CVPR #01

Final Project: Machine Learning algorithm for Stock value prediction

CS 584 - Machine Learning - Fall 2022

Project Proposal

1. Introduction

1.1. Team members

The team will be made up of two members. On the one hand, ***Julen Ferro Bañales*** CWID:A20512110, currently studying the M.Sc. in Computational Decision Science & Operations Research. On the other hand, ***Eneko Gonzalez*** CWID: A20520157, currently studying the M.Eng. Computer Science.

1.2. Description of the problem

This project will deal with a really known problem in Finance called **'Stock prediction'**. Nowadays, the business environment is quite unstable due to different types of crisis that the world is suffering and the new context that has been given by the globalization. Therefore, the enterprises are forced to find different methods to fight against this uncertainty, in which there are not so used to work compared to the previous 'blue oceans' of economic competitiveness in which they have lived so far.

In order to sum up, this project will try to develop a tool valid for stock prices analysis and prediction. This way, the software that will be developed will try to train an **Artificial Intelligence** algorithm which tries to give an stock price output, based on past stock prices inputs. The insights given by the output of the software may be used as guidance for the consultants of the finance enterprises, due to the fact that the information got by the software will be in terms of likelihood and nothing will be assured.

1.3. Work done so far & project proposal

The authors of the project have just started to get information about the project. It has been a couple of weeks of research until finding an attractive topic and then another week gathering interesting links and bibliography related to the topic [7] [4] [6]. After having delivered the initial report for the project proposal, authors will start with the programming of the algorithm.

In relation to data, in the first instance the idea is to try to make use of systems for downloading and updating data in real time. However, for the early stages of development data-sets that are available in Kaggle [5] [3] will be used.

The election of the topic has been motivated by the idea of prioritizing the development of a hugely important and attractive project. Nevertheless, in order to create a more complex and useful project, originality, creativity, and innovation will be added in form of a more practical and real-use oriented implementation. The details of this aspect will be develop as progress is made.

1.4. Milestones

Oct 1, 2022: Initial Proposal
Oct 20, 2022: Algorithm programming finished
Nov 1, 2022: Intermediate project report (3-5 pages)
Nov 9, 2022: Tentative Final project presentation (5-8 minutes)
Dec 2, 2022: Final project report (8-10 pages)

2. Used data set

First of all, the data that has been used for the project must be described. The data set has been imported from Kaggle and it is comprised of an evolution of the stock price of the well-known enterprise Apple from the year 1980 to the year 2020.

The data set in matter is composed of more or less ten thousands of rows, that is to say, ten thousands of samples. It is quite a lot of information but the fact that almost forty years of stock price evolution is covered must be taken into account.

As attributes, the data set has eight different attributes described as follows: Date (the day in which the stock price is taken), Open (the price at which the first trade is made), High(the maximum price at which a trade is made in that period of time), Low(the minimum price at which a trade is made in the given period of time), Close(the price at which the last stock is trade in the given period of time) and the Adj Close (the same price as close, but after having paid the dividends).

As the most important price of the stock is the ***'Close Price'***, the Machine Learning model will be trained and tested with the ***'Close Price'*** data.

CVPR #01

3. Algorithm comparison

There are several valid Machine Learning models or algorithms for the well-known stock price prediction problem.

3.1. Simple Moving Average model

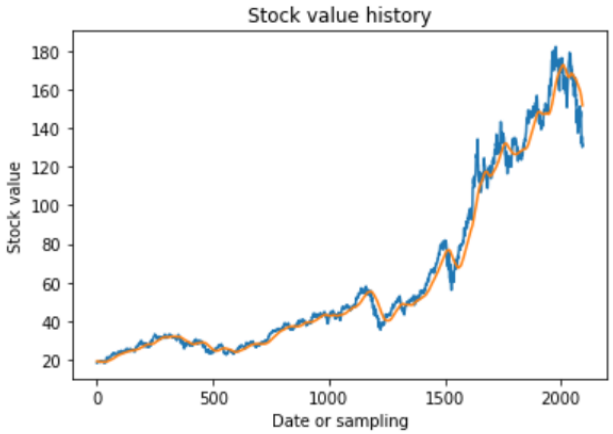
In the first place, the Simple Moving Average (SMA) algorithm, which is the simplest one and the most intuitive one. It gets a prediction of the stock price based on the weighting that has been given to the previous stock prices. The importance of the previous stock prices will be defined by the weighting that the previous events are given in the output or prediction formula [1] [**?**].

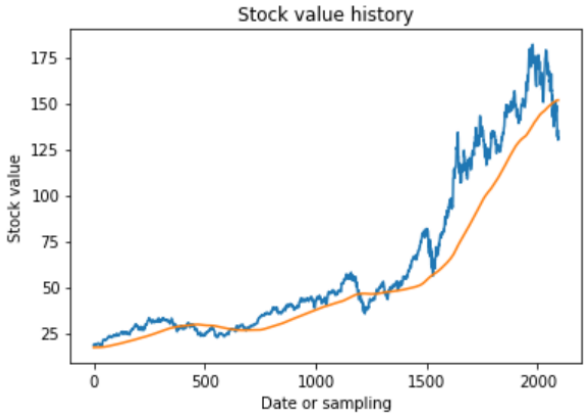
The predicted output will look like this equation.

$$SMA = \frac{P1 + P2 + ... + Pn}{N} \quad (1)$$

P1...Pn will be the values of the previous stock prices and N will be the window size. It must be taken into account that by varying the window size, more of the previous stock prices could be taken into account for the prediction calculation leading to a greater weighting of past events.

