Report For Stock Sentiment Analysis Using Machine Learning Techniques

KARTIK

Enrollment No: 22117065

Overview:

What is Stock Sentiment Analysis?

Stock sentiment analysis is the process of using natural language processing (NLP) and machine learning techniques to analyse textual data, such as news articles, social media posts, and financial reports, to gauge the sentiment or mood of the market towards a particular stock or the market as a whole. This analysis aims to classify text as positive, negative, or neutral, reflecting the general attitude of investors and market participants. By quantifying sentiment, analysts can identify trends and correlations between market sentiment and stock price movements. Sentiment analysis helps investors make informed decisions by providing insights into market psychology and potential future price actions based on prevailing sentiments.

Approach Of Project:

1. Data **Collection**:

- Gather historical stock prices from Yahoo finance over a historical period.
- Scrapped and Collect news headlines or articles related to the stocks from various sources such as Economic Times, MoneyControl etc.

2. Data **Preprocessing**:

- Clean stock price data, ensuring dates align and missing values are handled.
- Preprocess news headlines by removing special characters, converting to lowercase, and tokenizing.

3. Sentiment **Analysis**:

- Apply natural language processing (NLP) techniques to assign sentiment scores (positive, negative, neutral) to each news headlines.
- Use libraries or pre-trained models for sentiment analysis.

4. Feature **Extraction**:

- Convert pre-processed news headlines into numerical vectors using techniques like TF-IDF or word embeddings.

- Combine sentiment scores with stock price data to create a feature set for modelling.

5. **Labelling Data**:

- Define target labels based on stock price movements (e.g., increase, decrease, no change).
- Align news data with corresponding stock price changes to create labelled datasets.

6.Train**-Test Split**:

- Split the dataset into training and testing subsets to evaluate model performance.
- Ensure data is shuffled and split in a manner that avoids lookahead bias.

7. Model **Training**:

- Train multiple machine learning models (e.g., logistic regression, SVM, random forest) on the training data.
- Optimize hyperparameters using techniques like grid search or cross-validation.

8. Model **Evaluation**:

- Evaluate models using metrics like accuracy, precision, recall, F1 score, and ROC AUC.
- Select the best-performing model based on comprehensive evaluation.

9. Predict **and Evaluate on New Data**:

- Apply the trained model to new, unseen data to predict stock price movements.
- Assess the model's practical utility using trading metrics like Sharpe Ratio and win ratio.

10. Model **Refinement and Deployment**:

- Refine the model based on evaluation results, incorporating additional features or adjusting parameters as needed.
- Deploy the model for real-time prediction and continuously monitor its performance.

Defining Factors:

Stock_data Historical Period = 1 Year

Stocks in Model:

- TATASTEEL (TATASTEEL.NS)
- ZOMATO (ZOMATO.NS)
- HDFCBANK (HDFCBANK.NS)

Machine Learning Models used for Analysis:

- Logistic Regression
- Support Vector machines (SVM)
- Random Forests
- Neural Networks

For Evaluation and Judging the Performance:

• Accuracy: The ratio of correctly predicted instances to the total instances in the dataset.

• Precision: The ratio of correctly predicted positive observations to the total predicted positives.

• Recall: The ratio of correctly predicted positive observations to all the actual positives.

• F**1-Score**: The harmonic mean of precision and recall, providing a balance between the two metrics.

• ROC **AUC**: The area under the receiver operating characteristic curve, measuring the model's ability to distinguish between classes.

