FINE 695
Group Project
Description

Introduction:

The goal of this challenge is similar to the individual assignment:

1. Find machine learning (ML) methods that work well for building portfolios.
2. Identify key financial factors that help pick better-performing stocks.
3. Choose stocks and decide how much to invest in each one to create an optimized portfolio and predict its future success.
4. Run a back-test—this means checking how your portfolio would have performed in the past based on your stock picks.

The coding language of this competition is Python.

Please use the following style guide for your codes: https://google.github.io/styleguide/pyguide.html and be explicit in your commenting.

You may choose to be working on the Lightning AI Studio platform, which provides access to both CPUs and GPUs. This user-friendly platform allows you to code directly from your laptop while seamlessly utilizing their powerful computational resources through an API.

Data:

You now have access to a larger cross-section of stocks to work with. These are small- to mid-cap U.S. publicly traded firms, with the number of stocks per month ranging from 2,500 to 4,000. If you choose to use the dataset as provided, no further adjustments to the date or calendar time are needed. However, it's crucial to note that there are two future-looking variables in the dataset: **stock excess returns** (stock_exret) and any variables related to **earnings-per-share** (with the prefix eps_). The **date**, **year**, and **month** fields in the dataset represent the month when the returns you are predicting are actually realized. For instance, if the date is '20240830', the year is 2024, and the month is 8, this means the returns refer to **August 2024**. To predict these returns, the right-hand-side variables (listed in factor_char_list.csv) come from **July 2024**. This adjustment is already made in the dataset to ensure that no forward-looking data is used to predict future returns.

Similarly, for all earnings-related variables, including analyst estimates and actual earnings, the data comes from the **month in which you are predicting returns** (e.g., August 2024 in the example).

It's also important to note that not all firms have analysts' forecasts available from I/B/E/S. This is often because these are smaller firms that analysts typically don't cover. However, these firms can have the **highest growth potential** — though they also come with a higher risk of failure.

IMPORTANT:

If you plan to add more data, ensure that you do not simply merge datasets using the available year-month from the original data. Otherwise, you risk including **concurrent** (future-looking) information. For example, if you're predicting returns for rows where the date is '20240830', your **new data** must come from **July 2024** to avoid using any information from the future.

You will need to adjust for this yourself, and we will rigorously verify that no future data has been used in your estimations. The simplest approach is to create a new **date_lag** variable as part of your data handling process. In Python:

```
import pandas as pd
```

```
df['date_lag'] = df['date'] - pd.DateOffset(months=1)
```

The stock identifiers are the same as before. Depending on your data sources, you can identify the stocks by their PERMNO or CUSIP.

Each month, your portfolio should include *between 50 and 100 stocks (long only position)*. Your goal is to select stocks for your investment strategy using big data and machine learning (ML). Note, your portfolio should not exceed 100 stock names and be at least 50 names at all times.

As before, for each month and each stock, you have access to the same 147 firm-specific characteristics or signals as during the first challenge (their names are already imputed into a csv file which you used in our python codes: factor_char_list.csv). You can always consult Appendix A.

Note that all predictors are lagged by one month from time t, while all returns, the predicted variable (stock_exret), are from time t+1. Thus, this is a truly predictive exercise. Note, stock_exret is monthly access return over risk free rate. This is different from gross returns which also include risk free rate.

Here's an overview of the key suggestions and rules for the Group Project:

Identify Key Fundamentals: Your first task is to determine which fundamental factors are most influential for stock performance. These fundamentals should serve as the basis for your strategy.

Use ML/AI for Prediction: Next, employ machine learning, AI, and big data techniques to make predictions about these fundamentals. Instead of focusing directly on stock returns—which can be highly volatile and difficult to forecast— you can also focus on predicting stable, underlying fundamentals that drive returns. This approach often leads to more accurate and actionable insights.