The simple average model has been implemented in a python program and has been simulated with the twenty percent of the data set. During this testing the next figures have been plotted respectively for the cases of window size = 50 and window size = 300.





By comparing to each other, it can be seen that the Simple Moving Average model with the lowest window number has fitted better into the data than the one with a window size of 300. Hence, it can be seen that with a lower number of window size the model becomes more flexible and it adapts better to the high variance scenario of the stock prediction issue. Finally, large numbers of window sizes clearly make the prediction curve smoother because averages a longer array of numbers leading this way to a flat profile.

3.2. Exponential Moving Average model

In addition to this, it also exists the Exponential Moving Average (EMA). This model is based on the same mathematics as the Moving Average, but it will slightly vary taking into account that in this model, the closest samples to the recent dates are weighted to a greater extent. That is to say, the events that occurred long time ago but within the window size will lose importance in this model, compared to the Simple Moving Average model. While in the Simple Moving Average model, only the previous N (window size) stock price values where taken into account at the same weight, in the Exponential Moving Average model, all the previous stock prices are taken into account but being weighted to a greater extent the closer the data gets to the predicted element [2].

In this model, the formula that gives the predicted value of the model is the next one.

$$EMA = P_t * k + EMA_(t-1) * (1 - k) \quad (2)$$

Where Pt is the price at time point t and N is the number of time points in EMA.

The weighting factor is calculated as follows:

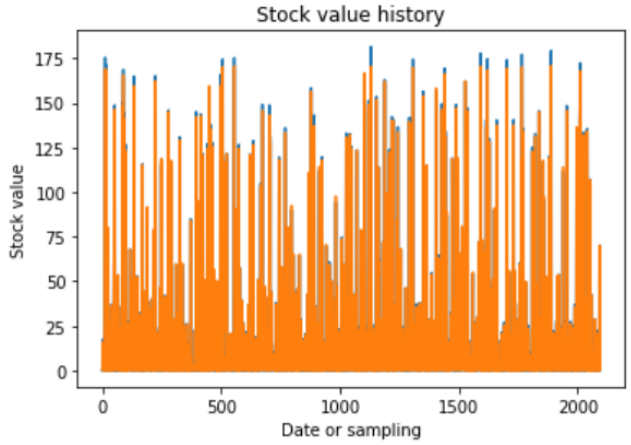
$$k = \frac{2}{N+1} \quad (3)$$

4. Long Short Term Memory

EXPLANATION OF THEORY

CVPR #01

PLOTTED GRAPH



5. Analysis of results

The chosen algorithm

```
66/66 [==============================] - 0s 3ms/step
THE MOVING AVERAGE MODEL with window size = 50
MAPE: 5.8
RMSE: 5.82


THE MOVING AVERAGE MODEL with window size = 300
MAPE: 16.19
RMSE: 15.94


THE LONG SHORT TERM MEMORY MODEL with window size = 50
MAPE: 61.57
RMSE: 1.18
```

6. Deriving theoretical properties of the algorithm

7. Conclusions

After having analysed how to predict the stock price values for the Apple company by using three different algorithms, several conclusions may be drawn.

In the first place, it has been demonstrated that the flexibility of the Long Short Term Memory Machine Learning algorithm has been the winner. It is able to fit the data better than other two methods, probably due to the fact that this method takes more into account the whole experience given by the data from the past. In the Moving Average methods, not the whole past events are taken into account, and even thought if sometimes are taken into account, the weighting gives more importance to the most recent ones. Therefore, one of the reasons of the Moving Average methods resulting in a worst solution, could be the seasonal character of the stock prices, due to the fact that Moving Average methods do not take this into account whereas the LSTM method represents better this kind of influence due to its flexibility.

In the second place, the Exponential Moving Average fits into the data better than the Simple Moving Average due to the fact that weights the used input data for the calculations by giving greater importance to the most recent samples. Anyway, if the Simple Average Method must be chosen, the window size number is a crucial parameter to be taken into account. From the analysis aforementioned, it has been drawn that smaller window size's adapt in a easier way to the curve, therefore withstanding high variance scenarios such as stock prediction. Finally, for flatter time series profiles, larger window size numbers can be used.

References

[1] Stock price prediction in finance. 2

[2] Stock price prediction with python. 2

[3] Sourav Banerjee. Nifty-50 stocks dataset, Jul 2022. 1

[4] Faryarmemon. S&P500 stock analysis for beginners, Jul 2022. 1

[5] Larxel. S&P 500 stocks (daily updated), Oct 2022. 1

[6] Liheng Zhang (University of Central Florida);Charu Aggarwal (IBM T. J. Watson Research Center);Guo-Jun Qi (University of Central Florida). Stock price prediction via discovering multi-frequency trading patterns. 1

[7] DataFlair Team. Stock price prediction - machine learning project in python, Aug 2021. 1