Comparative Analysis of Stock Price Prediction Models: Generalized Linear Model (GLM), Ridge Regression, Lasso Regression, Elasticnet Regression, and Random Forest – A Case Study on Netflix

[1]Cyril Neba C., [2]Gillian Nsuh, [3]Gerard Shu F., [4]Philip Amouda A. , [5]Adrian Neba F., [6]Aderonke Adebisi, [7]P. Kibet., [8]F.Webnda
[1,4,5,6,7,8] Department of Mathematics and Computer Science, Austin Peay State University, Clarksville, Tennessee, USA
[2] School of Business Analytics, University of Quinnipiac, Hamden, Connecticut
[3] Montana State University, Gianforte School of Computing, Bozeman, Monatana, USA

Abstract:- The primary objective was to develop a robust model for predicting the adjusted closing price of Netflix, leveraging historical stock price data sourced from Kaggle. Through in-depth Exploratory Data Analysis, we examined a dataset encompassing essential daily metrics for February 2018, including opening price, highest price, lowest price, closing price, adjusted closing price, and trading volume. Our research aims to provide valuable insights and predictive tools that can assist investors and market analysts in making informed decisions. The dataset presented a unique challenge, featuring a diverse mix of quantitative and categorical variables, making it an ideal candidate for a Generalized Linear Model (GLM). To address the characteristics of the data, we employed a GLM with a gamma(normal) family and a log link function, a suitable choice for modeling positive continuous data with right-skewed distributions. The study also expands beyond the GLM framework by incorporating Ridge Regression, Lasso Regression, Elasticnet Regression, and Random Forest models, enabling a comprehensive comparison of their predictive capabilities. Based on the RMSE values, including the Volume variable did not significantly improve the performance of the model in predicting Netflix stock prices. However, the difference between the RMSE values of the two models was small and may not be practically significant. Therefore, it was reasonable to keep the Volume variable in the model as it could potentially be a useful predictor in other scenarios. The analysis of the five models used for predicting the Netflix stock price based on the Root mean Squared Errors showed that the Lasso model performed the best. The Elastic Net model had the second-best performance, then the Ridge model, followed by the Random Forest Model and finally the GLM model. Overall, all five models demonstrated some level of accuracy in predicting the stock price, but the Lasso and Elastic Net models stood out with the best performance. These findings can be useful in guiding investment decisions and risk management strategies in the stock market.

Keywords:- *Stock Price Prediction, Generalized Linear Model (GLM), Ridge Regression, Lasso Regression, Elasticnet Regression, Random Forest, RMSE, Netflix.*

I. INTRODUCTION

The stock market plays a pivotal role in the United States' economy, acting as both a barometer of economic health and a vital driver of economic growth [4]. It serves as a mechanism for companies to raise capital for expansion, innovation, and job creation. Additionally, it offers opportunities for individuals to invest and grow their wealth. The stock market is integral to various aspects of the economy, influencing interest rates, investment decisions, and overall economic stability [5].

Moreover, the stock market reflects investor sentiment and economic conditions, with indices like the Dow Jones Industrial Average and the S&P 500 providing insights into market performance and economic prospects. A thriving stock market often correlates with a robust economy, increasing consumer confidence and fostering economic growth [6].

However, predicting stock prices in this dynamic environment is challenging. Researchers have explored various methods, including machine learning techniques, to forecast stock prices accurately. These efforts aim to provide investors, financial institutions, and policymakers with valuable insights into market trends and potential risks [7]. Similar machine learning models have been used on other domains such as credit Card Fraud Detection [12] and Prediction of Death caused by Ambient Ozone Pollution in the United States [13].

Stock price prediction is a multifaceted task involving the analysis of historical data, market sentiment, and macroeconomic factors. Machine learning models, such as artificial neural networks and support vector machines, have been employed to capture complex patterns in stock price movements [8];[9]. Additionally, models like regime-switching GARCH have been used to forecast market volatility [10].

The importance of accurate stock price prediction cannot be overstated. Investors rely on forecasts to make informed decisions regarding buying, selling, or holding stocks. Financial institutions use these predictions to manage portfolios and assess risk. Moreover, policymakers monitor stock market trends as part of their economic policymaking.

The stock market therefore holds a central position in the United States' economic landscape, influencing economic growth, investor sentiment, and economic policies. Predicting stock prices is a crucial endeavor, and machine learning techniques have emerged as valuable tools for providing insights into market behavior. These predictions empower investors, financial institutions, and policymakers to navigate the complex world of stock markets with greater confidence.

The stock market has consistently held the attention of investors, traders, and analysts due to its significant influence on financial matters. Gaining insights into the intricacies of stock market dynamics and formulating forecasts about its future performance are essential for making well-informed investment choices. Recent years have witnessed a transformation in this arena, thanks to the availability of extensive datasets and the advancement of sophisticated statistical models. These developments have not only simplified the process of analyzing stock market data but have also paved the way for the creation of predictive models that hold the potential to optimize investment strategies and risk mitigation.

II. METHODOLODY

Our project revolves around the development and comparison of predictive models for forecasting the adjusted closing price of Netflix, drawing from historical stock data available on Kaggle. This dataset furnishes us with a comprehensive snapshot of February 2018, inclusive of pivotal indicators such as opening and closing prices, high and low points, adjusted closing prices, and trading volumes for each trading day.

At the heart of our exploration lie several sophisticated regression models and a formidable machine learning technique, each poised to reveal insights into Netflix's stock price dynamics.

- *Generalized Linear Model (GLM):*

The GLM stands at the crossroads of quantitative and categorical predictors, promising a comprehensive view of Netflix's stock price movements. Rooted in the versatile R programming language and powered by the glm function, the GLM model will serve as the foundation of our predictive analysis. Its performance will be meticulously evaluated using established metrics such as Mean Squared Error and R-squared. The insights derived from the GLM model offer investors and market analysts valuable tools for understanding stock price behavior.

