**REPORT**: Sentiment analysis using News Headlines

**Introduction**

In this project, we explore the relationship between news sentiment and stock price movements using machine learning techniques. Specifically, we focus on Apple Inc. (AAPL) stock and analyze how news headlines affect its daily trading signals and overall portfolio performance. By leveraging sentiment analysis on scraped financial news headlines and applying machine learning models, we aim to predict AAPL stock movements and simulate a trading strategy based on these predictions.

Flow of Project

1. Data Collection
2. Data Preprocessing and Analysis
3. Feature Engineering and Model Training
4. Model Evaluation and Performance Metrics
5. Trading Strategy Simulation
6. Conclusion and Insights
7. References

**Data Collection**

**Web Scraping:**

* **Process**: Implemented web scraping using Python's requests and BeautifulSoup libraries.
* **Source**: Scraped financial news headlines related to Apple from the Financial Times website.
* **Methodology**:
  + Constructed URL format (https://www.ft.com/search?q=Apple&page={}) to iterate through multiple pages of search results.
  + Utilized requests.get() to fetch HTML content and BeautifulSoup for parsing and extracting relevant data.
  + Identified headline elements (<div class="o-teaser\_\_content">) and associated date tags (<time class="o-teaser\_\_timestamp-date">).
  + Extracted headline text and publication dates, ensuring data integrity and chronological order.
* **Purpose**: Acquired comprehensive dataset of Apple-related news, structured into a format suitable for subsequent analysis and modeling.
* def get\_ft\_headlines(keyword, pages=40):
* headlines = []
* base\_url = f"https://www.ft.com/search?q={keyword}&page={{}}"
* headers = {
* "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36"
* }

Used Financial Times website to scrap the data.

**Data Preprocessing and Analysis**

**Date Sorting and Grouping**:

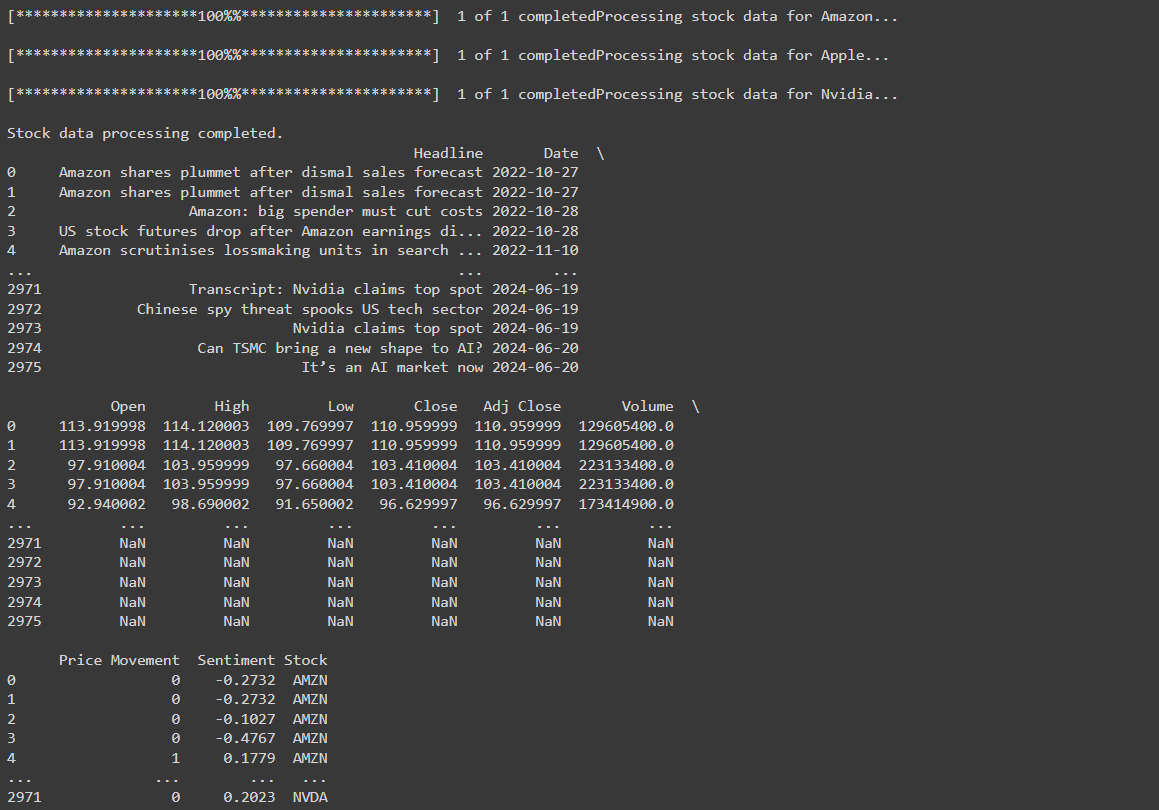
* **Purpose**: Organize headlines chronologically.
* **Process**: Sort headlines by date and group them to consolidate multiple headlines per day into a single record.
* **Importance**: Ensures temporal sequence for meaningful analysis, aligns with stock price movements.