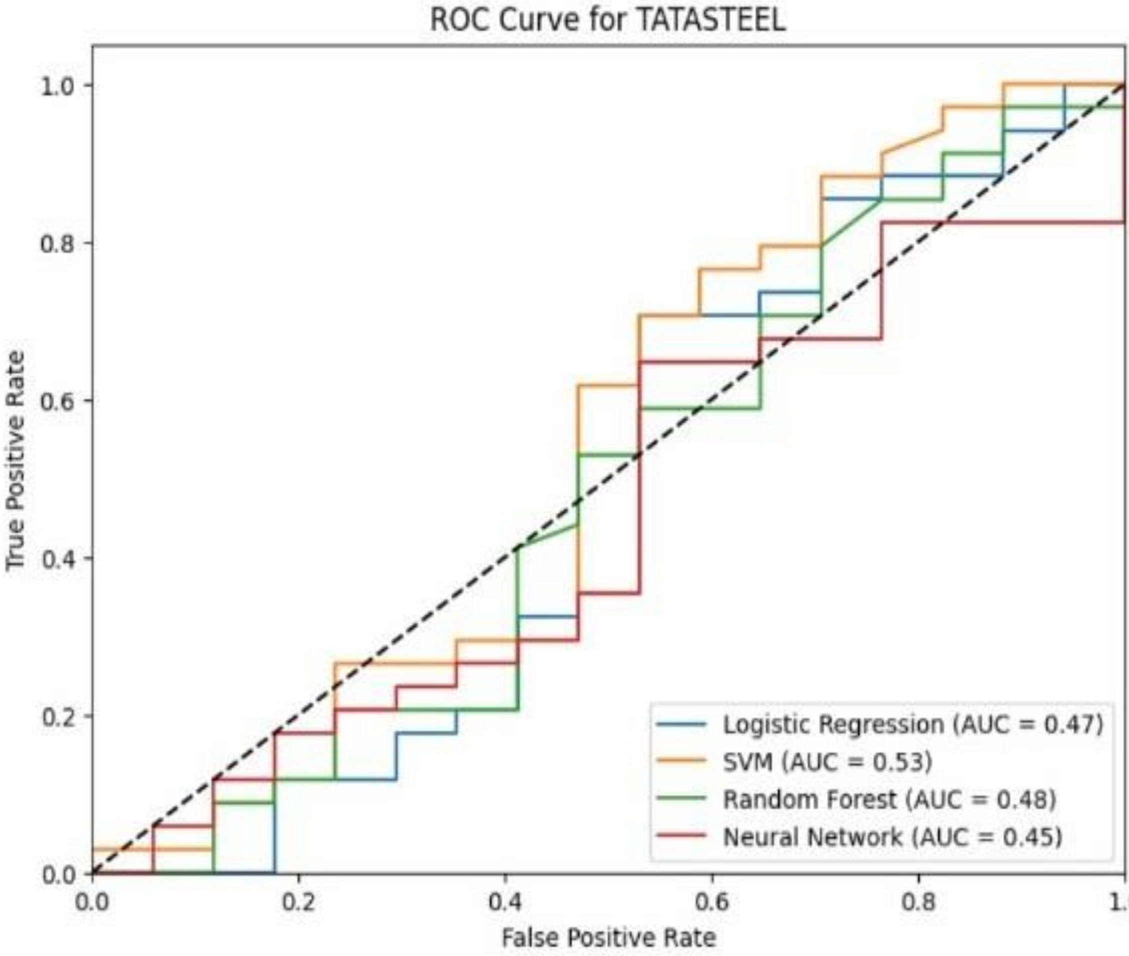
• Sharpe **Ratio**: The average return earned in excess of the risk-free rate per unit of volatility or total risk.

