Data Scientist Capstone

Build a Market Price Indicator

Adel Idjeraoui

Mars 1, 2021

Introduction

Project Overview

Stock market have always been difficult to predict and even the most brilliant mind of all times had difficulties with the markets

> Newton allegedly said that he could "calculate the motions of the heavenly bodies, but not the madness of people." [1]

Stock Market prices are not only defined by companies fundamentals. Human behaviour can alter the true value of a company. The valuation of a company can be understood in two well defined kind of analyses the fundamental and the technical.

For our project we will focus on technical analysis and explore daily data provided by Yahoo about companies and try to predict the Adj Close.

Problem Statement

We want to build a stock price predictor that takes daily trading data over a certain date range as input, and outputs projected estimates.

By using Machine Learning algorithm the system will predict the Adjusted Close price.

We want to understand ML predictions on time series

Metrics

As we are working with time series and it's a regression problem. We will use the Root Mean Squared Error (RMSE) to measure our model capacity to predict the price.

Our final model should minimize the RMSE.

$$\mathrm{RMSE}_{fo} = \left[\sum_{i=1}^{N} \left(z_{f_i} - z_{o_i} \right)^2 / N \right]^{1/2}$$

Analysis

Data presentation

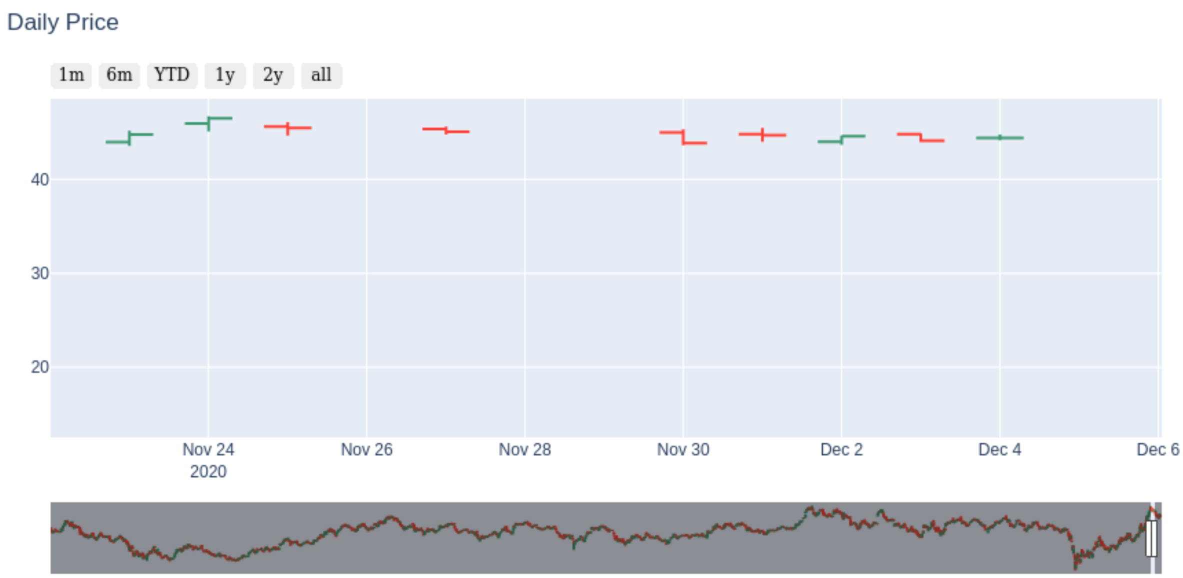
Our source for our data sets will be Yahoo Finance.