**Sentiment Analysis (VaderSentiment)**:

* **Purpose**: Assess sentiment polarity (positive, negative, neutral) of headlines.
* **Method**: Utilizes VaderSentiment library, which quantifies sentiment using lexical approach and sentiment lexicon.
* **Application**: Generates sentiment scores for each headline, capturing emotional context and its potential impact on stock market dynamics.
* **Outcome**: Provides numerical representation of sentiment, aiding in understanding news sentiment trends influencing stock price movements.

3.5

* , headlines\_df):
* start\_date = headlines\_df['Date'].min()
* end\_date = headlines\_df['Date'].max()



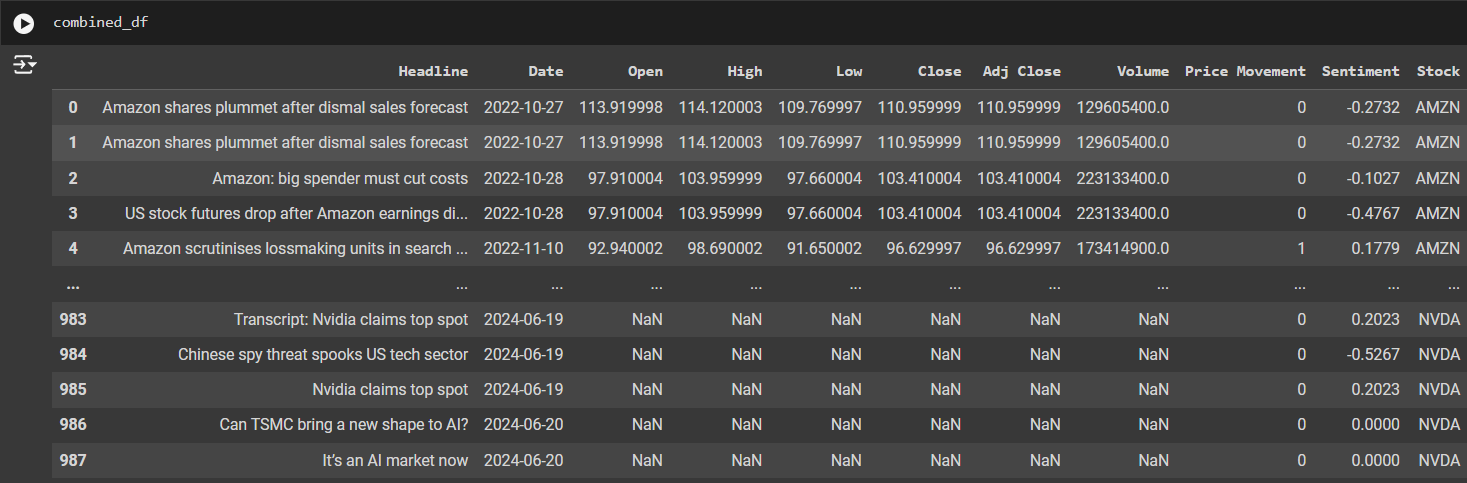
**Feature Engineering and Model Training**

**Text Vectorization (TF-IDF Vectorization)**:

* **Purpose**: Convert headlines into numerical features using TF-IDF, quantifying term importance.
* **Process**: Tokenization, count vectorization, TF-IDF transformation.
* **Importance**: Selects key terms for predicting stock movements, reduces feature space.