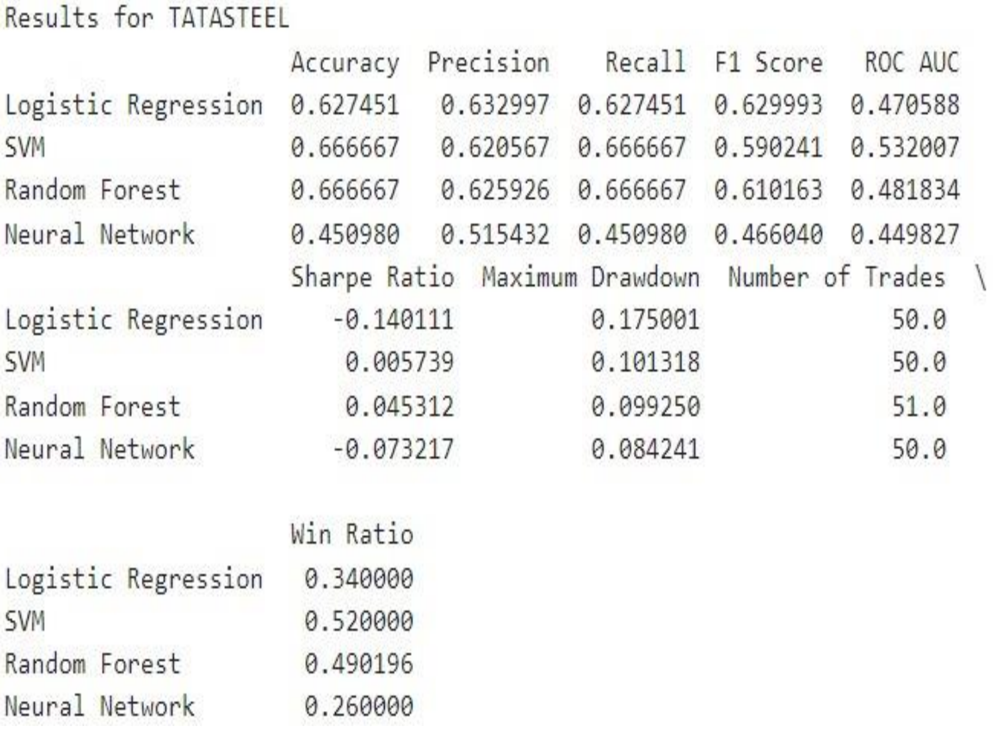
• Maximum **Drawdowns**: The maximum observed loss from a peak to a trough of a portfolio, before a new peak is attained.

• Number **of Trades Executed**: The total count of trades executed over a period.

• Win **Ratio**: The proportion of profitable trades to the total number of trades executed.

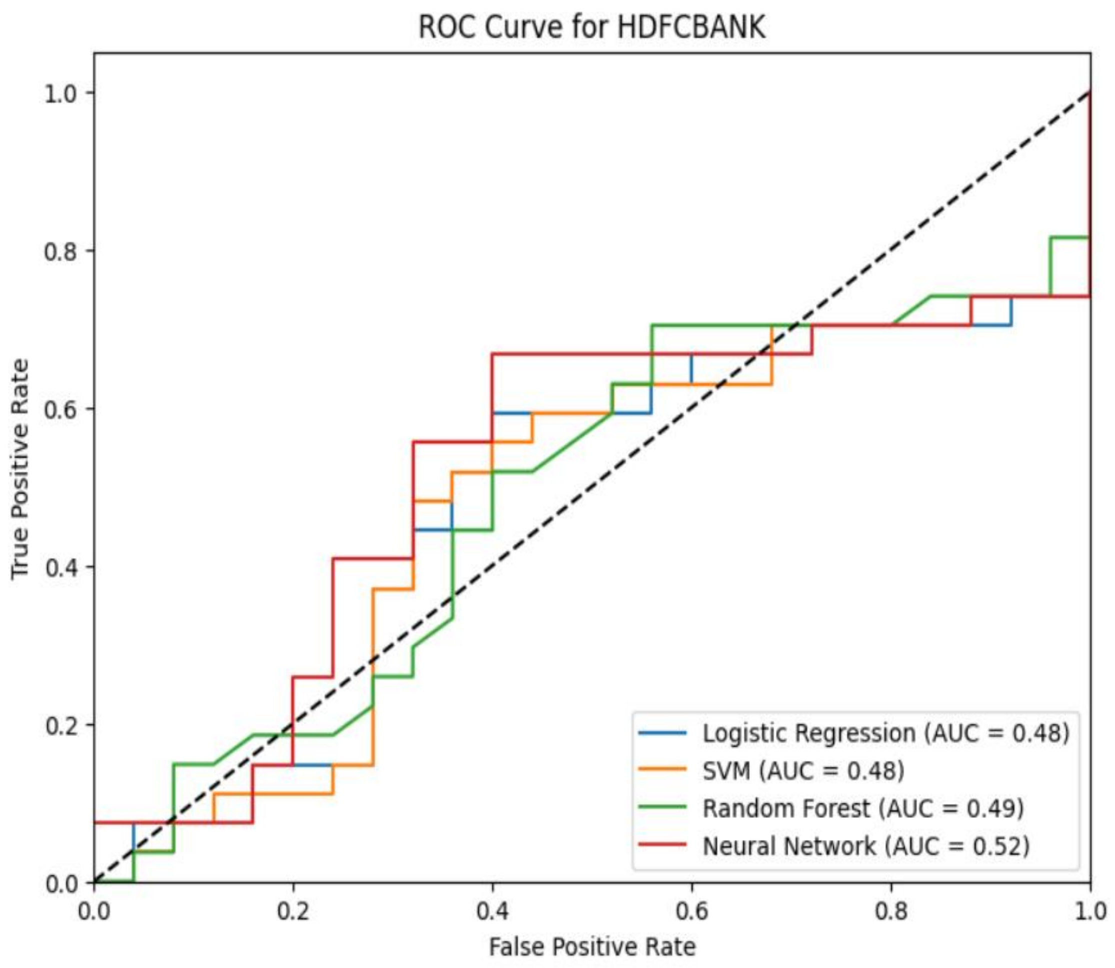
Model Visualization and matrices results For TATASTEEL STOCK

