REPORT FOR

STOCK SENTIMENT ANALYSIS USING MACHINE LEARNING TECHNIQUES

**PARAG CHHALOTRE**

**Enrollment No: 21112081**

What is Stock Sentiment Analysis?

Stock sentiment analysis is a technique used in finance and investing to gauge the overall sentiment or emotions of market participants towards a particular stock or the market as a whole. It involves analyzing textual data such as news articles, social media posts, earnings call transcripts, and financial reports to assess whether the sentiment expressed is positive, negative, or neutral. The goal of stock sentiment analysis is to understand how public opinion and emotions influence stock prices and market trends. By analyzing sentiment, investors and traders aim to gain insights into market sentiment shifts that may impact stock prices.

*Flow of Project:*

1. Collect Textual Data:
   * Collect a large dataset of textual data related to stocks, including news articles, social media posts, earnings reports, and analyst reports from various sources.
2. Preprocess the Data:
   * Preprocess the textual data by removing noise, tokenizing text into words or phrases, and applying techniques such as stemming and lemmatization to standardize text representations.
3. Extract Sentiment Score:

* Extract features from the textual data, such as word frequencies, sentiment scores, and topic modeling representations, to represent the text in a format suitable for machine learning algorithms.

1. Collect Stock Data:

* Gather historical stock data. You'll need daily or monthly closing prices for a reasonable historical period.
* Data is taken from Yahoo Finance.

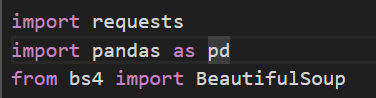
1. Label the Textual Data:

* Label the textual data with corresponding stock price movements (e.g., increase, decrease, or no change) over a specified time horizon to create a labeled dataset for supervised learning.

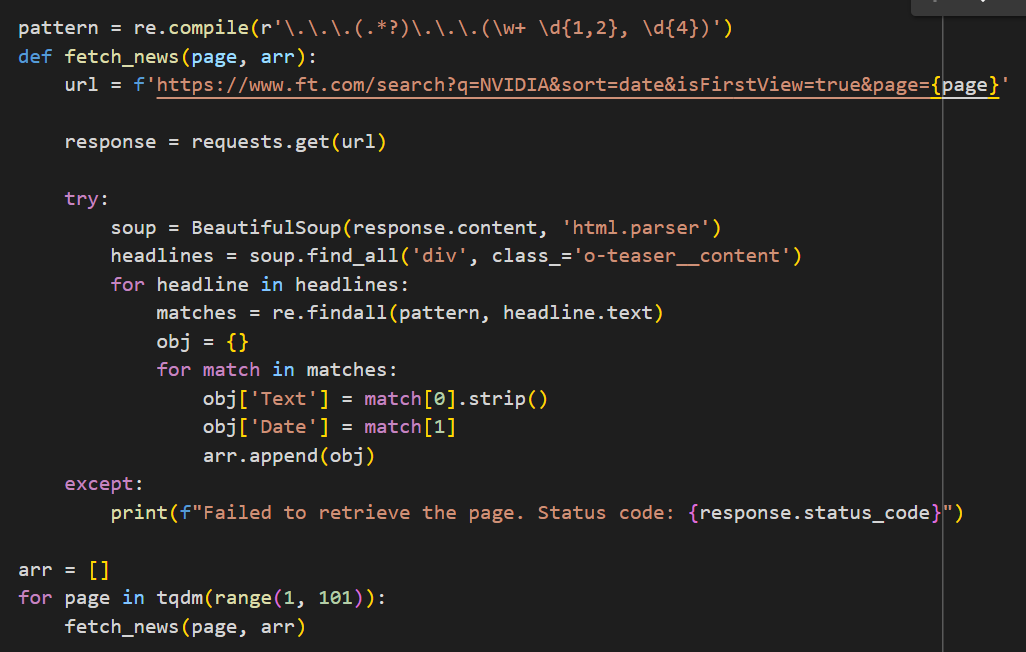
1. Train, Evaluate, and Choose the Best Model:

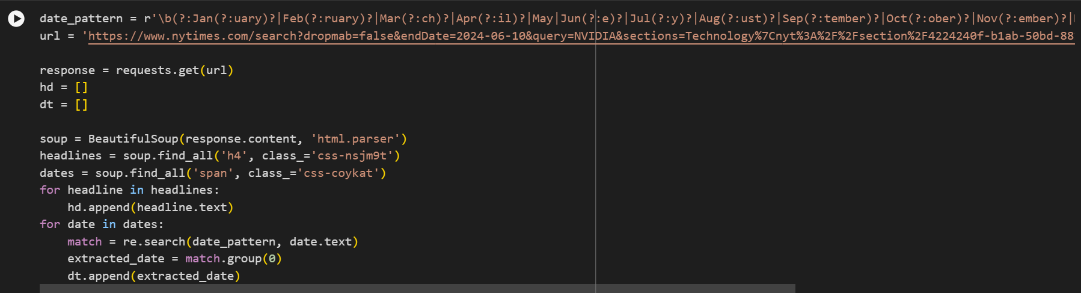
* Train and evaluate machine learning models, including classification algorithms such as logistic regression, support vector machines (SVM), random forests, and neural networks, to predict stock price movements based on textual sentiment.
* Perform model evaluation using appropriate metrics, such as accuracy, precision, recall, F1-score, and receiver operating characteristic (ROC) curve analysis, to assess the performance of the sentiment analysis model.

1. Visualization:
   * Matplotlib is used to plot the required data for analysis.
   * Visualization helps to give a better insight into the simulated data.
2. Analysis & Risk Calculations:
   * Final Portfolio and Total Returns
   * Daily mean and Daily Standard Deviation of Prices
   * Sharpe Ratio
   * Maximum drawdowns
   * Number of trades
   * Win Ratio
3. Collect Textual Data:

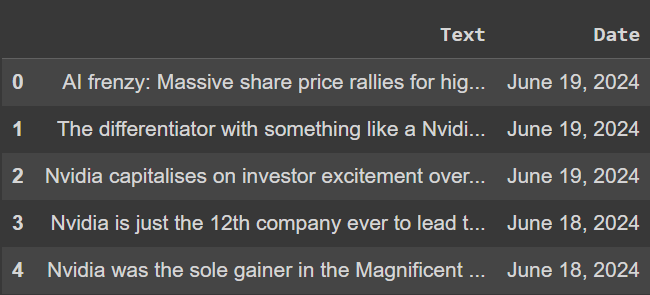


Import requests, BeautifulSoup from bs4 for web scraping.



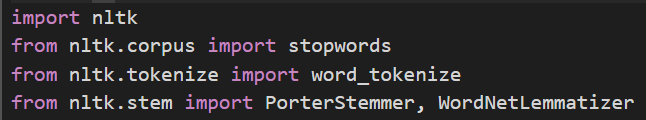




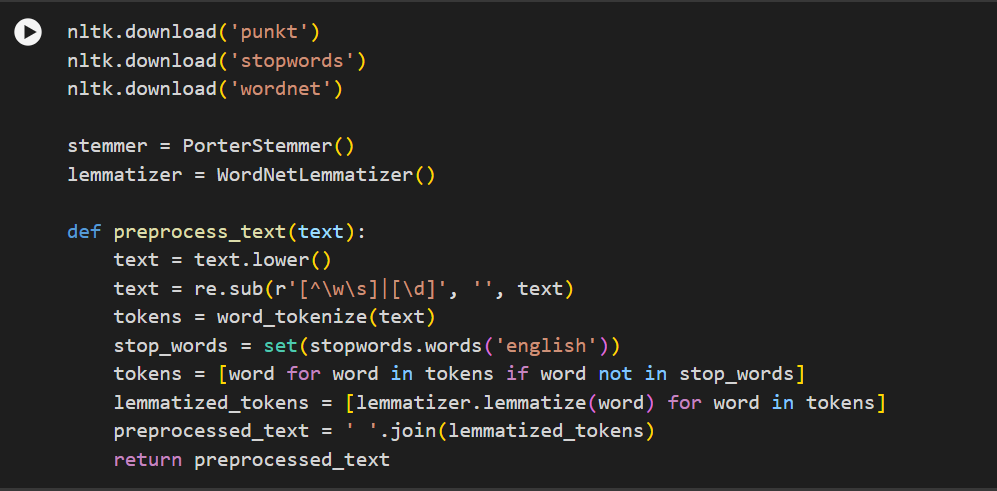


Collect textual data from various sources using web scraping.

