Conestoga College

School of Applied Computer Science & Information Technology

SENG8080 - Case Studies Big Data

Stock Price Prediction

Jayakrishnan Sunil Kumar
Kizhakepura Velayudhan Geedhu
Paulraj Jemima
Santhiravathanan Delaxshana

November 29, 2023

Abstract

People are Investing so much money on Stocks. Both individual investors and organizations must consider stock price forecasts while making investment choices. The aim is to forecast the stock prices of three major technological companies in this project namely American multinational technology company (AAPL), Google (GOOGL), and NVIDIA Corporation (NVDA), and also compare the ability of several machine learning models, such as Linear Regression, Moving average and Long Short-Term Memory (LSTM) neural networks. Data is collected utilizing Jupyter Notebooks and the yfinance library.

Multiple stages are taken in this study. First, historical stock price information for AAPL, GOOGL, and NVDA from January 2018 till the present date is gathered. The required training and testing datasets are created by preprocessing and transforming this data. Following that, we set the machine learning models into execution, tune their parameters, and train them on historical price data.

Each model's performance is evaluated using applicable evaluation metrics. The root mean squared error (RMSE) is determined for Linear regression LSTM and Moving average. These measures offer insightful information about the capacity of forecasting of every model.

We aim to determine the most effective approach for predicting stock prices among the three companies by comparing the models and their performance. To make wise financial judgments while investing, choosing the right model is important. The outcomes of this research will assist investors and financial professionals in choosing an accurate stock price forecast strategy and making strategic investment decisions.

Table of contents

Contents

Introduction 4
Data Research and Integration 5
Data Collection 5
Data Storage and Maintenance 6
Data Analysis and Visualization 7
Extension 12
Proposed Allocation Project Team Roles 12
Project Timeline 13
References 13

Introduction

The stock price forecast plays an important role in making well-informed investment decisions for individual investors and businesses in the current investment environment. The stock prices of American multinational technology companies (AAPL), Google (GOOGL), and NVIDIA Corporation (NVDA) are the three main technology businesses considered as part of this research. Stock price data of these companies is gathered using the data science platform Jupyter Notebooks and the Yahoo Finance API.

The study's objective is to compare the capabilities of predicting stock prices by different machine learning models like Linear Regression, Moving average, and Long Short-Term Memory (LSTM) neural networks. By leveraging these models, the project aims to determine the most effective approach for predicting stock prices among the three companies. This evaluation is crucial for investors and financial professionals who seek accurate stock price forecast strategies to guide their investment decisions.The project follows a systematic approach, encompassing multiple stages to ensure robust analysis. First, the historical stock price data for AAPL, GOOGL, and NVDA is gathered and preprocessed to create the required training and testing datasets4. Subsequently, the machine learning models are implemented, and their parameters are tuned using the historical price data. This enables training the models and evaluating their performance against appropriate evaluation metrics.

For the Logistic Regression evaluation metrics such as accuracy, precision, recall, and F1-score are calculated. On the other hand, the performance of the LSTM model and and Moving average model are assessed using the root mean squared error (RMSE). These evaluation measures offer valuable insights into the forecasting capabilities of each model and aid in the comparative analysis.

The outcomes of this research hold significant implications for making wise financial judgments and strategic investment decisions. By identifying the most effective model for stock price prediction among the three companies, investors and financial professionals can enhance their ability to anticipate market trends and optimize their investment strategies.

By leveraging machine learning models and an extensive evaluation process, this project aims to provide actionable insights and accurate stock price forecast strategies. These findings will empower investors and financial professionals in their pursuit of profitable investment decisions, ultimately contributing to their financial success and growth.

Data Research and Integration

The project leverages the use of yfinance an open-source library that accesses the financial data available on Yahoo Finance[1]. Yfinance API was chosen, because of certain reasons including user-friendly API, Free access, and ease to download from start to end date etc., The major intention is to collect the historical stock price data for the top three growing companies namely Apple (AAPL), Google (GOOGL) and NVIDIA (NVDA) and integrate all three seamlessly for further analysis.

