Stock Data Analysis And Prediction

Fall 2023 CSGY-6513-D, Big Data

Ujjwal Narain Asthana (una207),
Vivek Nayak (vgn2004),
Prasanna Sai Puvvada (pp2689)

Github Code:

FrontEnd: https://github.com/pprasannasai/StockPredictionDashboards
BackEnd: https://github.com/unasthana/StockAnalysisAndPrediction

Project Statement

In the contemporary landscape of financial markets, the ability to make timely and informed investment decisions is a critical challenge. The global stock market, characterized by its dynamic nature, is influenced by an intricate web of factors, such as economic conditions, political events, corporate earnings, market sentiment and so on. These variables, along with the inherent volatility of stock prices, make it exceptionally challenging for a potential investor to make investment decisions in the stock market.

Objective

- Develop dashboards for visualizing, grouping and analyzing stocks based on different metrics.
- Implement predictive models to predict future stock prices.
- Help users diversify their portfolio by clustering stocks into different categories based on different degrees of risk and return.

Data Source

The dataset employed for this project is titled "S&P 500 Stock Data."

It encompasses the daily trading information of all the stocks included in the S&P 500 Index Fund, with data entries spanning from February 2013 to February 2018.

1) Size: Approximately 58MB when uncompressed.

2) Number of Rows: Approximately 619,000.

3) Number of Columns: 7.

4) Dataset link: https://www.kaggle.com/datasets/camnugent/sandp500/

The columns are:

1) Open Price
2) Close Price
3) High Price
4) Low Price
5) Volume of shares traded
6) Stock Ticker
7) Date

Technology - Model Architecture and Storage

Machine Learning (K-Means Clustering)

- For clustering stocks together according to varying degrees of Market Risk and Stock Returns.

Deep Learning (Tensorflow)

- For predicting future stock prices, an LSTM based Neural Network is trained for each stock.
- Tensorflow offers HDF5 model saves that support weight compression, thereby leaving a smaller memory footprint.

MongoDB

- LSTM model's HDF5 files were around 0.5 MB each (500 of them).
- Small file sizes implies we do not need to use object-storage solutions like S3.
- 250 MB total size, fit well within M0 limit of 500 MB.

GridFS

- GridFS allows us to stream model files from MongoDB
- Metadata can be used to tag the ticker name, which helps in quick search

Technology - Backend

Django

- Django Admin dashboard, serializers and ORM layer are useful for quick setup.
- For complex data wrangling, pandas was used; integrates well with other Python code.
- Python backend plays well with other python libraries required for this project, like tensorflow.

PostgreSQL (prod), SQLite (dev)

- On local, SQLite ensures quick development and easy data sanity checking on the admin dashboard.
- In production, Heroku integrates well with PostgreSQL (plug-in) and does not play well with SQLite.
- PostgreSQL is more performant and scalable than SQLite.

Heroku

- Free tier allows apps with size upto 500 MB.
- Used ‘tensorflow-cpu’ instead of ‘tensorflow’ to stay within memory bounds.

Technology - Frontend

Streamlit

- Rapid Prototyping: Allows for quick development and iteration.
- Ease of Use: Intuitive syntax and low learning curve, ideal for data scientists
- Interactive Visualizations: Seamless integration with libraries like Matplotlib and Plotly for data representation.

Analysis

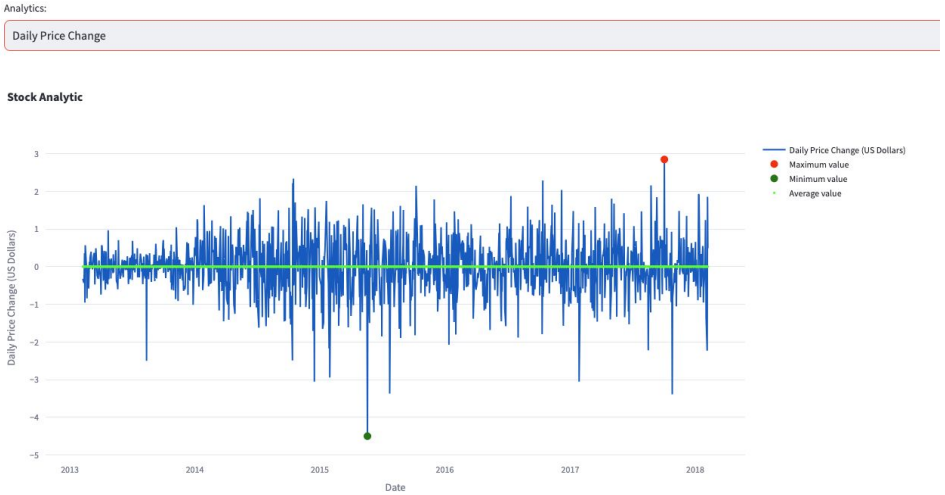
Daily Returns of given Stock:

- Find Daily Returns for each Stock with respect to close price and for a particular time frame.
- Daily Returns is the percentage change in the close price of a stock in successive periods.
- It helps model performance comparison, volatility measurement and so on.

