**Project 1:**

***Sector Focus: Exclude Financials, Real Estate, Utilities sectors in your sample***

You are an expert Data Analytics and Research team called DARES in a large financial investment company. The company management would like to launch an “AI-driven” stock investment strategy product. Last year, they hired a consulting company named AI-Cons that consists a team of data analysis experts with MMA diplomas. The AI-Cons with MMA degrees did not deliver any convincing results.

Some of the competing investment management firms have already launched AI-driven products. The market pressure is mounting to deliver your own version. DARES team have to deliver a good analysis and provide advisory on how to implement such a product.

You (the DARES Team) think that the historical data on stocks should reveal patterns of successful indicators that predict the high-performing, or “winning”, portfolios. Big data may reveal what works in the stock market.

If your ML model is shown to be successful, you will be able to implement it as a real portfolio, sell it as an investment product to your clients, and potentially bring a significant amount of revenue into the company.

With an exciting prospect for a good bonus and professional fame, the team has already started to build an approach. As the first step, a large historical data set is constructed. The data set consists of stock returns and many company features driven from financial accounts such as valuation ratios, profitability ratios etc.

Data collection is already done. The team is ready to gather around a table to discuss what to do. There are a number of points to discuss and decide initially:

* Data files are downloaded from different sources. They need to be merged/cleaned/organised.
* What are the particular issues related to ‘cross-sectional, time-series’ data?

**Step 1:**

Organize the data set, prepare the variables for standard ML algorithms. Make sure the data set is organised and prepared (scaling, normalization, cleaning, etc. done) after a careful EDA process.

The details of the analysis of the model and the related code are to be decided via team discussion. You are free take executive decisions to select variables to drop from the data set if you do not see much use or value for the ML analysis.

Your aim is to forecast the set of stocks that are likely to be the best (and worst) FUTURE 3-Month performers at a point in time based on their features measured and recorded prior to the observed stock returns. If your model is able to pick the FUTURE WINNERS and avoid the FUTURE LOSERS, your portfolio will outperform the other possible portfolios that are constructed with a passive decision (i.e., hold all stocks in equal weight, or buy the largest market-cap stocks).

By using the historical data available at a point in time, your model will (hopefully) forecast the WINNERS of the next 3-months, and you will create an equal-weighted portfolio with the predicted WINNERS.

*Analogy: If the stock returns in each time interval is assumed to be the results of a horse race, your objective is to find an ML model to predict the winning horses by using the information available prior to the horse race. The historical data set allows you to Train the models and Test them to check the prediction success achieved in the past.*

As a part of the required output, your team has to demonstrate the performance of your ML-driven stock picking strategy with a back-test. That is to show the historical success of your method as an investment strategy to answer the following: If we had actually used our ML methods to construct portfolios in the past with the information that was available back then, and repeated the prediction process over time for many “horse races” what would be our performance?

**Target Variable:** Forward 3-Month return WINNERS (best performing stocks in the next 3-month period)

**Features:** Financial Ratios, Past Returns, Sector or Industry Group etc. available in the provided data sets

**Questions to discuss and answer:**

How do we define the WINNERS and/or LOSERS at each period? (Rank them?)

How shall we design the prediction model? Is this a Classification problem or a Regression problem?

How shall we select TRAIN, TEST and VALIDATION sets? Should we use a moving-window approach?

Are there redundant features that we can discard? How do we decide?

**Step 2:**

* Run the ML method(s) of your choice to identify the features that have the greatest importance (and significance) to predict future returns.
* Present the results of your ML approach along with a description of your method.
* How do your results change when you use different train-test samples? Run your ML procedure over moving windows of time (such as moving 3-year window for the train sample, and subsequent period for the test sample).
* Do the “useful features” change over time with the moving sub-samples? Show how the set of “important features” (if any) change over moving-window samples. Show if there are any consistently useful features to predict the future Winner and Loser stock groups.
* Are there any features that you would propose to use consistently for the new AI-driven product? Are the ML model results statistically or economically convincing? Briefly explain.

**Step 3:**

For each train/test sub-sample and the associated model, create the basket of Top 30 and Bottom 30 stock return predictions. Compare the PREDICTED (i.e. predict (my\_model, test\_sample)) and ACTUAL (actual return data in the test sample) of the PREDICTED TOP 30 basket and PREDICTED BOTTOM 30 basket. Show if the model and methods you employ predict the future WINNER and LOSER stocks consistently over different train-test samples.

By using the returns of the PREDICTED TOP 30 and PREDICTED BOTTOM 30 stocks, show the overall performance of your strategy during your sample period.

In other words, collect the portfolio returns that you would have achieved over the sample period if you had applied your methods repeatedly as a portfolio strategy tool in the past. See how you would have performed if you implemented your own predictive machinery.

**Step 4.**

Show your stock portfolio that you would hold based on your ML-based model forecasts by the end of 2023. Which stocks would you select (Top 30 [or 40] Predicted Winners) to be in your portfolio by 2023-December?

Assume you created an equal-weighted portfolio of Forecasted Top 30 Winner Stocks in 2023-Dec. Collect the stock price/return data for the 2023-Jan to 2023-Mar period and show how your portfolio has performed compared to a market benchmark index. What would be your over/under-performance in the first quarter of 2023?

**Step 5.**

Apply a different AI strategy.

1. Ask GPT what to buy among your stock universe by the end of 2023 based on the information available by the end of 2023. See how GPT’s portfolio would have performed relative to your ML-based portfolio and relative to the market benchmark in the first 3 months of 2023.
2. Repeat (i) for the portfolios constructed by the end of 2022.

**Make sure to record your GPT prompts and the responses that you used for portfolio construction.**

Based on (i, ii, iii) who wins? GPT, or your ML model?

Do you think LLMs can help with portfolio strategy? Why or why not?

-----------------------------------------------------------------------------------------------

A simple example of preliminary GPT prompt for Step 5:

“*I will give you a list of stock tickers. Make a guess or a forecast about the percent return of these stocks over the next 3 months by suggesting a predictive model or by creating a story to justify your stock preferences. Use the latest data, news and information. Suggest a portfolio of 4 stocks based on your model and/or story. Here are the tickers: AMZN, MSFT, XOM, JJ, C, JPM, PFE, AA, K, HD, NVDIA, TSLA*”

…..

*“Now, assume you have information and historical data only and only up to 2022-January. You do not know anything beyond the calendar year of 2022. Eliminate all information that is dated beyond the year 2022. What would be your portfolio decision then?*

Try similar prompts and variants for your stock universe. Use different iterations, back-and-forth prompt conversations to decide on the useful form of prompting the LLM model.