Safaricom Stock Return Prediction

B.M Njuguna

2023-04-15

Contents

Introduction

Safaricom PLC, a Kenyan mobile operator, was established in 1997 as a fully-owned subsidiary of Telkom Kenya. In 2000, Vodafone Group PLC, a UK-based telecommunications company, acquired a 40% stake in the firm and assumed management responsibilities. In 2008, the Kenyan government made 25% of Safaricom's shares available to the public via the Nairobi Securities Exchange, leading to its initial listing on the Nairobi Stock Exchange (NSE) in June of that year under the "SCOM" ticker symbol. As the largest telecommunications provider in Kenya and one of the most profitable companies in East and Central Africa, Safaricom is most well-known for its mobile banking SMS-based service, **MPESA**.

Objective

The objective of this study is to use a machine learning algorithm to predict, the direction of Safaricom PLC's stock returns. By leveraging machine learning techniques and analyzing historical data on the company's stock prices and associated financial indicators, the study aims to identify patterns and trends that can predict future stock returns.

This study will also compare the different machine learning algorithms and choose the one with the highest accuracy. The data analysis was done in R.

Data Preprocessing

A glimpse of the first few rows of SCOM stock prices data is as shown below;

Date	Open	High	Low	Close	Volume
2023-04-14	18.00	18.00	17.50	17.80	2620100
2023-04-13	18.35	18.40	17.25	17.85	2163500
2023-04-12	18.60	19.00	18.20	18.35	2908300
2023-04-11	18.50	19.30	18.35	18.45	5052300
2023-04-06	18.75	19.45	18.25	18.60	4375000
2023-04-05	18.20	18.50	18.00	18.05	24080000

The data contains 6 variables and 3,717 rows. The rows contains the different stock prices from June 9, 2008 to April 14, 2023. The 6 variables or columns contains the date, Open, high, low, close and Volume of the stock prices as can be seen above. The summary statistics are shown below;

	mean	variance	skewness	kurtosis
Open	17.234	1.3860e+02	0.36028	1.9286
High	17.443	1.4194e+02	0.35329	1.9176
Low	17.007	1.3447e+02	0.36737	1.9472
Close	17.197	1.3779e+02	0.36473	1.9432
Volume	11187316.540	2.3672e+14	8.71636	163.5540

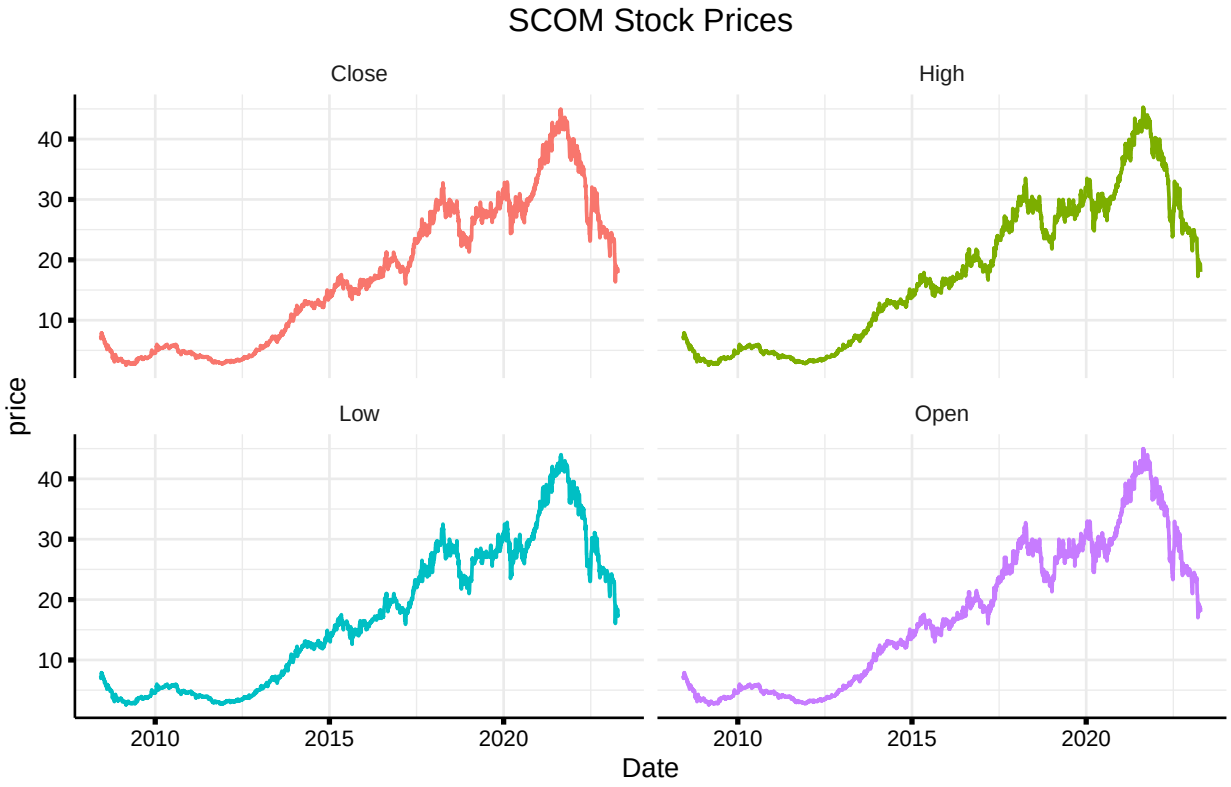
The mean is the average value of each variable (Open, High, Low, Close, and Volume) in the data. Variance is a measure of how spread out the values are from the mean. A larger variance means that the values are more spread out, and a smaller variance means that the values are more tightly clustered around the mean. For example, the variance of the Open price is 138.60, meaning that the Open price varied quite a bit over the period of time covered by the data. Generally, the Open, high, low close and volume vaired over the time period evidences by the high variances.

skewness is the measure of the asymmetry of the distribution of values. A skewness of 0 means that the distribution is perfectly symmetrical, while positive or negative skewness means that the distribution is skewed to the right or left, respectively. For example, the skewness of the Open price is 0.36028, indicating a slight right skew.

kurtosis is the measure of the "peakedness" of the distribution of values. A kurtosis of 0 means that the distribution has a normal (bell-shaped) curve, while positive kurtosis means that the distribution is more peaked than normal, and negative kurtosis means that the distribution is flatter than normal. For example, the kurtosis of the Volume variable is 163.5541, indicating a very high level of peakedness in the distribution.