For each ticker we will have the daily data

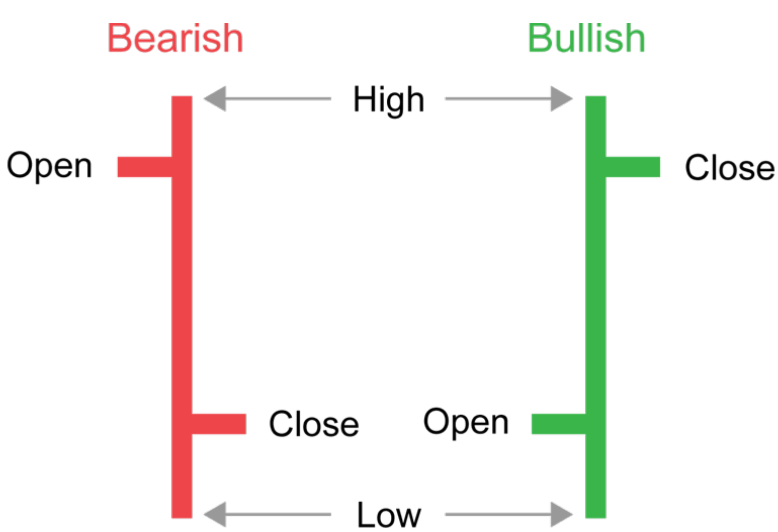
- Date – the date for each tradable day
- High – The highest price for that day
- Low – The highest price for that day
- Open – The price at the beginning of the trading session
- Close – The price at the end of the the trading session
- Volume - the amount of an asset or security that changes hands over some period of time, often over the course of a day.[3]
- Adjusted Close – The close price that take into consideration corporate actions (splits, dividends)

	Date	High	Low	Open	Close	Volume	Adj Close
0	2010-01-04	136.610001	133.139999	136.250000	133.899994	7599900	133.899994
1	2010-01-05	135.479996	131.809998	133.429993	134.690002	8851900	134.690002
2	2010-01-06	134.729996	131.649994	134.600006	132.250000	7178800	132.250000
3	2010-01-07	132.320007	128.800003	132.009995	130.000000	11030200	130.000000
4	2010-01-08	133.679993	129.029999	130.559998	133.520004	9830500	133.520004

We can visualize the data associated to the price with a Open-High-Close-Low(OHCL) chart.



Each day is represented by the 4 data points in the OHCL.



Missing values

Yahoo data’s don’t have missing values but we can see that we don’t have all the calendar days.

The main reasons are
- The markets are open on week days
- The markets are closed for holidays
- For exraordinary ocasions the markets can be closed

In general we can assume to have 252 trading day in a year.

Correlation

We don’t have a correlation between the volume and the price data points.

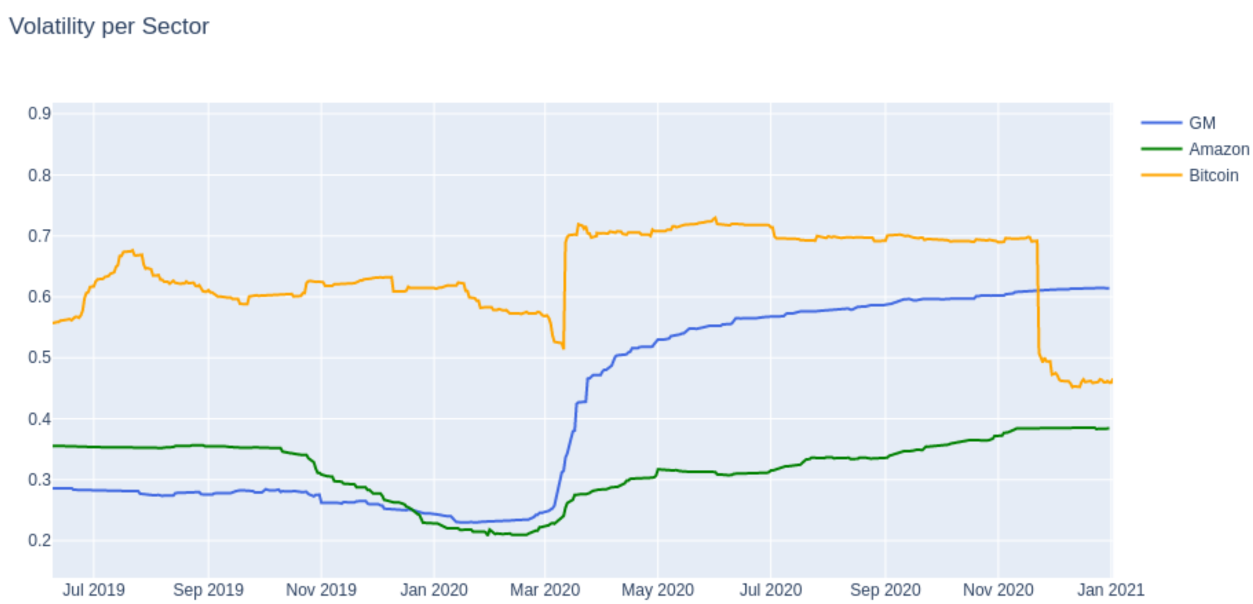
	High	Low	Open	Close	Volume	Adj Close
High	1.000000	0.997572	0.998279	0.998294	0.058785	0.918071
Low	0.997572	1.000000	0.997981	0.998273	0.031721	0.912532
Open	0.998279	0.997981	1.000000	0.996381	0.047877	0.913518
Close	0.998294	0.998273	0.996381	1.000000	0.043946	0.916404
Volume	0.058785	0.031721	0.047877	0.043946	1.000000	-0.019975
Adj Close	0.918071	0.912532	0.913518	0.916404	-0.019975	1.000000

