

# Financial Data Analysis Lab

## Final Task

January 10, 2023

Develop a trading strategy for the stocks that compose the S&P 500 index and implement it using Python. Use historical data for the SP 500 index to train a machine learning model that can identify potentially profitable investing opportunities based on past stock performance and other financial data if you choose to. The system should automatically choose the stocks based on the likelihood of the specifically chosen portfolio increasing in value during the next three months

## Part 1

### Statistical Analysis (40%)

- For instance if AAPL changed its price during the target period in +10% then the value for AAPL will be 10
- Use a features-target (6Months-3Months) split based on the template from Task 1, for each of the stocks calculate the change in price during the target period in percent, and give the label 1 if that change is above the median change or 0 if it is below (same calculation as the target variable in Task 1)
- Use the data that is labeled as features that contain 6 Months of daily OHLCV for each of the stocks and summaries that period into a single variable for each stock ( Same as the features in Task 1)
  1. Use the features data to calculate the correlation and regression between the feature data (a single value per stock extracted from the 6M labeled as features) and the change in percent of the stock during the 3 months of the labeling period
  2. Determine a threshold for that feature above which all of the stocks that received higher value in the feature period (first 6M) will be predicted as 1 and below it, they will be predicted as 0
    - Calculate the statistical difference between the values of the group of stocks that received 1 and the group that received 0, use box plot, histogram, mean, median, 20% percentiles value, skew, and kurtosis
    - Use the threshold base prediction to determine the confusion matrix, precision, and recall

3. Do the same as on task1.2 but instead of using the actual feature value use the predicted value from the regression in part 1.1
4. Summaries your findings, what does it means in terms of the three different algorithms' ability to predict high-performing stocks?

### **Part 1 Grading (40%)**

- Quality of the chosen feature (15%)
- Completeness of the answer (20%)
- Summarizing of the results (5%)

## **Part 2**

### **Statistical Analysis (40%)**

1. Create a model same as the model in task 1 using 10 features extracted from the feature part (6M) to predict the target (3M)
2. After training the model, test the same model (without retraining of the model) on 12 consecutive months
  - For instance if 2019 and 2020 were not used to train the model
  - First run: Extract the features from 1-1-2019 to 1-6-2019 and test on the period between 1-6-2019 to 1-9-2019
  - Second run: Extract the features from 1-2-2019 to 1-7-2019 and test on the period between 1-7-2019 to 1-10-2019
  - repeat for 12 steps and calculate for each the confusion matrix, precision, and recall
  - Use the model to choose a portfolio and invest in 300 - 50 stocks from 2022-06-01, you can train the model on any period you choose, what made you choose that period?
  - Calculate the Sharpe ratio and compare your portfolio and the S&P500 in terms of Sharpe ratio, Volatility, Max Draw-down

### **Part 2 Grading (40%)**

- Completeness of the answer (10%)
- Quality of the chosen features (15%)
- Quality of the final portfolio results in terms of (10%)
- Originality and Overall Quality of the task(5%)

## **Part 3**

### **Using a Scanner + Trader hybrid (20%)**

1. Use the portfolio that was used in part 2.3 to pick a stock to run a simple trading algorithm based on a single indicator
2. Calculate the performance in terms of Sharpe Ratio, Volatility, and Max Draw-Down of the hybrid solution
3. Discuss the difference between the results of part 2 and of part 3

### **Part 3 Grading (20%)**

- Completeness of the answer (15%)
- Conclusions from the comparison (5%)