Time series Plot

The time series plot of the Open, High, Low, and close are as shown below;



Based on the time series plot, it appears that there was a discernible upward trend in the stock prices until approximately 2021, after which point the prices began to decrease. This suggests that some significant event or factor may have had an impact on investor sentiment, leading to a decline in demand for the stock. However, without further information about the specific stock, its industry, and the broader economic and political context in which it operates, it is difficult to draw definitive conclusions regarding the underlying causes of this trend. A more comprehensive analysis is necessary to fully understand the dynamics at play.

Feature Engineering

The Dependent Variable

The dependent variable utilized in this study was the directional movement of stock returns. Specifically, if the returns of a given day were negative, this was denoted as "Down", while if the returns were positive,

this was recorded as "Up". However, in instances where the returns were exactly equal to zero, these were also classified as "Down" for the sake of simplicity.

To provide further clarification, the designation of "Up" for positive returns is intended to convey the notion of profit, signifying a gain in value for the stock or asset being analyzed. Conversely, the classification of "Down" for negative returns denotes a loss or decline in value. The plot for the stock prices for the last 60 days is as shown below;

The directional dependent variable in the data was shifted forward by one day, which means that the direction value for each day now indicates the direction of the return movement for the following day.

For example, suppose the original Direction value for Day 1 was "Up". By shifting the Direction column forward by one day, the "Up" direction value for Day 1 now corresponds to the returns movement for Day 2.

So, the Direction column is now indicating the stock price movement for each day, based on the direction value for the following day. This shift was made to align the Direction values with the returns movements, making it easier to analyze and interpret the data.

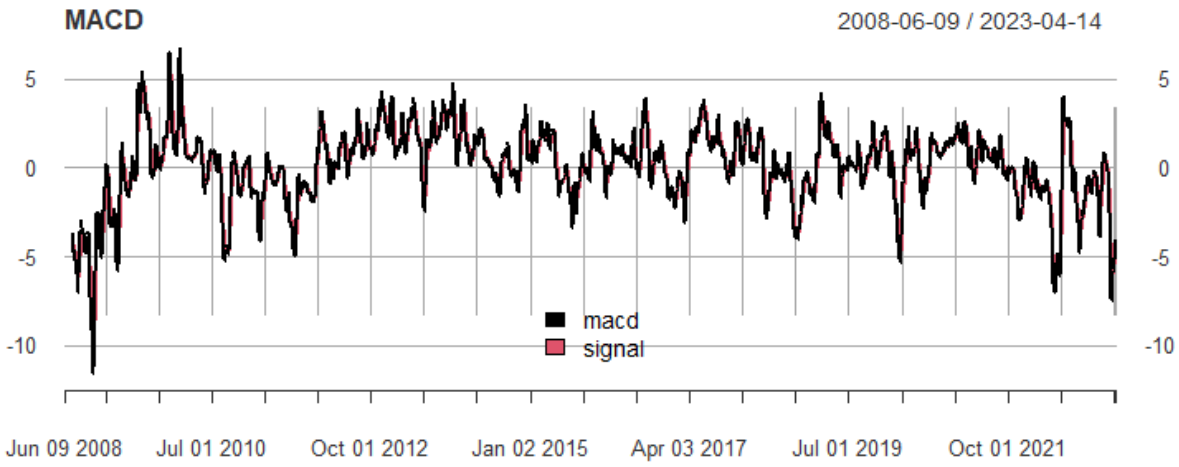
The Independent Variables.

The following **technical indicators** were used as the independent variables.

Moving Average Convergence Divergence (MACD)

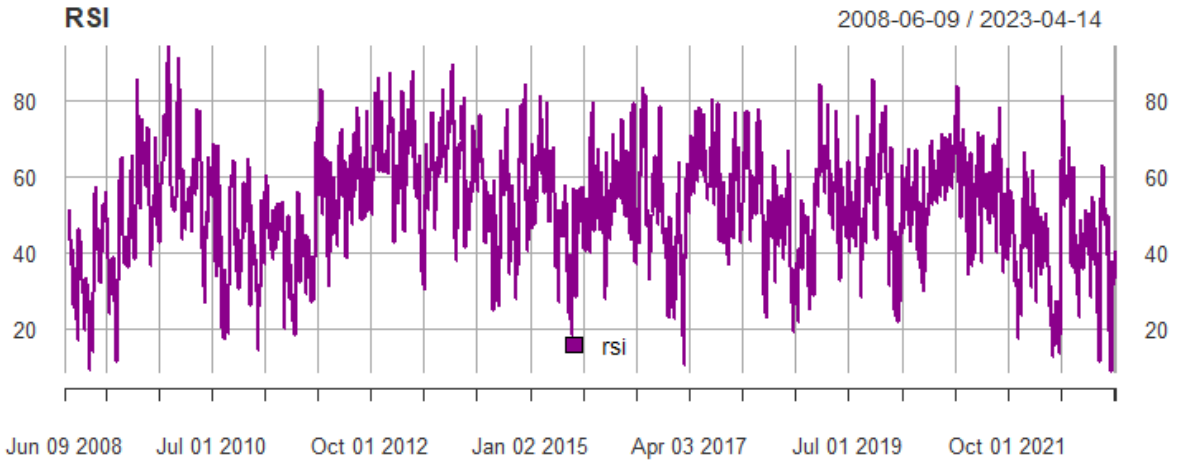
The MACD is a trend-following momentum indicator that shows the relationship between two exponential moving averages. It is calculated by subtracting the 26-period Exponential Moving Average from the 12-period Exponential Moving Average.