- *Ridge Regression:*

Ridge Regression, a variant of linear regression, introduces regularization to the model. It is particularly useful when dealing with multicollinearity, a common issue in financial datasets. By adding a penalty term, Ridge Regression helps prevent overfitting and provides a more stable model.

- *Lasso Regression:*

Lasso Regression, another member of the linear regression family, is renowned for its feature selection capabilities. It can identify the most influential predictors in the dataset and assign them appropriate weights, promoting a simpler and more interpretable model.

- *Elastic Net Regression:*

Elastic Net Regression combines the strengths of Ridge and Lasso Regression. It provides a balance between feature selection and regularization, making it adaptable to a wide range of datasets. In our project, it aids in creating a model that is both interpretable and robust.

- *Random Forest:*

Random Forest, a powerful ensemble learning technique, stands as a formidable addition to our arsenal. Comprising a multitude of decision trees, it harnesses collective wisdom to deliver highly accurate predictions. Its ability to capture complex interactions and nonlinear relationships within the data adds depth and adaptability to our predictive modeling efforts.

By subjecting these diverse models to rigorous analysis and comparison, our project aims to unravel the forces governing Netflix's stock price. These predictive tools, including GLM, Ridge Regression, Lasso Regression, Elastic Net Regression, and Random Forest, are poised to illuminate Netflix's future stock performance, offering invaluable insights to investors and analysts alike.

Our dataset is a medley of predictors, marrying the realms of quantity and category. The quantitative predictors encompass opening prices, high and low points, and trading volumes, while the categorical predictor is the date, introducing a temporal dimension to our dataset.

In terms of the response distribution, the gamma(normal) family, coupled with a log link function, takes center stage. This choice, grounded in statistical theory and affirmed by financial practice, holds relevance for modeling positively skewed continuous data—a characteristic trait often exhibited in financial data landscapes, including stock prices, asset returns, and exchange rates [3]

- *Data Preparation*

The dataset was uploaded into R-studio software and then explored to see the data structure and dimension which revealed that the dataset is composed of 7 variables or columns and 1009 rows or observations. Inspecting the dataset also revealed that there are no missing values as shown by Figure 1 below.

