Mikailalsys Journal of Mathematics and Statistics

e-ISSN : 3030-816X
p-ISSN : 3030-8399

Index : Harvard, Boston, Sydney University, Dimensions, Lens, Scilit, Semantic, Google, etc

https://doi.org/10.58578/MJMS.v3i2.5927

Modeling Stock Data Using Multiple Linear Regression and LASSO Regression Analysis

Adashu Jacob Daniel, Musa Dahiru Ibrahim, Anule Aondulum Josaphat
Federal University Wukari, Taraba State, Nigeria
adashu@fuwukari.edu.ng; dahirumusa1996@gmail.com

Article Info:

Submitted:	Revised:	Accepted:	Published:
Apr 19, 2025	May 15, 2025	May 28, 2025	Jun 2, 2025

Abstract

This study evaluates and compares the model fitting and predictive performance of Multiple Linear Regression (MLR) and Least Absolute Shrinkage and Selection Operator (LASSO) regression in the context of stock price prediction for four leading Nigerian companies. A dataset comprising 1,300 observations from 2019 to 2025 was obtained from Yahoo Finance and Investing.com. Multicollinearity assessment using the Variance Inflation Factor (VIF) revealed substantial collinearity among certain predictors, particularly for the variables "Open" (Honeywell: 55.45; Zenith: 920.30) and "Low" (Oando: 621.81), indicating the need for variable selection or dimensionality reduction. Comparative analysis based on model selection criteria, including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) demonstrated the superior performance of LASSO over MLR across all companies. For example, Honeywell's LASSO model recorded an AIC of –12,112.64 and an

MSE of 0.000021, compared to MLR's AIC of –2,690.54 and MSE of 0.00998. LASSO regression also identified key predictors such as "High" price, which exhibited strong statistical significance for Oando ($z = 18.991$, $p < 0.001$) and Zenith ($z = 7.066$, $p < 0.001$), whereas trading volume generally showed weak predictive power. The study concludes that LASSO provides a more parsimonious and accurate predictive model for financial time-series data. It is recommended for use in financial forecasting and investment analysis, particularly when dealing with multicollinear datasets and high-dimensional predictor variables.

Keywords: LASSO Regression; Multiple Linear Regression; Stock Price Prediction; Model Selection; Financial Data Analysis; Multicollinearity

INTRODUCTION

The prediction of stock prices has emerged as a crucial research domain because of the intricate and volatile nature of financial markets. Scholars around the world have delved into various methodologies, striving to enhance forecasting accuracy by using both traditional statistical techniques and advanced models. A stock price signifies the current market value of a single share in a publicly traded company, shaped by the forces of supply and demand. Historically, stock prices have been recognized as both a catalyst and a reflection of economic activity, serving as indicators of public sentiment according to Seethalakshmi (2018). A robust stock market is often perceived as a mirror of a nation's economic health and growth.

Technical analysis has long been a cornerstone in stock price forecasting, relying on historical price data and technical indicators to project future market movements. It also enables feature selection which is pivotal in managing data with multiple variables by identifying the most relevant predictors for modeling purposes. Mishra et al. (2020) applied backward elimination in multiple linear regression to optimize feature selection for predicting stock indices. Regression techniques have been extensively employed over the years for modeling financial data with multiple variables due to their capacity to handle large sets of predictors efficiently. With stock prices constantly fluctuating based on a multitude of factors—ranging from historical price trends to macroeconomic influences—

accurate forecasting is an increasingly vital tool for investors, portfolio managers, and financial analysts. This study aim to address the challenges of predicting stock prices, with an emphasis on high market prices and their impact on prediction accuracy. The study will evaluate the predictive performance of Multiple Linear Regression against LASSO regression in modelling stock data and also check the efficiency of their performance in the presence of multicollinearity.