Are some sectors more predictable ?

For the project we will analyses data for multiple type of ticker available from Yahoo Finance. To be able to understand the differences between the predictions of our models.

Our ticker of interest

- General Motors (GM), an industrial company
- Amazon (AMZN), a tech company that is in a volatile industry
- Bitcoin (BTC-USD), a cryptocurrency in a highly volatile market



Results

Model Evaluation and Validation

We will start with an LSTM model.

Trying to use a different optimizer didn’t improved the model predictions. Ada optimizer is the most suitable optimizer for our time-series predictions.

We optimized on the number of epoch and the number of hidden dimensions

hidden dimensions, epochs	Train Score (RMSE)	Test Score (RMSE)
32,250	26.01	31.49
64,250	22.02	25.64
64 , 500	**20.13**	**23.52**
64 , 1000	19.68	23.90
128,500	19.68	24.72
128,1000	19.71	23.60

Increasing the number of epoch to 1000 and the number of hidden dimension over 64 give us no gain on the test data. We were overfitting.

Let’s see the result on different tickers for the **LSTM** model

Ticker	2015-2018	2015-2019	2015-2020
GM (General Motors)	Train Score: 0.44 RMSE Test Score: 0.70 RMSE	Train Score: 0.51 RMSE Test Score: 0.57 RMSE	Train Score: 0.51 RMSE Test Score: 0.93 RMSE
AMZN (Amazon)	Train Score: 15.93 RMSE Test Score: 53.76 RMSE	Train Score: 21.77 RMSE Test Score: 25.39 RMSE	Train Score: 33.09 RMSE Test Score: 154.03 RMSE
BTC-USD (Bitcoin)	Train Score: 282.68 RMSE Test Score: 324.10 RMSE	Train Score: 297.83 RMSE Test Score: 352.02 RMSE	Train Score: 381.18 RMSE Test Score: 962.07 RMSE