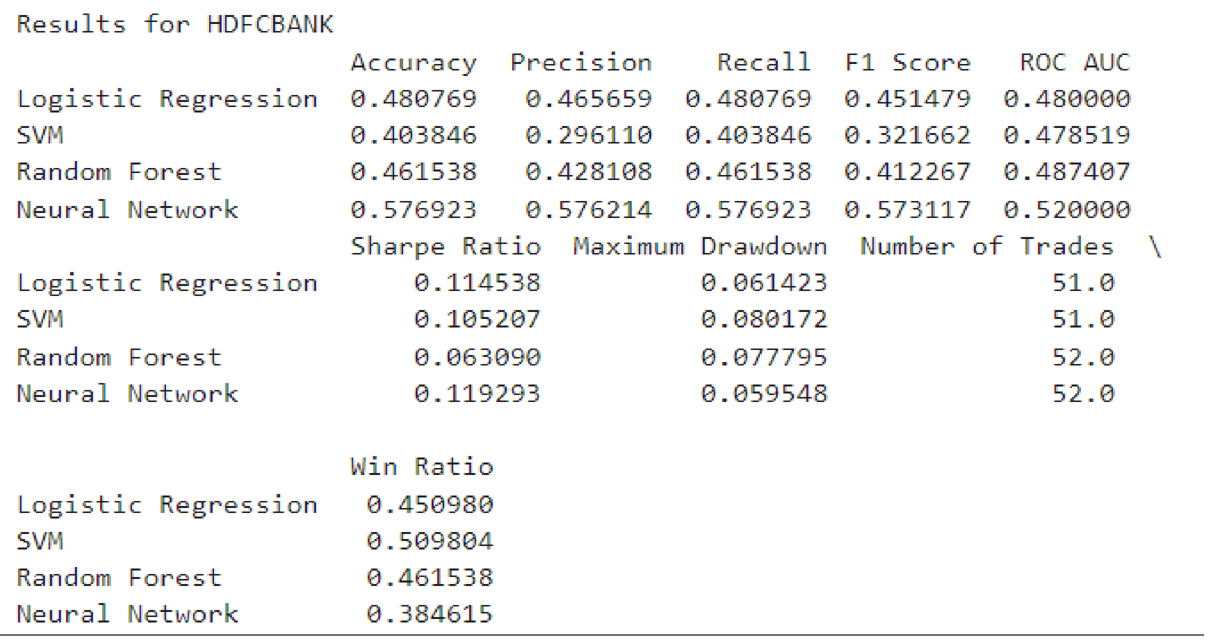




```
Results for TATASTEEL
                     Accuracy  Precision    Recall  F1 Score   ROC AUC
Logistic Regression  0.627451   0.632997  0.627451  0.629993  0.470588
SVM                  0.666667   0.620567  0.666667  0.590241  0.532007
Random Forest        0.666667   0.625926  0.666667  0.610163  0.481834
Neural Network       0.450980   0.515432  0.450980  0.466040  0.449827
                     Sharpe Ratio  Maximum Drawdown  Number of Trades  \
Logistic Regression     -0.140111          0.175001              50.0
SVM                      0.005739          0.101318              50.0
Random Forest            0.045312          0.099250              51.0
Neural Network          -0.073217          0.084241              50.0

                     Win Ratio
Logistic Regression   0.340000
SVM                   0.520000
Random Forest         0.490196
Neural Network        0.260000
```