1. Preprocess the Data:



These imports set up NLTK for common text preprocessing tasks like removing stop words, tokenizing text into words, and reducing words to their base forms using stemming or lemmatization.



 Converts the text to lowercase

 Removes non-alphanumeric characters and digits.

 Tokenizes the text into individual words.

 Removes stopwords from the tokens.

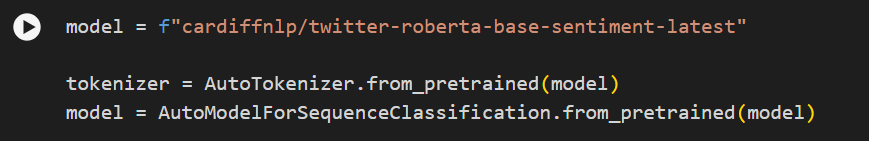
 Lemmatizes each token using the WordNetLemmatizer.

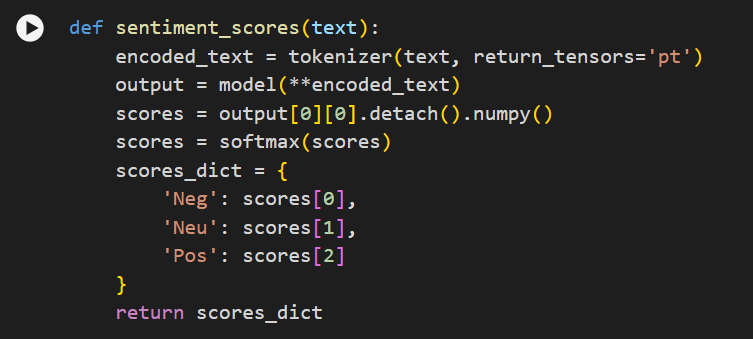
 Joins the lemmatized tokens back into a preprocessed string.

 Returns the preprocessed text.

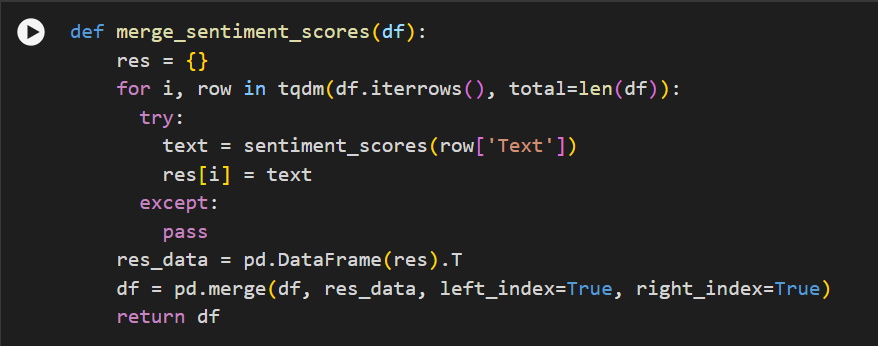
1. Extract Sentiment Scores:



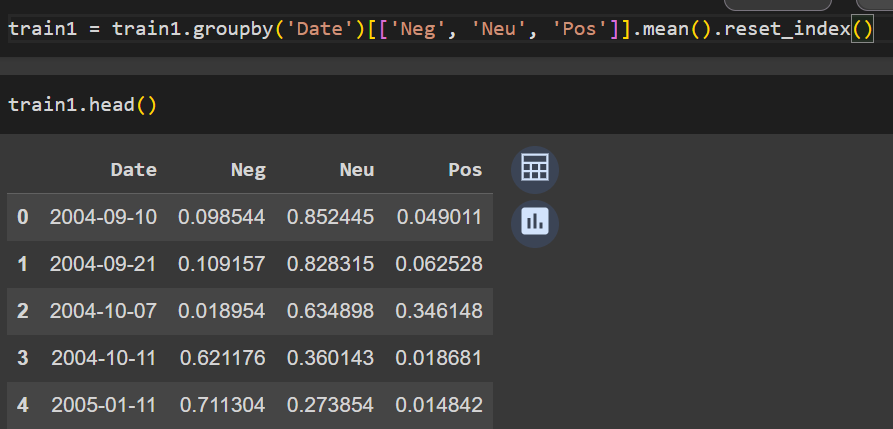




Use the pre-trained transformer model Roberta for extracting sentiment scores from the text. The AutoTokenizer helps in tokenizing text appropriately for the model, while AutoModelForSequenceClassification loads the model itself, and softmax can be used to interpret the model's outputs as probabilities after classification.



