The S&P 500 Index and Selective Stocks Data Project

1. Problem definition:

1.1 Topic background

The S&P 500 is one of the most important financial barometers to indicate the overall status of the macro economy. Compared to the S&P 500, the Dow Jones Industrial Average is so selective as it comprises only 30 stocks. So the S&P 500 is more representative of the economy as a whole than others.

1.2 The problem we are trying to solve

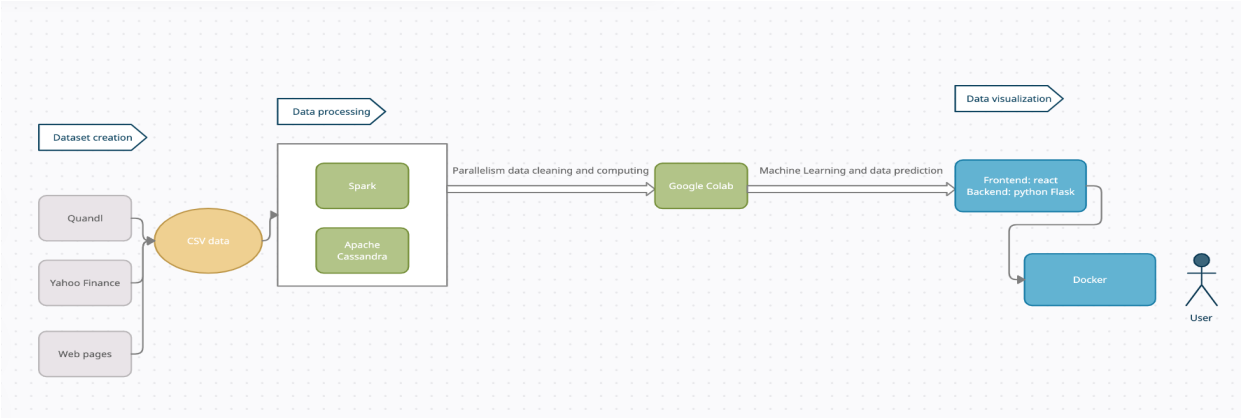
S&P 500 is a good macroeconomic indicator for investors. We are trying to build several figures from history data(2012-2022) to show historical trends and predict future trend forecast for both S&P 500 index and 5 stocks(AAPL:Apple, MSFT: Microsoft, GOOGL: Google , AMZN: Amazon, NVDA: Nvidia)

We are trying to develop a useful tool to support investment decision making for users. We provided 7 different kinds of charts to visualize the historical data of S&P 500 as a benchmark for the market, and we also provided a good future data prediction for both S&P 500 and 5 selective company stocks mainly focused on the technology industry.

We are collecting (over 100MB data) from various sources including Quandl API(Quandl | Nasdaq Data Link), Yahoo Finance(yfinance package) and data retrieved from Web pages(XML data from curl result). We aggregated our data from different sources into CSV files for better data processing and data availability.

This project is designed for investment beginners, which helps them understand how to analyze and predict the S&P 500 index, how to analyze top S&P 500 companies, and how to predict the selected high-quality stocks.

2. Methodology



2.1 Data Collection
The data we need for this project is from an authoritative financial institution, Yahoo Finance.

We are collecting (over 100MB data) from various sources including Quandl API(Quandl | Nasdaq Data Link), Yahoo Finance(yfinance package) and data retrieved from Web pages(XML data from curl result). We aggregated our data from different sources into CSV files for better data processing and data availability.

2.2 ETL
Duplicate data handling before loading to Cassandra. We choose Cassandra because it is a distributed database.

2.3 Algorithm
We used multiple methods to analyze the historical data, including time series analysis and time series decomposition. In time series analysis, we compared several typical indexes, including simple moving average, exponential moving average, relative strength index, moving average convergence divergence. For time series decomposition, we decomposed using three different frequencies, which is daily, monthly, yearly.

For prediction, we tried and compared predictions with Random Forest and LSTM from framework Keras based on neural networks, and chose the LSTM Algorithms as our model for future prediction methods based on their performance.