Analysis

Daily Price Change of given Stock:

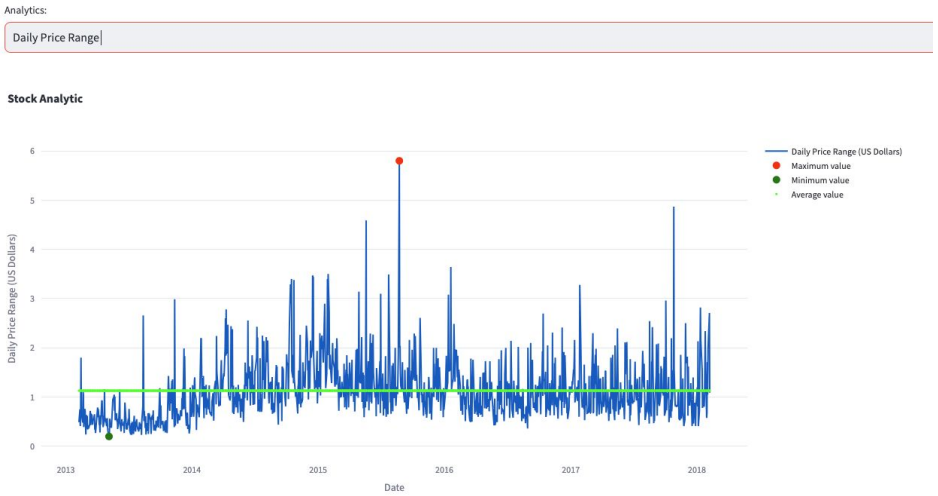
- Find daily price change for each Stock for a particular time frame.
- Daily Price Change = Close Price - Open Price
- It helps model market direction, volatility measurement among other things.



Analysis

Daily Price Range of given Stock:

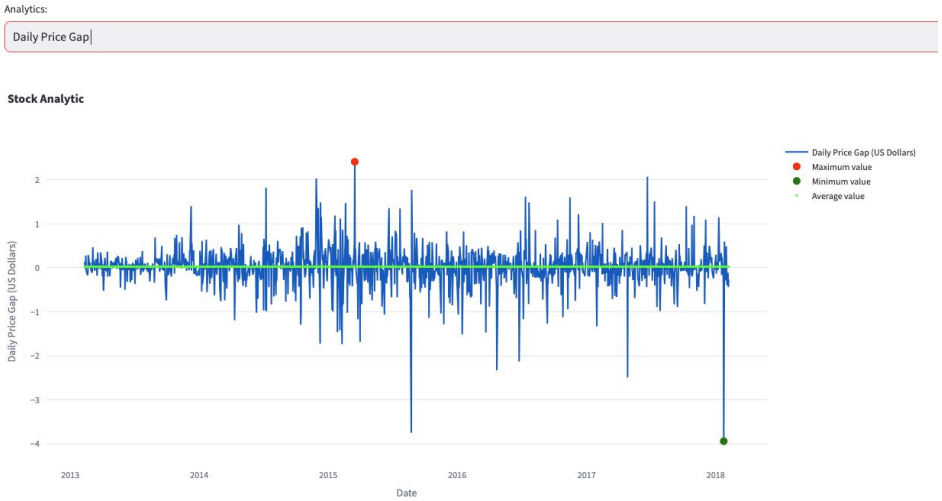
- Find daily price range for each Stock for a particular time frame.
- Daily Price Range = High Price - Low Price
- Again, this comes in very handy in volatility analysis.



