Springboard--DSC

Capstone Project 3:

Distributed Computing and Auto-ML Applied to Time Series Forecasting of Stock Price

Final Report

By Pallavi Bothra

04-2022

1.Introduction:

Stock markets are where individual and institutional investors come together to buy and sell shares in a public venue. These exchanges exist as electronic marketplaces. Supply and demand help to determine the price for each security or the levels at which stock market participants, investors and traders are willing to buy or sell. The stock market is very unpredictable. Any geopolitical change can impact the trend of stocks in the share market. Therefore, it is very difficult to do a reliable trend analysis. However, with the advent of modern technologies, computational resources, advanced algorithms it is possible to have a data-driven well informed decision before making any investment.

In Capstone Project 2, we used time series forecasting models for stock market trend analysis and future prediction for ~1600 companies. We predicted future values (30 days) based on previously observed daily values (5 years).

Capstone project 3 is split into two parts. In the first part, we will convert the time series dataset to supervised learning dataset. Then, after computing additional features, the AutoML[1] framework will be used to train and forecast the stock price. Finally, error metrics will be calculated for each company and compared with the PMDARIMA[2] and PROPHET[3] model (results obtained from Capstone Project 2). In the second part, we will deploy the entire workflow of Capstone Project 2 in a Cloud platform and develop a Web client application.

This report presents a summary of how the project was developed. Jupyter notebooks can be browsed through the following link:

https://github.com/Pallavi43/Springboard/tree/main/projects/interactive_time_series_forecasting.

[1] https://supervised.mljar.com/

[2] https://pypi.org/project/pmdarima/

[3] https://pypi.org/project/fbprophet/

2.Approach:

2.1 Data Acquisition and Wrangling

The present project is a continuation of Capstone Project 2. So, basically we used the same dataset which was pulled in the previous project. However, to run in the Cloud environment, we transformed previous notebooks so that they can utilize Dask Dataframes[4] which are partitioned to do computations in parallel fashion so that system cores can be effectively utilized.

2.2 Storytelling and Inferential Statistics

While exploring the data, we asked the following questions among others with respect to the different sectors:

1.Does the market follow a trend across all the sectors?

2. Which sector is more volatile in 6 years (2012-2017)?

3. Which sector is more stable in 6 years (2012-2017)?

[4] https://docs.dask.org/en/stable/dataframe.html

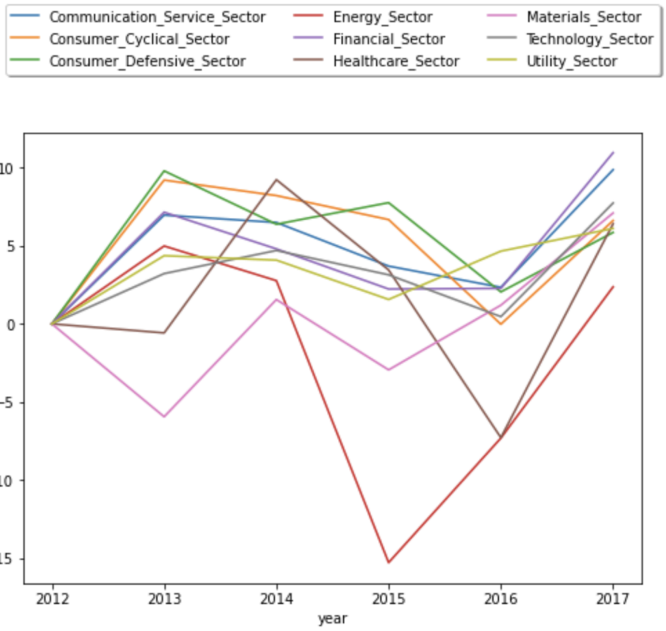


Figure 1. Summary of the sectors with the average stock price trend across 6 years

Figure 1. summarizes the trend in average stock prices of all the sectors in 6 years (2012-2017). It can be seen from the figure that except 2 or 3 sectors all the other sectors follow a similar trend. On average, the market goes up and down together. Energy sector has crashed the most among all the sectors. Utility sector seems to be most stable in the 6 years of this period.

