

Stock Market Prediction Proposal

Requirements

- The project's **domain background** — the field of research where the project is derived;
- A **problem statement** — a problem being investigated for which a solution will be defined;
- The **datasets and inputs** — data or inputs being used for the problem;
- A **solution statement** — the solution proposed for the problem given;
- A **benchmark model** — some simple or historical model or result to compare the defined solution to;
- A set of **evaluation metrics** — functional representations for how the solution can be measured;
- An outline of the **project design** — how the solution will be developed and results obtained.

Domain Background

There has been a lot of interest in stock trading lately that's been triggered by COVID and lockdown amongst other reasons. Many analysts and traders that I talk with feel that the market is more unpredictable mpw since many of the new traders do not follow the normal routine used by professionals e.g. researching financial statements, technical analysis ..etc. and their trading strategies are driven by either passion, gut feeling, hype, or sometimes following the crowd as in the example of Game Stop.

There are many interesting research papers on stock prediction covering Machine Learning, Deep. Learning and other techniques that inspire this project.

- Predicting the direction of stock market prices using Random Forest <https://arxiv.org/pdf/1605.00003.pdf>
- Stock Price Forecasting in Presence of COVID-19 Pandemic and Evaluating Performance of Machine Learning Models for Time Series Forecasting <https://arxiv.org/pdf/2105.02785.pdf>
- A Comparative study of Different Machine Learning Regressors For Stock Market Prediction <https://arxiv.org/pdf/2104.07469.pdf>
- Accurate Stock Price Forecasting Using Robust and Optimized Deep Learning Models <https://arxiv.org/pdf/2103.15096.pdf>
- Forecasting with Deep Learning (S&P500) <https://arxiv.org/pdf/2103.14080.pdf>
- Support Vector Regression Parameters Optimization using Golden Sine Algorithm and its application in Stock Market <https://arxiv.org/pdf/2103.11459.pdf>
- Stock Price Forecasting with Deep Learning <https://arxiv.org/pdf/2103.14081.pdf>

Problem Statement

Can statistical models and machine learning predict future stock prices? The project will investigate and attempt to answer this question to determine if stock data contains enough signals to allow for accurate predictions.

Datasets and Inputs

I will use an API to pull stock data, more specifically Pandas DataReader. The data contains daily stock prices in the form of opening price, closing price, adjusted closing price, high trading price of the day, low trading price of the day, and volume.

The datasets will cover data from 2016-2021 split into training and testing. Training from 2016-2019 to predict 2020, then another round including 2020 in the training to predict 2021.

- Strong Performers (2020) portfolio = ['TSLA', 'ZM', 'MRNA', 'AMZN', 'NFLX', 'NVDA', 'AAPL', 'GME']
- Poor Performers in 2020 portfolio = ['CCL', 'MRO', 'UAL', 'SLB', 'OKE', 'FARM', 'GLBS']
- S&P500

The idea is to focus on those that went up and stayed up during and post COVID and those that plummeted. This is to determine if the impact of COVID was more of an accelerator meaning those who were doing good and destined to rise it speed it for them, and those who were destined to fail just failed much faster during COVID.

Lastly, the goal is to forecast forward 10-15 trading days (2-3 weeks).

Solution Statement

In this project I will use classical time series forecasting approaches such as ARIMA and GARCH to predict volatility and price. Additionally results will be compared across different approaches towards forecasting including the use of Machine Learning and Deep Learning.

Benchmark Model

A benchmark model would be a **Linear Regression**, given how simplistic it is that it would be an ideal choice. This will be used to compare the performance of other models against this benchmark (based on the evaluation metrics described below).

Evaluation Metrics

Ideally, I will train on time slices like 2016-2019 (maybe extend to 2020), then attempt to forecast into 2020 or 2021 (testing). The metrics that I will use for evaluation will be:

- Root Mean Square Error (RMSE)
- R Squared
- and Mean Absolute Percentage Error (MAPE)
- Forecast Skill (SS)

Complexity

Stock market prediction is not an easy task and is an advanced topic of research. But I am hoping this attempt would be a baseline or foundation for further investigation and fine tuning.

Process

For stock market prediction/forecasting I will use:

- Data Collection using Pandas DataReader and AlphaVantage API
- Data Preprocessing to prepare the data depending on the model used. Some models like ARMA expect the data to be stationary, while others do not. Also, standardization will need to occur as well.
- The data will split between training and testing. Training data will contain data from 2016-2019 then use the model to forecast 2020. Even though 2020 is an exception given the COVID situation it will give a good baseline model. Then change the training data to include 2020 in the training set and test against 2021 stock prices.

Models that I expect to test:

- Time Series: ARMA, ARIMA, and GARCH
- ML: Random Forest and other ensemble tree-based regressors
- DL: A RNN/LSTM for Time Series data
- Finally test Facebook Prophet library as a forecasting tool

Model Deployment

I will deploy the selected model using Streamlit so users can interact and pick a stock (symbol) and then have the model create predict future prices (1 week or 2 week look ahead forecast).