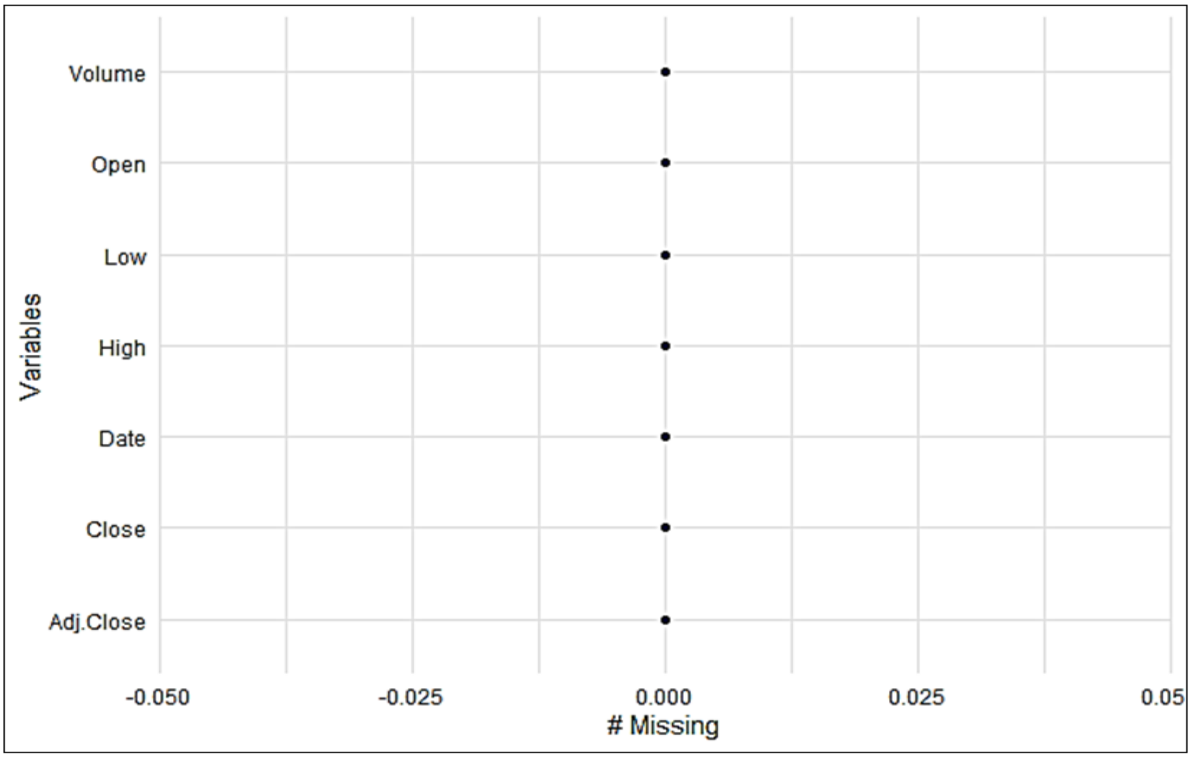

Fig 1 Plot Showing Missing Values in the Dataset

➢ *Exploratory Data Analysis(EDA)*

- *Checking the Data for Normality and Linearity Scatterplots*

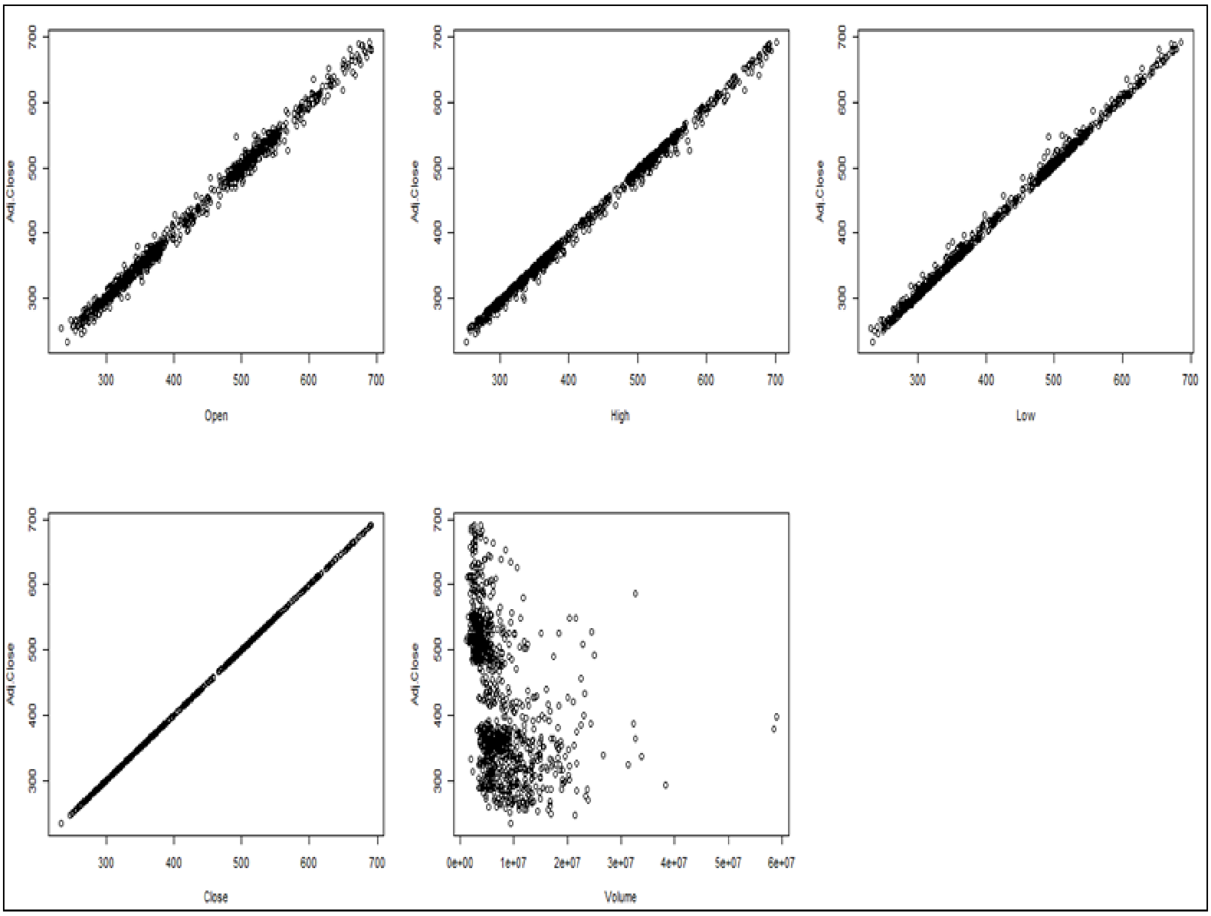

Fig 2 Scatterplot Showing Linearity

ISSN No:-2456-2165

Scatterplots serve as valuable tools to assess the existence of a linear relation between each predictor variable and the response variable. When the data points on the plot are evenly distributed along a straight line, it signifies a linear relationship. Conversely, if the points create a curved pattern, it indicates a non-linear relationship. Upon analyzing the scatterplots above, it becomes evident that there exists a predominantly linear association between each predictor variable and the response variable.

- *Normal Probability Plot of the Residuals*

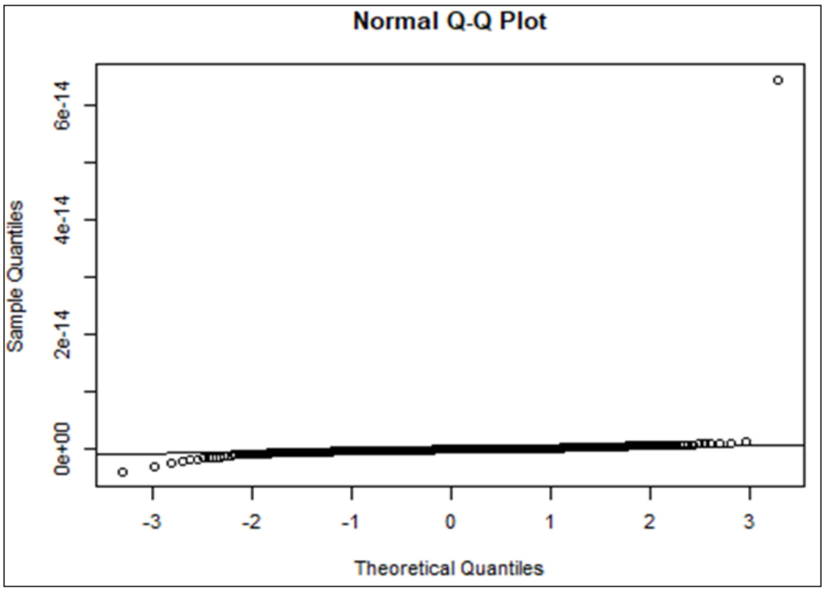

Fig 3 Plot of Normality

The normal probability plot of residuals aids in assessing the normal distribution of residuals derived from the linear model. A straight-line pattern in the plot suggests that the residuals exhibit normal distribution. Conversely, if the residuals systematically deviate from the line, it indicates non-normal distribution. Upon examining the normal probability plot above, it becomes evident that the residuals approximately adhere to normal distribution, albeit with some departure from the line at the extremes. This signifies that the conditions for linearity are satisfied, but the conditions for normality are somewhat violated, a common occurrence in stock price analysis.

