PERSONAL PROJECT

FORECASTING AND ANALYZING THE FINANCIAL MARKET

Developer: Marcos Matheus de Paiva Silva

Barra Mansa - RJ

May 2022

Contents

1 Objective

Several factors have influenced the stock market such as the news, the quotes themselves, the reports that companies carry out, among others. Although these questions are extremely relevant, nowadays other conditions are influencing the stock market, such as data analysis and machine learning algorithms.

Thus, in the current project we fully contemplate the stages of a science and data analysis project. Therefore, from the information consumed from the Data Reader api, which is collected on the Stooq and from data visualization we can probe some questions that would help users, companies and shareholders to understand about:

- The forecast of the value of the shares on the aple stock exchange;
- In which month the stocks are higher and what would that value be;
- On which day did the highest share prices occur;
- Which day to watch out for to maximize the probability of stock trading.

In addition to these information that we can obtain with the project, it is possible to extend such questions to other stock markets using the Data Reader. In the end, after creating the model, we did not implement it, because one of the good practices of implementing temporary series is every time you have new data, you must train the model, as the hosting we usually use is free with limitations we chose not to implement the model in the cloud or even serialize it.

2 Proposed Solution

This project was developed following the steps in (GéRON, 2019), and also based on some phases in (MüLLER; GUID, 2016). Among other references, we can mention the articles (SAK; SENIOR; BEAUFAYS, 2014) and (ZAREMBA; SUTSKEVER; VINYALS, 2014). To find the answers to the questions raised in the previous section, we first used the Google Colaboratory so that any individual can run the code in python and also better understand all the theoretical argument, insights and methodology used in the project.

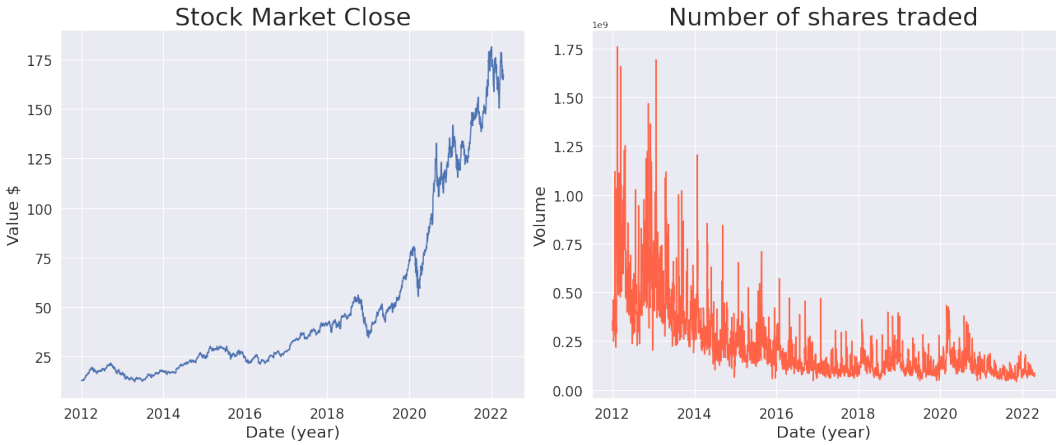
The first step of the project was to collect data from the Stooq website through the Data Reader api. After this procedure, a jupyter notebook was used in which the libraries to operate with pre-processing, data analysis, graphics, machine learning and statistical inference were:

- scikit-Learn, imblearn and pandas-datareader;
- pandas, statsmodels and numpy;
- matplotlib;
- Keras and Tensorflow.

Furthermore, after creating some machine learning models in Google Collaboratory we compared them to understand which one had better accuracy and better performance.