The project uses a platform like Jupyter workbench in GCP for the complete process to be accomplished which includes data collection, analysis, model building, and prediction of the stock prices.

Data Collection

The three leading companies (Apple, Google, and NVIDIA) Equity prices for the past five years from the current date are collected using the yfinance library which collects the shareholder values of each company from the Yahoo finance web page. A separate column named “Stock” is created for identifying the ticker symbol which is the company name is added to the data frame in addition to the columns that will be extracted from the Yahoo finance web page for the identification of the organization. The data frame will contain eight columns namely “Open”, “High”, “Low”, “Close”, “Adj Close”, “Volume” and “Stock” (A column that is added to identify the equity shareholder). There are 4380 rows and 8 columns in our dataset.

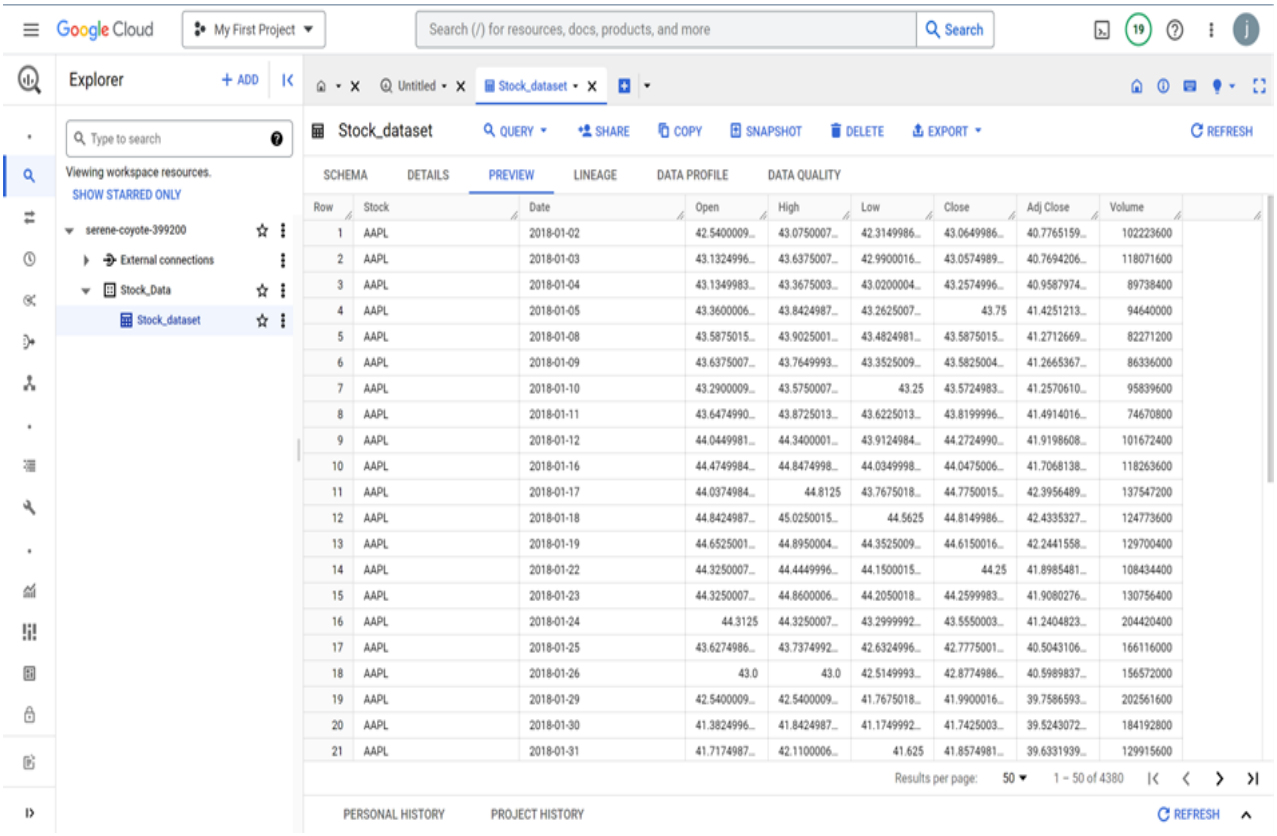
Description of the feature :