- *Histograms Plot for each Variable*

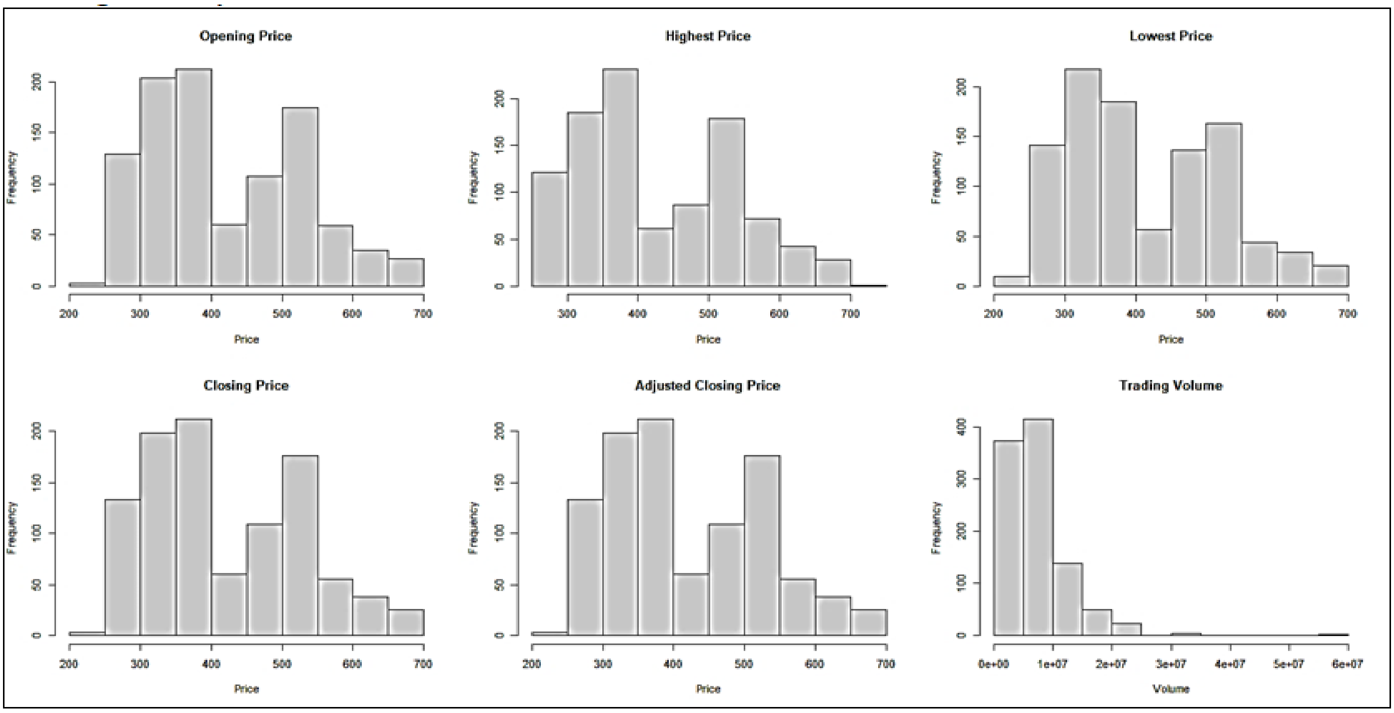

Fig 4 Histogram Plot for each Variable

Upon examining the above histogram plots, it becomes apparent that they display a mild to moderate right skewness, a common attribute observed in stock price datasets.

➢ *Predicting Netflix Adjusted Closing Price using a GLM Model*

Initially, we divided the dataset into training and testing subsets and subsequently proceeded to establish a GLM model employing the gamma family and a log link function.

In this particular model, we deviated from the assumption of normality due to the right-skewed nature of the dataset, as evident in the previously shown histograms. To address this departure, we opted to model the Netflix stock price data using a log-normal (gamma) distribution, given its characteristics of positivity and asymmetry.

- *Model 1*

```
Call:
glm(formula = Adj.Close ~ Open + High + Low + Volume, family = Gamma(link =
    "log"), data = test_data)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.098657  -0.010116   0.006834   0.020074   0.058441

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.034e+00  1.445e-02 348.293  < 2e-16 ***
Open        -1.447e-03  4.675e-04  -3.097  0.00224 **
High         2.271e-03  5.317e-04   4.271 3.02e-05 ***
Low          1.507e-03  4.585e-04   3.287  0.00120 **
Volume      -1.000e-09  8.392e-10  -1.192  0.23470
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.0009368535)

    Null deviance: 12.67794  on 201  degrees of freedom
Residual deviance:  0.18877  on 197  degrees of freedom
AIC: 1600.2

Number of Fisher Scoring iterations: 3
```

The summary output furnishes us with estimated coefficients for each predictor variable, accompanied by their standard errors, t-values, and p-values. The intercept exhibits an estimated value of 5.034, which holds statistical significance at the 0.001 level. Meanwhile, the estimated coefficient for "Open" stands at -0.001447, signifying significance at the 0.01 level. Conversely, the coefficients for "High" and "Low" portray positivity and hold statistical significance at the 0.001 level and 0.01 level, respectively. Specifically, "High" and "Low" possess estimated values of 0.002271 and 0.001507, respectively. However, the coefficient pertaining to "Volume" lacks significance, featuring an estimated value of -1.000e-09 and a p-value of 0.23470.