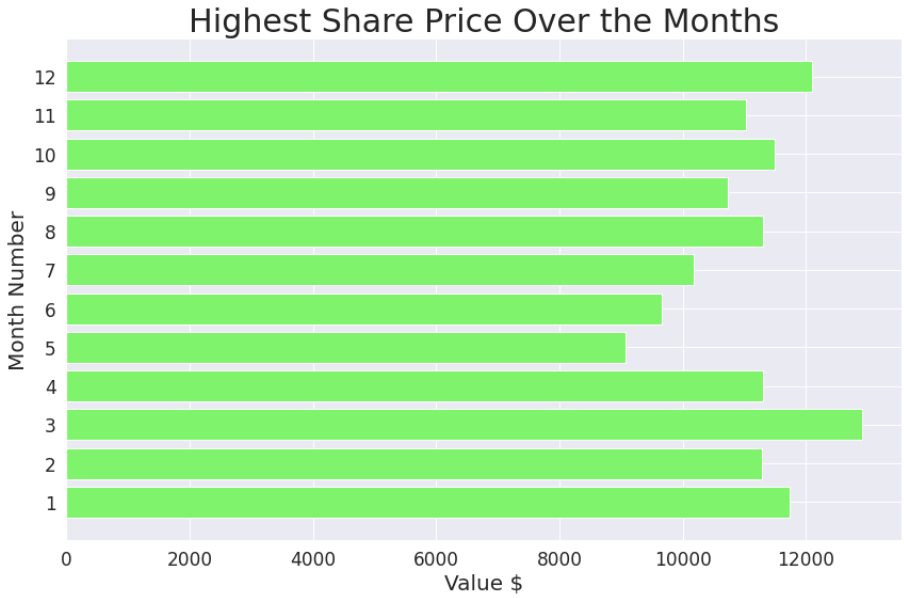
3 Results

To answer business questions, in the data visualization stage we plot a graph:



Source: Author

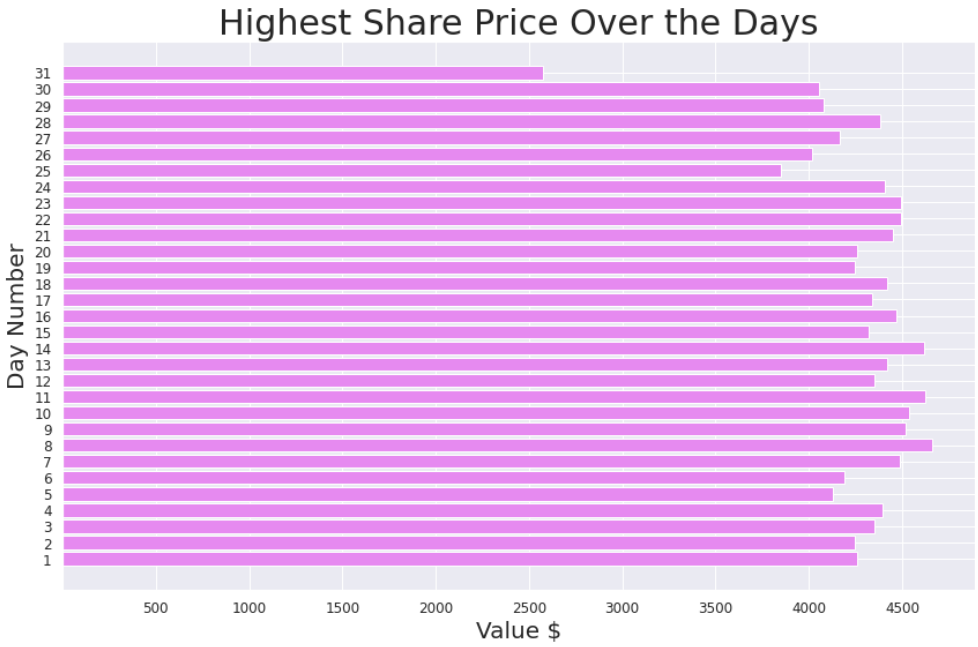
Thus, in these two graphs we see that the value of the closing of a deal grows over time, while the volume of business decreases over time. To know in which month the stocks have the biggest gains, we draw the following chart:



Source: Author

Based on this chart, we understand that the month in which the shares had the highest value was March, whose value is $ 12911, 249, followed by December. The graphs of the other columns in the dataset are very similar. We also found an exorbitant March turnover of $ 61588607525.