Stock: Corresponding Stock symbol of the organization.
Date: The date at which the trading activity is recorded.
Open: The price at the beginning of the trading day.
High: Highest price during the trading day.
Low: Lowest price during the trading day.
Close: The price at the end of the trading day.
Adj Close: It is the adjusted close price that accounts for corporate actions like dividends and stock splits, providing a more accurate historical performance measure.
Volume: Total number of shares traded during the trading session.

Data Storage and Maintenance

Once the data collection is done, it is important to do data storage. It is vital to guarantee its long-term storage for usage in the future thereafter. Google Cloud Platform is used to achieve this as it is a highly reliable and secure cloud-based platform that provides efficient data storage, easy access, maintenance, and recovery capabilities. So we chose to store the data in Big Query (GCP).

The Table is created in Big Query with the corresponding schema for the features. Once the table creation is done, The Big Query client is called and using the Biq Query client the data is pushed from the jupyter workbench (GCP) to Big Query (GCP).

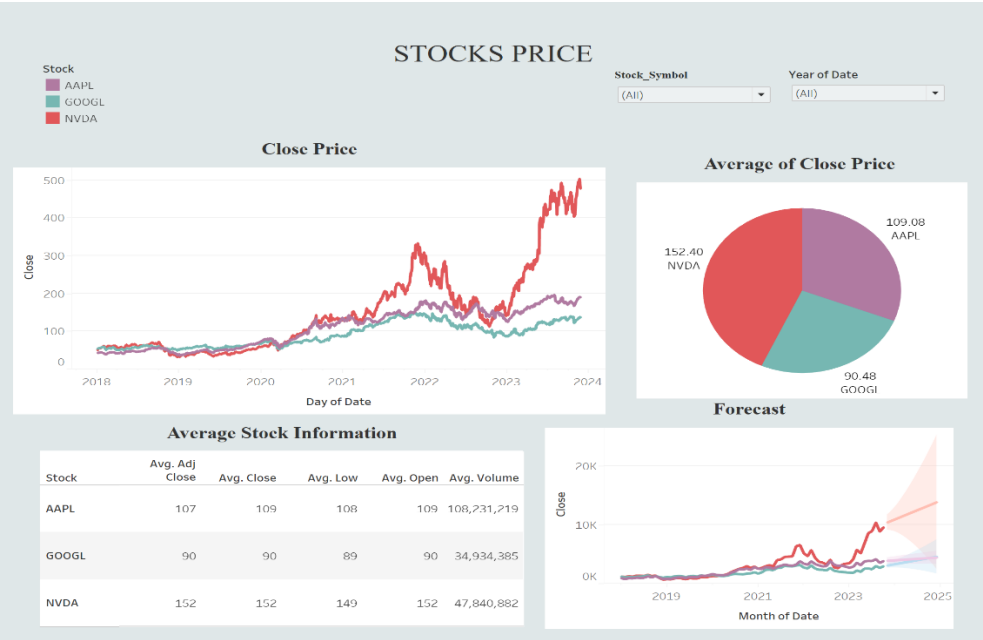


Data Analysis and Visualization

The data visualization is done in Tableau. Tableau is the most popular tool for data analysis and visualization. The Big query and the Tableau is connected using GCP service account. Various visualizations in shown in a dashboard which shows several information including the stocks close price of the selected three companies with respect to the trading day, Average closing price of the three companies with respect to year, Forecasting the close price of the three companies foe next few months. Then finally the comparison between the average open price, average close price, Average lower price, Average Adj closed price, average volume for three companies with respect to year.

To make the visualization more interactive, we have used the filters for stock_symbol and the Year. It helps us to compare the stock price with respect to each year. From the observation we found that Nvidia is growing rapidly compared to others.