Literature Review

Over the years, many researchers have tried to make accurate predictions of stock market returns, especially much emphasis has been given to identify the best predictors. However, most bivariate OLS models fail to beat the historical average forecast according to Welch and Goyal (2008). This problem remains even when all predictors are included in a multiple linear regression model for forecasting. This is likely due to overfitting and having too many parameters, which often leads to poor out-of-sample predictions. This overfiting is as a result of multiple variables in the dataset according to Fan, et al., (2014) which also leads to the problem of multicollinearity. Advanced models, such as the Partially Protected LASSO by Yaman et al., (2024), offer solutions by selecting pertinent predictors while minimizing sensitivity to data noise. This model strives to balance theoretical significance with forecasting accuracy, presenting a promising tool for navigating the unpredictable terrain of stock market returns as observed by Rapach and Zhou (2013).

The LASSO regularization method from Tibshirani (1996) is one of such methods. The strength of the LASSO lies in its ability to create a parsimonious and selecting the most important variables from high-dimensional data, contributing to better out-of-sample predictions. This advantage is further emphasized in this paper.

Based on review literatures, studies shows very limited the use of feature selection and to predict stocks price. Hence in this project we seek to investigate the application of LASSO regression to predict stock and compare its efficiency of the top five stocks in Nigerian Stock Market.

METHODOLOGY

Five top stocks from Nigerian Stock Market was used for the study, the data were collected from yahoo finance and Investment.com. The data includes Honeywell, Monsard, Oando,

and Zenith Bank. The variables used are:

Price: The daily stock price at the close of trading.

Open Price: The price at which a stock opens on a particular trading day.

Low Price: The lowest price recorded by a stock during the trading day.

High Price: The highest price reached by a stock during the trading day.

Change: The difference between the current stock price and its previous closing price.

Volume: The number of shares traded during a specific period, typically within a trading day.

Method of Data Analysis

Multiple Linear Regression (MLR) and LASSO (Least Absolute Shrinkage and Selection Operator) regression were used both for modelling and prediction.

Multiple Linear Regression (MLR)

The general form of MLR model is given:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Where

Y is the dependent variable.

β_0 is the intercept term.

$\beta_1, \beta_2, \cdots, \beta_p$ are the coefficients representing the impact of each independent variable.

$X_1, X_2, \cdots, X_p$ are the independent variables.

ε is the error term, accounting for variability not explained by the model.

In this study, the dependent variable y_i is the price while independent variables x_i include open price, high price, low price, change in price and volume.

Estimation of the Parameters in Linear Regression Models

The method of least squares is typically used to estimate the regression coefficient in a multiple linear regression model. The general regression equation is given by:

$$y_i = \beta_0 + \beta_1 x_{i1} + + \beta_2 x_{i2} + ... + + \beta_k x_{ik} + \varepsilon_i \tag{1}$$

Eqn. (1) can also be written as:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1,2,...,n \quad j = 1,2,...,k$$

(2)

The eqn. (1) may be written in matrix notation as

$$y = X\beta + \varepsilon \tag{3}$$

Where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \text{and} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In general, y is an $(n \times 1)$ vector of the observations, X is an $(n \times p)$ matrix of the levels of the independent variables, β is a $(p \times 1)$ vector of the regression coefficients, and ε is an $(n \times 1)$ vector of random errors. We wish to find the vector of least squares estimators, $\hat{\beta}$, that minimizes

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon = (y - X\beta)'(y - X\beta) \tag{4}$$

Note that L may be expressed as

$$L = y'y - \beta' X' y - y' X\beta + \beta' X' X\beta \tag{5}$$

Because $\beta' X' y$ is a (1×1) matrix, or a scalar, and its transpose $(\beta' X' y)' = y' X\beta$ is the same scalar. The least squares estimators must satisfy

$$\left. \frac{\partial L}{\partial \beta} \right|_{\hat{\beta}} = -2X' y + 2X' X\hat{\beta} = 0$$

Which simplifies to

$$X' X\hat{\beta} = X' y \tag{6}$$

Eqn (6) is the matrix form of the least squares normal equations. To solve the normal

equations, multiply both sides of eqn (6) by the inverse of $X'X$. Thus, the least squares estimators of β is

$$\hat{\beta} = \left(X'X\right)^{-1} X'y \qquad (7)$$

Assumptions of MLR:

The MLR works under the following assumptions.