Construct a Long-Only Portfolio: Based on your fundamental predictions, build a long-only trading strategy. Your strategy should involve selecting stocks and constructing a portfolio that aligns with your insights on fundamental factors.
This is just a suggestion as it involves 3 independent steps. You can do it all in one step or skip one of the steps. *Note, the we will be judging your investment process, and how much sense it makes, besides the overall portfolio performance.*

Important Rule: You cannot use forward-looking data to identify or predict fundamentals. For example, if your training sample spans from January 2000 to December 2009, you can only use signals and fundamentals available within that timeframe. Data from 2010 onwards should not influence or inform your model training, as this would mean the model has access to future information.

The data, as before, will keep the variable identifying earnings per share reported in the next quarter (eps_actual), the consensus of analyst forecasts for the earnings-per-share next quarter are available in the column eps_meanest (mean of analyst forecasts) or eps_medest (median of analyst forecasts). eps_stdevest provides the standard deviation of analyst forecasts, which tells you how disperse the forecasts are. Note that both analyst forecasts and actual earnings are already from the future (just like stock returns, stock_exret),

you can directly use them as left-hand side variables without any adjustments.

As before you can us other data from WRDS (https://wrds-www.wharton.upenn.edu/), or from EDGAR, or from any other sources to merge them with our data set. You can merge the data by the following stocks' identifiers: *permno* or *cusip*, and date (common practice is to use the year-month pair). The stock's *permno* or *cusip* are unique firm identifier in the data, you just need to control for the date to have them correctly aligned in calendar time. Aside from key identifiers, dates, company information, and stock characteristics, there are other data which were used during the data construction process. You can safely ignore them if they are not specifically mentioned in this note.

The provided dataset already contains investment strategies designed to potentially outperform the market portfolio. This means there are valuable signals and strategies embedded within it, which could help you strive to outperform benchmarks like the S&P 500. However, the actual magnitude of outperformance you can achieve remains uncertain and open-ended.
Consider, for example, the well-known performance of Buffett's portfolio. While he's widely recognized, there are other managers (like this one for example) who have outperformed him—but they simply aren't as prominent in public media. Moreover, the best-performing human portfolio manager has yet to be rigorously compared against an AI-driven portfolio manager. This emphasizes that, in today's evolving landscape, there are no definitive limits or standards for what level of performance is achievable.
As you work on your strategies, remember that this is a unique opportunity to explore the boundaries of what AI can accomplish in asset management. Good luck, and we're excited to see the innovative approaches you'll bring to the table!

Choice of ML algorithms & Training:

As before – we advise you to have at least the first 10 years as the first training sample, and the choice of ML algorithm is also yours. We are no longer providing sample codes as you already developed your own algorithms and approaches. We still encourage you to use training and tuning/validation sample.

Trading Strategy Portfolio Evaluations:

While the statistical performance (*OOS* R^2, or MSE) provides some preliminary ideas about the accuracy of the predictions, investors care more about the economical benefits that the model provides. That is, whether we can use the predictions to construct a portfolio that generates superior returns. Your portfolio is an investment product and your evaluation criteria will follow common industry reporting statistics.

Estimation of portfolio Alpha, Beta, and Tracking Error:

We continue to define the alpha as the intercept from a simple linear regression where you regress your portfolio excess over risk free rate returns on the excess over risk free rate returns of S&P500. That is

$$R_{p,t} - r_{f,t} = \alpha + \beta(R_{SP500,t} - r_{f,t}) + \epsilon_t$$

where $R_{p,t}$ is your strategy portfolio monthly returns, r_f is risk free rate, and $R_{SP500,t}$ is monthly returns on S&P500 index. Since the returns you predict are pre-adjusted by subtracting the risk rate (*exret* stands for excess return), you can directly use your portfolio return as the left-hand side. The data file *mkt_ind.csv* contains the monthly data for risk free rate and S&P500 returns.

The intercept is Alpha or risk adjusted return, and the slope coefficient is the beta. When you want to annualize Alpha (as an output from the regression it is monthly), you can multiply it by 12.

Tracking error is the standard deviation of residuals (ϵ).