These visualizations helped to understand the stock price situation of a particular company and helped to see if the company's stock price would rise or decline in upcoming months.



Models

The stock prices are forecasted using three algorithms Linear Regression, Moving Average and Long Short Term Memory (LSTM).

Linear Regression

The linear regression model returns an equation that determines the relationship between the independent variables and the dependent variable. The linear regression algorithm is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The algorithm aims to minimize the difference between predicted and actual values, adjusting coefficients iteratively through gradient descent.

The algorithm initializes coefficients, defines a hypothesis function to predict values, calculates a cost function to measure the prediction error, and uses gradient descent to update coefficients iteratively.

Moving Average

A quantitative method of forecasting or smoothing a time series by averaging each successive group (number of observations) of data values. The term “Moving” is used because it is obtained by summing and averaging the values from a given number of periods, each time deleting the oldest value and adding a new value.

A moving average is a commonly used statistical calculation that is used to analyze data points by creating a series of averages of different subsets of the full data set. It is particularly useful for smoothing out short-term fluctuations and highlighting longer-term trends or cycles. The window size can impact the level of smoothing and the sensitivity of the moving average to changes in the data. Smaller window sizes provide more responsiveness to short-term changes, while larger window sizes result in a smoother curve that may highlight longer-term trends.

LSTM

The LSTM model is implemented to predict the stock price since LSTM works well with sequence data. The LSTM model is trained with 10 epochs with a window size of 60. The loss function used is mean squared error and the optimizer is adam optimizer. The forecast is done for the next 100 days.

LSTM cell has several components like cell state, hidden state, and gates(input, output, forget). Cell state transfers the relevant information throughout the sequence chain. Gates helps to find which information is relevant and which is not, and based on that It decides which one to forget and which one to keep.

Forget Gate:

It decides which one to forget and which one to keep. Information from previous hidden state (ht-1) and the information from the current input xt is passed through the sigmoid function. Values come out between 0 to 1. The closer to 0 means to forget and closer to 1 means to keep.

Input gate:

It is used to update the cell state. In this layer, there are 2 parts (sigmoid and tanh). In the sigmoid function, it decides which value to let through by transforming the values between 0 to 1. 0 indicates it is not information and 1 indicates it is important. Tanh function gives weightage to the values which are passed deciding their level of importance by converting the value between -1 and 1.

Then it multiplies the tanh output with the sigmoid output, then the sigmoid output will decide which information is important to keep from the tanh output. Now the cell state need to be updated, output from the forget gate and input gate is utilized to do calculation and cell state gets updated.

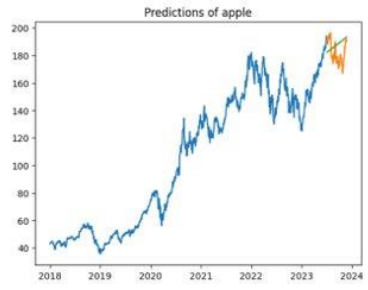
Output gate:

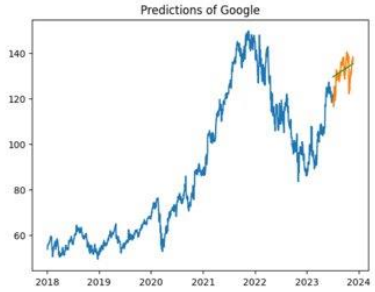
It decides on what the next hidden state should be. Hidden state contains the information on previous inputs and hidden states are used for predictions. The previous hidden state which is ht-1 and current input xt is passed into a sigmoid function, then the newly modified cell state is passed to the tanh function, then it multiplies the tanh output with the sigmoid output to decide what information the hidden state should carry. New hidden state are then carried out to next timestep.