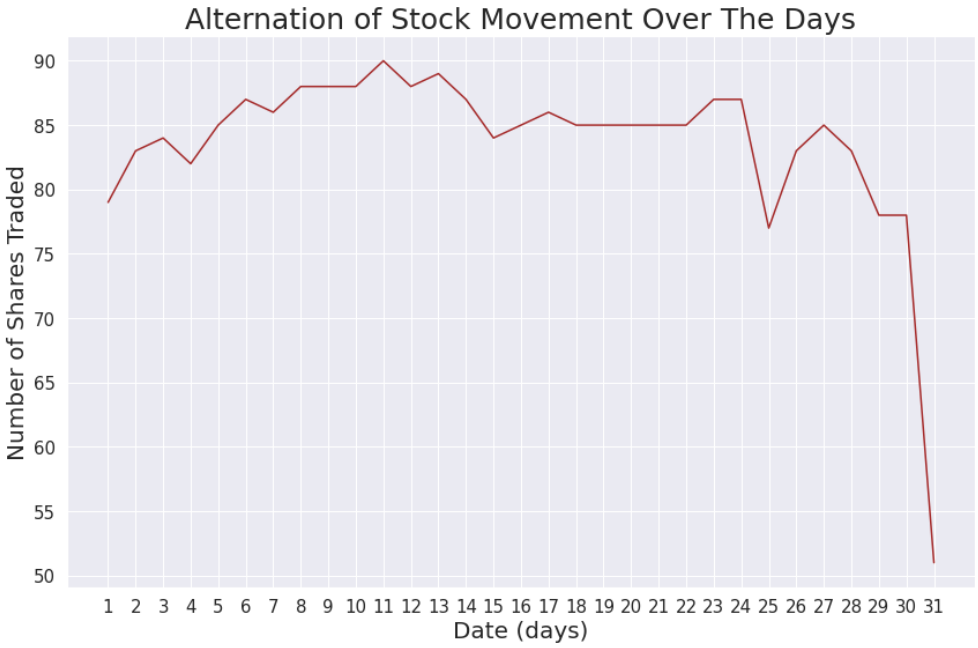
At the end of the data visualization we seek two more answers. First we try to understand on which day the stock values were highest with the chart:



Source: Author

We understand that the day on which the highest trading value of a share was obtained was the 8th, this value was above $4500.

At the end of the data visualization stage, we would like to know which day we should maximize the probability of trading a stock:



Source: Author

This last chart was created by grouping trading days and volume, peak trading volume occurred between the 8th and 13th, peaking on the 11th, so it is extremely important for shareholders to be aware during this range of days. It is worth mentioning that the trading volume was high in this period, which means that you can have lows and highs of stocks on these days.

After going through several stages of the project, we showed the autocorrelation of the series, seasonality and among other factors, the most important of these processes was a statistical inference that we performed, called the ADF test. Basically the ADF test will tell us if a series is stationary or not, from our test we found that our series is not stationary, meaning that the statistical properties of the series are not preserved, we only used this information later in the arima model.

For the end, after having done all the treatment on the data, we trained three models from neural networks or from statistical methods, so the three models were, respectively, SimpleRNN, LSTM, ARIMA. SimpleRNN served as a baseline mainly for the LSTM model, as arima uses another configuration. Of these models, the two best were the LSTM and ARIMA, due to their accuracy in the test data, although the Arima has better accuracy and performance, it seems to overfit, I do not discuss values here, because as in the case of the LSTM, such accuracy changes from Sometimes. Although the ARIMA model seems to be overfitted, it seems that the arima method is the best found, such method is statistical and uses the most elementary AutoRegressive Moving Average extension that includes the concept of integration.

4 Conclusion

In the course of this project, we developed relevant concepts in science and data analysis, in order to suggest the solution to know what will be the value of the shares of an investment exchange at the end of the day. In addition, we seek to discover some market information through data visualization. Thus, in future applications it will be possible to carry out the implementation and compare other stock exchanges if we have adequate accommodation, helping customers and shareholders to find investing in what they want.

We went through several precise steps to extract as much information as possible from our dataset. As a result, initially we understand the behavior of the market that stops the closing price whose time series had an upward behavior and we also understand that the trading volume has been decreasing over the years. From there, we used the grouping method to obtain information from the data, with this we saw that the highest increases in shares took place in March and the highest trading volume as well. Using a similar procedure to find out which day had the highest increase, we found a value of $ 4500 on the 8th. We also saw that the peak in business volume occurred on the 8th and 13th with the peak on the 11th.

In summary, we performed a statistical inference to find out if the time series is non-stationary and this statement was true. With that in hand, after training two models of neural networks, we arrived at the most effective and highest performing method, called ARIMA. The fact that this happens probably happens because in the arima method we must specify parameters that mitigate the trend and seasonality, which harm the method making the time series stationary, which is much easier to work with.

Bibliography

GéRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2ª. ed. GCanada: O'Reilly, 2019. Citado na página 3.

MüLLER, A. C.; GUID, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. 1ª. ed. United States of America: O'Reilly, 2016. Citado na página 3.

SAK, H.; SENIOR, A.; BEAUFAYS, F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv, https://arxiv.org/abs/1402.1128, 2014. Citado na página 3.

ZAREMBA, W.; SUTSKEVER, I.; VINYALS, O. Recurrent neural network regularization. arXiv, https://arxiv.org/abs/1409.2329, 2014. Citado na página 3.