Merge sentiment scores in the data frame.

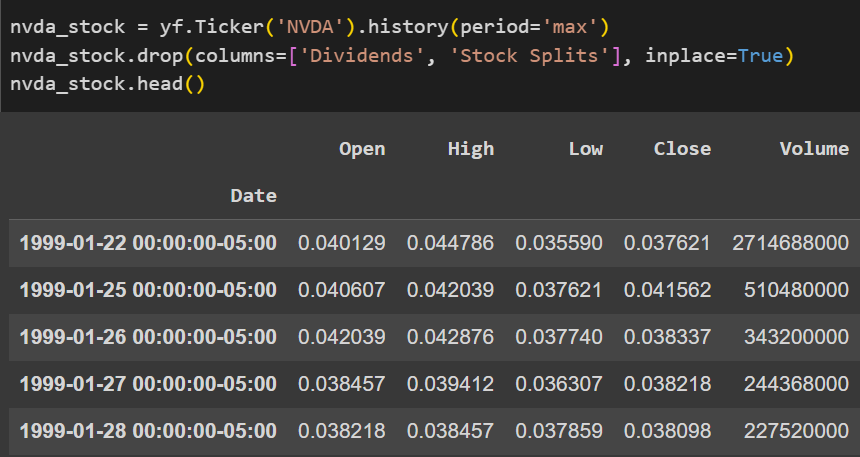


Calculate mean sentiment scores for each day.

1. Collect Stock Data:

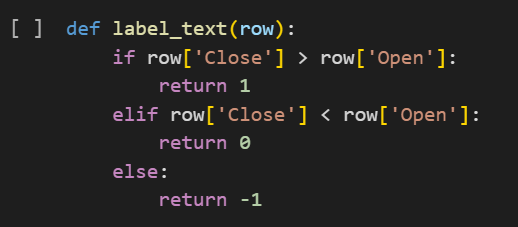


Import Yahoo Finance for stock data.



Collect stock data for various companies.

1. Label the Data:

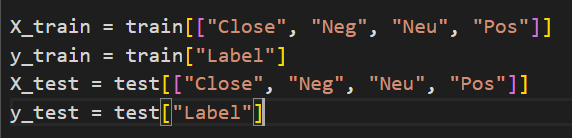


Close price is greater than the Open price, then label it as 1, Close price is smaller than the Open price, then label it as 0 and equal then as -1.

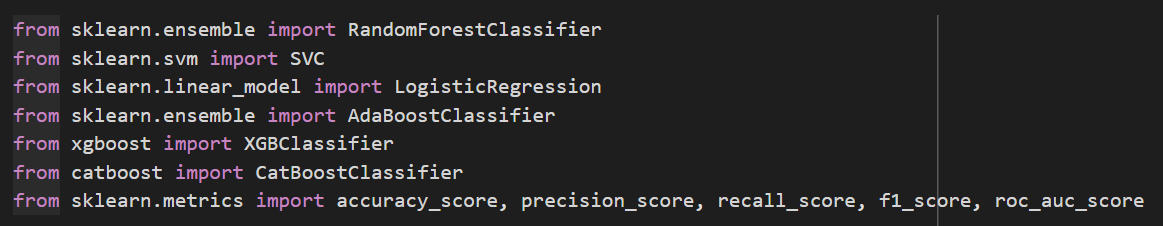


Merge the stock data with textual data on the Date columns.