**Machine Learning Model Selection (RandomForestClassifier)**:

* **Purpose**: Predict AAPL stock movements (Increase, Decrease, No Change) from headlines.
* **Model**: RandomForestClassifier aggregates decision trees for robust predictions.
* **Training**: Data split, model fit on TF-IDF features, hyperparameter tuning.
* **Outcome**: Predicts stock movement based on headline content and sentiment scores.
* # Combine all data into a single DataFrame
* combined\_df = pd.concat(all\_data)
* print("Combined data creation completed.")

****

**Model Evaluation and Performance Metrics**

**Evaluation Metrics:**

* **Metrics Used**: Employed key metrics including accuracy, precision, recall, F1-score, ROC curves, and AUC (Area Under the Curve).
* **Purpose**: These metrics were utilized to comprehensively evaluate the RandomForestClassifier's performance in predicting AAPL stock price movements. Accuracy provided an overall measure of correct predictions, precision and recall assessed the model's ability to correctly classify each class (Increase, Decrease, No Change), while F1-score balanced precision and recall. ROC curves and AUC visualized the classifier's ability to distinguish between classes across different thresholds.
* def evaluate\_classifiers(X\_train, X\_test, y\_train, y\_test):
* classifiers = {
* 'Logistic Regression': LogisticRegression(max\_iter=1000),
* 'Support Vector Machine': SVC(probability=True),
* 'Random Forest': RandomForestClassifier(),
* 'Neural Network': MLPClassifier(max\_iter=500)
* }

**Visualization:**

* **ROC Curves and AUC**: Visualized ROC curves plotted sensitivity (True Positive Rate) against 1-specificity (False Positive Rate) for various decision thresholds. The AUC summarized the ROC curve's overall performance, indicating how well the model differentiated between stock movement classes.

**Trading Strategy Simulation**

**Signal Generation:**

* **Model Predictions**: Leveraged machine learning predictions to generate buy and sell signals based on anticipated AAPL stock movements (Increase, Decrease, No Change).
* **Strategy Execution**: Simulated a trading strategy with an initial capital of $100,000, adjusting positions dynamically in response to predicted price changes.

**Portfolio Management:**

* **Simulation Overview**: Managed portfolio positions dynamically based on predicted signals, calculating portfolio values over time.
* **Visual Representation**: Plotted AAPL stock prices alongside buy and sell signals to visually assess the trading strategy's performance and profitability.

**Conclusion and Insights**

**Insights:**

* **Sentiment Impact**: Discussed insights into how sentiment analysis of news headlines influenced AAPL stock movements, highlighting correlations between sentiment scores and subsequent price changes.
* **Model Effectiveness**: Evaluated the RandomForestClassifier's effectiveness in generating accurate trading signals based on sentiment and textual data.