2.3 Data Visualization
We chose plotly, matplotlib and Apache Echarts to make data visualization.
For Web UI design for massive plot display, we used tabs to split and fold figures into 5 themes, so users can easily choose and check the figures they are interested in.
For App build and deployment, we used docker to containerize our application for easy build and deployment to any platform(local machines, virtual machines or deployed to the cloud). Our application is portable and it is isolated from other containers.

3. Problems
At the beginning, we tried to use Cassandra as our database and complete all machine learning tasks in the cluster. However, there is technology limitation in the cluster(ERROR MSG: The TensorFlow library was compiled to use SSE4.1 instructions, but these aren't available on your machine.). Due to limitations of cluster hardware conditions, we changed our data analyzing approach from spark to jupyter notebook for framework picking flexibility. We finally combined spark and Google Golab to overcome this problem.

For the company dataset from S&P 500 index company list, some missing data made it difficult to conduct company importance analysis. However, we used Yahoo Finance open-source module to request the ticket to make up for missing data.

4. Results

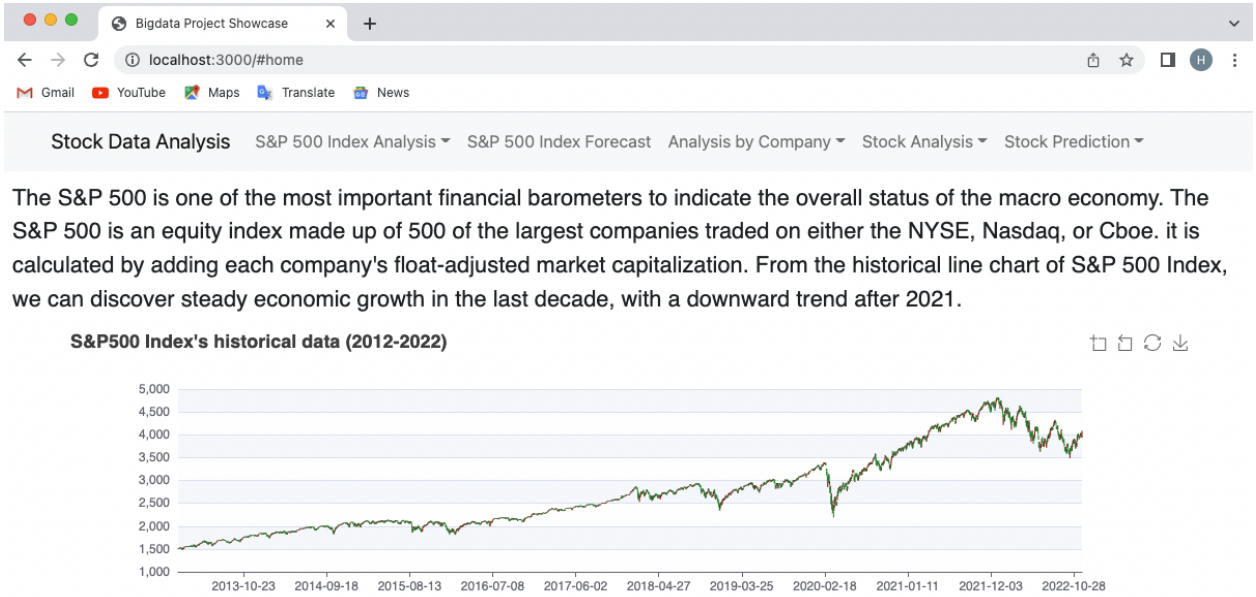
From the S&P 500 Index's historical data Chart, we can discover steady economic growth in the last decade, with a downward trend in 2020 and 2022. According to the forecast of the S & P 500 Index, it is still fluctuating but has a slightly upward trend, which means that the capital market has rebounded a little now and people can choose some Index Funds to make investment decisions.

We analyzed the top 10 selected companies in S&P 500 index company list. Based on company analysis chart, we can see that there are 50% high-tech companies in the top 10 companies. So, we further analyzed these 5 high-tech companies and made predictions. From these forecast charts, except AAPL, the predictions of other four stocks seem not promising as S&P 500 Index forecast, so it is not a good time to invest GOOGL, MSFT, AMZN, NVDA. Unlike other four stocks on a downtrend, AAPL has had a small rebound in the float, so AAPL is a potential stock to watch and invest in.