Linearity: The relationship between the dependent and independent variables is linear.

Independence: Observations are independent of each other.

Homoscedasticity: The variance of the error terms is constant across all levels of the independent variables.

Normality: The error terms are normally distributed.

Multicollinearity and Outliers

Multicollinearity occurs when independent variables in a regression model are highly correlated, leading to unreliable coefficient estimates.

Correlation

To calculate the Pearson correlation coefficient between two variables, X and Y, we use

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\left(x_i - \bar{x}\right)^2}\sqrt{\left(y_i - \bar{y}\right)^2}}$$

(08)

r_{xy} is the sample Pearson correlation coefficient.

n is the number of data pairs.

x and y are individual sample points.

Detecting Multicollinearity:

A common method to detect multicollinearity is by calculating the **Variance Inflation Factor (VIF)** for each predictor variable.

Variance Inflation Factor (VIF):

The VIF quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors. It is calculated using the formula:

$$VIF_j = \frac{1}{1-R_j^2} \tag{09}$$

Where:

- R_j^2 is the R-squared value obtained by regressing the j-th predictor variable against all other predictor variables.
- A VIF value greater than 10 indicates significant multicollinearity, suggesting that the corresponding predictor variable is highly collinear with others.

LASSO Regression (LR)

LASSO Regression performs both variable selection and regularization to enhance the prediction accuracy. It is particularly useful when dealing with datasets containing numerous variables, as it helps in identifying the most significant predictors by shrinking less important coefficients to zero.

Objective Function: LASSO aims to estimate the coefficients of a regression model by minimizing the following objective function:

$$\gamma_{LASSO} = \arg\min_{\gamma}\left(\sum_{i=1}^{N}\left(Y_i-\alpha-\gamma X_i\right)^2+\lambda|\gamma|\right) \qquad \lambda>0 \tag{10}$$

Where:

λ is a penalty parameter.

Consider the case where $|\gamma|=\gamma$. where, $\gamma_{LS}=\overline{Y}_{X=1}-\alpha$ is large.

For a known λ, we have the following first order condition:

$$-2\sum_i X_i\left(Y_i-\alpha-\gamma X_i\right)+\lambda=0$$

$$\sum_{i:X_i=1}\left(Y_i-\alpha-\gamma\right)-\frac{\lambda}{2}=0$$

$$N_1\left(\overline{Y}_{X=1}-\alpha\right)-\frac{\lambda}{2}-N_1\gamma=0$$

$$\gamma_{LASSO}=\overline{Y}_{X=1}-\alpha-\frac{\lambda}{2N_1}$$

Now, consider the case where $|\gamma|=-\gamma$. This case results in:

$$\gamma_{LASSO}=\overline{Y}_{X=1}-\alpha-\frac{\lambda}{2N_1} \tag{11}$$

Therefore, LASSO results in the following estimator:

$$\gamma_{LASSO}=\begin{cases}\overline{Y}_{X=1}-\alpha-\dfrac{\lambda}{2N_1}, & \overline{Y}_{X=1}-\alpha\geq\dfrac{\lambda}{2N_1}\\ 0, & -\dfrac{\lambda}{2N_1}<\overline{Y}_{X=1}-\alpha<\dfrac{\lambda}{2N_1}\\ \overline{Y}_{X=1}-\alpha+\dfrac{\lambda}{2N_1}, & Y_{X=1}-\alpha<-\dfrac{\lambda}{2N_1}\end{cases} \tag{12}$$

Penalty Term: The term $\lambda|\gamma|$ adds an L1 penalty to the regression, which has the effect of shrinking some coefficients exactly to zero. This results in a sparse model where only the most significant variables are retained, effectively performing variable selection.

