluation_of_deep_learning_models_for_multi-step_ahead_time_series_prediction.

[2] Bjoern Krollner, Bruce Vanstone, and Gavin Finnie. Financial time series forecasting with machine learning techniques a survey. *ESANN*, page 25–30, 2010. Core Available from https://core.ac.uk/download/196602604.pdf.

[3] Lyle Daly. Call vs. put options: What's the difference? Inestopedia, Available from https://www.fool.com/investing/how-to-invest/stocks/call-options-vs-put-options/#:~:text=You.

[4] Akhilesh Ganti. What is insider trading and when is it legal?, Jun 2023. Inestopedia, Available from https://www.investopedia.com.

[5] P. Graton. What is the stock market and how does it work?, 2024. Investopedia, Available from: https://www.investopedia.com/terms/s/stockmarket.asp#:~:text=It.

[6] Wikipedia contributors. Warren buffett — Wikipedia, the free encyclopedia, 2024. Wikipedia, Available from https://en.wikipedia.org/w/index.php?title=Warren_Buffett&oldid=1239102327.

[7] David Floyd. Buffett's bet with the hedge funds: And the winner is ..., Jun 2019. Inestopedia, Available from https://www.investopedia.com/articles/investing/030916/buffetts-bet-hedge-funds-year-eight-brka-brkb.asp.

[8] I. Kumar, K. Dogra, C. Utreja, and P. Yadav. A comparative study of supervised machine learning algorithms for stock market trend prediction. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, April 2018. Available from: https://ieeexplore.ieee.org/abstract/document/8473214.

[9] L.P. Chen. Using machine learning algorithms on prediction of stock price. *Journal of Modeling and Optimization*, 12(2):84–99, December 15 2020. Available from: https://core.ac.uk/download/pdf/328150951.pdf.

[10] Santosh Ambaprasad Sivapurapu. Comparitive study of time series and deep learning algorithms for stock price prediction. *International Journal of Advanced Computer Science and Applications*, 11(6), 2020. ResearchGate, Available from https://doi.org/10.14569/ijacsa.2020.0110658.

[11] O.E. Orsel and S.S. Yamada. Comparative study of machine learning models for stock price prediction. https://arxiv.org/abs/2202.03156, 2022. arXiv preprint arXiv:2202.03156.

[12] T.B. Shahi, A. Shrestha, A. Neupane, and W. Guo. Stock price forecasting with deep learning: A comparative study. *Mathematics*, 8(9):1441, August 27 2020. Available from: https://www.mdpi.com/2227-7390/8/9/1441.

[13] Nazish Ashfaq, Zubair Nawaz, and Muhammad Ilyas. A comparative study of different machine learning regressors for stock market prediction, Apr 2021. Arxiv, Available from https://arxiv.org/abs/2104.07469.

[14] Apple Inc. (aapl) stock price, news, quote & history. Yahoo Finance, Available from: https://uk.finance.yahoo.com/quote/AAPL/.

[15] JP Morgan Chase & Co. (jpm) stock price, news, quote & history. Yahoo Finance, Available from: https://uk.finance.yahoo.com/quote/JPM/.

[16] BlackRock Inc. (blk) stock price, news, quote & history. Yahoo Finance, Available from: https://uk.finance.yahoo.com/quote/BLK/.

[17] Amazon.com Inc. (amzn) stock price, news, quote & history. Yahoo Finance, Available from: https://uk.finance.yahoo.com/quote/AMZN/.

[18] Wikipedia contributors. Vanishing gradient problem — Wikipedia, the free encyclopedia, 2024. Wikipedia, Available from https://en.wikipedia.org/w/index.php?title=Vanishing_gradient_problem&oldid=1222680571.

[19] Rian Dolphin. Lstm networks — a detailed explanation, March 2021. TowardsScience, Available From https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9.

[20] Nvidia. What is xgboost? Nvidia, Available from https://www.nvidia.com/en-gb/glossary/xgboost/.

A Appendix

Git Repository

Table A.1: Models Descriptions Parameters

Models	Descriptions
LSTM	Epochs 50, Dropout 0.3, Activation 'relu' ,100 days for predictions, Units 50
CNN	Epochs 50, Time Steps 100, Filters 64 and 128, Activation 'relu'
XGBoost	Parameter grid: max depth: [2,3,4,5,6,7,8], # estimators: [1000,5000,800,300], learning rate: [0.001, 0.0003]

Figure A.1: Project's Gantt Chart