STOCK MARKET DIRECTION PREDICTION USING MACHINE LEARNING.

LOGISTIC REGRESSION & DECISION TREES CLASSIFICATION MODELS.

GICHUHI LEVIS KIMANI

Project Overview

Objective: To develop a machine learning model to accurately predict stock market direction for a quantitative hedge fund.

Data: A dataset containing opening, closing, high and low prices of the day for the SnP500.

Model: Logistic Regression and Decision Trees.

Data Preparation and Feature Engineering

1. **Data Cleaning:**
   - Handled missing values by dropping the rows as accuracy is key.
   - No need to address outliers as prediction is binary.
   - Removed irrelevant or redundant features.
2. **Feature Engineering:**
   - Created new features such as adjusted volume, today(change in daily volume), and lag_1 > lag_5 which represented the past 5 days movement.

Model Development and Training

1. **Model Selection:**
   - Evaluated various classification algorithms (Logistic Regression, Decision Trees).
   - The choice of algorithm will depend on the type of data, the problem being solved, and the desired level of accuracy.
   - Selected both logistic regression and decision trees for easier comparison as they both have their strengths and weaknesses.
2. **Hyperparameter Tuning:**

   Our hyperparameter tuning improved the model slightly which in prediction is positive. We can use all the help we can get. We used the gridsearchCV model on both algorithms we used to contrast performance.

3. **Model Training:**
   Trained the two classification models on the prepared dataset

Model Evaluation

Performance Metrics:

- **Accuracy:** 54%
- **Precision:** 54%
- **Recall:** 95%
- **F1-score:** 69%
- **ROC-AUC Score:** 0.51

Confusion Matrix:

Predicted	Actual Positive	Actual Negative
Positive	622	22
Negative	539	24

Interpretation:

- The model correctly predicted 54% of the direction correctly.
- It has a high recall of 97%, indicating its ability to identify most of the actual market direction cases and an F1 score of 69% which is a key metric where there is class imbalance.

Conclusion

The developed classification models demonstrate better than average performance in predicting market direction one day prior. It can be effectively used to influence key trading decisions and predict market sentiment but there remains room for improvement. Future improvements may include incorporating fundamental analysis techniques and exploring other algorithms however the model accuracy is not questionable.