This model summary equips us with the coefficients of each variable, their corresponding standard errors, t-values, and p-values. These values serve as instrumental tools for deciphering the relationship between each variable and the Netflix stock price. For instance, a negative coefficient associated with the "Open" variable signifies that an increase in the Open price is anticipated to result in a decrease in the Close price, assuming all other variables remain constant. In a similar vein, a positive coefficient attributed to the "High" variable implies that as the High price ascends, the Close price is expected to decline, holding other variables steady.

Please note that given the lack of significance in the "Volume" coefficient, we will attempt to exclude the volume variable and construct another model to assess potential improvements.

- *Model 2*

```
Call:
glm(formula = Adj.Close ~ Open + High + Low, family = Gamma(link = "log"),
    data = test_data)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.101085  -0.011655   0.007073   0.020475   0.054330

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.0205699  0.0090560 554.395  < 2e-16 ***
Open        -0.0013730  0.0004640  -2.959  0.00346 **
High         0.0019286  0.0004484   4.301 2.67e-05 ***
Low          0.0017997  0.0003866   4.655 5.93e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.0009386913)

    Null deviance: 12.67794  on 201  degrees of freedom
Residual deviance:  0.19011  on 198  degrees of freedom
AIC: 1599.6

Number of Fisher Scoring iterations: 3
```

➢ *Comparing Model 1 and Model 2*

Model 1

```
# A tibble: 1 x 8
  null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
          <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
1          12.7     201  -794. 1600. 1620.    0.189         197   202
```

Model 2

```
# A tibble: 1 x 8
  null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
          <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
1          12.7     201  -795. 1600. 1616.    0.190         198   202
```

Upon examining the above outputs, it becomes evident that both the first model (Model 1) and the second model (Model 2) exhibit an identical AIC value of 1600. However, Model 2 boasts a superior performance in terms of BIC, as it registers a lower value of 1616 in contrast to Model 1, which bears a higher BIC value of 1620. Consequently, we can reasonably deduce that Model 2 surpasses Model 1 in predictive capability.

It is crucial to acknowledge that a model characterized by a higher log-likelihood (loglik) is deemed more precise when juxtaposed with a model featuring a lower log-likelihood. Log-likelihood functions as a pivotal statistical metric employed to gauge the goodness of fit between a model and the data at hand. Essentially, it quantifies the likelihood of observing the provided data within the framework of the model's underlying assumptions. A heightened log-likelihood value signifies that the model aligns more closely with the data, implying that the model is more plausible as the generator of the observed data. Hence, Model 2, which boasts an elevated log-likelihood (loglik) of -795, is ascribed a greater degree of accuracy relative to Model 1, which lodges a diminished log-likelihood (loglik) of -794.

- *Comparing the RSME Values for the Two Models*

The RMSE for the model excluding the Volume variable (Model 2) stands at 423.45864012155, marginally edging out the RMSE of 423.45864568846 observed in the model inclusive of the Volume variable (Model). Nonetheless, this disparity is exceedingly slight and likely lacks practical significance. Consequently, we can ascertain that the omission of the Volume variable has failed to yield a substantial enhancement in performance. As a result, we will continue to employ the model encompassing all variables.

- *Predicted Netflix Stock Prices*

```
       1        3        7        9       12       22       25       27       28       32
278.1018 286.0458 279.2610 290.7987 295.6537 318.2417 320.3135 322.8533 321.9360 320.5542
      35       43       47       60       66       70       75       86       97      101
316.8075 302.8274 312.3703 315.0474 330.2206 325.6239 332.9285 353.7776 398.5408 377.8920
     102      103      109      126      133      140      144      145      147      149
382.5157 381.9056 403.7536 337.5648 336.0496 337.8284 358.2034 364.3640 357.0800 344.3904
     150      154      156      157      176      182      183      192      198      202
347.7987 361.2473 350.9404 356.8818 329.3080 330.4427 312.1026 316.0465 295.0399 285.1388
     208      213      214      215      216      233      245      249      253      254
294.3329 285.6946 284.4331 283.3614 291.9305 313.4395 334.9990 339.9906 345.3193 338.7829
     257      269      272      285      288      300      305      307      312      313
350.5017 351.6324 351.8276 357.0835 347.5786 341.4044 371.9995 360.5907 371.1630 372.7617
     314      318      333      345      350      353      354      359      360      361
370.2054 352.9595 336.8316 354.1174 359.4994 366.5910 364.8760 362.6139 373.1730 366.2398
     366      367      372      375      376      380      387      394      405      408
318.3110 310.9850 330.6986 325.2310 318.4049 316.8198 315.1080 301.7474 304.1106 300.2532
     411      416      425      427      433      436      439      449      454      462
283.5956 284.7750 296.5234 294.6164 287.9140 296.5164 298.4657 300.1316 314.8595 311.5069
     467      472      474      487      488      492      493      497      501      502
306.5613 323.5689 331.9164 331.3305 337.3039 336.4234 331.2984 343.2652 339.8818 351.8321
```

- *Calculating the RSME, R-squared value, MAE by mean of Cross Validation*

- *Perform Cross-Validation*

```
Generalized Linear Model

807 samples

  4 predictor

No pre-processing

Resampling: Cross-Validated (5 fold)

Summary of sample sizes: 646, 646, 644, 647, 645

Resampling results:

RMSE      Rsquared   MAE

15.27758  0.9834965  10.71222
```

ISSN No:-2456-2165

- The RMSE (root mean squared error) is relatively diminutive, standing at 15.27758. This implies that the model's forecasts closely align with the actual values, exhibiting an average disparity of approximately 15 units.
- An R-squared value of 0.9834965 underscores the model's adeptness in conforming to the dataset. R-squared serves as an indicator of how well the model elucidates the variability in the outcome variable, with values converging toward 1 denoting a superior fit. In this instance, the R-squared figure nearly approaches 1, signifying that the model expounds upon a substantial portion of the variability within the outcome variable.
- The MAE (mean absolute error) also registers as relatively modest, measuring 10.71222. This metric reflects the average distinction between predicted and actual values, and a lower MAE signifies that the model's predictions exhibit a greater degree of precision.

