Stonks

By: Pranay and Dom

Project Description

The Goal:

- Determine the **best stock market sector** to invest till March 2023 based on past performance.

The Why:

- Investing in the stock market can be risky, especially investing into a single company. However, this risk can be mitigated by investing into a sector instead of an individual stock.

The Sectors

1. Industrials
2. Financials
3. Health Care
4. Consumer Discretionary
5. Consumer Staples
6. Real Estate
7. Utilities
8. Materials
9. Communication Services
10. Energy
11. Information Technology

Similar Works

https://www.geeksforgeeks.org/stock-price-prediction-using-machine-learning-in-python/

https://towardsdatascience.com/predicting-stock-prices-using-a-keras-lstm-model-4225457f0233

They are predicting individual prices, we will predict sectors wise, by creating an index.

Obtaining the Data

Sources

1. List of S&P 500 companies and the sectors they belong to: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies
2. Outstanding shares for each company: https://companiesmarketcap.com/
3. Daily share price for each company: https://pypi.org/project/yfinance/

Extraction Methods

1. Downloaded CSV
2. Web-scraping with Selenium
3. Yahoo Finance API

Feature Engineering the Data

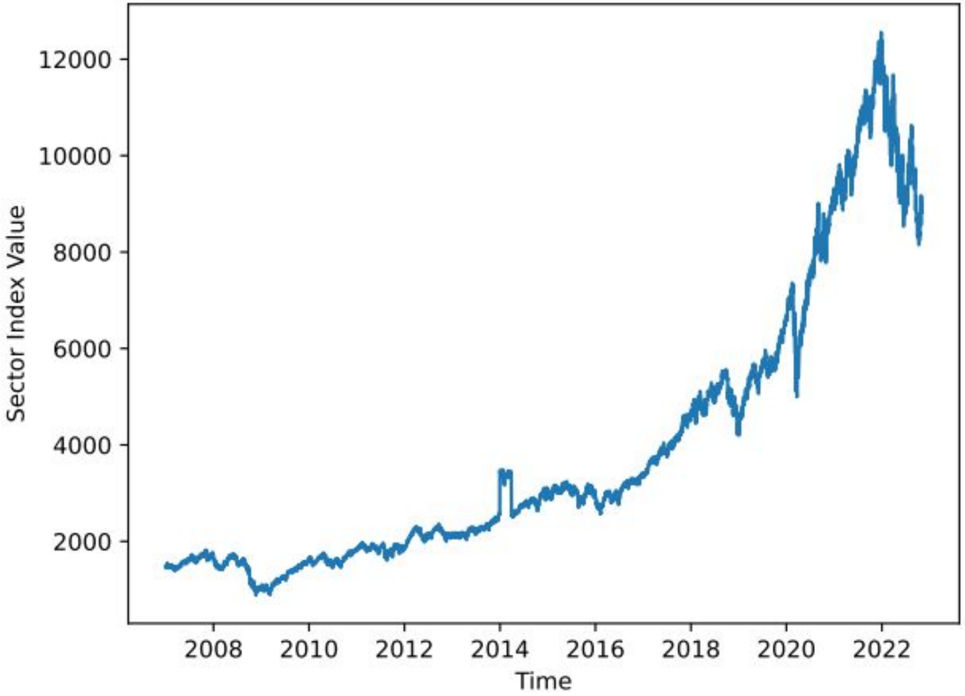
Using the data we gathered for the major companies in each sector, we built an **index** for each sector

** index = measures the price performance of a group of securities/companies*

Features we engineered to build the index:

- **Market Cap** = Closing price * Shares Outstanding
- **Total Market Cap** = sum(Market Cap) group by date
- **Index** = (Total Market Cap / 1000)
- **Index Weigh**t = (Market Cap / Total Market Cap) * 100
- **Volatility** = ((Close - Open) / Open) * 100