2.3 Conversion of Time Series Dataset to Supervised Learning Dataset :

In this section, we started converting time series dataset to supervised learning dataset. First of all, we created some new features by feature engineering for all the companies across various sectors. We computed statistical properties (mean, min, max, std) of 1 week and 2 weeks prior data and added them as independent variables. We also considered the previous day as a feature. Finally, we used the automated machine learning (AutoML) frameworks on supervised learning dataset for training and forecasting stock prices. AutoML trains the dataset with several ML algorithms and finally chooses the one with the best performance metric. In this case, since the predictor variable is continuous, AutoML considered regression algorithms (e.g., LinearRegression, SVM, RandomForest Regressor, GradientBoostingRegressor etc. In section 3, we will discuss the error metrics (Root Mean Squared Error (RMSE) and Mean Average Percentage Error (MAPE)) for different companies' stock price prediction using different methods (PMDARIMA, Prophet, AutoML).

2.4 Data Preparation Workflow for Distributed Computing:

The next task is to deploy the model in the Cloud platform. We uploaded all the stock data files to a Bucket in the AWS S3 object store. We spun off a 3-node EMR (Elastic Map Reduce) cluster with Hadoop Yarn[5] setup. We set up a dask-yarn cluster setup on the master node of the EMR cluster. Next, we executed various notebooks that read all the stock data files from AWS S3; performed wrangling, exploring and modeling in the EMR cluster and wrote back to the AWS S3 filesystem. Figure 2. represents the entire data preparation workflow used in this project.

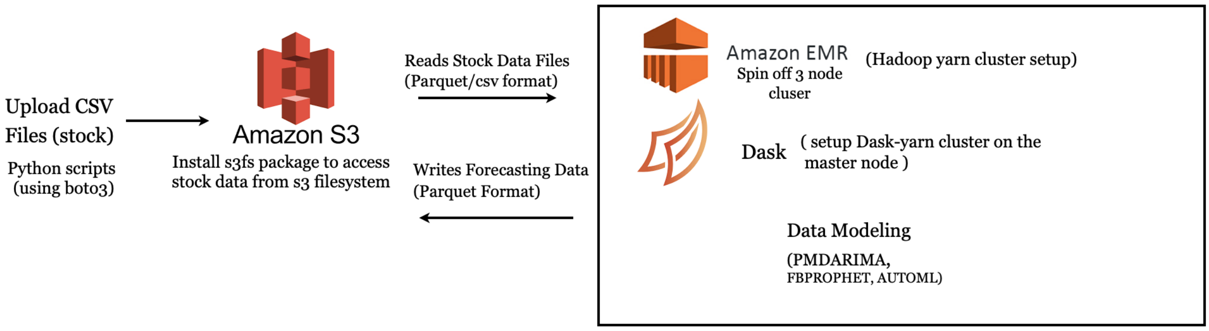

Upload CSV Files (stock)
Python scripts (using boto3)
Amazon S3
Install s3fs package to access stock data from s3 filesystem
Reads Stock Data Files (Parquet/csv format)
Writes Forecasting Data (Parquet Format)
Amazon EMR
Spin off 3 node cluser
(Hadoop yarn cluster setup)
Dask
(setup Dask-yarn cluster on the master node)
Data Modeling
(PMDARIMA, FBPROPHET, AUTOML)

Figure 2. Data Preparation Workflow for Distributed Computing

2.4 Visualization App using Plotly Dash:

[5] https://yarnpkg.com/

In the final step, an interactive web application will be developed using Plotly Dash[6]. Based on the user input from one of the inputs on the Dash application, the Plotly application fetches the stock data/model file based on the user input from Amazon S3 and runs the various models like PMDARIMA, Prophet and AutoML and creates figures that visualize the stock forecasts over 30 days with error metrics. Figure 3 describes the architectural diagram used in this project for visualization.