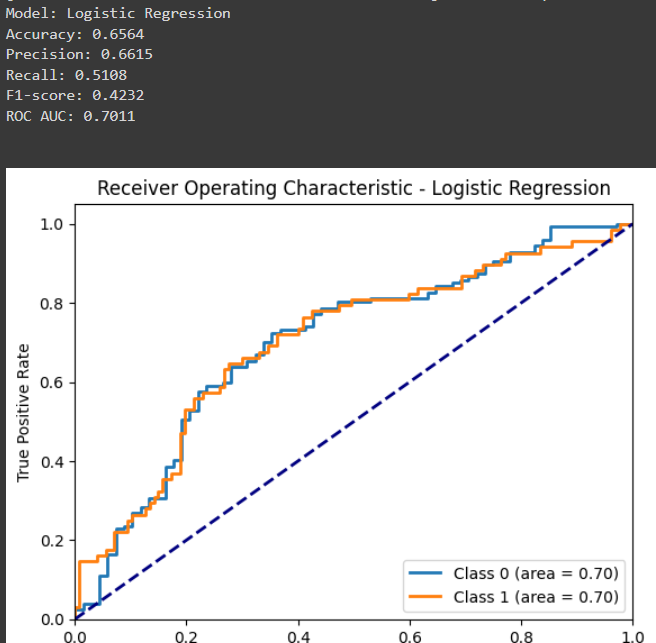
**Applications:**

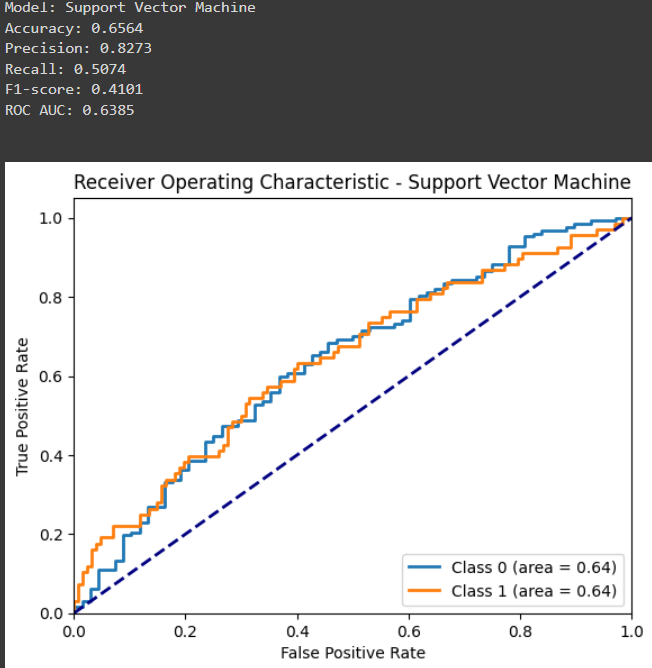
* **Practical Uses**: Explored practical applications of sentiment analysis and machine learning in financial forecasting and investment decision-making, emphasizing their role in enhancing trading strategies and decision support systems.

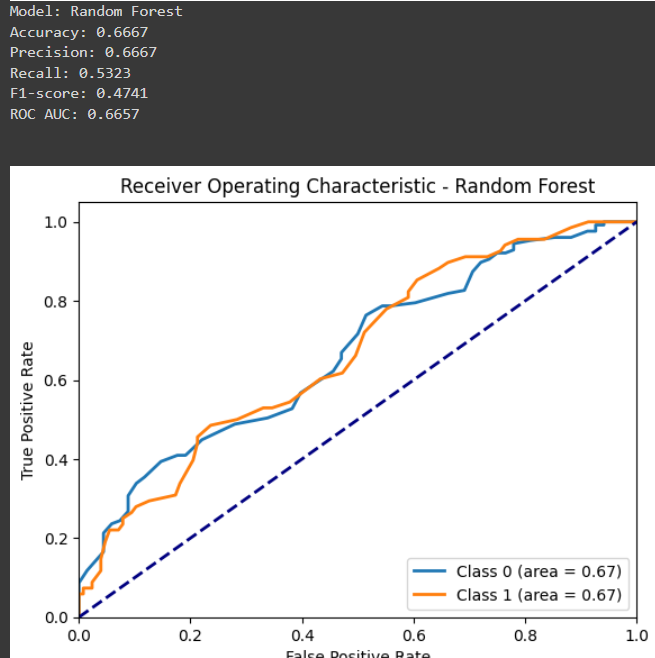
This comprehensive analysis and simulation provide valuable insights into the integration of machine learning with financial data, showcasing its potential to inform robust investment strategies.