Analysis

Daily Price Gap Analysis of given Stock:

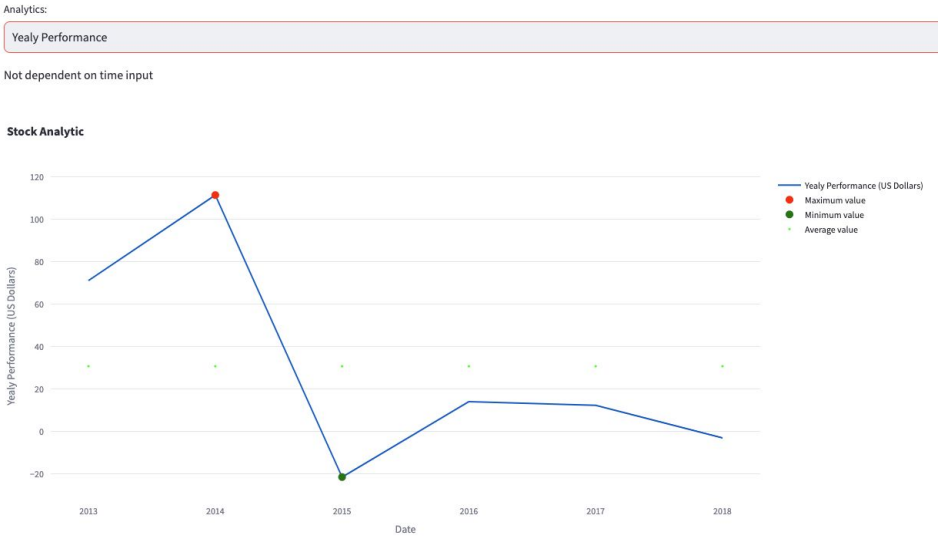
- Find Daily Price Gap for each Stock for a particular time frame.
- Daily Price Gap = Open Price - Previous Day's Close Price
- This helps in finding price momentum, overnight market sentiment, trend confirmation and so on.



Analysis

Yearly Performance Analysis of given Stock:

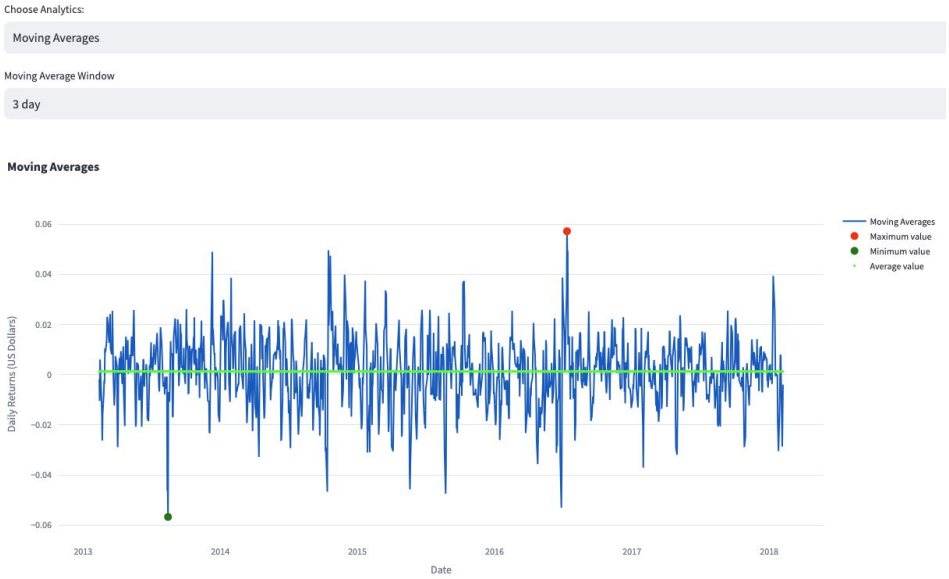
- Find yearly Performance for each Stock.
- Yearly Performance = Percentage Change in close price between the start and end of the year.
- Good metric to find investment returns, Performance comparison, etc.



Analysis

Moving Averages of given Analytic of a given Stock:

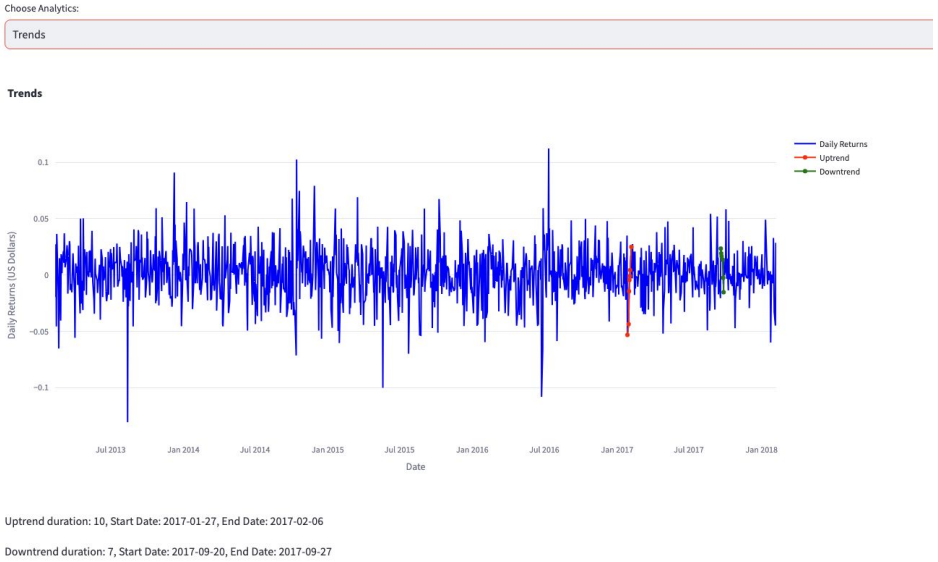
- Based on the analytics mentioned before what are moving averages of different stocks for a particular window of time.
- Very useful in smoothing out day-to-day volatile changes in different analytic measurements.
- Makes the trend emerge out in a more clear way.



Analysis

Trend Analysis of given Stock:

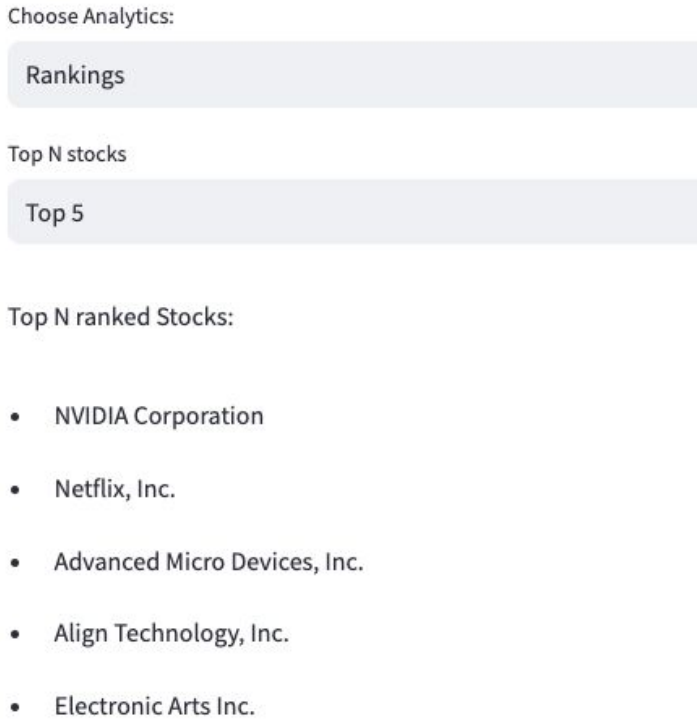
- When was the longest continuous uptrend or downtrend and the duration of that trend in different analytics mentioned above?
- Choose the analytic and moving average window to get the trends.



Analysis

Rank Analysis of given Stock:

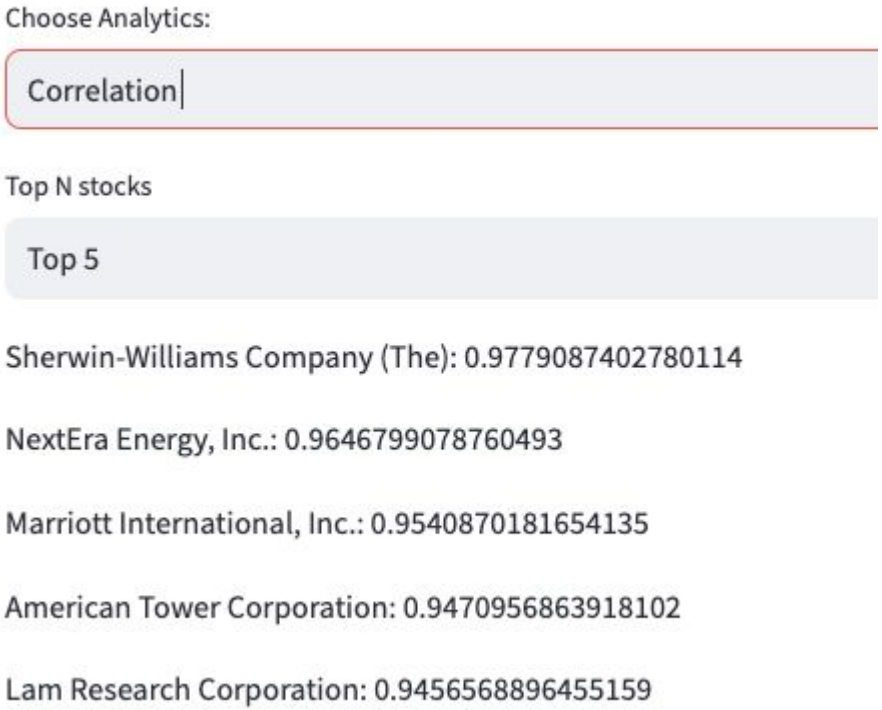
- What are the rankings of stocks based on different analytics, such as Daily Returns, Price Change Statistics, Volume Analysis, Performance, Price Range Statistics and Gap Analysis?
- Choose the analytic to get the rankings.