When the Moving Average Convergence Divergence (MACD) line crosses from below to above the signal line (9-period Exponential Moving Average), it is considered to be a bullish signal. Conversely, when the MACD line crosses from above to below the signal line, it is considered bearish. A bullish MACD crossover indicates that the stock prices may experience an upward trend, whereas a bearish crossover suggests a possible downward trend.



Relative Strength Index (RSI)

The RSI is a momentum oscillator that measures the strength and speed of a stock's price movements. It ranges from 0 to 100 and is calculated based on the average gains and losses over a specific time period. In simpler terms, it is used to measure the strength of stock price actions. It is calculated as the Ratio of Upward movement to Downward price movement over specified period of time.

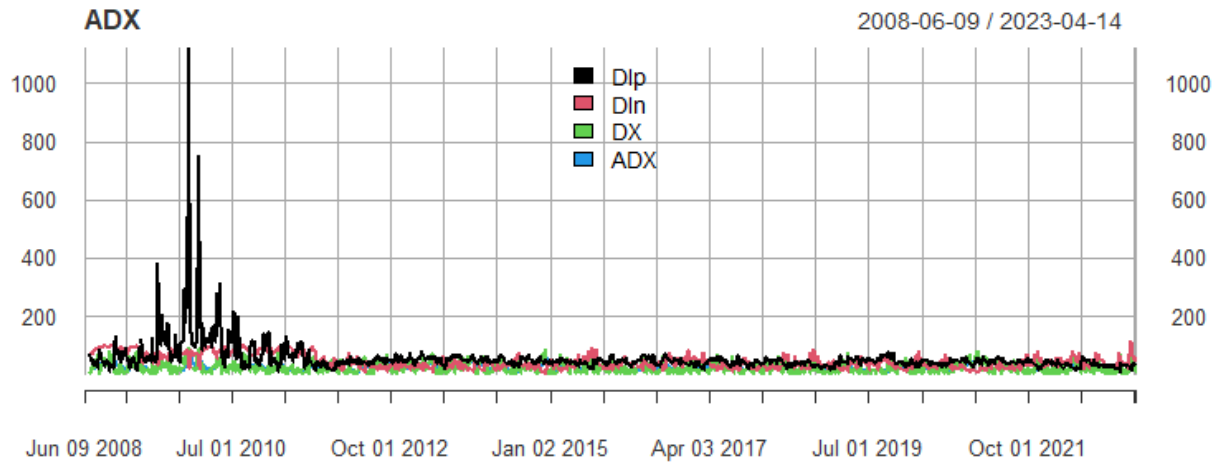


Average Directional Index (ADX)

The ADX indicator measures the strength of the trend in the stock prices. It's calculation results from;

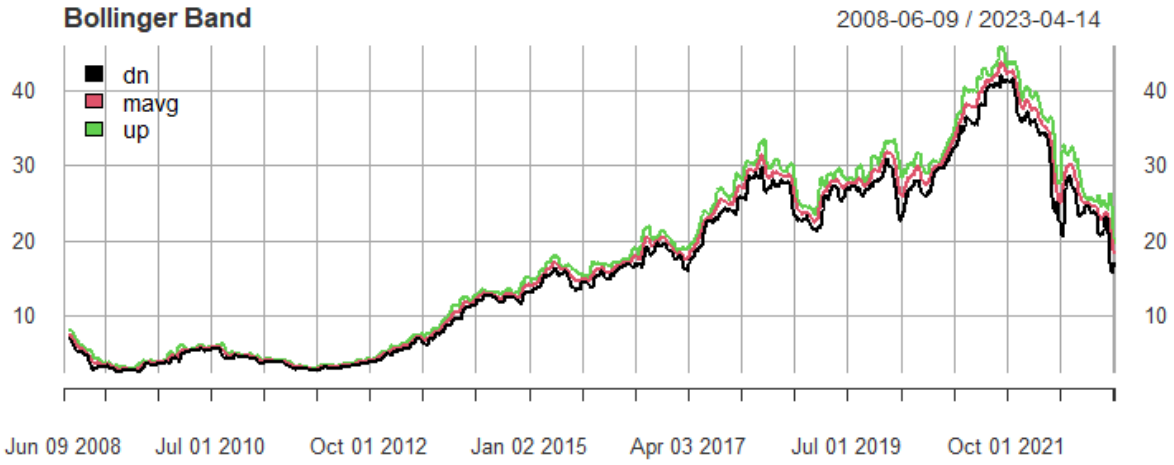
1. *DI+* - The +D measures the strength of upward price movement
2. *DI-* - The -D measures the strength of downward price movement
3. *DX* - The DX is the difference between the DI+ and DI-, divided by their sum, multiplied by 100.

The ADX indicator is calculated as the moving average of DX.