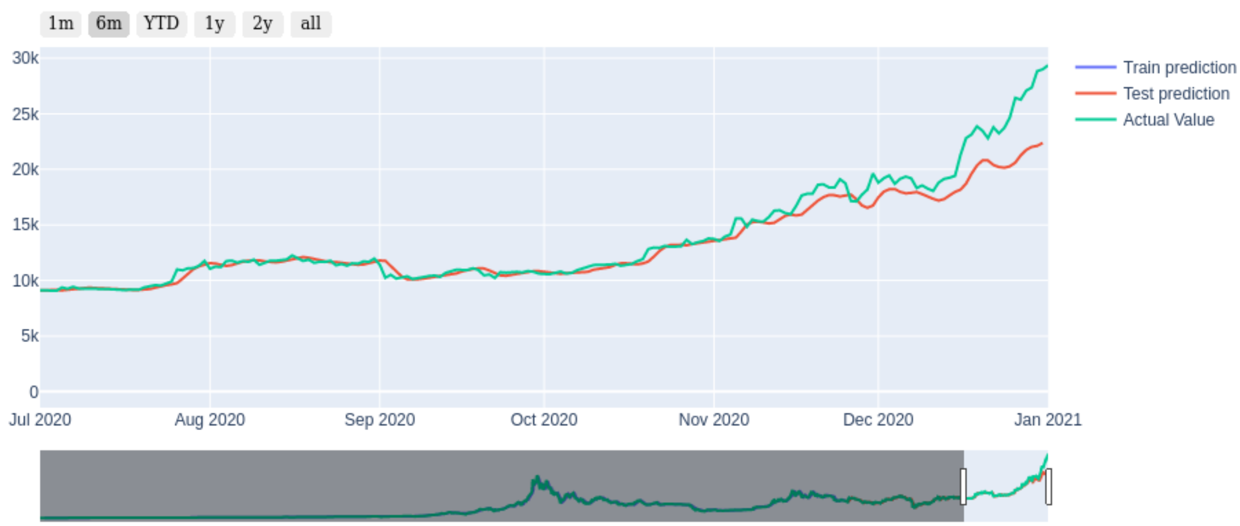
We will try to improve our predictions by using an **GRU** model

Ticker	2015-2018	2015-2019	2015-2020
GM (General Motors)	Train Score: 0.45 RMSE Test Score: 0.70 RMSE	Train Score: 0.50 RMSE Test Score: 0.55 RMSE	Train Score: 0.51 RMSE Test Score: 0.94 RMSE
AMZN (Amazon)	Train Score: 12.87 RMSE Test Score: 42.75 RMSE	Train Score: 21.18 RMSE Test Score: 25.01 RMSE	Train Score: 22.90 RMSE Test Score: 121.90 RMSE
BTC-USD (Bitcoin)	Train Score: 280.35 RMSE Test Score: 254.92 RMSE	Train Score: 269.55 RMSE Test Score: 329.19 RMSE	Train Score: 294.20 RMSE Test Score: 542.49 RMSE

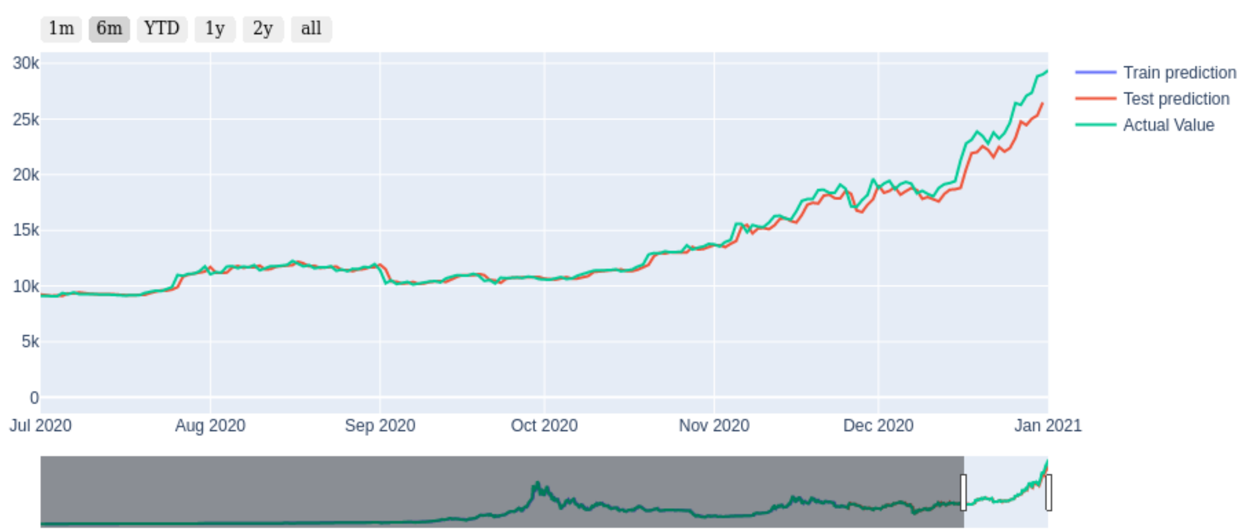
Justification

For GM the two models have similar results. The GRU model has better results for Amazon and Bitcoin for each period. But as the volatility increase we had extreme results for both.