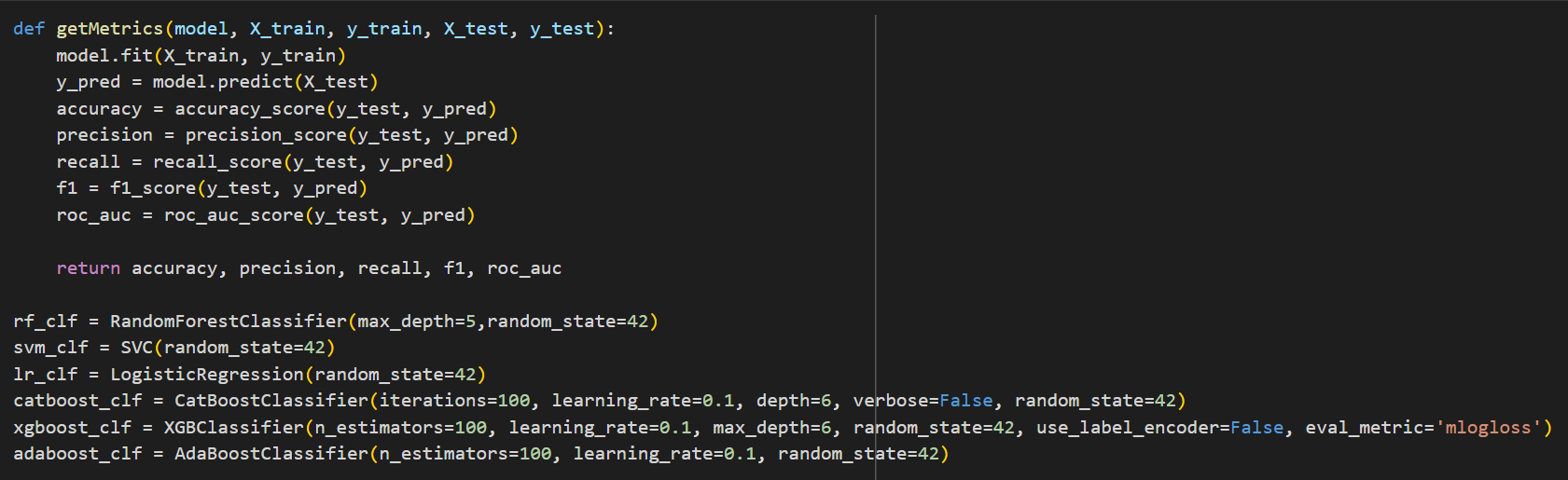
1. Train, Evaluate, and Choose the Best Model:



Separate the dataset into features (X\_train, X\_test) and the target variable (y\_train, y\_test) necessary for training and evaluating a machine learning model.