Analysis

Correlation Analysis of given Stock:

- How are stocks correlated with each other?



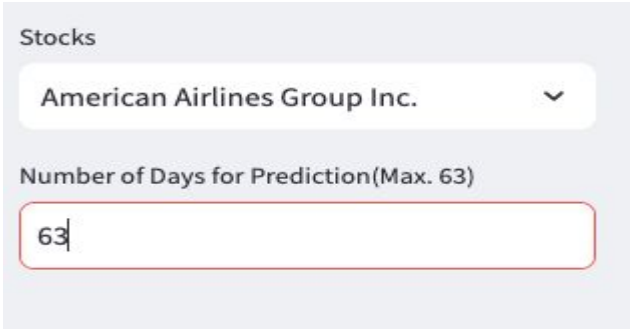
Stock Prediction

This dashboard would answer the following questions:

- What is the predicted price for the selected stock for the next ‘N’ days?
- Visualize the price movement in order to help the customer to make an informed trading

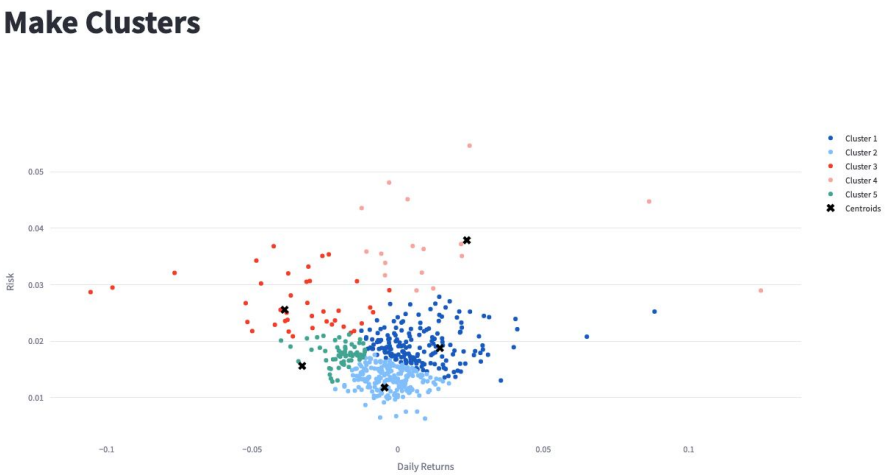
decision.

- LSTM Model Training for stock price prediction.



Cluster Prediction

- Stocks are grouped into 5 clusters based on Market Risk and Stock Return.
- Helps user diversify his portfolio.
- Cluster 1 represents some stocks which have high returns and low risk. (The Ideal Ones!)
- Cluster 3 represents some stocks negative returns and high risk. (The Bad Ones!)



Conclusion

Achieved a sophisticated level of analysis, by integrating various advanced statistical and machine learning techniques, each selected for its particular strength in handling the complexities of stock market data.

Lessons Learned

- Understanding Financial Markets: It's crucial to have a basic understanding of how financial markets work, including factors that affect stock prices.
- Use of Advanced Analytics and Machine Learning: Predicting stock prices often involves complex algorithms and machine learning models.
- User Experience and Interface Design: The application's interface should be user-friendly. Presenting complex stock data in a simple, understandable format is a key challenge.

Future Work

- Making front-end UI more intuitive for the end user.
- Using state-of-the-art Risk calculation methods for clustering stocks.
- Implement cluster number optimization.
- Implementing more efficient caches for different analytics.
- Design a Reinforcement Learning based Portfolio Optimizer.
- Implement Sector based analytics.

Demo

Thank you!

Q&A