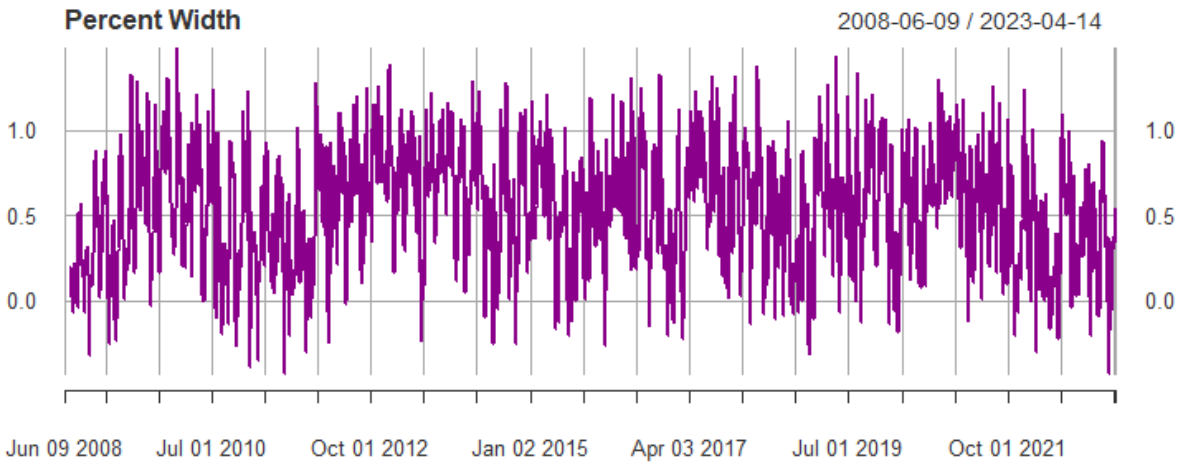


Bollinger Bands

Bollinger Bands represent a commonly used indicator that closely monitors volatility. This dynamic indicator features three distinct lines, with a moving average placed in the middle and an upper and lower band situated at two standard deviations from the moving average. The bands are designed to track fluctuations in the stock price and provide valuable insights into whether the stock is overbought or oversold. Specifically, Bollinger Bands are useful for identifying times when a stock is performing outside of its normal range. When the stock price climbs above the upper band, it indicates that it is overbought and may be due for a pullback.



Also from the bollinger bands, there is the **percent Bandwidth**. It is calculated by dividing the difference between the current price and the lower Bollinger Band by the difference between the upper and lower Bollinger Bands. In terms of volatility, the percent B (%B) indicator provides a measure of how volatile a market is. When %B is high, it indicates that the market is volatile and prices are moving away from the moving average. Conversely, when %B is low, it indicates that the market is less volatile and prices are closer to the moving average.



Stochastic Oscillator

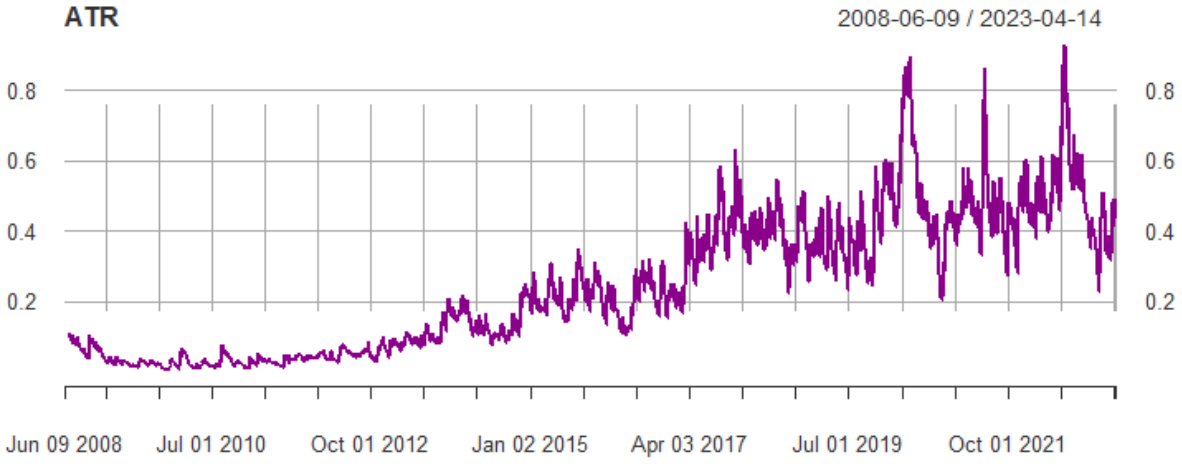
The stochastic oscillator is a tool that helps investors understand how a stock's price compares to its price range over a certain period of time. It measures momentum and can help identify when a stock is overbought or oversold. There are three key indicators:

1. *fastK* - which represents the stock's closing price compared to its range.
2. *fastD* - which smooths out fluctuations in fastK.
3. *slowD*- which further smoothes out the oscillator.

Values above 80 indicate an overbought condition, while values below 20 indicate an oversold condition.

