# University of New Haven

## Connecticut


Fall 2022

CSCI-6401-01 (Data Mining)

# Phase 5 – Data Modeling

**Submitted by**

Team BruteForce

Department of Computer Science

**Submitted to**

Dr. Shivanjali Khare

1) Team Name – **Team BruteForce**

   Team Members –

   1. Naga Anvesh Kunuguntla (nkunu1@unh.newhaven.edu)
   2. Karthik Kannamreddy (kkann2@unh.newhaven.edu)
   3. Surya Teja Chinigepalli(schin15@unh.newhaven.edu)

2) The datasets we've chosen date back five years and come from a legitimate website. We'll be dealing with data from the big-tech firms FAANG (Facebook, Apple, Amazon, Netflix, and Google). We would want to examine these five firms' stocks in order to comprehend how they have expanded rapidly over time and how they have impacted the globe since they are the pinnacle of technology and because we are technocrats ourselves. Additionally, we would want to research potential investment opportunities.

3) Modeling Techniques Used:

   Since the data in our dataset is numerical, we employed regression in our study. We have employed a covariance matrix in regression and a classifier known as logistic regression in the covariance matrix. Statsmodels are used to implement Logistic Regression. We also carried out this using the Sklearn package.

4) Parameters and Hyper parameters:

| | Hyper Parameter | Parameters |
|---|---|---|
| AAPL | random_state:[0] | const, Lag 1, Lag 2, Lag 3, Lag 4, Lag 5, Volume |
| NFLX | regressor.coeff_, regressor.intercept_, random_state:[0] | Mean Absolute Error, Mean Squared Error, Root Mean Squared Error |

5) Hardware:

   The Google Research Colaboratory, or "Colab," has been utilized. Colab is extremely useful for machine learning, data analysis, and teaching since it enables anybody to create and run arbitrary Python code through the browser. Technically speaking, Colab is a hosted Jupyter notebook service that can be accessed instantly and offers free access to computer resources, including GPUs. The open-source project on which Colab is based is Jupyter. Without having to download, install, or execute anything, Colab enables you to utilize and share Jupyter notebooks with others.

6) Outcomes of Data Mining Techniques from different perspectives using varied performance metrics:

Logistic Regression: (AAPL)

| | Date | Today | Lag 1 | Lag 2 | Lag 3 | Lag 4 | Lag 5 | Volume | Direction | |
|---|---|---|---|---|---|---|---|---|---|---|
| **6** | 2015-01-12 | -2.464063 | 0.107244 | 3.842227 | 1.402209 | 0.009429 | -2.817176 | 0.214798 | 0 |
| **7** | 2015-01-13 | 0.887879 | -2.464063 | 0.107244 | 3.842227 | 1.402209 | 0.009429 | 0.198603 | 1 |
| **8** | 2015-01-14 | -0.381053 | 0.887879 | -2.464063 | 0.107244 | 3.842227 | 1.402209 | 0.268368 | 0 |
| **9** | 2015-01-15 | -2.714042 | -0.381053 | 0.887879 | -2.464063 | 0.107244 | 3.842227 | 0.195826 | 0 |
| **10** | 2015-01-16 | -0.777016 | -2.714042 | -0.381053 | 0.887879 | -2.464063 | 0.107244 | 0.240056 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1757** | 2021-12-23 | 0.364389 | 1.531883 | 1.908684 | -0.812195 | -0.650173 | -3.926396 | 0.092135 | 1 |
| **1758** | 2021-12-27 | 2.297481 | 0.364389 | 1.531883 | 1.908684 | -0.812195 | -0.650173 | 0.068357 | 1 |
| **1759** | 2021-12-28 | -0.576724 | 2.297481 | 0.364389 | 1.531883 | 1.908684 | -0.812195 | 0.074920 | 0 |
| **1760** | 2021-12-29 | 0.050199 | -0.576724 | 2.297481 | 0.364389 | 1.531883 | 1.908684 | 0.079144 | 1 |
| **1761** | 2021-12-30 | -0.657832 | 0.050199 | -0.576724 | 2.297481 | 0.364389 | 1.531883 | 0.062349 | |

Logistic Regression: (NFLX)

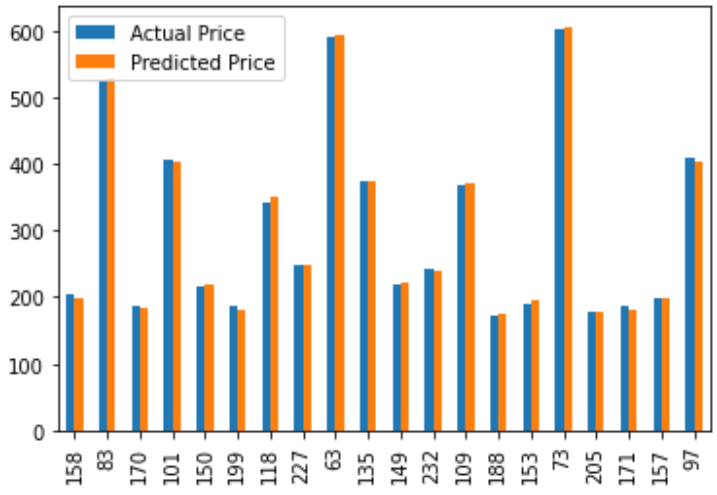
```
Mean Absolute Error:
3.026439592947956

Mean Squared Error:
14.653571143809746

Root Mean Squared Error:
3.8279983207689297
```

7) Visualization Techniques used:

To determine how many right and wrong guesses the classifier made, we employed a confusion matrix.
Logistic regression: (AAPL)

| Actual | Down | Up |
|---|---|---|
| Down | 144 | 677 |
| Up | 122 | 813 |

Logistic Regression: (NFLX)



8) Conclusion:

In this stage, a logit model based on the logistic regression technique is successfully developed to forecast stock market movement. The important technical indicators to forecast stock market movement based on historical data from year 2015 to 2021 have been established through the use of logistic regression. The future market movement exhibits positive movement for Apple and not so good for Netfilx stocks, according to the logit model.

**GitHub Repository Link:**

https://github.com/anvesh-lp/BruteForce