

# Capstone Project: Stock Prediction with Alpha vantage

15.06.2022

---

## 1. Domain Background and Motivation:

Stock market analysis enables investors to identify the intrinsic worth of a security even before investing in it. Studying and evaluating past and current data helps investors and traders to gain an edge in the markets to make informed decisions. Performing a research before making an investment is a must. It is only after a thorough research that one can

make some assumptions into the value and future performance of an investment. Prediction of stock performance depends on many factors. Data Science is a popular subject. Many times data is represented as numbers and these numbers can represent many different things. These numbers could be the amount of sales, inventory, consumers, and last but definitely not least — cash. Stocks, commodities, securities, and such are all very similar when it comes to trading. We buy, we sell, we hold.

My motivation is to try out whether machine learning and data science can help us to predict stock performance. I am working on a hobby project Financifical collaborated with others. (<https://www.financifical.com/>) where we try to predict the stock prices. Here we consider simple moving average, exponential moving averages as indicators. The images of stock patterns are the features and I use CNN as the model.

With the growth of data science and machine learning techniques the approaches to predict the performance of stock has improved. The training involves using selected data or a portion of the data to “train” a machine learning model.



## 2. Problem Statement:

As my interest lies in commodity-based companies like Eagle Pharmaceuticals, Energy Focus etc, I have considered these in my project. The problem statement is to analyze 2

biggest commodity-based companies that were active since 2014 and understand their behavior, correlations and associations and also to dive deeper into how different factors impact. Furthermore, exploratory data analysis on the data and to train models and compare how they perform. Hyperparameter Tuning and visualizations.

### 3. The Datasets and Inputs:

The data will be gathered using the Alpha Vantage Finance API which I used to extract the stock data. The stock data as per NASDAQ is updated here. Here is the article (<https://medium.com/codex/alpha-vantage-an-introduction-to-a-highly-efficient-free-stock-api-6d17f4481bf>)

We will gather data of the 2 biggest commodity-based companies in USA, that are:

- Eagle Pharmaceuticals Inc (EGRX): Eagle Pharmaceuticals is a specialty pharmaceutical company focused on developing and commercializing injectable products. Founded in 2007
- Energy Focus Inc (EFOI): the company is among the largest companies in Brazil, and is the largest producer of iron ore in the world. Founded in 1985

The data will be divided into 6 features for each day: low, high, open, closed and Adjusted closed and Change. It would be nice to start with data from 2014 until today. The data will be saved in a CSV and uploaded to S3.

### 4. Solution Statement:

The solution is to test two algorithms used for forecasting:

- a. The DeepAR, created by Amazon and available in the Estimator API;
- b. The LSTM approach. Long-term memory (LSTM) is a deep learning artificial recurrent neural network (RNN) architecture. Unlike traditional feed-forward neural networks, LSTM has feedback connections. It can handle single data points (such as pictures) as well as full data sequences. Other libraries required for constructing and visualizing the LSTM model outputs. We'll be using the Keras library from the TensorFlow framework for this. All modules are imported from the Keras library.

Having these two methods I can compare both and use the best one to deploy the solution.

## 5. Benchmark Model:

Since it is an academic student project the benchmark can be established from a classical machine learning model. I will use Random Forest as a regression algorithm because the type of output is continuous float values and to fit the data and to make predictions for the next days.

Using such a simple algorithm as a baseline is a good fit, we have a simple way to create a reference model and if the DeepAR or LSTM did not perform better than the Random Forest, they should be discarded away or at least tweaked or made necessary modifications.


## 6. Set of Evaluation Metrics:

The forecasting task works with continuous output values and the common metrics used in such cases are RMSE - Root Mean Square Error and MAE- Mean Absolute Error. There are other evaluation metrics which could be used like R Square/Adjusted R Square or MSE.

## 7. Project Design:

The flow of the project are as follows:

- a. Collection of the Data using Alpha vantage API
- b. Processing the data to Analyze
  - Cleaning the data, checking for missing values etc
- c. Data Exploration
  - Explore the risk and volatility of each stock(company)
  - Explore the correlation between the closing price
  - Explore the simple moving averages and trends
  - Feature engineering like binning of the variation or computing the difference between the lowest and highest price for the day.
- d. Distribution of data. Split into train and test
- e. Uploading the data to AWS S3 bucket
- f. Scaling the data for a better performance of the models;
- g. Create, Training and Testing sets a Random Forest Regressor for Baseline;
- h. Create, Training and Testing sets for DeepAR model;

- 
- i. Create, Training and Testing sets for LSTM model as a rolling basis model;
  - j. Fine tune of hyperparameters like number of LSTM layers, number of neurons, dropout rate, learning rate, etc.;
  - k. Comparing the results of the models;
  - l. Deploy the best one.

Since the nanodegree title is AWS Machine Learning Engineer and to get practical understanding of AWS and Sagemaker , the instances, data estimators and all the details related to the project like storing models, deploying endpoints will be on AWS. Cloudwatch logs have greatly helped me throughout the course to understand the bugs and errors and hence I will also be considering logging the errors in my scripts to be able to find the issue soon.