MODEL OUTPUT VISUALIZATION

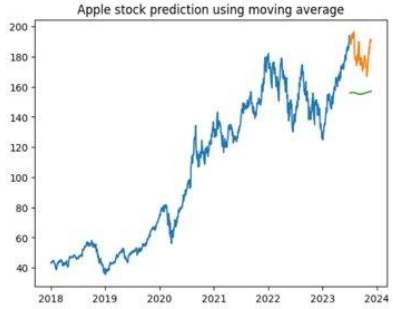
MODEL PREDICTION

Linear Regression



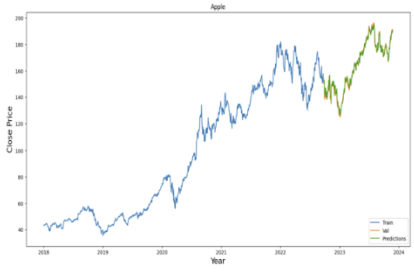


Moving Average

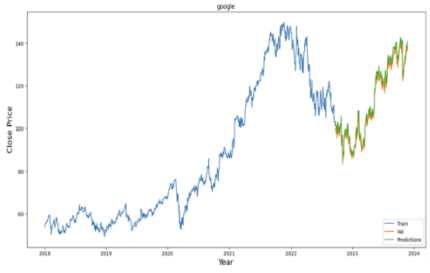


LSTM

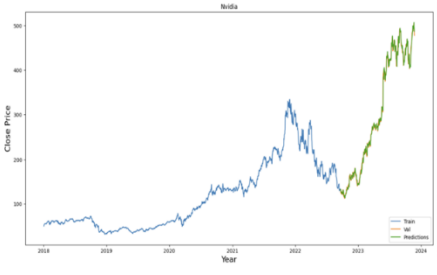
Apple



Google

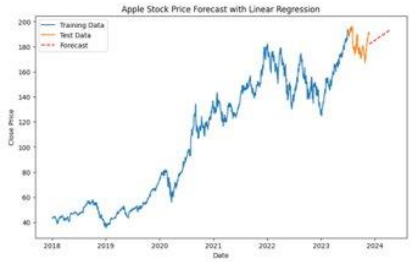


Nvidia

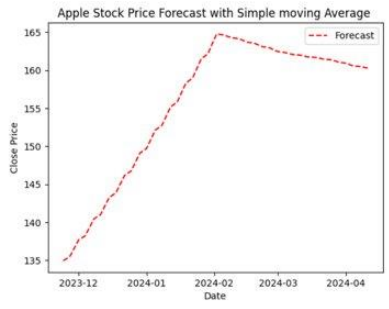


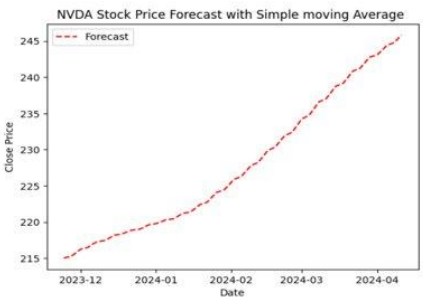
MODEL FORECASTING

Linear Regression

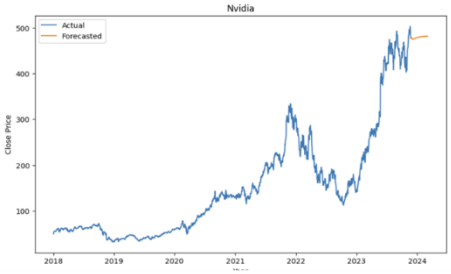


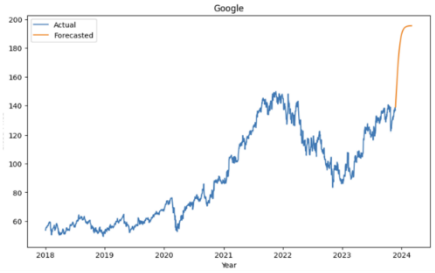
Moving Average





LSTM





PERFORMANCE EVALUATION

RMSE:
RMSE stands for Root Mean Squared Error. It is a commonly used metric to measure the accuracy of a predictive model. It is the average of the squared differences between the predicted and actual values.