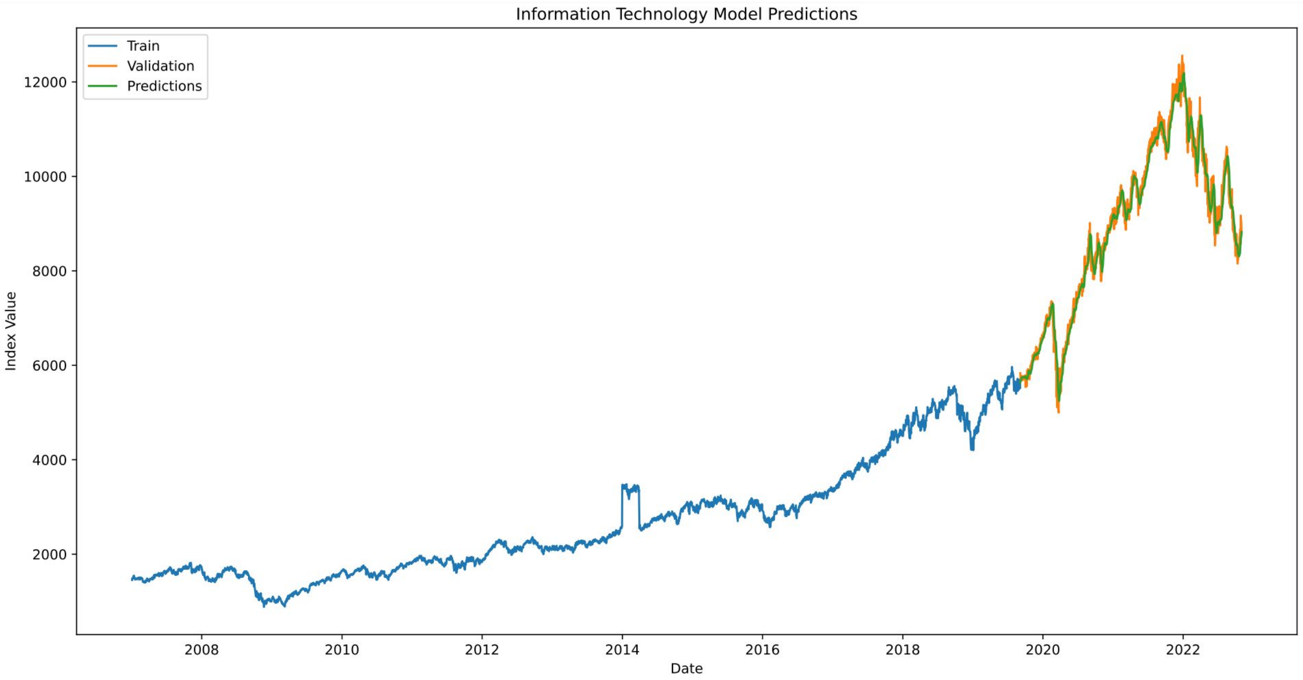
Information Technology Index



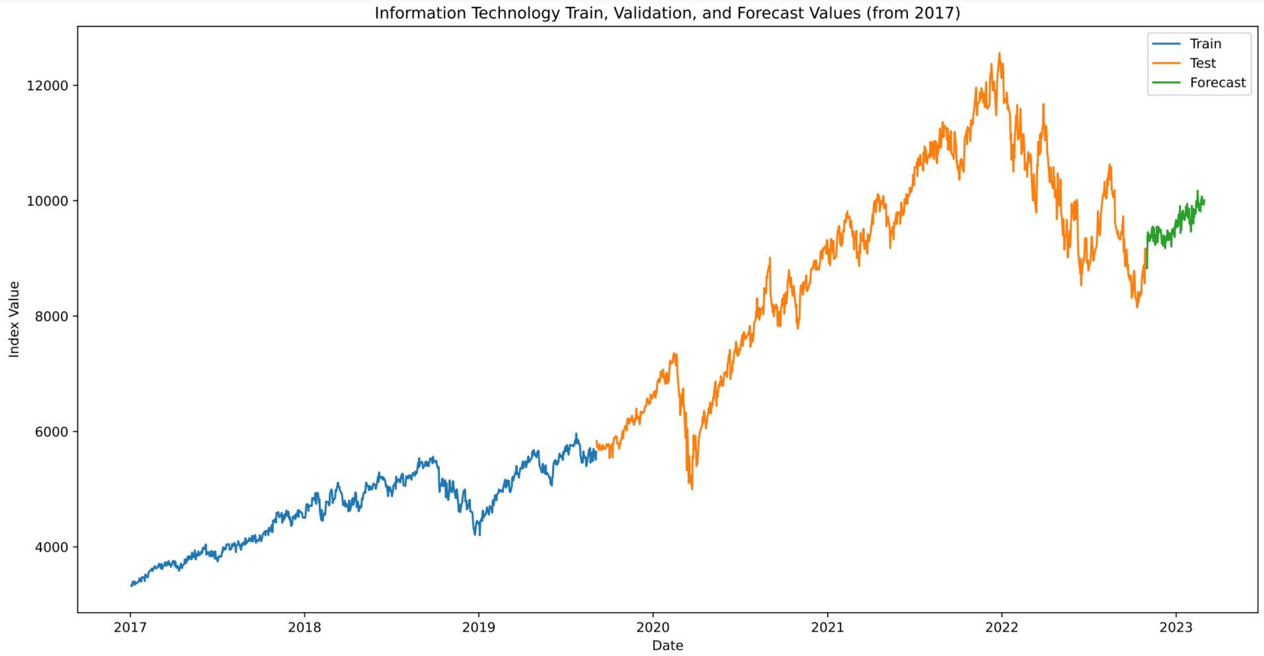
The Model

- **LSTM** with 3 layers and 5 epochs
    - 1st layer: 100 nodes
    - 2nd layer: 100 nodes
    - 3rd layer: 25 nodes
- Trained on the features of the index we built for each sector
    - **Train** and **test** datasets were split **80%** and **20%**, respectively
    - The splits were done according to date
        - **Train** set was from **2007 to 2019**
        - **Test** set was from **2019 to 2022**
- Target Variable: **Index**

The Model’s Performance (Train + Validation)



The Model's Future Predictions (till 31st March 2023)



The Model’s Performance (The Root Mean Squared Error)

* This is not the our model accuracy. It is a mean squared error on each of the sectors

Sector	**RMSE**
Communication Services	148.18
Consumer Discretionary	44.93
Consumer Staples	16.57
Energy	6.69
Financials	32.11
Health Care	130.01
Industrials	13.24
Information Technology	31.68
Materials	11.98
Real Estate	11.98
Utilities	11.94

The Model’s Ranking of the Sectors

Sector	Percent Change
Consumer Discretionary	25.48
Communication Services	15.77
Information Technology	13.31
Materials	6.26
Health Care	3.01
Industrials	2.6
Financials	-1.8
Utilities	-3.41
Real Estate	-4.51
Consumer Staples	-6.31
Energy	-13.26

Companies/Sectors to Invest in:

Consumer Discretionary:

- Amazon (AMZN)
- The Home Depot (HD)

Communication Services:

- T-Mobile US, Inc. (TMUS)
- AT&T (T)

Information Technology

- Microsoft (MSFT)
- Google (GOOGL)

Companies/Sectors NOT to Invest in:

Real Estate:

- Prologis (PLD)
- American Tower (AMT)

Consumer Staples:

- Procter & Gamble (PG)
- Coca Cola (KO)

Energy:

- Chevron (CVX)
- ExxonMobil (XOM)

Conclusion

- Version 2.0

  We want to create a dynamic interface where a user can input an amount he wants to invest eg. $10000 and we can say that he should invest 5% in Sector A, 12% in sector B etc. For that we need to calculate the risk in each sector. We can use softmax function for weights or we can use our curated dataset and run an LSTM on volatility, check the mean squared error on that and accordingly weigh each company for its risk.