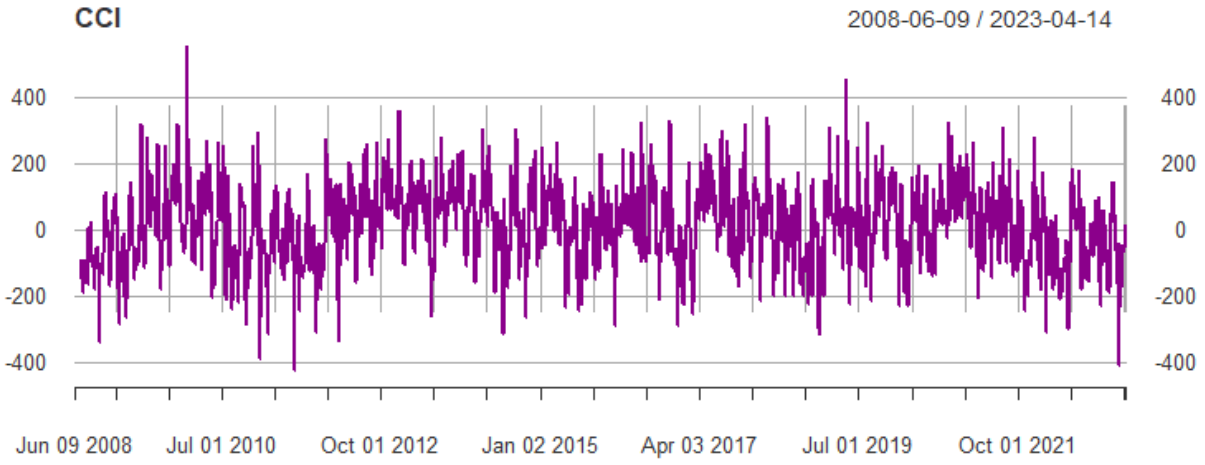
Average True Range (ATR)

Average True Range is a technical indicator commonly used in financial markets to measure the volatility of an asset. ATR is calculated using a formula that takes into account the price range of an asset over a specified period of time, and provides a measure of the average movement of the asset's price on a day-to-day basis.



Commodity Channel Index (CCI)

Commodity Channel Index (CCI) is a technical indicator used to identify cyclical trends in securities. It is often used to identify overbought and oversold levels. The CCI measures the difference between the current price of a security and its average price over a given period of time, adjusted for normal deviations in price. The indicator oscillates above and below zero. A positive value suggests that the price is above the average, while a negative value suggests that the price is below the average.



Rolling Closing Price

This variable was created by taking the rolling standard deviation of the closing price (Cl) over a window of 10 periods. This means that it captures the volatility of the stock over the past 10 periods, as measured by the standard deviation of the price movements. The rolling standard deviation is used instead of the standard deviation of the entire dataset to provide a more current estimate of volatility that takes into account recent price movements. By including this variable in the analysis, we can better understand how the volatility of the stock is related to its returns movement.

Signal-to-Noise Ratio of the High-Low-Close

The SNR is a measure of the strength of the signal (the trend of the price movements) relative to the noise (the random fluctuations in the price). A higher SNR indicates a stronger trend in the price movements.

The Volume Weighted Average Price (VWAP)

The Volume Weighted Average Price (VWAP) was calculated and used as one of the independent variables in the analysis. VWAP is a measure of the average price at which a stock has traded throughout the day, weighted by the volume of shares traded at each price. It provides insight into whether a stock is being bought or sold at a good price and is often used by institutional investors to make trading decisions. The VWAP was calculated using a 5-period window and was included in the analysis as an independent variable.

Thus, the selected independent variables were;

1. Moving Average Convergence Divergence (MACD)
2. Relative Strength Index (RSI)

3. Average Directional Index (ADX)
4. The percent Bandwidth from the Bollinger Bands
5. *fastK*[1] from the Stochastic Oscillator.
6. The Average True Value (ATR)
7. Commodity Channel Index (CCI)
8. Signal-to-Noise Ratio
9. The Volume Weighted Average Price
10. Rolling standard deviation of the closing price (Cl) over a window of 10 periods [2]. The complete data set was as follows;

MACD	RSI	ADX	VOL	SNR	fastK	ATR	CCI	pctB	VWAP	Direction
-4.1342	18.018	16.591	0.18439	3.43636	0.00000	0.10051	-117.819	0.11866	-0.00930	Down
-4.7167	14.906	19.798	0.19214	1.15849	0.00000	0.10034	-125.584	0.09700	-0.01598	Up
-4.4337	35.370	12.958	0.19896	0.38327	0.15000	0.11690	-79.997	0.23905	-0.01080	Down
-4.3017	32.214	14.663	0.20521	3.55382	0.11111	0.09460	-78.232	0.24177	-0.00409	Down
-4.2668	29.046	16.383	0.18567	4.87901	0.06667	0.07973	-76.928	0.24548	-0.00458	Down
-4.6368	21.348	21.375	0.17670	9.65697	0.00000	0.10315	-97.355	0.17959	-0.00506	Down

Methodology

The Machine Learning Algorithm

In this study, three machine learning algorithms were used to predict the movement of SCOM stock returns: Random Forest, Lasso Regression, and Ridge Regression.

Random Forest is a powerful ensemble learning method that builds multiple decision trees based on a random subset of the predictors and observations, and aggregates their predictions to improve accuracy and reduce overfitting. Random Forest is known for its ability to handle complex interactions and nonlinearities in the data, making it a good choice for this study where technical indicators were used as predictors.

Lasso Regression and Ridge Regression are both linear regression techniques that use regularization to control the complexity of the model and prevent overfitting. Lasso Regression performs L1 regularization, which shrinks the coefficients of less important predictors to zero, effectively selecting a subset of the most relevant predictors. Ridge Regression performs L2 regularization, which penalizes large coefficients and encourages small, smooth coefficients. Both techniques are commonly used in feature selection and variable importance analysis.

The selection of these three models was likely based on their strengths and suitability for the task at hand. Random Forest was chosen for its ability to handle complex data and interactions, while Lasso and Ridge Regression were chosen for their ability to perform variable selection and regularization. By using three different models, the study was able to compare and validate the results, and choose the best model based on accuracy and other performance metrics.

[1] In the stochastic oscillator, fastK is the first line of the stochastic calculation, which represents the current closing price relative to the recent range of prices. It is typically calculated over a period of time, such as 14 days, and is used to determine the momentum of the stock or asset being analyzed.