Choosing the Regularization Parameter (λ): The parameter λ determines the extent of regularization applied. A larger λ imposes a greater penalty, leading to more coefficients being shrunk to zero. Conversely, a smaller λ allows more variables to remain in the model. Selecting an appropriate value for λ is crucial and is typically done using techniques like cross-validation or information criteria such as AIC or BIC.

Advantages of LASSO Regression:

- **Variable Selection:** By shrinking some coefficients to zero, LASSO helps in identifying and retaining only the most significant variables, simplifying the model and enhancing interpretability.

- **Handling Multicollinearity:** LASSO can effectively deal with multicollinearity by selecting one variable among correlated predictors and shrinking the others, thus improving model stability.

RESULTS

Check for Multicollinearity

In Table 1 shows VIF values for independent variables across five companies—Honeywell, Monsard, Oando, and Zenith—are presented. Variables with VIF values exceeding 10 are generally considered to exhibit significant multicollinearity, though some researchers adopt more conservative thresholds, such as 5.

Table 1: VIFs for the independent variables.

Company	Interpretation
Honeywell	Open (55.4532), Low (62.8827): Both variables have VIFs well above the 10 threshold, suggesting strong multicollinearity. High (10.4461): Just below the 10 threshold, indicating moderate multicollinearity. Volume (1.0321), Change (1.1166): Low VIFs, implying minimal multicollinearity.
Monsard:	Open (34.5265), High (11.7521), Low (49.9076): VIFs above 10, pointing to substantial multicollinearity. Volume (1.0160), Change (1.0027): Minimal multicollinearity indicated by low VIFs.
Oando:	Open (660.8080), High (600.7262), Low (621.8136): Extremely high VIFs, signifying severe multicollinearity. Volume (1.0286), Change (1.2178): Low VIFs, suggesting negligible multicollinearity.
Zenith	Open (920.2981), High (489.4860), Low (418.9664): Very high VIFs, indicating critical multicollinearity issues. Volume (1.0818): Low VIF, reflecting minimal multicollinearity. Change (2.1053): A VIF above 2, suggesting slight multicollinearity.

For companies like Oando and Zenith, where VIFs for certain variables are exceedingly high, multicollinearity poses a significant concern. This can lead to unreliable regression coefficients and challenges in determining the individual effect of each predictor. Addressing multicollinearity may involve removing highly correlated variables, combining similar variables, or employing dimensionality reduction techniques. In contrast, variables

with low VIFs across all companies suggest minimal multicollinearity, allowing for more reliable interpretation of regression coefficients.

Table 2: Multiple linear regression results for five companies

Companies		Estimate	Standard Error	t-value	P-value
Honeywell	Intercept	0.0085	0.00526	1.630	0.1030
	Open	0.8885	0.01038	85.613	0.0000
	High	0.0013	0.00407	0.334	0.7380
	Low	0.1092	0.01091	10.011	0.0000
	Volume	-0.0000007	0.0000087	-0.083	0.9340
	Change	2.667	0.06810	39.165	0.0000
Monsard	Intercept	-0.0103	0.00576	-1.786	0.0744
	Open	0.9373	0.00821	114.061	0.0000
	High	-0.0033	0.00469	-0.705	0.4808
	Low	0.0709	0.01013	7.004	0.0000
	Volume	-0.000016	0.000011	-1.448	0.1479
	Change	0.3015	0.03927	7.677	0.0000
Oando	Intercept	0.0229	0.01873	1.226	0.220
	Open	-0.1211	0.01812	-6.685	0.0000
	High	0.6056	0.01698	35.660	0.0000
	Low	0.5121	0.01798	28.490	0.0000
	Volume	0.000052	0.000058	0.899	0.3690
	Change	4.3600	0.3830	11.383	0.0000
Zenith	Intercept	-0.0499	0.0207	-2.408	0.0162
	Open	0.2286	0.0217	10.535	0.0000
	High	0.3943	0.0156	25.176	0.0000
	Low	0.3788	0.0148	25.497	0.0000
	Volume	0.0001	0.0001	1.232	0.2181
	Change	11.5284	0.3642	31.650	0.0000

