University of New Haven

Connecticut

Spring 2022

CSCI-6674 (Data Mining)

Phase 5 – Data Modeling

Submitted by   **Submitted to**

Team Data Diggers   Dr. Shivanjali Khare

Department of Computer Science

Date – 04/01/2022

1) Team Name – **Team Data Diggers**

   Team Members –

   1. Anil Kumar Reddy Allikepalli (aalli2@unh.newhaven.edu)
   2. Kunal Dhanaitkar (kdhan1@unh.newhaven.edu)
   3. Koushik Akkenapally (kakk1@unh.newhaven.edu)

2) The datasets we have selected are from the **past few years** and are from an authenticated website**.** We'll be working with datasets from the Facebook, Apple, Amazon, Netflix, and Google (FAANG) big-tech companies. Since these five companies are technological epitome and being technocrats ourselves, we would like to study their stocks and understand how they have exponentially grown over the years and made an impact on the world. Also, we would like to study on which company we can invest in the near future.

3) Modeling Techniques Used:

   We have used Regression in our project because the data in our dataset is numerical.

   In Regression we have used Covariance matrix and in covariance matrix we use classifier as Logistic Regression. Logistic Regression is implemented using statsmodels. We have also performed this using sklearn library.

4) Parameters and Hyper parameters:

   For Logistic Regression: (AAPL)

   Hyper Parameter – random_state:[0]

   Parameters – const, Lag 1, Lag 2, Lag 3, Lag 4, Lag 5, Volume

   For Logistic Regression: (NFLX)

   Hyper Parameter – regressor.coeff_, regressor.intercept_, random_state:[0]

   Parameters – Mean Absolute Error, Mean Squared Error, Root Mean Squared Error

5) Hardware:

We have used Google Research's Colaboratory or 'Colab'. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs. Jupyter is the open-source project on which Colab is based. Colab allows you to use and share Jupyter notebooks with others without having to download, install, or run anything.

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

Logistic Regression: (AAPL)

<div align="center">

Logit Regression Results

Dep. Variable:	Direction	No. Observations:	1001
Model:	Logit	Df Residuals:	994
Method:	MLE	Df Model:	6
Date:	Sat, 02 Apr 2022	Pseudo R-squ.:	0.007289
Time:	02:47:56	Log-Likelihood:	-685.02
converged:	True	LL-Null:	-690.05
Covariance Type:	nonrobust	LLR p-value:	0.1222

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2695	0.156	1.726	0.084	-0.036	0.575
Lag 1	-0.0780	0.032	-2.441	0.015	-0.141	-0.015
Lag 2	0.0107	0.032	0.337	0.736	-0.052	0.073
Lag 3	-0.0143	0.032	-0.450	0.653	-0.077	0.048
Lag 4	0.0401	0.032	1.262	0.207	-0.022	0.102
Lag 5	0.0344	0.031	1.097	0.273	-0.027	0.096
Volume	-0.7469	1.126	-0.663	0.507	-2.954	1.461

</div>

Logistic Regression: (NFLX)

  Mean Absolute Error: 3.1013126719607764

  Mean Squared Error: 19.399699385083693
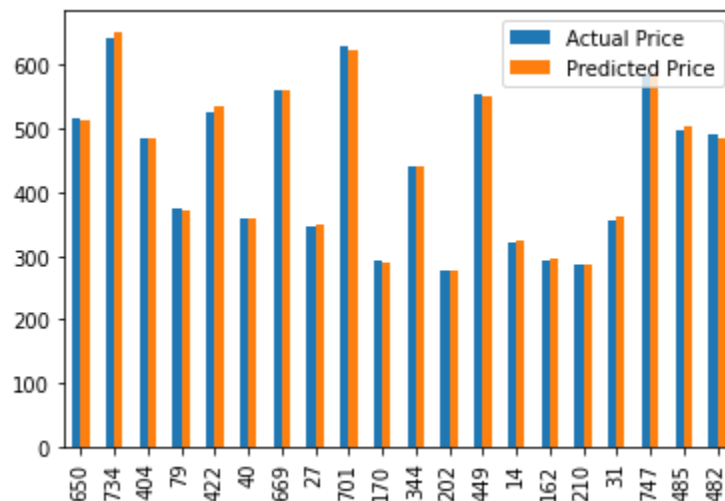
  Root Mean Squared Error: 4.404508983426381

7) Visualization Techniques used:

  We have used confusion matrix to calculate the number of correct and incorrect predictions made by the classifier.

Logistic regression: (AAPL)

Predicted	Down	Up
Actual		
Down	74	383
Up	66	478

Logistic Regression: (NFLX)

8) Conclusion:

This phase effectively formulates a logit model based on the logistic regression method to predict the stock market movement. By using logistic regression, the significant technical indicators to predict the stock market movement based on historical data from year 2019 to 2021 has been identified. Based on the logit model, the future market movement shows positive movement.

GitHub Repository Link:

https://github.com/aanilkumarreddy/DataDiggers