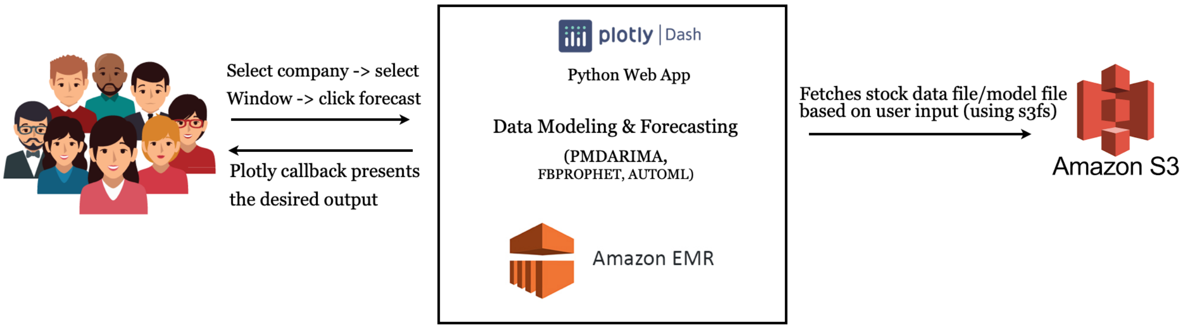

Select company -> select Window -> click forecast
Plotly callback presents the desired output
plotly | Dash
Python Web App
Data Modeling & Forecasting
(PMDARIMA, FBPROPHET, AUTOML)
Amazon EMR
Fetches stock data file/model file based on user input (using s3fs)
Amazon S3

Figure 3. Architectural Diagram for data visualization

3. Findings:

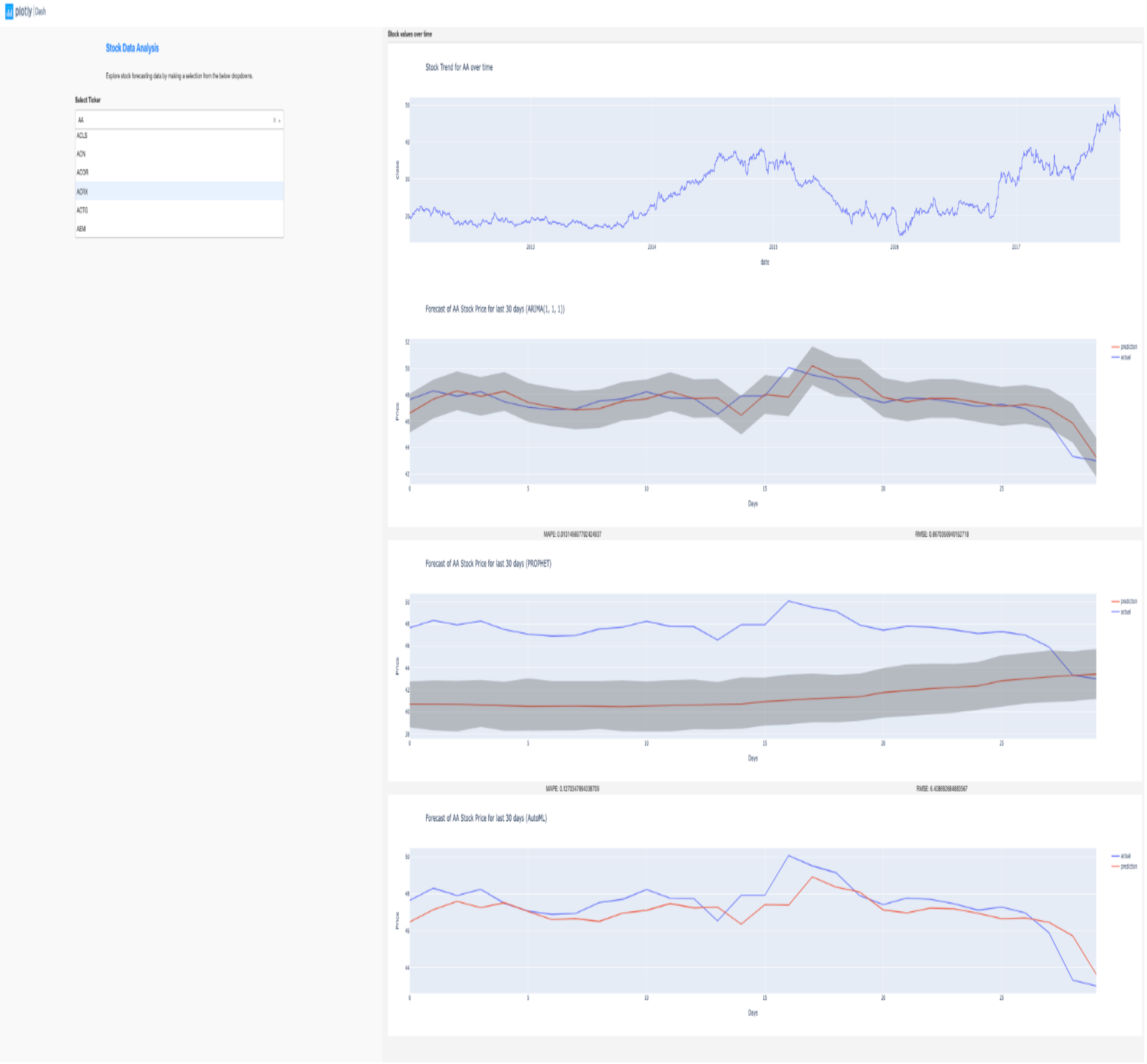
Table 1 shows the error metrics (RMSE and PMDARIMA) of five companies (please note that these 5 companies are randomly selected). Lower MAPE/RMSE values indicate a better model in terms of reduced variability and less bias in the result. It can be seen that PMDARIMA and AutoML outperforms the PROPHET model.