Discussion of Multiple linear regression models for the five companies

Table 2 shows results from the multiple linear regression analysis for the five companies show. **Honeywell**: The intercept is not statistically significant with a p-value of 0.103, suggesting that it doesn't significantly contribute to the model. However, the variable "Open" is highly significant (p = 0.0000) with a large coefficient of 0.8885, meaning the

opening price has a strong positive effect on the stock price. "Low" is also significant (p = 0.0000), with a coefficient of 0.1092, indicating that the lowest price of the day positively affects the stock price. On the other hand, "High" (p = 0.7380) and "Volume" (p = 0.9340) are not significant, indicating that these variables don't have a strong influence on the stock price for Honeywell. "Change" is highly significant with a large coefficient of 2.667, suggesting that changes in stock price are a significant driver for the company's price.

Monsard: For Monsard, the intercept is marginally significant with a p-value of 0.0744, but it's close to the threshold for significance. "Open" is highly significant (p = 0.0000), and "Low" (p = 0.0000) also has a positive effect on stock price. However, "High" (p = 0.4808), "Volume" (p = 0.1479), and "Change" (p = 0.0000) are significant, but their influence is not as strong as for the other companies.

Oando: The intercept is not significant (p = 0.220), indicating its lesser impact. "Open" (p = 0.0000), "High" (p = 0.0000), and "Low" (p = 0.0000) have large coefficients and are highly significant, suggesting that these factors significantly drive Oando's stock price. "Volume" (p = 0.3690) is not significant, and while "Change" is highly significant (p = 0.0000), its effect is slightly less impactful compared to other variables.

Zenith: For Zenith, the intercept is significant (p = 0.0162), suggesting its importance in the model. "Open" (p = 0.0000), "High" (p = 0.0000), "Low" (p = 0.0000), and "Change" (p = 0.0000) are all highly significant, with strong positive coefficients. "Volume" (p = 0.2181) is not significant, indicating that trading volume does not affect the stock price in this case.

Therefore, "Open," "Low," and "Change" consistently emerge as significant predictors for the stock prices of these companies, while "Volume" and "High" generally have less impact. These findings suggest that price dynamics and changes in stock values are critical factors in determining the stock prices for these companies.

Discussion of LASSO regression models for the five companies

Table 4 presents the results of LASSO regression analyses conducted on the five companies: Honeywell, Monsard, Oando, and Zenith. For each company, the table provides the estimated coefficients (Estimates), standard errors (Std. Error), z-values, and

p-values for various predictors, including Intercept, Open, High, Low, Volume, and Change.

The following were observed:

- **Honeywell:** The 'High' predictor has a significant z-value of 3.256 and a p-value of 0.0011, suggesting a strong positive relationship with the dependent variable. The 'Volume' predictor has a negative coefficient (-0.000002) with a p-value of 0.10542, indicating a weak negative relationship, though not statistically significant at the 0.05 level.
- **Monsard:** The 'Open' predictor shows a positive relationship with a z-value of 2.916 and a p-value of 0.00355. The 'High' and 'Low' predictors have positive coefficients with z-values around 1.655 and 1.595, respectively, but their p-values (0.09798 and 0.11081) suggest these relationships are not statistically significant at the 0.05 level.
- **Oando:** The 'High' predictor has a notably high z-value of 18.991 and a p-value of 0.0000, indicating a very strong positive relationship. The 'Open' and 'Low' predictors also show positive relationships with significant z-values and p-values below 0.05. The 'Volume' predictor has a negative coefficient with a z-value of -1.580 and a p-value of 0.1142, suggesting a weak negative relationship, though not statistically significant.
- **Zenith:** The 'High' predictor exhibits a strong positive relationship with a z-value of 7.066 and a p-value of 0.0000. The 'Open' and 'Low' predictors also show positive relationships with significant z-values and p-values below 0.05. The 'Volume' predictor has a coefficient close to zero with a z-value of -0.004 and a p-value of 0.99688, indicating an insignificant relationship.