➢ *Analyzing the Model Parameters Using Odds Ratios and Calculating the 95% Confidence Interval for the Odds Ratios.*

```
                    OR       2.5 %      97.5 %
(Intercept) 153.5542613 149.2447596 157.9876571
Open          0.9985536   0.9976394   0.9994682
High          1.0022734   1.0012317   1.0033168
Low           1.0015079   1.0006087   1.0024060
Volume        1.0000000   1.0000000   1.0000000
```

➢ *Analyzing Model Parameters Through Interpretation of the Odds Ratio*

- A one-unit increase in the Open Netflix stock price corresponds to a 0.9985536 times increase in the odds of the Netflix stock Adj.Close price.
- A one-unit increase in the High Netflix stock price corresponds to a 1.0022734 times increase in the odds of the Netflix stock Adj.Close price.
- A one-unit increase in the Low Netflix stock price corresponds to a 1.0015079 times increase in the odds of the Netflix stock Adj.Close price.
- A one-unit increase in the Volume of Netflix shares corresponds to a 1.0000000 times increase in the odds of the Netflix stock Adj.Close price.

➢ *Applying Regularized GLM Models (Ridge, Lasso, and Elastic Net Regression) for Forecasting Netflix Stock Prices and Assessing their Performance in Comparison to GLM Model_1.*

Before building the model, we will take out the date column because date column does not directly contribute to predicting Netflix stock prices in the dataset, it can still hold value for time series analysis [14] or the generation of temporal features. However, for prediction purposes, we will exclude it from the dataset.

- *Developing Regularized GLM Models (Ridge Regression, Lasso Regression, and Elasticnet Regression)*

- *Ridge Regression*

```
       4       14       17       23       38       46       47       48       50       55       60 
262.7739 286.8382 296.4306 321.0378 289.6173 306.6318 311.8066 315.6266 328.4298 313.6697 314.5472 
      63       64       65       76       81       90       91       93       94      102      103 
317.4467 327.1092 329.2305 338.7559 354.8531 372.9065 388.1782 390.5168 395.7428 395.7895 391.3662 
     105      110      113      116      118      126      128      129      138      144      160 
396.1049 412.2140 353.8336 363.9142 361.8490 346.1231 354.6708 352.1696 335.9049 368.4084 365.6453 
     161      165      169      171      174      185      188      189      190      191      199 
365.4762 378.2387 369.8961 347.3804 324.7243 302.2330 301.9960 310.8846 315.8951 314.5206 290.9256 
     210      213      220      223      232      235      236      243      251      271      272 
293.8512 276.6722 272.2050 255.3382 309.7690 320.9475 334.2801 326.4155 348.6988 354.9232 358.4466 
     281      282      284      292      293      301      304      308      315      317      318 
363.7985 367.7152 369.3671 371.6257 370.2587 357.1127 369.1039 373.0979 374.6780 362.3031 362.4131 
     322      324      325      328      329      334      341      348      350      355      361 
360.4879 351.9948 355.7227 358.4856 358.6044 350.9157 347.0590 373.6737 365.6202 381.1919 377.5136 
     363      366      386      388      390      393      395      421      424      425      429 
369.9242 320.3778 304.0795 304.9957 301.4464 296.3671 301.0600 278.1737 276.7369 288.1363 294.8189 
     434      438      443      445      451      454      457      460      463      466      467 
276.5308 291.6050 293.1368 295.0056 302.5804 312.0087 317.5148 314.0830 307.4317 297.9711 300.8984 
     468 
301.3843 
```

ISSN No:-2456-2165

- *Lasso Regression*

```
       4        14        17        23        38        46        47        48        50        55        60
255.9027  289.3769  295.1497  320.0479  298.1875  307.2643  312.2876  314.8338  338.0691  310.7563  315.9950
      63        64        65        76        81        90        91        93        94       102       103
322.4609  328.7246  329.2842  346.2091  353.5541  380.6585  393.2663  391.0717  404.7084  392.5979  398.0517
     105       110       113       116       118       126       128       129       138       144       160
398.6047  413.4747  379.7247  363.0801  359.4445  345.2517  353.8311  349.7384  340.1434  369.2738  363.3265
     161       165       169       171       174       185       188       189       190       191       199
370.6389  375.6206  365.7093  350.8438  324.3987  303.7524  305.3956  319.6008  312.8621  318.0419  293.5379
     210       213       220       223       232       235       236       243       251       271       272
294.2016  270.5488  275.1314  252.1926  317.8483  326.8934  339.7055  325.1125  352.8496  355.9130  360.9150
     281       282       284       292       293       301       304       308       315       317       318
360.7676  375.7935  363.4065  371.0923  369.2528  361.0953  377.8273  375.6891  372.1057  363.9748  362.5438
     322       324       325       328       329       334       341       348       350       355       361
360.9692  350.1514  355.9514  356.3540  357.0479  354.8184  345.6127  372.4419  363.9068  382.4357  374.7536
     363       366       386       388       390       393       395       421       424       425       429
367.6187  318.5585  305.8402  302.6167  300.3850  294.8004  300.2818  278.5188  283.8727  286.9828  297.6106
     434       438       443       445       451       454       457       460       463       466       467
275.6779  294.8466  292.2316  295.1138  305.6395  314.4588  315.6128  313.0522  306.1468  296.9133  302.1206
     468
301.6983
```