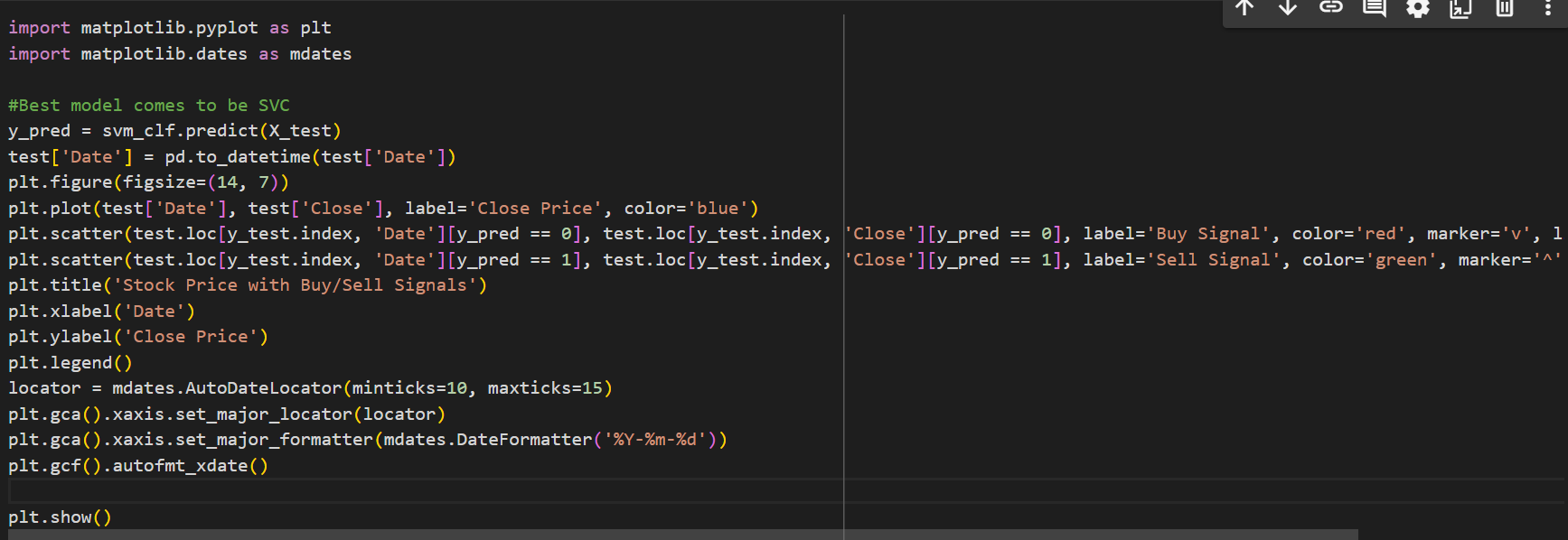


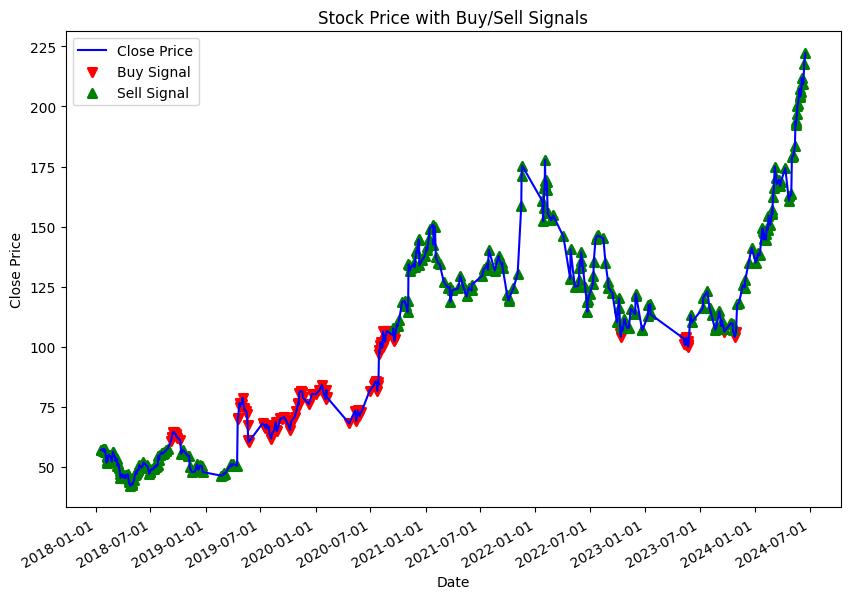
Import various models like RandomForest, SVC, Logistic Regression, AdaBoost, XGBoost, and CatBoost.



Train the models and calculate various metrics for finding the best model.

