Capstone Proposal: Stock Price Forecasting using SageMaker

Pedro Henrique Couto Correa Camargos
Machine Learning Engineer Nanodegree

1. Domain Background:

The stock price market might seems chaotic at a first glance. There are a lot of information that must be taken into account in order to decide whether to buy or not a stock. Some say that the Fundamental Analysis [1] is the right way to decide whether or not to buy a stock, and some say that the Technical Analysis [2].

Based on this lack of convergence about which method should be used, and once that we can find a giant amount of data, a data science project could be a nice fit to predict the behaviour of the prices in the next day. Therefore we could use these predictions to guide the decisions.

Metrics like opening and closing prices, max price in the day and moving average are common data collected for every company present in a market. This types of analysis, the basic knowledge of the market and these metrics, provide a strong foundation to start the project.

2. Problem Statement:

Knowing that predict stock prices is a so challenging topic, that the prize is great and that there are lots of data available, the problem to be solved is **to create a pipeline with SageMaker that ingest data and return a forecast for the next 2 and 4 weeks to be used to guide the decisions of buying or selling stocks.**

3. The Datasets and Inputs

The data will be gathered using the Bloomberg APIs and the Yahoo! Finance API.
Since it is the Machine Learning Engineer Nanodegree, I intend to follow the whole SageMaker workflow. It means the data will be collected and transformed into one unique CSV file that can be uploaded to S3 and then, the data can be used as input for training with the SageMaker API.

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. Create my on Neural Network using LTSM (Long Short Term Memory) method.

Having these two method we can compare both and use the best one to deploy the solution.

5. Benchmark Model

As benchmark will be used this notebook by Fares Sayha and this notebook by Nagesh Singh Chauhan, both found on Kaggle.

6. Evaluation Metrics

Because this type of forecasting works with continuous output that can be a big range of values, the metrics used in the project will be **MSE (mean square error)**, **RMSE (root mean square error)** and if needed **MAE (mean absolute error).**

7. Project Design:

The steps the project will follow are:

a. Gathering the Data;
b. Cleaning and Exploring the Data;
c. Selection of some companies to explore in-depth;
d. Uploading the data to S3;
e. Create, Training and Testing a DeepAR model;
f. Create, Training and Testing and customized LSTM model;
g. Comparing the results of the models;
h. Deploy the best one.

Since this is the capstone project for Machine Learning Engineer Nano degree, all work will be done using AWS and will focus more on working with SageMaker whether to build extremely great models.

REFERENCES

[1] https://en.wikipedia.org/wiki/Fundamental_analysis#The_two_analytical_models

[2] https://en.wikipedia.org/wiki/Technical_analysis