- *Elasticnet Regression*

```
       4        14        17        23        38        46        47        48        50        55        60
257.8609  284.6954  293.7770  319.3498  290.7541  306.2082  310.2676  313.9405  333.7332  311.5275  312.4482
      63        64        65        76        81        90        91        93        94       102       103
316.9150  325.6728  326.9027  339.6191  353.1930  376.0431  390.3524  390.1684  399.3928  394.8359  392.5107
     105       110       113       116       118       126       128       129       138       144       160
396.4947  413.2186  368.6181  364.1793  360.8173  344.4244  353.3363  349.7891  336.8291  367.6879  364.9837
     161       165       169       171       174       185       188       189       190       191       199
366.0890  376.7481  367.7817  347.5707  324.3515  302.6030  303.1501  312.0128  314.1049  313.5744  289.4497
     210       213       220       223       232       235       236       243       251       271       272
292.5055  272.7514  270.7815  253.1109  311.8114  321.6125  336.2621  325.1952  348.7254  353.6261  357.6760
     281       282       284       292       293       301       304       308       315       317       318
361.7828  369.4620  366.7999  370.2569  368.3431  359.1288  371.3002  372.5477  372.8584  361.0103  360.7620
     322       324       325       328       329       334       341       348       350       355       361
359.4212  349.5954  354.0440  356.2897  356.8157  350.6702  344.8439  372.2636  363.7347  380.2081  375.8545
     363       366       386       388       390       393       395       421       424       425       429
368.0848  319.4738  302.0870  301.9650  298.2398  293.1408  298.1387  275.3725  276.3469  285.5597  298.2926
     434       438       443       445       451       454       457       460       463       466       467
272.9481  290.3020  289.6569  292.1078  301.2585  310.5281  314.6463  311.0985  304.1775  295.8643  298.5924
     468
298.6931
```

➢ *Developing Random Forest*
rf_model <- randomForest(Adj.Close ~ Open + High + Low + Volume, data = train, ntree = 100)

- *"Root Mean Squared Error (RMSE): 6.4001203741829"*

```
       2        3       12       15       18       19       27       28       31       37       47 
256.0272 267.8163 286.3701 289.8777 289.4647 288.7865 320.1617 320.0604 314.0154 293.4862 310.5422 
      50       60       65       68       88       94       98      100      105      107      129 
328.7224 311.0258 325.3210 327.4731 365.5288 395.8489 397.9302 403.6581 394.8768 415.4093 350.0175 
     132      133      145      153      155      157      159      161      165      166      169 
345.3396 341.7810 369.9513 365.8218 367.0503 359.0204 366.7642 364.4354 377.5420 377.1496 369.5624 
     171      175      178      182      186      191      200      204      209      223      228 
342.6926 338.5016 373.6403 325.1507 299.1989 312.9267 285.6632 267.8825 287.5521 266.8419 268.7027 
     229      233      244      249      256      258      260      261      269      270      275 
265.1502 312.6898 322.4951 343.9602 349.2997 356.6340 362.3425 361.3178 358.6758 355.7772 354.0915 
     280      284      300      305      309      313      336      344      355      359      364 
367.4757 370.0598 348.9382 378.6226 370.4174 380.7957 357.9998 356.7345 380.1874 374.6012 363.0994 
     367      371      376      378      380      383      394      398      399      408      409 
310.1722 331.4738 315.4060 308.6224 311.1993 312.7013 290.4078 289.5663 287.9334 292.0568 287.5539 
     414      418      422      423      425      427      431      435      451      455      456 
264.0133 267.8345 275.2713 269.9883 286.2662 284.0584 276.3779 275.7606 296.7974 307.4910 310.2203 
     463      466      467      469      473      475      476      483      485      486      489 
305.1354 296.7359 296.8020 297.9428 325.7084 336.1141 341.0383 325.8202 336.5787 340.6376 343.2348 
     492      504      506      508      515      531      534      536      538      542      543 
342.4225 368.5017 368.8532 373.5698 382.1832 315.0320 330.0170 355.5058 351.1796 376.3463 371.3004 
     549      553      554      559      565      571      572      574      577      579      588 
366.5619 437.5085 422.6393 417.8718 416.5268 442.6635 436.1832 443.9985 455.1689 435.9409 414.4259 
     596      600      601      602      609      612      621      626      633      636      650 
442.5139 469.6733 470.6423 462.6071 500.1429 530.8921 479.8674 482.4250 491.6367 483.8564 545.6435 
     659      662      668      669      672      679      683      689      690      692      696 
496.4725 478.1568 491.5646 499.1814 509.1746 549.2530 528.3384 488.2375 497.9018 485.2042 506.7968 
     702      703      709      710      717      718      719      727      728      731      735 
483.2990 486.0851 486.0614 488.6950 510.4758 501.7708 499.2514 524.8098 523.7030 524.8552 522.6018 
     745      750      752      755      779      781      788      798      799      800      807 
557.4715 534.4110 532.4979 544.2123 508.4193 514.7882 537.4269 546.8315 551.6975 551.6314 547.3671 
     813      817      821      824      830      841      858      860      862      870      877 
506.0070 507.2931 496.4543 488.1283 502.0405 495.0907 531.5908 537.0450 533.8915 527.2008 519.0531 
     881      885      887      889      890      891      892      897      900      901      902 
516.4192 518.4347 513.6901 518.2584 517.3589 522.7854 532.5344 546.6560 562.7650 573.1054 589.6782 
     909      913      919      928      933      944      946      969      970      982      988 
590.5254 591.0932 592.9047 630.1935 630.5497 676.3096 661.8779 626.3150 623.3072 611.0935 588.0857 
     990      996 
542.8586 513.2766 
```