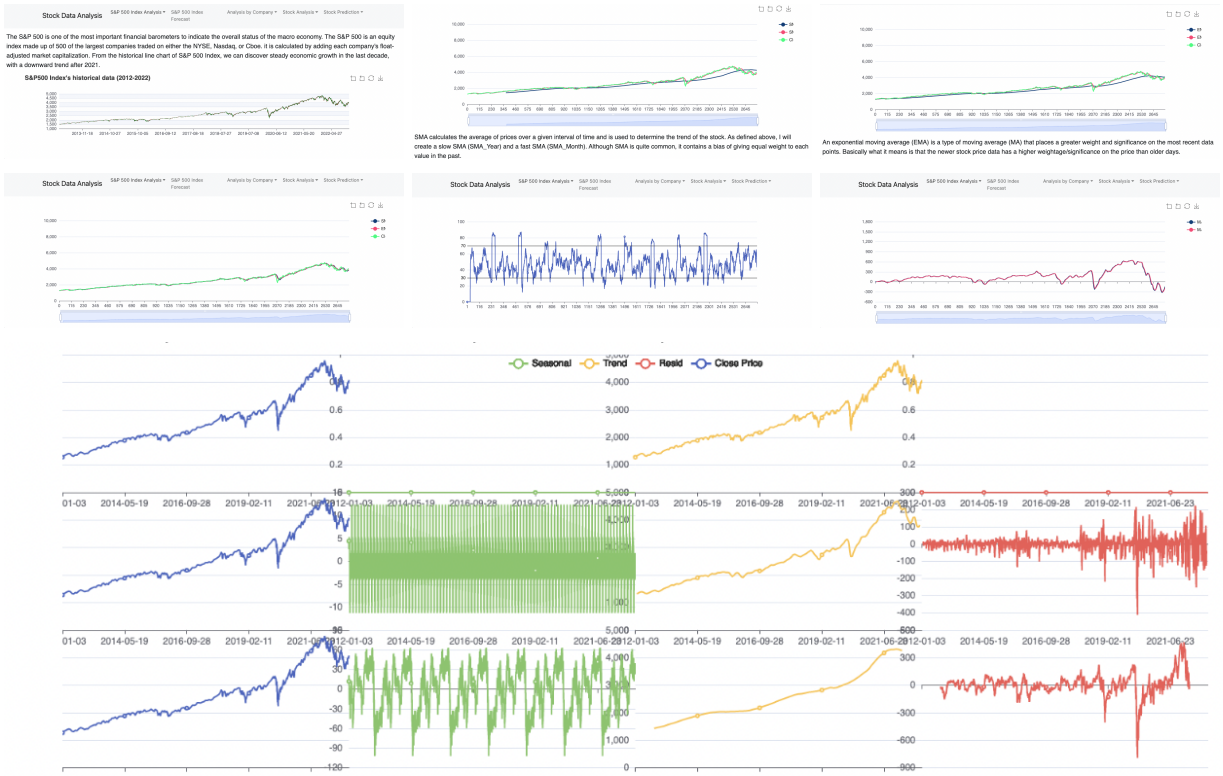
The main goal of this project is to help rookie investors better understand the capital market from the Macro Index, such as the S&P 500 index. From the selected stocks in S&P 500, we made further analysis to show them how to choose good-quality stocks. According to our outcomes, our project is a meaningful tool to let people understand the correlation of macro index and the economy, the trend and forecast of stocks, and how to choose the right time to invest. Also our UI provides a smooth user experience, we created a clean layout with intensive data display and easy for users to zoom in and out to check each point of data in more details.