Company	PMDARIMA (MAPE)	PROPHET (MAPE)	AutoML (MAPE)	PMDARIMA (RMSE)	PROPHET (RMSE)	AutoML (RMSE)
Google	0.007	0.03	0.009	11.02	37.16	9.11
Apple	0.008	0.04	0.01	1.95	6.92	1.30
ABM Ind	0.005	0.03	0.01	0.30	1.75	0.40

[6] https://plotly.com/dash/

Alcoa Corp	0.01	0.04	0.01	0.20	0.56	1.12
Future Fintech Grp	0.04	0.32	0.03	0.09	0.66	0.51

Table 1. Performance Metrics for five companies with the error metrics



Stock Data Analysis

Explore stock forecasting data by making a selection from the below dropdowns.

Select Ticker

AA

ACLS
ACN
ACOR
ACRX
ACTG
AEM

Stock values over time

Stock Trend for AA over time

Forecast of AA Stock Price for last 30 days (ARIMA(1, 1, 1))

MAPE: 0.013146607792424937

RMSE: 0.8670356940152718

Forecast of AA Stock Price for last 30 days (PROPHET)

MAPE: 0.1270347904338709

RMSE: 6.408692684883567

Forecast of AA Stock Price for last 30 days (AutoML)

Figure 4. Interactive Web Application developed using Plotly Dash

Figure 4 shows a screenshot of the interactive web application created using Plotly Dash. Clients can select any company and forecasting of 30 days stock price using PMDARIMA, PROPHET and AutoML will be displayed.

4. Conclusions and Future Work:

In conclusion, we converted the time series dataset to a supervised learning dataset. AutoML framework has been used to train the dataset and forecast the stock price for next 30 days. In the next step, the complete workflow has been deployed in a Cloud platform and finally interactive web client applications have been developed. Finally, performance metrics (MAPE and RMSE) computed by PMDARIMA, PROPHET and AutoML for some of the companies have been tabulated.

In the future, we are planning to use deep learning algorithms to check if the current performance improves or not. We are also planning to do sentiment analysis to capture the sentiments of the investor and have an idea about the future of the market.

5. Recommendations for the Clients:

a. Figure 1 gives an overview of stock market trends across the sectors. The comparison among the sectors will give an overall overview of stock price trends across the companies which will help to make an informed decision before investing in any particular sector.

b. The performance metrics for different trained models of each company explains the reliability of the model. This will also help to forecast the stock price. Also, interactive web applications will give clients a visualization about the past and future trends. So, we believe that this thorough analysis using advanced models will help both existing and new investors to understand and make a data driven decision to invest in the share market.

6. Consulted Resources

1.Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia.

Web reference: https://otexts.com/fpp2/

2. https://machinelearningmastery.com/time-series-forecasting/

3. https://supervised.mljar.com/