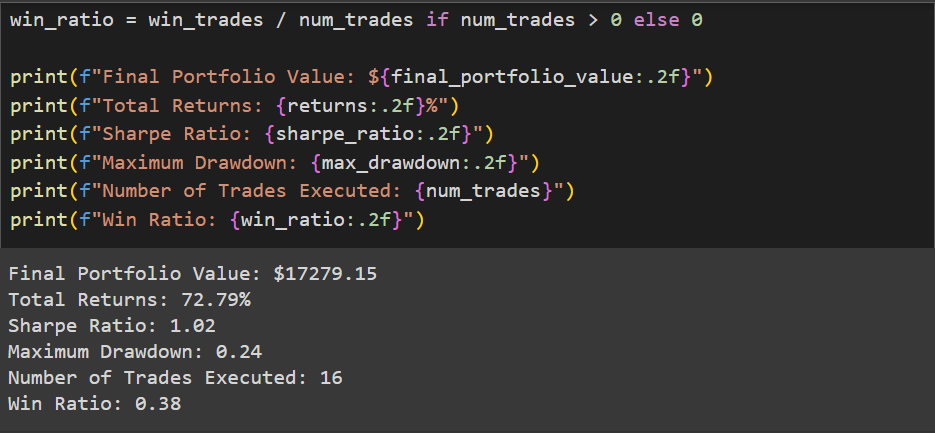
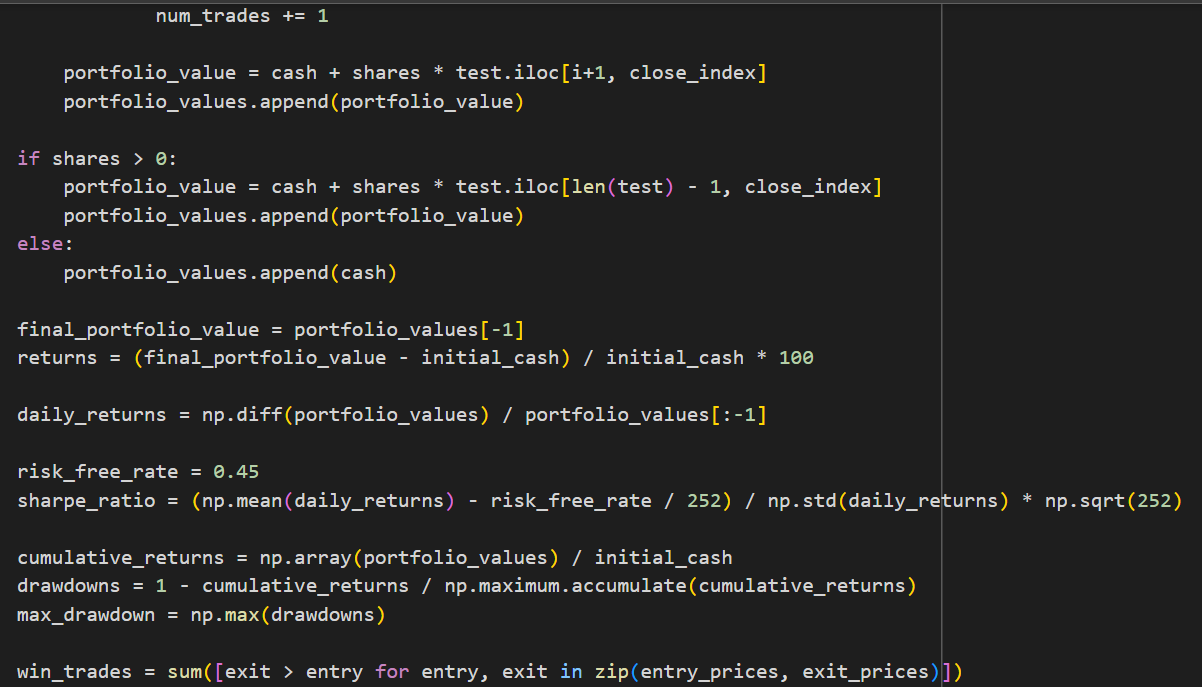
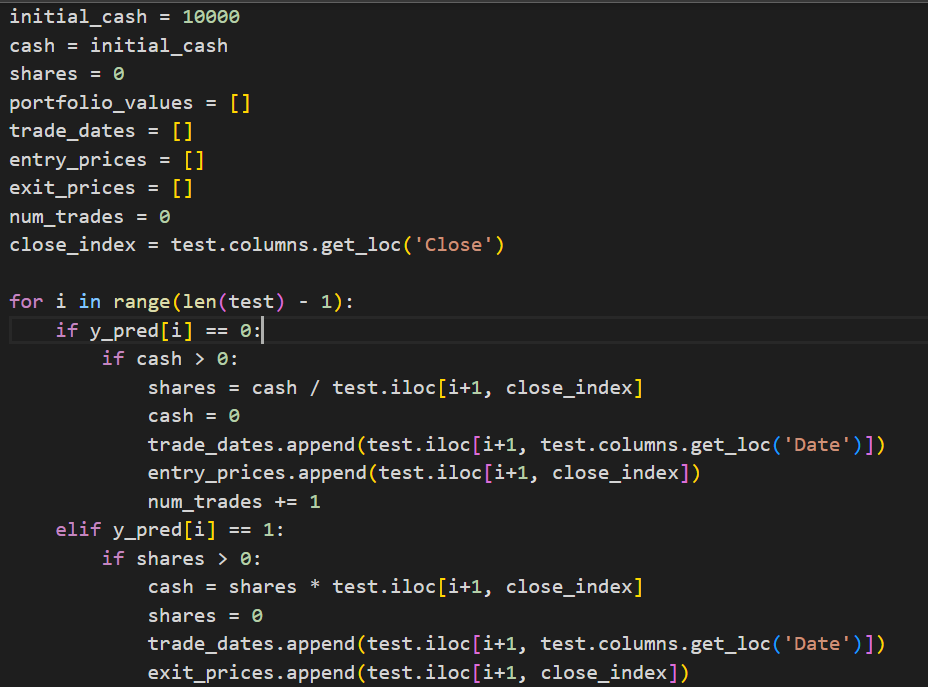
1. Visualization:





Graph showing buy and sell points.

1. Analysis & Risk Calculations:



**Note**:

The best model, Final Results, and Graph may change as I web scrape news websites, and news is being updated regularly.

This project is originally done by me.