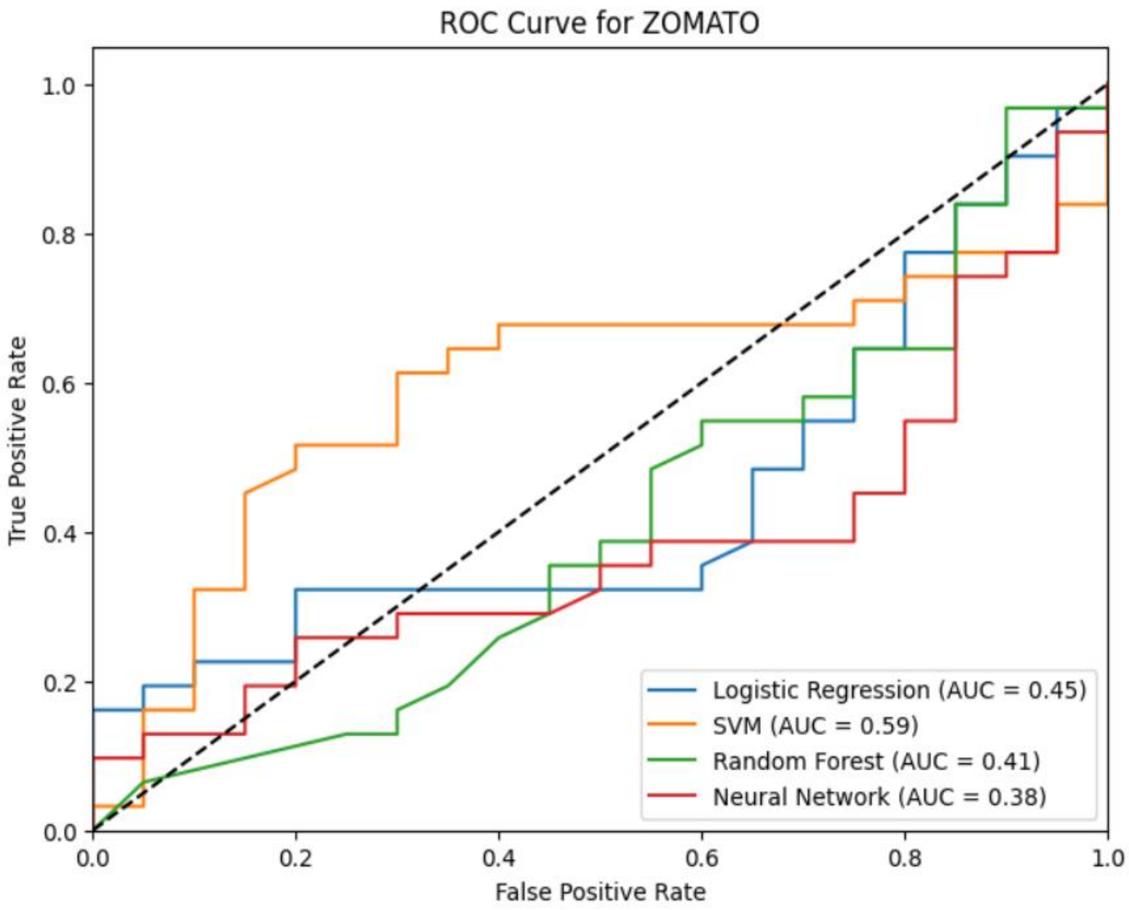
Model Visualization and matrices results For HDFCBANK STOCK