Table 4: LASSO regression results for five companies

Companies		Estimate	Standard Error	Z-value	P-value
Honeywell	Intercept	0.007061	0.001181	5.981	0.0000
	Open	0.000608	0.000229	2.658	0.0078
	High	0.000662	0.000203	3.256	0.0011
	Low	0.000611	0.000224	2.728	0.00637
	Volume	-0.000002	0.000001	-1.619	0.10542

	Change	0.000000	0.009884	0.000	1.00000
Monsard	Intercept	0.037360	0.005325	7.016	0.0000
	Open	0.002460	0.000844	2.916	0.00355
	High	0.001329	0.000803	1.655	0.09798
	Low	0.001365	0.000856	1.595	0.11081
	Volume	-0.000005	0.000007	-0.751	0.45262
	Change	0.000000	0.022740	0.000	1.00000
Oando	Intercept	0.008666	0.000968	8.957	0.0000
	Open	0.000062	0.000026	2.426	0.0153
	High	0.009911	0.000522	18.991	0.0000
	Low	0.000057	0.000026	2.201	0.0278
	Volume	-0.000004	0.000002	-1.580	0.1142
	Change	0.000000	0.012219	0.000	1.0000
Zenith	Intercept	0.024719	0.002938	8.412	0.0000
	Open	0.000154	0.000056	2.759	0.00579
	High	0.002853	0.000404	7.066	0.0000
	Low	0.000133	0.000057	2.354	0.01858
	Volume	0.000000	0.000012	-0.004	0.99688
	Change	0.000000	0.018931	0.000	1.00000

Model Selection

Table 5 compares the performance of Multiple Linear Regression (MLR) and Least Absolute Shrinkage and Selection Operator (LASSO) regression models across five companies: Honeywell, Monsard, Oando, and Zenith. The evaluation metrics used include the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

AIC and BIC Analysis:

Both AIC and BIC are metrics used for model selection, balancing model fit with complexity. Lower values of AIC and BIC indicate better-fitting models. In this comparison, LASSO consistently outperforms MLR in terms of both AIC and BIC across all five companies. Honeywell's LASSO model has an AIC of -12,112.64 and a BIC of -12,086.53, significantly lower than Honeywell's MLR model, which has an AIC of -

2,690.541 and a BIC of -2,653.214. This trend suggests that LASSO provides more parsimonious models with better fit than MLR.

MSE and RMSE Analysis:

MSE measures the average squared difference between observed and predicted values, while RMSE provides the square root of this average, offering error magnitude in the original units. Lower MSE and RMSE values indicate better model performance. LASSO models generally exhibit lower MSE and RMSE values compared to their MLR counterparts. Honeywell, the LASSO model has an MSE of 0.000021 and an RMSE of 0.004593, whereas the OLS model has an MSE of 0.00998 and an RMSE of 0.09992. This pattern is consistent across the other companies, highlighting LASSO's superior predictive accuracy.

Company-Specific Observations:

- Honeywell: LASSO significantly reduces both AIC and BIC compared to MLR, with much lower MSE and RMSE values, indicating a better-fitting model.
- Monsard: LASSO shows a considerable decrease in AIC and BIC, along with lower MSE and RMSE, emphasizing its effectiveness.
- Oando: The LASSO model outperforms MLR with lower AIC, BIC, MSE, and RMSE values, indicating enhanced fit and predictive accuracy.
- Zenith: LASSO achieves significantly lower AIC and BIC, and its MSE and RMSE are much lower than those of MLR, reinforcing its superiority.