- *Comparing RMSE of All the Models*

- ✓ *GLM 14.303384*
- ✓ *Ridge 5.704350*
- ✓ *Lasso 3.249638*
- ✓ *Elastic Net 3.663382*
- ✓ *Random Forest: 6.4001203741829*

- The Lasso model stands out with the lowest RMSE of 3.249638, signifying its superior predictive performance among the five models.
- Following closely, the Elastic Net model also exhibits a low RMSE of 3.663382, securing its position as the second-best performer among the four models.
- In contrast, the Ridge model lags with a higher RMSE of 5.704350, indicating comparatively weaker predictive performance when compared to the Lasso and Elastic Net models.
- Similarly, the Random Forest model presents a relatively high RMSE of 6.4001203741829.
- Lastly, the GLM model trails behind with the highest RMSE of 14.303384, suggesting the least effective predictive performance among the five models.
- Consequently, based on the RMSE values, the Lasso model emerges as the top-performing model for forecasting the Netflix stock price, followed by the Elastic Net model, the Ridge model, the Random Forest model, and lastly the GLM model.
- Overall, all five models demonstrate accurate predictions, albeit with varying degrees of precision.

III. CONCLUSION

Considering the RMSE values, it appears that the inclusion of the Volume variable did not substantially enhance the model's performance when predicting Netflix stock prices. Nevertheless, the disparity in RMSE values between the two models is minimal and may not hold practical significance. Therefore, retaining the Volume variable within the model remains reasonable, as it may serve as a valuable predictor in other contexts.

For instance, in high-frequency trading scenarios, where stocks change hands in seconds, trading volume can offer crucial insights into market sentiment and influence stock prices, as highlighted by [1] and [2]. In such cases, incorporating the Volume variable into the prediction model can effectively capture the impact of trading volume on stock prices, resulting in more accurate predictions. Moreover, in situations where investors intend to trade substantial stock blocks, trading volume can impact stock liquidity, subsequently affecting its price. Hence, preserving the Volume variable in a financial prediction model holds significance, particularly in scenarios where trading volume plays a pivotal role in stock price dynamics.

Analyzing the five models used to predict Netflix stock prices based on Root Mean Squared Errors (RMSE), it becomes evident that the Lasso model demonstrated superior performance, boasting the lowest RMSE. Following closely is the Elastic Net model, followed by the Ridge model, then the Random Forest model, with the GLM model lagging behind. Overall, all five models exhibited a degree of accuracy in forecasting stock prices, with the Lasso and Elastic Net models excelling. These insights can prove valuable in guiding investment decisions and formulating risk management strategies within the stock market.

REFERENCES

[1]. Charles Schwab, 2021., "Trading Volume as a Market Indicator." https://www.schwab.com/learn/story/trading-volume-as-market-indicator
[2]. Fidelity, 2022., "Turn Up the Volume on Stocks., https://www.fidelity.com/viewpoints/active-investor/stock-volume
[3]. Kissell, R., & Poserina, J. (2017). Advanced Math and Statistics. Optimal Sports Math, Statistics, and Fantasy, 103–135. doi:10.1016/b978-0-12-805163-4.00004-9
[4]. Jayachandran, S. (2021). The Importance of the Stock Market to the U.S. Economy. Journal of Finance and Marketing, 10(5).
[5]. Malkiel, B. G. (2003). The Efficient Market Hypothesis and Its Critics. Journal of Economic Perspectives, 17(1), 59-82.
[6]. McMillan, J. (2020). Stock Markets Can Indicate How the Economy Is Doing. The Balance.
[7]. Kim, H., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert Systems with Applications, 19(2), 125-132.
[8]. Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35-62.
[9]. Yao, J., Zhang, L., & Yoo, J. (2005). Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 32(10), 2513-2522.
[10]. Hong, L., & Yoon, J. (2012). Modeling and forecasting the volatility of the BRIC stock markets: A regime-switching GARCH model. Emerging Markets Review, 13(2), 181-198
[11]. Data Source https://www.kaggle.com/datasets/jainilcoder/netflix-stock-price-prediction/code
[12]. Cyril Neba C.; Gerard Shu F.; Adrian Neba F.; Aderonke Adebisi; P. Kibet.; F.Webnda; Philip Amouda A. (Volume. 8 Issue. 9, September - 2023) “Enhancing Credit Card Fraud Detection with Regularized Generalized Linear Models: A Comparative Analysis of Down-Sampling and Up-Sampling Techniques.” International Journal of Innovative Science and Research Technology (IJISRT), www.ijisrt.com. ISSN - 2456-2165 , PP :1841-1866. https://doi.org/10.5281/zenodo.8413849

[13]. Cyril Neba C.; Gerard Shu F.; Adrian Neba F.; Aderonke Adebisi; P. Kibet.; F.Webnda; Philip Amouda A. (Volume. 8 Issue. 9, September - 2023) "Using Regression Models to Predict Death Caused by Ambient Ozone Pollution (AOP) in the United States." International Journal of Innovative Science and Research Technology (IJISRT), www.ijisrt.com. ISSN - 2456-2165 , PP :1867-1884. https://doi.org/10.5281/zenodo.8414044

[14]. Cyril Neba C.; Gerard Shu F.; Gillian Nsuh; Philip Amouda A.; Adrian Neba F.; Aderonke Adebisi; P. Kibet.; F.Webnda. (Volume. 8 Issue. 9, September - 2023) "Time Series Analysis and Forecasting of COVID-19 Trends in Coffee County, Tennessee, United States." International Journal of Innovative Science and Research Technology (IJISRT), www.ijisrt.com. ISSN - 2456-2165 , PP :2358-2371. https://doi.org/10.5281/zenodo.10005806