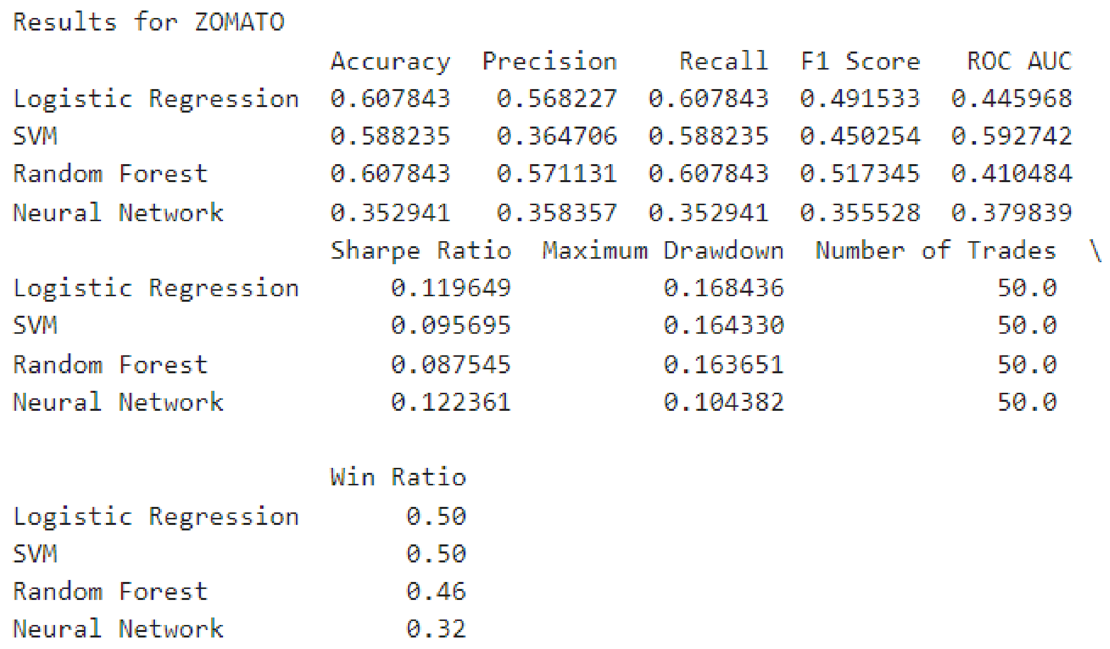




```
Results for HDFCBANK
                     Accuracy  Precision    Recall  F1 Score   ROC AUC
Logistic Regression  0.480769   0.465659  0.480769  0.451479  0.480000
SVM                  0.403846   0.296110  0.403846  0.321662  0.478519
Random Forest        0.461538   0.428108  0.461538  0.412267  0.487407
Neural Network       0.576923   0.576214  0.576923  0.573117  0.520000
                     Sharpe Ratio  Maximum Drawdown  Number of Trades  \
Logistic Regression      0.114538          0.061423              51.0
SVM                      0.105207          0.080172              51.0
Random Forest            0.063090          0.077795              52.0
Neural Network           0.119293          0.059548              52.0

                     Win Ratio
Logistic Regression   0.450980
SVM                   0.509804
Random Forest         0.461538
Neural Network        0.384615
```