	Linear Regression	Moving Average	LSTM
AAPL	11.342	26.804	2.658
GOOGL	5.187	28.890	2.821
NVDA	188.026	223.928	9.436

Extension

The prediction and forecasting of stock prices are limited to only three stocks currently in this project which are Apple (AAPL), Google (GOOGL) and NVIDIA (NVDA). This project can be scaled up to other stocks as well. The forecasting of prices is calculated for the next 100 days for each stock. This forecasting period could be scaled up in the future. The dataset size is 974KB and it includes past 5 year's data. So the storage need not to be scaled up in the upcoming future unless we include many other company stocks in this project.

Proposed Allocation Project Team Roles

Delaxshana researched a few stock market APIs and their advantages and discussed them with our team members and the team members agreed upon the API after several research and she provided the code for downloading the dataset using yfinance API.
Geedhu implemented the complete code for downloading the data for these companies and created a code for pushing the data to big query.
Sunil Created a table in Big Query (GCP) with the corresponding schema and he is managing the data in big query.

Jemima Connected the Big Query to Tableau using a service account and create a dashboard which shows several information including the stocks close price of the selected three companies with respect to the trading day, Average closing price of the three companies with respect to year, Forecasting the close price of the three companies foe next 3 months

Data preprocessing is done by Sunil. Three algorithms were implemented for forecasting stock prices. Linear regression was implemented by Geedhu, Moving average by Delakshana and LSTM by jemima. Final adjustments made and checked by sunil.

Project Timeline

Date	**Deliverable**	**Responsible**
Oct 15	Data Collection API Finalized	Delaxshana
Oct 16	Code Implementation for data collection	Geedhu
Oct 17	Big Query Table Creation	Sunil
Oct 18	Tableau dashboard creation for visualization	Jemima
Oct 21	Final Adjustments made and checked in Report 1	Jemima, Geedhu
Oct 25	Data Preprocessing	Sunil
Nov 01	Linear Regression	Delaxshana
Nov 08	Moving Average Algorithm	Geedhu , Sunil
Nov 15	LSTM	Jemima
Nov 29	Report	Sunil, Delaxshana, Jemima, Geedhu
Nov 29	PowerPoint Presentation	Sunil, Delaxshana, Jemima, Geedhu

References

1. Bland, G. (2020, November 3). *yfinance Library – A Complete Guide*. AlgoTrading101 Blog. https://algotrading101.com/learn/yfinance-guide/

2. (2023, May 1). *Stock prices prediction using machine learning and Deep Learning*. Analytics Vidhya. Retrieved October 1, 2023, from https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/

3. A.Moghar, M.Hamiche,"Stock Market Prediction Using LSTM Recurrent Neural Network", Procedia Computer Science,Volume 170,2020,Pages 1168-1173

4. S. Kumar S, C. D and S. Rajan, "Stock price prediction using deep learning LSTM (long short-term memory)," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1787-1791

5. Gianluca Malato(2020, June 23). *An algorithm to find the best moving average for stock trading.* *https://towardsdatascience.com/an-algorithm-to-find-the-best-moving-average-for-stock-trading-1b024672299c*

6. Debasish Kalita (March 24th, 2022). *An Overview on Long Short Term Memory (LSTM).* *https://www.analyticsvidhya.com/blog/2022/03/an-overview-on-long-short-term-memory-lstm/#:~:text=Structure%20Of%20LSTM&text=The%20flow%20of%20information%20into,time%20series%20of%20uncertain%20duration*.

7. Dr. Suresh Subramanian1 , Sai Saketh Reddy, "STOCK MARKET PREDICTION SYSTEM USING ML",JOURNAL OF EDUCATION: RABINDRA BHARATI UNIVERSITY, Vol: XXIV, No. :1(II), 2022, Pages 155-159.

8. (n.d.). Stock Price Prediction & Forecasting with LSTM Neural Networks in Python. Youtube. https://www.youtube.com/watch?v=CbTU92pbDKw