**STRATEGY SELECTION PROJECT**

**DOCUMENTATION**

**INTRODUCTION**

The idea behind the project is that it tests the performance of 3 different day trading strategies i.e., moving average crossover, volume surge, and range trading, on a security given by the user, and based on the historical performance, recommends a strategy that should be used by the user for trading, the next day.

**LIBRARIES INCLUDED**

1. SCIKIT-LEARN- The scikit-learn library is a machine learning library of python, providing numerous features for supervised machine learning. Several modules and functions of the sklearn library have been used in the project:

* sklearn. linear\_model- This is used to import the machine learning model- Linear Regression.
* sklearn. pipeline- This module contains the Pipeline() function which allows the main events i.e., Preprocessing of data, and instantiation of the model, to take place one after another.
* sklearn.metrics- This module allows to import functions for model performance metrics. Mean squared error function has been imported from this module.
* sklearn.ensemble – This module contains functions that help enable ensemble techniques. It has been used for importing the Random Forest Regressor .
* sklearn.preprocessing- This module provides functions for preprocessing data. Using this module, the following functions have been imported:

1. StandardScaler- This function allows to scale the data such that mean is 0 and standard deviation is 1.

* Sklearn.model\_selection- This module is used to import two important functions:

1. Train\_test\_split- this allows us to divide the dataset into training and testing data.
2. GridSearchCV- this function allows to find the best model of a hyperparameter, along with cross validation.
3. JOBLIB- The joblib library has been used to download and re-use the machine learning models created for each time the code is run.
4. NUMPY – This is the python library for performing numerical computations.
5. PANDAS – This is the python library used for performing operations related to data.
6. YFINANCE- This library has been used to access OHLCV data from yahoo finance.
7. TALIB – This library provides numerous functions to perform technical analysis of stocks. It provides functions for various indicators, oscillators, and candlestick patterns.
8. DATETIME – This library is used for working with and manipulating dates and times.
9. DATEUTIL – This library is an extension of the datetime library and extends its functionalities.

**FUNCTIONS**

1. mainfunc

This is the core function of the project. The ticker of the required security and current date is taken as input from the user after which the date of one year back is calculated. The ticker, current date, and date one year back are then given as arguments to the three functions- dfmac , volsurge, and range.

This dataframe is the one being given to the regression models with all columns except the ‘tomorrow\_buying\_return’ column as features, and the latter as target. A copy of this dataframe is made, and according to the returns, a position column is added, and this dataframe is the one being given to the classification models, with the position as target, and other columns as features.

These functions return dataframes, which are then sent to the random forest regression model which predict the return for the next day using each strategy. After this, OHLCV data for the current day, along with the 50-day moving average, 20 day moving average, average high, low, and average and standard deviation of the volume and open price as of the current day’s end is taken as input from the user.

Finally, an ML model predicting the return using each strategy acts on the data, and the strategy giving the maximum return for the next day is found and returned by the function, along with the return.

1. mac

The mac function is used to create a dataframe that shows the implementation of the moving average crossover strategy. A long position is taken when the short term moving average crosses over the long term moving average, and a short position is taken if the opposite happens. It takes ticker symbol of the security, start date, and end date as input. It then downloads the OHLCV data from yahoo finance and then long (50 day) and short (20 day) term moving average at the end of the day is calculated. Based on this, another column for the position (1 for long, -1 for short) is created, and the buying return is calculated.

1. volsurge

The volsurge function is used to create a dataframe that shows the implementation of the volume surge strategy. The average and standard deviation of the volume and open price for each day is calculated. Accordingly, the position and return for the next day is calculated.

1. range

The range function shows the implementation of the range trading strategy, which trades on the basis of whether the price is near the average of the high or average of the low of the past 20 days. The average of the high and low for the past 20 days is calculated. Position for the next day is decided on the basis of whether the close of the current day is next to the average high or low price and return for buying the next day is also calculated.

1. **forest**

The forest function is used to train a random forest regressor model. It splits the dataset into features and targets and then further divides into training and testing data. After this, a pipeline for standardizing the data and instantiating the model is created. This is followed by finding the optimal value of the hyperparameter and thus the best model, which is returned by the function.

**WORKING**

The mainfunc function is first called, and the user is prompted to enter the ticker symbol and enter the current date. The date for one year back is then calculated. These 3 details are then sent as arguments to the mac, volsurge, and range functions. These functions then project the performance of the respective strategies over a one-year period and return a dataframe each which is then sent to the random forest regression model. The model trains how to predict the performance of the strategy for the next day using the data and is then returned to the mainfunc function. Finally, the strategy that gives the highest return for the next day is returned, along with the return.

**LIMITATIONS**

1. USER INPUT DEPENDANCE – The code requires a lot of input from the user. A function can be made to make all these calculations and find the value of the inputs.
2. LACK OF MODEL EVALUATION – No metrics for evaluating the performance of the regression model have been used. In addition, only one model for regression i.e., Random Forest Regressor has been used.