Trading Criteria:

1. You should have a ***long only*** position in min 50 stocks or max 100 stocks (between 50 to 100 holdings).

2. **Portfolio turnover:** not to exceed 25% per month. This is already aggressive, as on 100 names holdings you can flip 25 names per month. If you are an institution managing 10 bln portfolio – this turnover will be prohibitively high as market frictions will deteriorate your total portfolio performance. You can see a good article about price impacts here.
3. **Leverage**: None, you cannot use any form of leverage.
4. **Risk**: Normally institutions have a limit on ex-post tracking errors. Tracking error is the excess of volatility of your portfolio over benchmark. Your benchmark is still S&P500. We suggest the cap on excess volatility of 500 bps, or 5%. That is if for example the monthly volatility of S&P500 is 4%, the volatility of your portfolio should not exceed 9%. Note, lower volatility of your portfolio will be even more appreciated. This criteria applies for the testing, out of sample period for your portfolio vs. S&P 500.
5. **Max investment in one single name (stock) is** not to exceed 10%. That is the average weight in one single name should not exceed 10%.

Guidelines for the presentations and the Decks.

While the technology is important, your innovative ideas, your investment process, and the future further potential of your investment product will definitely play a very significant role. Sometimes, as you are measuring ex-post portfolio performance, and you break limits on turnover or volatility – it is OK. Just report that you broke the constraints and as long as you provide an economic reasoning why this is happening, and what further work you would do if you had more time.

Here is the minimum to include in the deck:

1: Executive Summary.

Summarize your strategy, ML algorithm(s) chosen and portfolio performance vs S&P500

2: Describe your investment strategy in detail: What predictive signals you use to form the strategy? Why and how you chose these signals? Present your top 10 holdings on average over out-of-sample testing period, 01/2010 to 08/2024, and their average weights. Plot cumulative performance returns of your trading strategy vs S&P500 for OOS testing period, 01/2010 to 08/2024

3. Data and Methodology: describe the data if you use alternative extra data and main methodology. Try to justify why you chose a specific ML approach but do not go too deep unless you need to connect finance fundamental to your specific technological adaptations.

4. Portfolio Performance statistics for out-of-sample testing period, 01/2010 to 08/2024 for your portfolio vs S&P 500

At the very least you have to report the following portfolio performance statistics (their computation is provided in portfolio_analysis_mma.py) vs corresponding statistics of S&P500 for the same time period:

- Average annualized portfolio returns
- Annualized portfolio standard deviation
- Annualized Alpha (market risk-adjusted return, for your portfolio only)
- Sharpe Ratio (annualized)
- Information Ratio (annualized, for your portfolio only)[1]
- maximum drawdown,

[1] Sharpe and Information ratios are already annualized in the code provided, portfolio_analysis_hackathon.py. Th annualize standard deviation – you need to multiply it by $\sqrt{12}$. To annualize Alpha or average return, multiply it by 12.

- maximum one-month loss
- Tracking Error (annualized for you portfolio only)
- *Portfolio Turnover* (for your portfolio only)

5. The discussion of your strategy. Did it perform the way you trained it and did it meet your expectations? What are the main fundamental signals contributing to the performance of your portfolio? What are the most profitable positions (stocks) that drove the performance, and why. What are the macro-economic events that contributed to the performance. Potential improvements that you could make to this strategy

Grading Criteria	**Max points**
Ideas How original from above description and innovative are the ideas and strategies?	**25**
Investment Process Is the investment process well thought-out, structured, and consistent with the stated objectives?	**25**
Use of Technology (ML/AI)	**25**
Performance vs S&P 500	**15**
Potential for Real Life Implementation	**10**

Appendix A.