4.1 Outcomes and Web UI

4.1.1 Home Page



4.1.2 S&P 500 Index Analysis

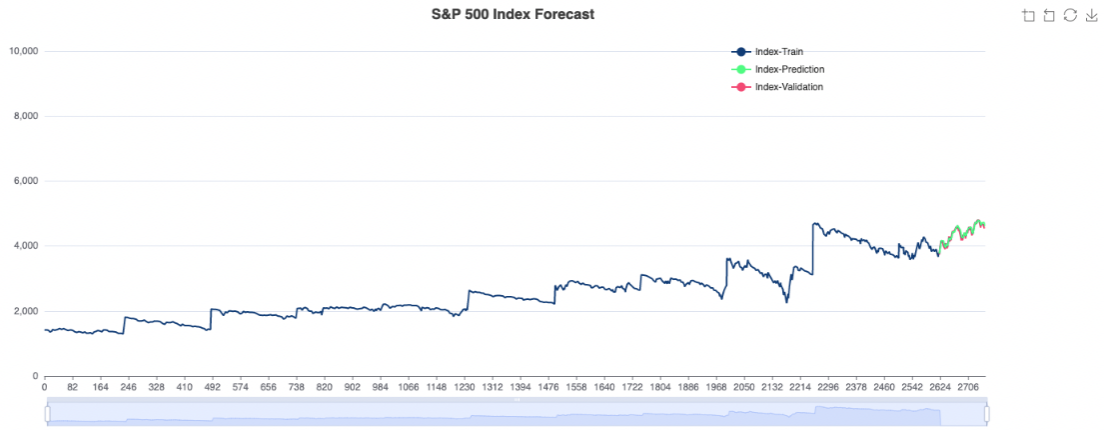


There is clearly an upward trend in the above plot. You can also see the uniform seasonal change. Non-uniform noise that represents outliers and missing values.

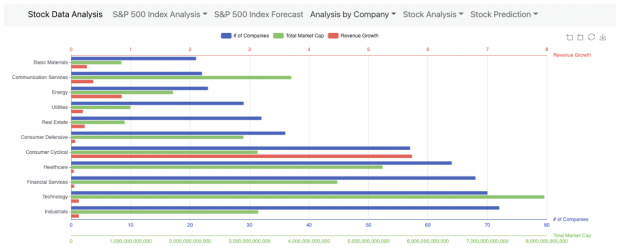
4.1.3 S&P 500 Index Forecast

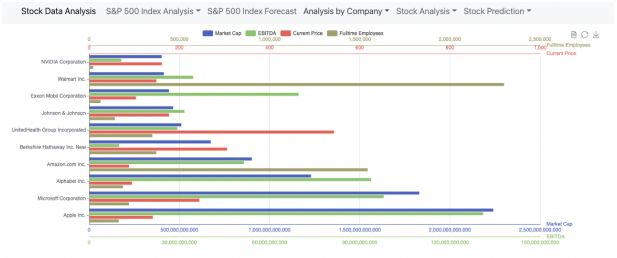
Stock Data Analysis | S&P 500 Index Analysis | S&P 500 Index Forecast | Analysis by Company | Stock Analysis | Stock Prediction

Use LSTM model to make prediction:

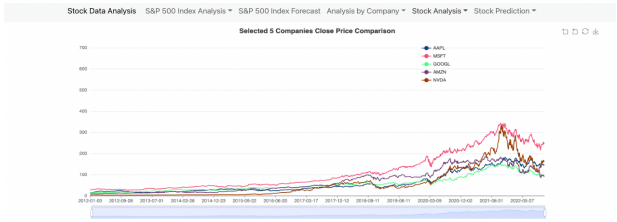


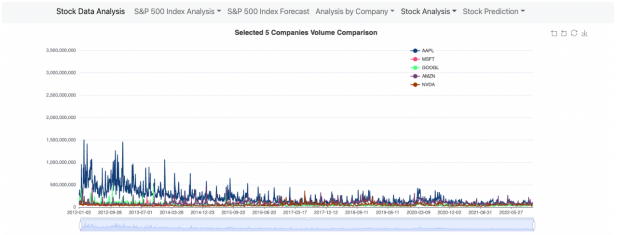
4.1.3 Analysis By Company



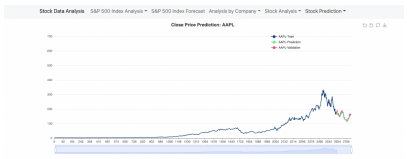


4.1.4 Individual Company Stock Analysis and Prediction





Those stocks are from Top 10 stocks of company analysis. From the chart, we can see that the trends of these 5 stocks are basically



Project Summary:

Categories	Mark
Getting the data	1
ETL	1
Problem	3
Algorithmic work	4
Bigness/parallelization	3
UI	3
Visualization	4
Technologies	1
TOTAL	20