Therefore, across all five companies, LASSO regression consistently demonstrates superior performance over MLR regression, as evidenced by lower AIC, BIC, MSE, and RMSE values. This suggests that LASSO not only provides more parsimonious models but also enhances predictive accuracy.

Model Fitting

Table 5. Comparing MLR and LASSO regression for five (5) Companies

Criteria	Honeywell		Monsard		Oando		Zenith	
	OLS	LASSO	OLS	LASSO	OLS	LASSO	OLS	LASSO

AIC	-2690.541	-1211 2.64	-1948.877	-5326.777	-1961.45	-7190.392	-107.4518	-1041 2.2
BIC	-2653.214	-1208 6.53	-1912.899	-5305.896	-1997.544	-7165.494	-70.1251	-1038 8.61

Table 6. Model Prediction for five (5) Companies

Criteria	Honeywell		Monsard		Oando		Zenith	
	OLS	LASSO	OLS	LASSO	OLS	LASSO	OLS	LASSO
MSE	0.00998	0.000021	0.012345	0.000851	0.26745	0.000213	0.054079	0.000064
RMSE	0.09992	0.004593	0.11110	0.02917	0.51716	0.01459	0.23254	0.008012

CONCLUSION

The analysis compared MLR and LASSO regression models for five companies: Honeywell, Monsard, Oando, and Zenith. Correlation analysis revealed strong relationships between stock price variables, particularly between opening, closing, high, and low prices across all companies. Multicollinearity was detected using Variance Inflation Factors (VIF), with severe issues in Oando and Zenith for opening, high, and low price variables.

The findings indicate that "Open," "Low," and "Change" are consistently significant predictors across the companies, highlighting their substantial influence on stock prices. Conversely, "High" and "Volume" generally exhibit less impact, with many instances of statistical insignificance. These results suggest that price dynamics and fluctuations play a critical role in determining stock prices, while trading volume and daily highs may have a lesser effect.

LASSO regression consistently outperformed MLR across all companies based on AIC, BIC, MSE, and RMSE metrics for all the five comapnies. LASSO provided more parsimonious models with lower prediction errors due to its regularization technique that shrinks irrelevant coefficients to zero and enhances model interpretability by reducing overfitting and multicollinearity issues.

Recommendations

Form the findings of this study, we recommend that LASSO regression should be adopted dealing with data with Multiple independent variable, given its consistent outperformance over MLR across all companies studied. Also regular model evaluations and updates should be carried out to account for changing market dynamics and maintain predictive accuracy.

REFERENCES

David E. Rapach and Guofo Zhou (2013). Forecasting Stocks returns. Chapter 6 in Handbook of Economic Forecasting 2013, vol 2, pp328-383.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, 1(2), 293-314.

Ivo Welch and Amit Goyal (2008). A comprehensive look at the empirical Performance of Equity Premium prediction

Mishra, C., Mohanty, L., Rath, S., Patnaik, R., & Pradhan, R. (2020). Application of backward elimination in multiple linear regression model for prediction of stock index. In *Intelligent and Cloud Computing: Proceedings of ICICC 2019* (Vol. 2, pp. 543–551). Springer Singapore.

Robert Tisbhrani (1996). Regression Shrinkage and selection via the LASSO. Journal of the Royal Statistical Society. Series B Vol. 58.No. 1 (1996), pp 267-288.

Seethalakshmi, R. (2018). Analysis of stock market predictor variables using Linear Regression. *International Journal of Pure and Applied Mathematics*, 119 (15), 369-378.

Yaman, Y., Ozturk, M., & Yilmaz, M. (2024). Partially Protected LASSO: A novel approach to variable selection in high-dimensional data. *Journal of Statistical Planning and Inference*, 220, 1-15.