Model Visualization and matrices results For ZOMATO STOCK



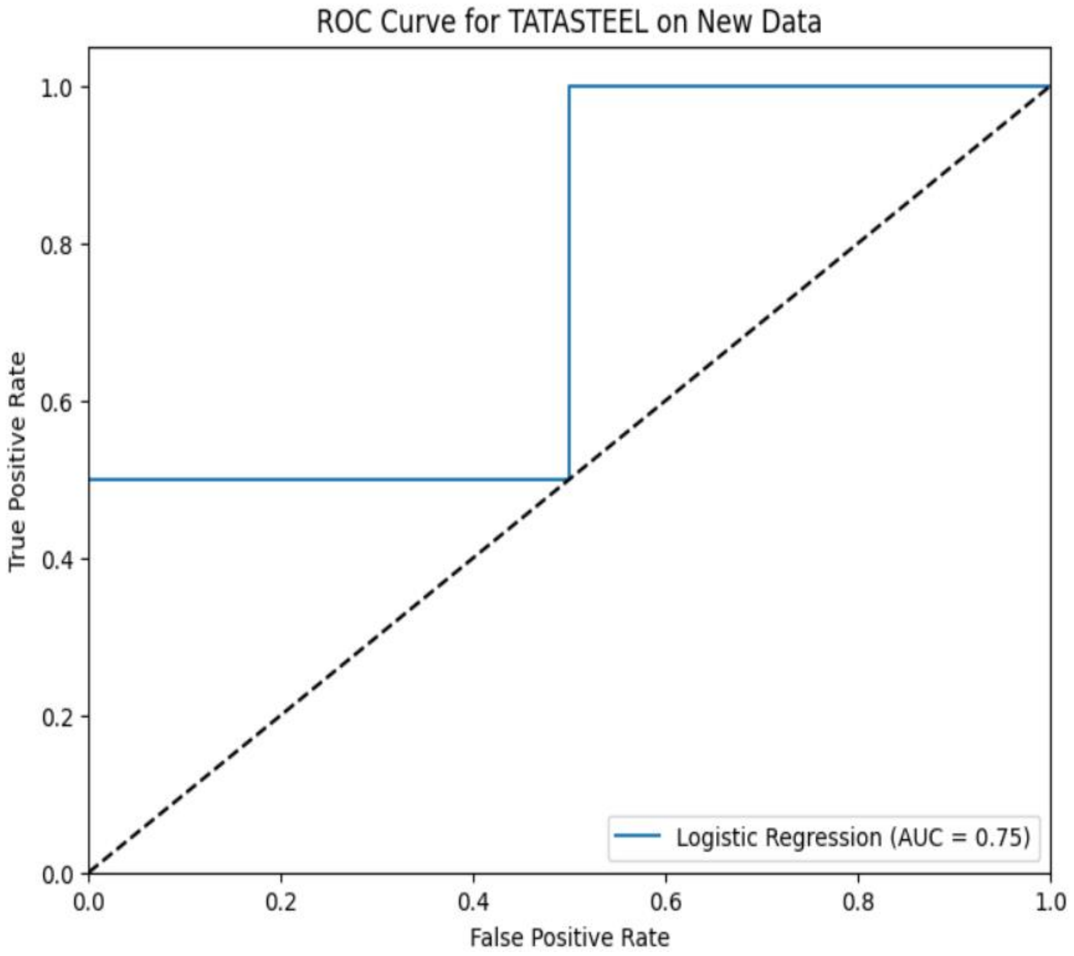


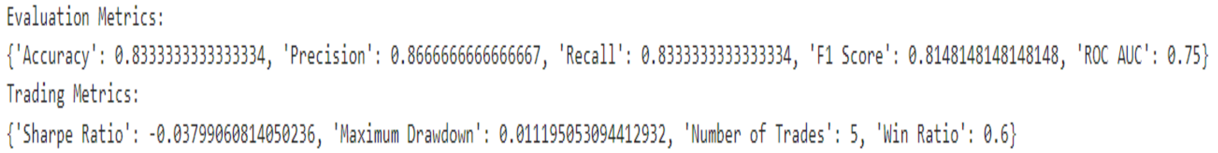
```
Results for ZOMATO
                    Accuracy  Precision    Recall  F1 Score   ROC AUC  
Logistic Regression 0.607843   0.568227  0.607843  0.491533  0.445968  
SVM                 0.588235   0.364706  0.588235  0.450254  0.592742  
Random Forest       0.607843   0.571131  0.607843  0.517345  0.410484  
Neural Network      0.352941   0.358357  0.352941  0.355528  0.379839  
                    Sharpe Ratio  Maximum Drawdown  Number of Trades  \
Logistic Regression     0.119649          0.168436              50.0   
SVM                     0.095695          0.164330              50.0   
Random Forest           0.087545          0.163651              50.0   
Neural Network          0.122361          0.104382              50.0   

                    Win Ratio  
Logistic Regression      0.50  
SVM                      0.50  
Random Forest            0.46  
Neural Network           0.32  
```