[2] named as VOL in the data set

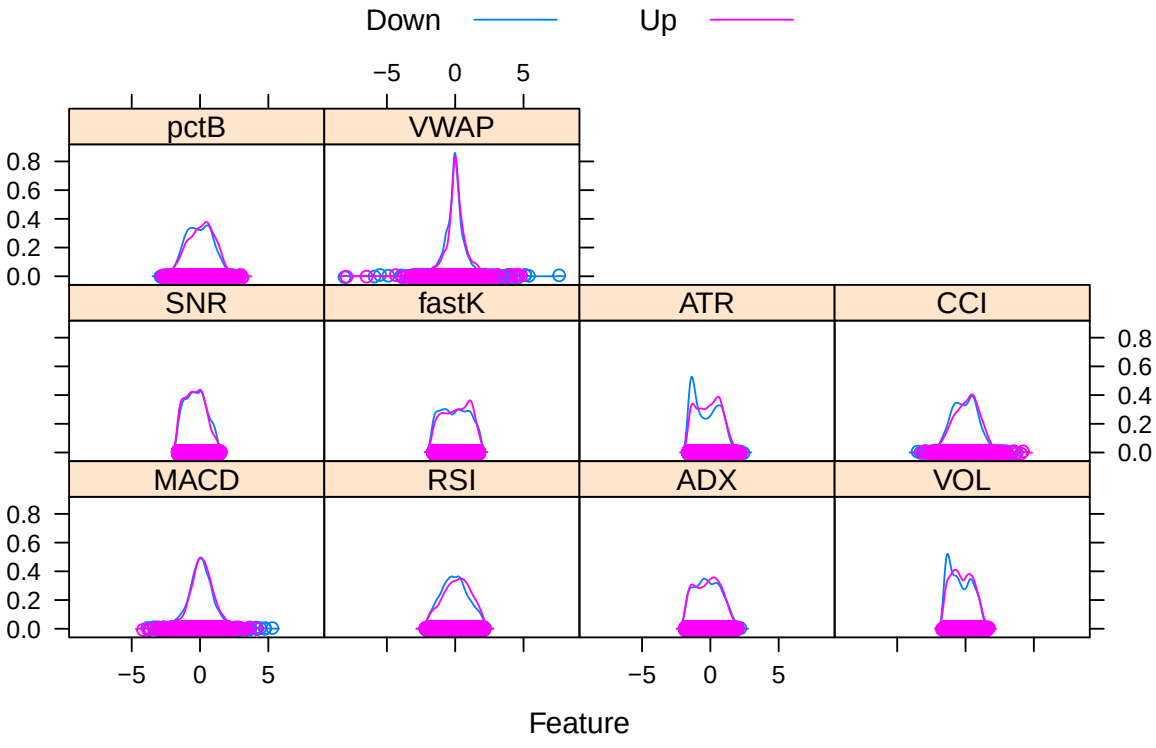
Data Splitting

After performing feature engineering, the dataset consisted of 3,689 rows and 8 variables (columns), which includes the dependent variable. To train the model, 70% of the data was used, while the remaining 30% was reserved for testing and prediction.

Data Analysis and Results.

Before training the model, centering and scaling were applied to the technical indicators to bring them to a common scale. This is important because some of the technical indicators could have been measured on different scales, which could lead to some variables having a higher influence on the model than others. By centering and scaling the variables, they are transformed to have a mean of 0 and a standard deviation of 1, which ensures that all variables have the same weight in the model.

Additionally, the Yeohjon's method was used to transform the variables to have a more symmetric distribution. This is because many statistical models, including the ones used in this study, assume that the variables are normally distributed. The Yeohjon's transformation can be applied to variables that have skewness or kurtosis, which means that they deviate from a normal distribution. By transforming the variables to be more normally distributed, the statistical models used in the study can be applied more reliably. The transformed variables were as follows;



Model Training

In order to train the predictive model, k-fold cross-validation was employed. The dataset was divided into k ($k = 10$) subsets (or folds) of approximately equal size. The model was then trained on k-1 subsets, and the remaining subset was used as a validation set to evaluate the model's performance. This process was

repeated k times, with each subset being used as the validation set once. The results from the k iterations were then averaged to give an estimate of the model's performance on new, unseen data. This method helps to prevent overfitting and provides a more reliable estimate of the model's predictive ability.

Random Forest

From the 10-fold cross validation training process, the results were as follows;

```
Random Forest

2956 samples
  10 predictor
   2 classes: 'Down', 'Up'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2660, 2660, 2661, 2661, 2660, 2660, ...
Resampling results across tuning parameters:

  mtry  Accuracy  Kappa
  2     0.60082   0.10108
  3     0.60150   0.10608
  4     0.60725   0.11804
  6     0.59981   0.10707
  8     0.59911   0.10547

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 4.
```

From the results, the threshold that gave the highest **Youden's J statistic** [3] was chosen. In practice, Youden's J is often used to determine the optimal cutoff point for a binary classifier. The cutoff point is the probability threshold above which the classifier assigns a positive label, and below which it assigns a negative label. The optimal cutoff point is the one that maximizes the Youden's J score, as this corresponds to the point where the true positive rate is maximized relative to the false positive rate. It was as displayed below;

	mtry	prob_threshold	Sensitivity	Specificity	Accuracy	J
233	8	0.62245	0.49493	0.60187	0.53721	0.0968

The Youden's J is usually calculated as $(Sensitiviy + Specificity) - 1$. For the Random Forest algorithm, the highest J obtained was 0.12841. A Youden's J score of 0.12841 indicates that there is a slight imbalance between the true positive rate and the false positive rate of a binary classification model. This means that the model is not very effective at distinguishing between positive and negative cases. The positive class in this case was "Down" level.