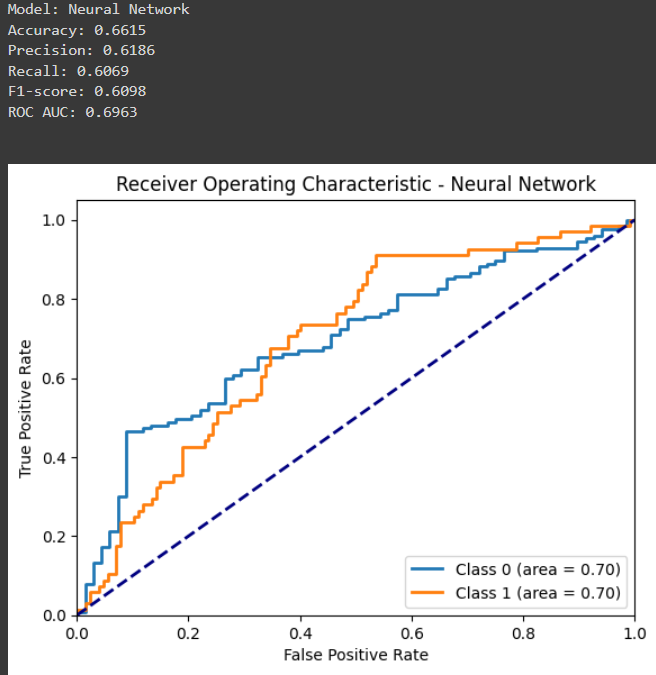
Results

ROC/AUC