A.1 Stocks-Specific Features

Table 1: Stock-specific Features

Feature	Acronym	Reference
Firm age	age	Jiang Lee and Zhang (2005)
Liquidity of book assets	aliq_at	Ortiz-Molina and Phillips (2014)
Liquidity of market assets	aliq_mat	Ortiz-Molina and Phillips (2014)
Amihud Measure	ami_126d	Amihud (2002)
Book leverage	at_be	Fama and French (1992)
Asset Growth	at_gr1	Cooper Gulen and Schill (2008)
Assets-to-market	at_me	Fama and French (1992)
Capital turnover	at_turnover	Haugen and Baker (1996)
Change in common equity	be_gr1a	Richardson et al. (2005)
Book-to-market equity	be_me	Rosenberg Reid and Lanstein (1985)
Market Beta	beta_60m	Fama and MacBeth (1973)
Dimson beta	beta_dimson_21d	Dimson (1979)
Frazzini-Pedersen market beta	betabab_1260d	Frazzini and Pedersen (2014)
Downside beta	betadown_252d	Ang Chen and Xing (2006)
Book-to-market enterprise value	bev_mev	Penman Richardson and Tuna (2007)
The high-low bid-ask spread	bidaskhl_21d	Corwin and Schultz (2012)
Abnormal corporate investment	capex_abn	Titman Wei and Xie (2004)
CAPEX growth (1 year)	capx_gr1	Xie (2001)
CAPEX growth (2 years)	capx_gr2	Anderson and Garcia-Feijoo (2006)
CAPEX growth (3 years)	capx_gr3	Anderson and Garcia-Feijoo (2006)
Cash-to-assets	cash_at	Palazzo (2012)
Net stock issues	chcsho_12m	Pontiff and Woodgate (2008)
Change in current operating assets	coa_gr1a	Richardson et al. (2005)
Change in current operating liabilities	col_gr1a	Richardson et al. (2005)
Cash-based operating profits-to-book assets	cop_at	
Cash-based operating profits-to-lagged book assets	cop_atl1	Ball et al. (2016)
Market correlation	corr_1260d	Assness, Frazzini, Gormsen, Pedersen (2020)
Coskewness	coskew_21d	Harvey and Siddique (2000)
Change in current operating working capital	cowc_gr1a	Richardson et al. (2005)
Net debt issuance	dbnetis_at	Bradshaw Richardson and Sloan (2006)
Growth in book debt (3 years)	debt_gr3	Lyandres Sun and Zhang (2008)
Debt-to-market	debt_me	Bhandari (1988)
Change gross margin minus change sales	dgp_dsale	Abarbanell and Bushee (1998)
Dividend yield	div12m_me	Litzenberger and Ramaswamy (1979)
Dollar trading volume	dolvol_126d	Brennan Chordia and Subrahmanyam (1998)
Coefficient of variation for dollar trading volume	dolvol_var_126d	Chordia Subrahmanyam and Anshuman (2001)
Change sales minus change Inventory	dsale_dinv	Abarbanell and Bushee (1998)
Change sales minus change receivables	dsale_drec	Abarbanell and Bushee (1998)
Change sales minus change SG&A	dsale_dsga	Abarbanell and Bushee (1998)
Earnings variability	earnings_variability	Francis et al. (2004)
Return on net operating assets	ebit_bev	Soliman (2008)
Profit margin	ebit_sale	Soliman (2008)
Ebitda-to-market enterprise value	ebitda_mev	Loughran and Wellman (2011)
Hiring rate	emp_gr1	Belo Lin and Bazdresch (2014)