Best model is SVM with ROC AUC: 0.59

Testing Analysis Model with Best Machine Learning Model:

- Took new_Textual_Data for TATASTEEL Stock as an input.
- Took Stock_data of Last 10 days as an input.





```
Evaluation Metrics:
{'Accuracy': 0.8333333333333334, 'Precision': 0.8666666666666667, 'Recall': 0.8333333333333334, 'F1 Score': 0.8148148148148148, 'ROC AUC': 0.75}
Trading Metrics:
{'Sharpe Ratio': -0.03799060814050236, 'Maximum Drawdown': 0.011195053094412932, 'Number of Trades': 5, 'Win Ratio': 0.6}
```

RESULT ANALYSIS:

Evaluation Metrics:

1. **Accuracy (0.833)**:
   - The model correctly predicted 83.33% of the instances in the dataset.
   - This is a strong indication of the model's overall performance in classification tasks.
2. **Precision (0.867)**:
   - The model's precision of 86.67% indicates that a high proportion of the predicted positive movements were correct.
   - This is important in scenarios where false positives need to be minimized, such as in trading.
3. **Recall (0.833)**:
   - With a recall of 83.33%, the model successfully identified 83.33% of all actual positive movements.
   - This reflects the model's ability to capture most of the true positive cases.
4. **F1 Score (0.815)**:
   - The F1 Score, which balances precision and recall, is 81.48%, indicating a well-rounded performance.
   - This metric is especially useful when there is an uneven class distribution.
5. **ROC AUC (0.75)**:
   - An ROC AUC of 0.75 suggests that the model has a 75% chance of distinguishing between positive and negative classes.
   - This is a decent score, reflecting a good ability to differentiate between stock price movements.

Trading Metrics:

1. **Sharpe Ratio (-0.038)**:
   - The Sharpe Ratio is negative, indicating that the strategy did not generate returns higher than the risk-free rate.
   - This suggests that the model's predictions did not lead to a profitable trading strategy when adjusted for risk.
2. **Maximum Drawdown (0.011)**:
   - A Maximum Drawdown of 1.12% shows that the largest peak-to-trough decline was relatively small.
   - This is a positive indicator, suggesting that the trading strategy did not experience significant losses.
3. **Number of Trades (5)**:
   - The model executed 5 trades, indicating a relatively low trading frequency.
   - This could be beneficial in reducing transaction costs but might miss out on more trading opportunities.
4. **Win Ratio (0.6)**:
   - A Win Ratio of 60% shows that 60% of the trades were profitable.
   - This is a positive sign, suggesting that the majority of the trades resulted in gains, though the overall profitability needs to be considered alongside other metrics like the Sharpe Ratio.

Drawbacks:

The primary drawback of the project is the negative Sharpe Ratio, indicating that the trading strategy did not yield returns commensurate with the risks taken. Additionally, the low number of trades suggests a conservative approach that might miss out on more profitable opportunities. While the model has high accuracy and precision, the overall profitability is hindered by these factors, necessitating further refinement.

Recommendation:

To address these issues, refine the trading strategy to enhance the Sharpe Ratio by incorporating risk management techniques such as stop-loss orders. Increase the number of trades by adjusting the model to identify more opportunities without compromising accuracy. Regularly monitor and re-train the model to adapt to changing market conditions, and consider diversifying the portfolio to mitigate risks and stabilize returns. Adding More textual , relevant and perfectly labelled data will help to enhance the ability of the Model .