In predicting the test data, the results were as follows;

```
Confusion Matrix and Statistics

          Reference
Prediction Down  Up
```

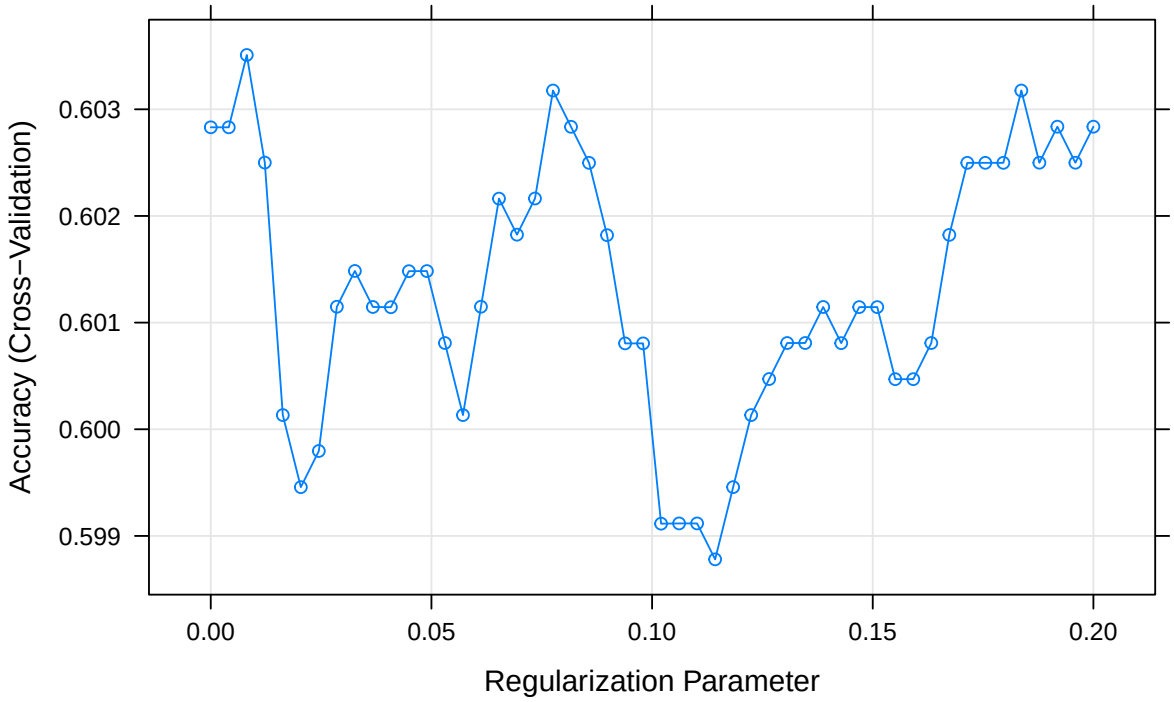
[3] Youden's J score is a statistic used in binary classification to measure the performance of a diagnostic test. It takes into account both sensitivity and specificity of the test, and ranges from 0 to 1. A score of 0 indicates that the test has no diagnostic accuracy, while a score of 1 indicates perfect accuracy. A score of 0.5 indicates that the test is no better than chance.

```
      Down  292 155
      Up    146 146

               Accuracy : 0.593
                 95% CI : (0.556, 0.628)
    No Information Rate : 0.593
    P-Value [Acc > NIR] : 0.516

                  Kappa : 0.152

 Mcnemar's Test P-Value : 0.645

            Sensitivity : 0.667
            Specificity : 0.485
         Pos Pred Value : 0.653
         Neg Pred Value : 0.500
             Prevalence : 0.593
         Detection Rate : 0.395
   Detection Prevalence : 0.605
      Balanced Accuracy : 0.576

       'Positive' Class : Down
```

The accuracy obtained was 58.7%. However, this study was interested in the Youden's J score. For the test data, the model achieved a Youden's J score of 0.138.

Lasso Regression

The results for lasso regression were as follows;



The Regularization (l1) parameter which gave the highest accuracy in the cross validation was $\lambda = 0.0040826$. Just as the Random Forest algorithm above, the threshold that gave the highest Youden's J statistic was as follows;

	alpha	lambda	prob_threshold	Sensitivity	Specificity	Accuracy	J
34	0	0	0.63878	0.39881	0.74739	0.53654	0.1462

For the lasso model, the highest J obtained was 0.153. The following results were obtained after predicting the test data;

```
Confusion Matrix and Statistics

          Reference
Prediction Down  Up
      Down  177 270
      Up     71 221

               Accuracy : 0.539
                 95% CI : (0.502, 0.575)
    No Information Rate : 0.664
    P-Value [Acc > NIR] : 1

                  Kappa : 0.137

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.714
            Specificity : 0.450
         Pos Pred Value : 0.396
```