LSTM



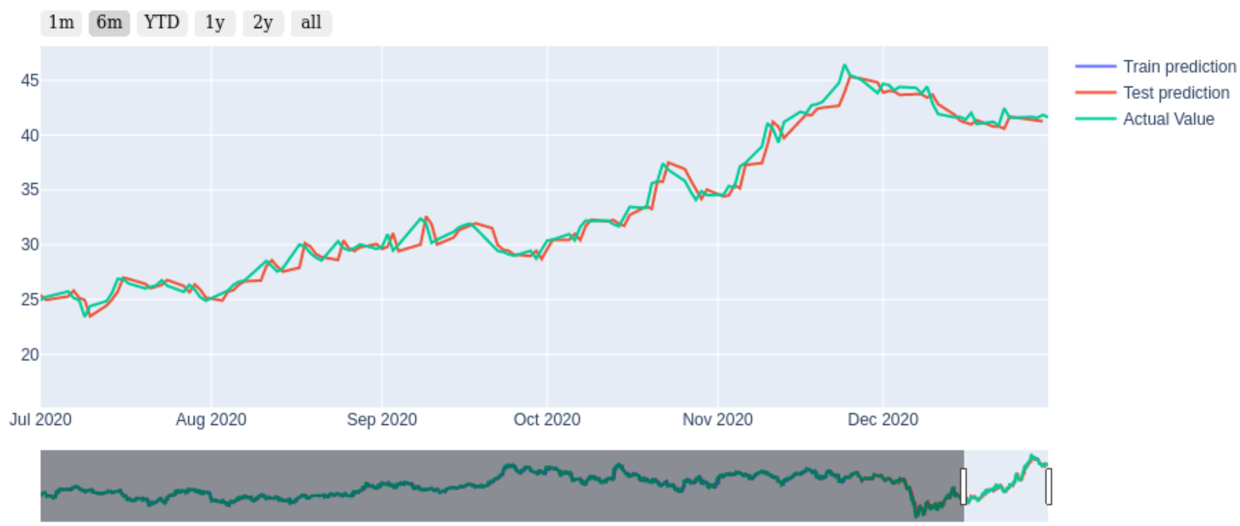
GRU



The model were not able to predict Bitcoin price as well as GM's price. The intrinsic attribute of Bitcoin makes it less predictable with only the price as input.

GM on the other hand is in a more stable industry, we had better results over all the period tested.

GRU – GM - 2015-2020



Amazon is between the two. It's a well established company, but is still in the tech industry.

Conclusion

Reflection

The stock market is an art more than a science. Patterns don’t seems to emerge form price data. We need to acknowledge some element of behavioural economic to understand the dynamic of the market.

We mainly focused on technical analysis to build our models and only used the price as input.

The financial domain is a difficult domain and a more accurate understanding of the domain would have helped use more metrics to build the models.

Improvement

Many improvements could be considered for the models we have built.

Taking into consideration multiple components from the technical analysis field. As example adding the volume and metrics derived from the volume. By building multiple features we could explore the impact of each one on the model.

To truly aims to predict the price we will also have to incorporate fundamental components to our model. We could explore to potential to use NLP to have a sentiment analysis on the companies declarations.

Trying to predict the price it self might be a difficult problem for ML. As Newton we might not be able to use science to predict the market.

References

1- Newtown Citation : https://physicstoday.scitation.org/doi/10.1063/PT.3.4521

2- Volume definition : https://www.investopedia.com/terms/v/volume.asp