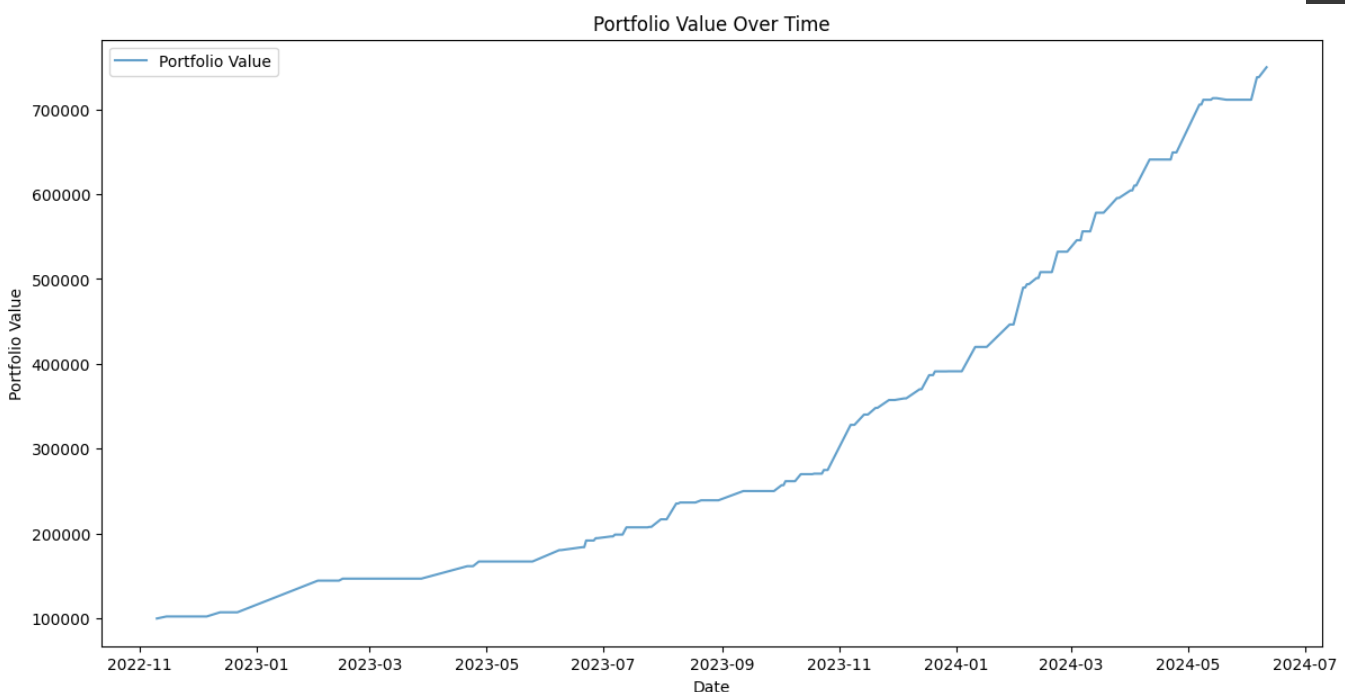
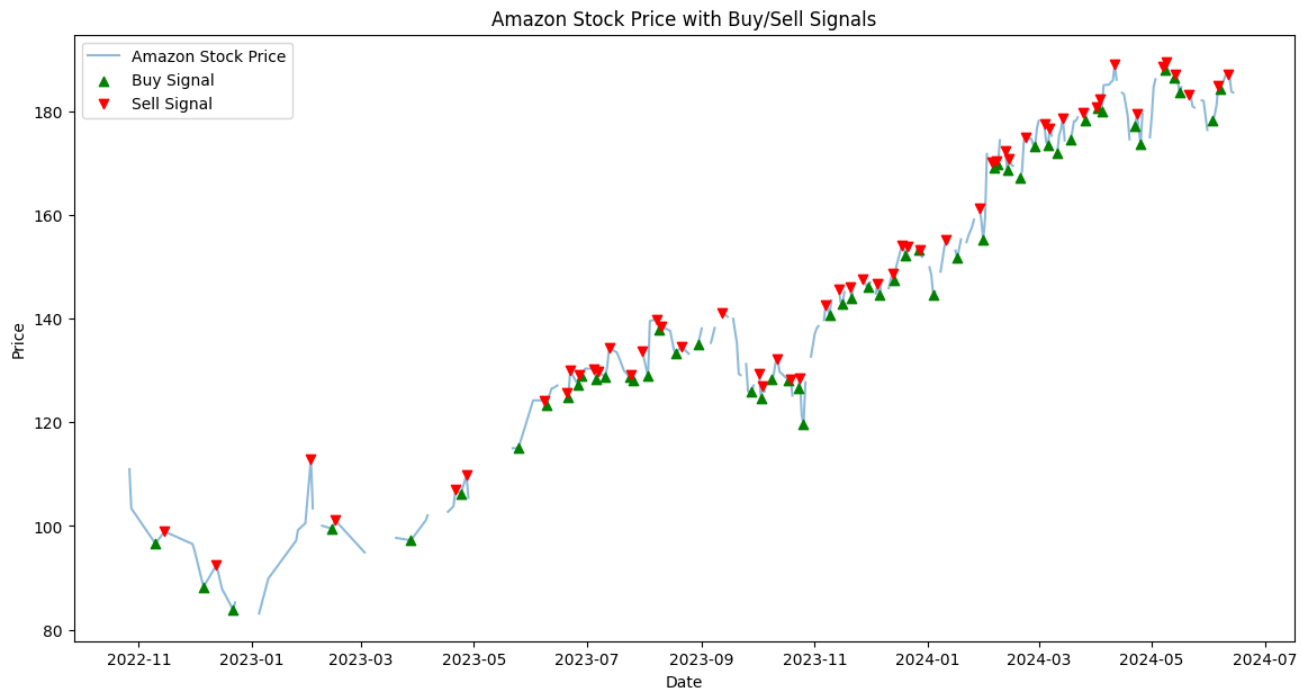