Table 1 continued from previous page

Feature	Acronym	Reference
Equity duration	eq_dur	Dechow Sloan and Soliman (2004)
Net equity issuance	eqnetis_at	Bradshaw Richardson and Sloan (2006)
Equity net payout	eqnpo_12m	Daniel and Titman (2006)
Net payout yield	eqnpo_me	Boudoukh et al. (2007)
Payout yield	eqpo_me	Boudoukh et al. (2007)
Pitroski F-score	f_score	Piotroski (2000)
Free cash flow-to-price	fcf_me	Lakonishok Shleifer and Vishny (1994)
Change in financial liabilities	fnl_gr1a	Richardson et al. (2005)
Gross profits-to-assets	gp_at	Novy-Marx (2013)
Gross profits-to-lagged assets	gp_atl1	
Intrinsic value-to-market	intrinsic_value	Frankel and Lee (1998)
Inventory growth	inv_gr1	Belo and Lin (2011)
Inventory change	inv_gr1a	Thomas and Zhang (2002)
Idiosyncratic skewness from the CAPM	iskew_capm_21d	
Idiosyncratic skewness from the Fama-French 3-factor model	iskew_ff3_21d	Bali Engle and Murray (2016)
Idiosyncratic skewness from the q-factor model	iskew_hxz4_21d	
Idiosyncratic volatility from the CAPM (21 days)	ivol_capm_21d	
Idiosyncratic volatility from the CAPM (252 days)	ivol_capm_252d	Ali Hwang and Trombley (2003)
Idiosyncratic volatility from the Fama-French 3-factor model	ivol_ff3_21d	Ang et al. (2006)
Idiosyncratic volatility from the q-factor model	ivol_hxz4_21d	
Kaplan-Zingales index	kz_index	Lamont Polk and Saa-Requejo (2001)
Change in long-term net operating assets	lnoa_gr1a	Fairfield Whisenant and Yohn (2003)
Change in long-term investments	lti_gr1a	Richardson et al. (2005)
Market Equity	market_equity	Banz (1981)
Mispricing factor: Management	mispricing_mgmt	Stambaugh and Yuan (2016)
Mispricing factor: Performance	mispricing_perf	Stambaugh and Yuan (2016)
Change in noncurrent operating assets	ncoa_gr1a	Richardson et al. (2005)
Change in noncurrent operating liabilities	ncol_gr1a	Richardson et al. (2005)
Net debt-to-price	netdebt_me	Penman Richardson and Tuna (2007)
Net total issuance	netis_at	Bradshaw Richardson and Sloan (2006)
Change in net financial assets	nfna_gr1a	Richardson et al. (2005)
Earnings persistence	ni_ar1	Francis et al. (2004)
Return on equity	ni_be	Haugen and Baker (1996)
Number of consecutive quarters with earnings increases	ni_inc8q	Barth Elliott and Finn (1999)
Earnings volatility	ni_ivol	Francis et al. (2004)
Earnings-to-price	ni_me	Basu (1983)
Quarterly return on assets	niq_at	Balakrishnan Bartov and Faurel (2010)
Change in quarterly return on assets	niq_at_chg1	
Quarterly return on equity	niq_be	Hou Xue and Zhang (2015)
Change in quarterly return on equity	niq_be_chg1	
Standardized earnings surprise	niq_su	Foster Olsen and Shevlin (1984)
Change in net noncurrent operating assets	nncoa_gr1a	Richardson et al. (2005)
Net operating assets	noa_at	Hirshleifer et al. (2004)
Change in net operating assets	noa_gr1a	Hirshleifer et al. (2004)
Ohlson O-score	o_score	Dichev (1998)
Operating accruals	oaccruals_at	Sloan (1996)
Percent operating accruals	oaccruals_ni	Hafzalla Lundholm and Van Winkle (2011)
Operating cash flow to assets	ocf_at	Bouchard, Krüger, Landier and Thesmar (2019)
Change in operating cash flow to assets	ocf_at_chg1	Bouchard, Krüger, Landier and Thesmar (2019)

Table 1 continued from previous page

Feature	Acronym	Reference
Asset tangibility	tangibility	Hahn and Lee (2009)
Tax expense surprise	tax_gr1a	Thomas and Zhang (2011)
Share turnover	turnover_126d	Datar Naik and Radcliffe (1998)
Coefficient of variation for share turnover	turnover_var_126d	Chordia Subrahmanyam and Anshuman (2001)
Altman Z-score	z_score	Dichev (1998)
Number of zero trades with turnover as tiebreaker (6 months)	zero_trades_126d	Liu (2006)
Number of zero trades with turnover as tiebreaker (1 month)	zero_trades_21d	Liu (2006)
Number of zero trades with turnover as tiebreaker (12 months)	zero_trades_252d	Liu (2006)