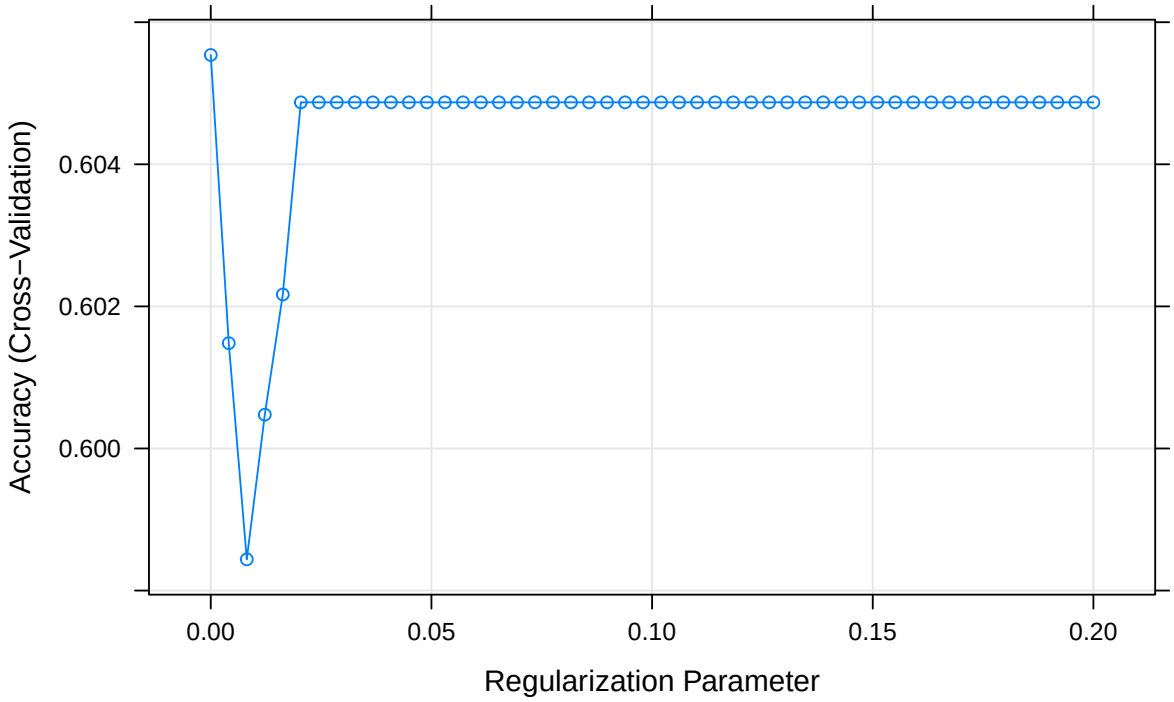
```
         Neg Pred Value : 0.757
             Prevalence : 0.336
         Detection Rate : 0.240
   Detection Prevalence : 0.605
      Balanced Accuracy : 0.582

       'Positive' Class : Down
```

The model achieved an accuracy of 55.3% and a Youden's J score of 0.143 when predicting the test data. It's worth noting that there was a probability threshold that gave a higher accuracy during cross-validation, but the metric of interest in this case was the Youden's J score. If the probability threshold that yielded the highest accuracy during cross-validation was used, the accuracy would have been 61.7%, as follows:

```
Confusion Matrix and Statistics

          Reference
Prediction Down  Up
      Down  398  49
      Up    235  57

               Accuracy : 0.616
                 95% CI : (0.58, 0.651)
    No Information Rate : 0.857
    P-Value [Acc > NIR] : 1

                  Kappa : 0.096

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.629
            Specificity : 0.538
         Pos Pred Value : 0.890
         Neg Pred Value : 0.195
             Prevalence : 0.857
         Detection Rate : 0.539
   Detection Prevalence : 0.605
      Balanced Accuracy : 0.583

       'Positive' Class : Down
```

Ridge Regression

The results of the Ridge regression was;



The l2 Reguralization parameter which yielded to the highest accuracy was $1 \times 10^{-}6$. while the probability threshold which yielded to the highest Youden's J score was as follows;

	alpha	lambda	prob_threshold	Sensitivity	Specificity	Accuracy	J
27	1	0	0.52449	0.86744	0.19096	0.60014	0.0584

On predicting the test data, the following results were obtained;

```
Confusion Matrix and Statistics

          Reference
Prediction Down  Up
      Down  177 270
      Up     73 219

               Accuracy : 0.536
                 95% CI : (0.499, 0.572)
    No Information Rate : 0.662
    P-Value [Acc > NIR] : 1

                  Kappa : 0.131

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.708
            Specificity : 0.448
         Pos Pred Value : 0.396
         Neg Pred Value : 0.750
             Prevalence : 0.338
         Detection Rate : 0.240
```

```
Detection Prevalence : 0.605
   Balanced Accuracy : 0.578

     'Positive' Class : Down
```

The model achieved an accuracy of 53.6% and a Youden's J score of 0.156. As a result, among the three models, Ridge regression was selected as the optimal model for predicting the stock return direction based on the Youden's J score. It is important to note that if prediction accuracy had been used as the comparative metric, the results might have been different.

Conclusion

After conducting a comparative study of ridge regression, random forest, and lasso regression, it was found that ridge regression yielded the highest Youden J statistic, indicating that it has the best balance of sensitivity and specificity for our classification task. It should be noted, however, that if accuracy was used as the sole metric for comparison, the results may have been different. Therefore, the choice of evaluation metric should be carefully considered based on the specific goals and requirements of the task. Overall, I recommend the use of ridge regression for similar classification tasks in the future.

The low accuracy and Youden's J statistic can be attributed to the violation of several assumptions in machine learning algorithms, including:

1. Violation of independence of observations: Stock price data often exhibits serial correlation, meaning that the return on a given day is likely to be related to the returns on previous days. This violates the assumption of independent observations, which many machine learning algorithms rely on.
2. Violation of homoscedasticity: Many machine learning algorithms assume that the variance of the errors is constant across all levels of the predictors. However, stock returns often exhibit heteroscedasticity, meaning that the variance of the returns may change over time, which can lead to biased predictions.
3. Violation of stationarity: Many machine learning algorithms assume that the data is stationary, meaning that the statistical properties of the data do not change over time. However, stock prices are often non-stationary and exhibit trends and other forms of temporal dependence, which can lead to inaccurate predictions if not properly accounted for.