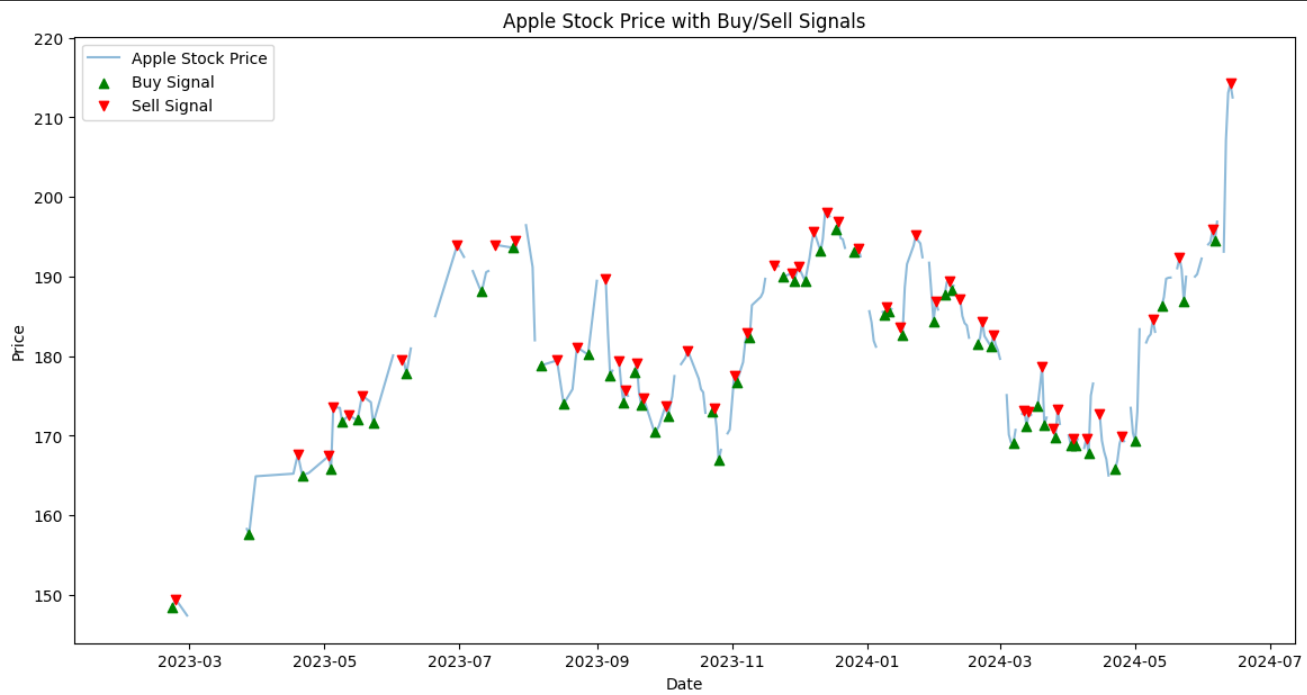


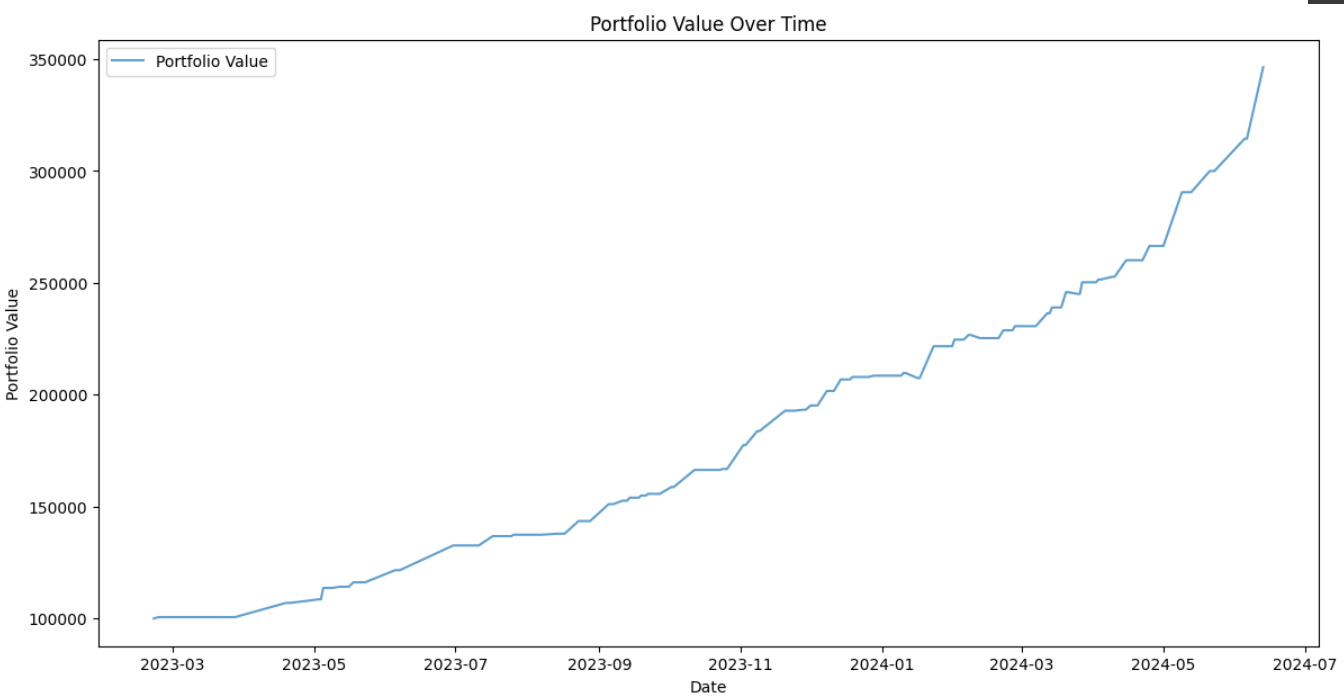
Buy and Sell:

Amazon



Apple:





Nvida:

