**FT5005: Machine Learning for Finance   
Financial Accounting Forecasting of USA Public Firms**

**Deadline:** April 21th Midnight 11:59pm

**Final Group Project:**

1. The tentative deadline is April 21th Midnight 11:59pm. You upload slides, reports, and recorded videos of your presentation to Canvas.
2. You need to decide on ONE target industry and focus on building a forecasting model that works well for public firms in that industry. See Appendix A for more information about the list of candidate industries.
   * Before Feb 26 Wed. Please choose 3 candidates of target industries in Appendix A. The reason is we need to prepare datasets for you. We will NOT prepare datasets in the industry no group or only few groups are interested in.
3. **The project’s context and objective**. You pretend you are a human financial analyst who recommend stocks in that industry. Most analysts rely purely on human intelligence. This project asks you to use ML to enhance human intelligence. So you should not blindly throw in all data into an ML model to predict. You should try to learn how a human anlyst make prediction in your target industry, collect the same data and use domain knowledge to select and create features that make sense to predict Y. This part will be graded as you read from the guidelines later.
4. Target variables for prediction. Choose **2 out of 3** (revenue, EBITA, stock’s cumulative abnormal return around earnings announcement) of the Y variables in Appendix B. You could propose one variable of your choice, pending on approval.
5. Datasets: one important requirement is you need to collect one more dataset and do features engineering. See Appendix A for more details. The default datasets are: <https://drive.google.com/drive/folders/1UaZQTNwb3pfNig_YfER8wN8ysFchtGVp?usp=sharing> (you do not need all datasets. Include only the variables and datasets related to the Y in your target industry.)
   * 1. Compustat => USA public firms’ accounting information
     2. Compustat – industry specific information: companies may disclose financial variables that are unique to an industry.
     3. IBES => Analysts forecasts, but it is difficult to merge with Compustat.
     4. CRSP => USA stock prices and other securities prices
     5. CBOE Volatility Index
     6. (To Be Uploaded) Macroeconomics statistics.
     7. (To Be Uploaded) News.
     8. (To Be Uploaded) Abnormal Return around earnings announcement.
     9. (To Be Uploaded) Earnings conference call scripts.
     10. Exchange rate datasets are provided. But you can get those easily from Yahoo Finance.
     11. You can also get many more commodity prices or ETF of those commodity prices from Yahoo Finance.

**Final Group Project’s First report:   
Deadline: midnight March 31st 11:59pm**

**Target region: New York City**

**Rui: Collect geographical information from different datasets based on the requirement from Korea Paper**

**Hokiya: draft the report for section2 requirement 1 and 2**

**Alya: draft the report and find more papers**

Page limit is 8 pages excluding references and exploratory data analysis charts but including everything else. 12 font size. Normal margin. A4 size. Single-spacing is acceptable but is not required if you meet the page limit.

* + **Section 1 (about 1 page): Overview**
    - Clearly explain your target industry.
    - Clearly state your research objective: what is the additional dataset and the features engineering (conceptually) to improve the prediction performance. As a simple example to explain this requirement, your objective could be Twitter sentiment can increase the accuracy of predicting stock return (Twitter data is collected by you). In the first report, you do not need to explain the concrete formula of how to construct the sentiment. But you need to decide Twitter sentiment is one of the key features you will construct.
    - Please briefly justify why your new dataset and new features engineering may be useful to predict Y variables.
    - Provide supporting evidence for your claims in this section. Credible sources include statistics from the government, analysts’ reports, famous news sources, academic papers, or famous consulting firms’ reports (IDC, Gartner …etc.). Bloggers are not preferred but are acceptable and are better than your own words without any supporting evidence.
  + Section 2 (about 3-4 pages): Existing Literature
    1. **Requirement 1**: You are required to identify some “explanatory theoretical papers” that shows your target feature engineering may affect the target Y.
    - The minimum requirement is you need to find at least ONE paper that shows X in section 1 causally affects your Y either theoretically or empirically.
      * + Factor X is something you can construct based on your additional data (your features engineering).
        + Y can be any of the three target variables.
    - This requirement is to help you build domain knowledge and guide you to use domain knowledge to do features engineering.
    - This part is opened, more papers and more potential X are better for getting higher marks.
    1. **Requirement 2**: Find some existing papers, blog articles, or data competition results about the same or similar prediction problem. Please create a table to list three things (1) the best method, (2) **performance metrics and values,** and (3) **important features** used in those papers. Important features could be defined by importance scores OR by conceptual arguments. If the authors do not disclose that, you can leave it blank. This is for you to compare your own performance with other similar prediction problems conducted by someone else.
    - An academic paper is preferred. But if you cannot find any academic paper, you can cite blog articles or credible data competition results.
    1. For searching academic papers, you can use Google Scholar or Web of Science via NUS library system. You should be able to download papers from almost all decent journals via Web of Science system. If you cannot download that paper, please make a list of references you plan to download and email me. I can help you or show you how to download those papers.
  + Section 3 (about 2 pages including tables): Data
    1. Explain the source of your additional dataset and how you collect those data. Include your data collection codes as part of the zipped submission.
    2. **Requirement 3**: Present the summary statistics of the processed variables before scaling. It is required to include your additional data into summary statistics. I am not asking for the statistics of raw data. I am asking for the summary statistics to after your data preprocessing. The data should not be scaled at this step because the key point is for you and for me to read the maximum and minimum to detect any data errors. For your model training, you need to scale your data.
    - Remember to report both DV (Y) and input features (X).
    - For numerical input features, about summary statistics, you need to report (mean, median, max, and min) values of numerical variables. For values of categorical variables, you only need to report selected variables which you think they are important. For categorical variables, you only need to report the number of types and “roughly” the proportion of each type. “Roughly” is when there are many types of one categorical variable Say if there are 100 types, then you only need to report the proportion of top types of a categorical variable. This depends on your data because you may have too many variables to report or too many types of one categorical variable.
    1. **(Optional) Exploratory Data Analysis**: Provide some tables, charts, graphs that visualize the data characteristics of your dataset. Put these into appendix and this is not counted within page limit. **This part is not graded** but this step is something you should do to learn more about the relationship between X and Y. See the Appendix C for more information.
    2. Explain the sample size and how you plan to construct your training and validation datasets clearly. This is important for times series (or panel data) forecasting problems.
    - **Please keep 2021 to 2024 as the holdout test samples** and report the prediction performance metrics in these 4 years of data.
    - For the other dataset, you can decide training and validation set. But try to use as many samples as possible.

**Final Group Project’s Second report:**

**Deadline: Same as everything else** April 21th Midnight 11:59pm

* + Page limit is 20 excluding references but including everything else. Formatting is the same as those in the first report. The page limit is longer for you to include charts and figures, not text. Quality is more important than quantity. I don’t grade by the number of pages. I may count the number of useful features constructed.
  + For completeness, you can include revised Sections 1-3. Also, you can update those sections, so the report is consistent. This does not count under 20 pages limit.
  + Section 4: Features Engineering
    - **Requirement 5: Features Engineering.** This is open-ended. You are required to create **AT LEAST 3** new features that can improve your prediction performance.
      * **At least one is based on your literature review in the first report.**
      * You must have at least **one other** input feature constructed from “advanced methods”. The definition of “advanced” is a bit vague. Advanced could mean complicated formula based on finance, or by advanced text mining methods such as topic modelling or a supervised learning approach. For example, a new feature could be by simple formulas. Call a python function, or existing dictionary approach (especially positive/negative words), then it does not meet this requirement. But if you somehow create your own dictionary manually or have a novel idea to create the dictionary, then it meets the requirement.
      * Explain how you construct these 3 features clearly.
      * Explain the domain knowledge behind the idea to create these 3 features.
      * You are allowed and highly encouraged to come up with more than 3 new features. The more the better and this is one focus of my grading. The grading about this part includes your efforts, creativity, use of your domain knowledge in that problem, and the performance improvement from features engineering. For example, you simply create a lot of ratios that do not make sense may not be a strong case of features engineering.
  + Section 5: Explain Your Methods Clearly.
    - In the final report, you are allowed to use any prediction method.
    - Methods not covered in this module are also fine.
    - You do not need to report all numerical results of all cases you have tried. You can mention what you have explored and only report the best model of your finding and some benchmarking cases (see the next bullet point about results).
    - **Requirement 6 Stacking:** You should try at least one case of Stacking, and you can discuss whether stacking helps.
  + Section 6: Results.
    - **Explain your main performance metrics clearly**. R-square is the default for numerical Y and ROC-AUC is the default for a binary Y.
    - **Start with simple benchmarking performance, report the following.**
      * Use Y(t-1) to predict Yt without any model.
      * Benchmarking cases from your first report.
      * I may ask my team to provide more benchmarking cases.
      * Any simple benchmarking cases decided by your group (OLS, ARIMA…etc.).
    - This part is also open-ended. You can decide how to report various techniques for improving the performance. You only need to compare your final best model with selected other case to show which part of your efforts (e.g., features, model, or preprocessing) contribute to the improvement of prediction performance. For example, (1) (required) in your best model, with or without features engineering, how is your performance? (2) (required) with or without stacking, how is the performance? (3) (optional) In your final model, with or without features selection or dimension reduction, how much is the performance improvement…etc. (4) (optional) other things.
    - Grading will depend on how much you can improve the prediction performance, your efforts, and the techniques used throughout the project.
  + Section 7: Discussions and Conclusion.

**Grading and Submission**

1. Only one student of each group submits ONE zipped file to Canvas. The zipped file includes one (PDF or Word) file of your report, Python codes for the final prediction modeling part, and the data file for running those codes in CSV format. Create another zip file and share with TA online that includes raw data and data pre-processing codes leading to the final input data file.
   1. Only Python is allowed for the machine learning parts.
2. For code and data files, make sure that is easy for the TA to verify. If TA cannot figure out how to make it work, you bear the risk of losing marks. We may or may not give you another chance to resubmit and likely you will lose some but not all marks for this part. You only need to provide the very last version of Python code and data file for verifying your results in the report.
3. Do not copy your Python codes from other groups. You will be heavily penalized if get caught. You can use the Python codes from the Internet, but it is better to provide a reference if you copy a large block of codes. You are allowed to use ChatGPT to create codes.
4. **Visualization in report can be created by any other software;** for example, Tableau is a famous one. Power BI could be the second most popular one. Both are not free. You don’t need to stick to ugly charts/figures by Python or Excel. Again, informative visualizations will be assessed during presentation. The brand of software you use won’t be graded directly.
5. My grading will be a balanced assessment of your efforts spent, technical difficulty (including assessment about errors or advanced techniques used in codes), the final report’s writing quality (including visualization of results and reporting of results), and your prediction performance over benchmarking cases.
   1. For this project’s learning objective, I prefer you try to replicate or use advanced packages than using your creativity to come up a new algorithm. In other words, if you choose to fully replicate a complicated project from previous winners in data competition, your overall grades may NOT be lower than a new idea of your own. Anything works well to improve the prediction performance will be rewarded, just like real world commercial solutions.

**About Your Presentation**

1. Each group is presented by one student. No bonus for that student.
2. The length of presentation is tentatively set at 12 minutes.
3. Please upload your presentation slides and the video file to Canvas before the deadline.
4. You can easily record your presentation by Zoom. Any other software is also fine.
5. Presentation contents include (1) what is your analytics problem, (2) data source (no need to cover details about data collection or data cleaning), (3) main DV for prediction and important useful features included in the dataset, and (4) baseline classification results: what is the proportion or distribution of Y without any classification? What is the performance metric from the easiest method or existing literature? (5) Your few best model’s classification model and results. (6) Summarize with “what kind of methods and features engineering perform the best?” in your project.

**Peer Assessment Report**

1. Option 1: The default is no peer report, and the same group will receive the same marks. You don’t need to do anything.
2. Option 2: If there is an unequal contribution and/or collaboration problem, please add an appendix at the end of the report that states the contribution of each student: I mean who did which part in the presentation, in conducting analytics, and in writing the report. I will give different grades to different students based on the quality and number of jobs done. This part is not within the page limit.
3. Option 3: For any reason that you don’t like to do Option 2, you can simply state clearly to me that you will email me peer report individually. Then each of your team member can email me separately about (1) who did what in the project? (2) you feel each member should receive higher, average, or lower marks in your group. In other words, I may also email each of you to gather more information about the contribution of team members in this kind of group.

**Appendix A: Target Industries**

1. The current list of industries is.
2. Airlines (SIC 4512)
3. Oil and Gas (SIC 13)
4. Utilities (SIC 49)
5. Retailers - Apparel & Footwear (SIC 56)
6. Retailers - General Merchandise Stores (SIC 53)
7. Eating & Drinking Places (SIC 58)
8. Real Estate Developers (SIC 6552)
9. General Building Contractors (Construction of Buildings) (SIC 15)
10. Banks (SIC 6021)
11. Life insurance firms (SIC 6311)

For broader industries, you are allowed to work on the subset of firms that are more homogenous in their core business.

* How to decide your target industry? Several considerations.
  1. Your group’s common interests in learning.
  2. You should choose the industry so that it is easier for you to meet the requirements of this project. (1) One challenge is you need to collect one more dataset and do features engineering. (2) You need domain knowledge to do features engineering so that you can predict *Y* in that industry well.
  3. The number of firms in the target industry should not be too small or too large, depending on your computational powers.
  4. When I create this list of industries, I try to think from your perspective and also pick the industry that is more homogeneous, and firms are more similar in their business. Also, I try to pick the industries that are not difficult to understand, except banking and insurance.

**Appendix B: Target Dependent Variables**

* 1. Firm-level quarterly revenue in one industry of your choice.
  2. Firm-level quarterly EBITA in one industry of your choice.
     1. For both revenue and EBITA, you can predict transformation of those variables as in A1 in your model. But you need to calculate and compare the prediction performance of the original values of Y.
  3. One case of abnormal stock return prediction around the earnings announcement. You decide the start date and end date of your stock return for prediction. We will share with AR and CAR. Do not use raw stock return.
     1. For example, t=0 is the earnings announcement date. You predict the return between t+10 closing price divided by t+1 closing price. Or, t+1 closing price divided by t-1 closing price is also fine.
        1. The first example implies you buy after earnings announcement and hold for 10 days.
        2. The second example means you buy right before earnings announcement and sold shortly after earnings announcement. In other words, if you can predict earnings, then you should trade like this case. In the first case, it is more about after earnings announcement, you predict what is the reasonable stock price.
  + You are allowed to replace one of the three by a financial variable in the datasets. In other words, you predict one variable of your own choice and one variable out of those three. **But you need to get my approval first**. The legitimate reason is for example
    1. You find a paper that predicts Y by an interesting dataset and/or interesting features engineering, and you want to replicate that paper.
    2. You find a Y of interests and also want to explore the value of an additional dataset and features engineering to predict that Y variable.
    3. The reason should not be you want to predict something easier.

**Appendix C: ChatGPT: how to do Exploratory Data Analysis**

### 1. Understand the Dataset

* **Objective**: Understand what you're working with and what you want to achieve.
* **Actions**:
  + Identify the number of features and observations.
  + Check the data types of each feature.
  + Understand what each feature represents, the units of measurement, and the ranges of values.

### 2. Clean the Data

* **Objective**: Ensure the quality of your data.
* **Actions**:
  + Handle missing values.
  + Correct data types if necessary (e.g., converting strings to dates, make sure numerical columns are not category ID).
  + Identify and address outliers.
  + Remove duplicates.
  + Perform data type conversions (e.g., categorical variables encoded as strings).

### 3. Univariate Analysis

* **Objective**: Examine each variable individually.
* **Actions**:
  + For continuous variables, use summary statistics and visualize with histograms, boxplots, or density plots.
  + For categorical variables, use frequency counts and visualize with bar charts or pie charts.

### 4. Bivariate Analysis

* **Objective**: Understand the relationship between two variables.
* **Actions**:
  + Use scatter plots and correlation analysis for continuous variable pairs.
  + For a continuous and a categorical pair, use boxplots or violin plots.
  + For two categorical variables, use cross-tabulation and associated chi-squared tests.

### 5. Multivariate Analysis

* **Objective**: Understand interactions between multiple variables.
* **Actions**:
  + Use heatmaps to visualize correlation matrices.
  + Employ pair plots to see bivariate relationships between combinations of variables.
  + For mixed data types, consider multiple correspondence analysis.

### 6. Identify Patterns and Anomalies

* **Objective**: Find meaningful insights that could inform subsequent analysis.
* **Actions**:
  + Look for trends over time.
  + Identify subgroup patterns.
  + Detect anomalies that may be of interest or that may affect later analysis.

### 7. Report Insights

* **Objective**: Document what you've found in a way that's accessible to stakeholders.
* **Actions**:
  + Create clear and informative visualizations.
  + Summarize key findings.
  + Outline potential limitations of your analysis.
  + Provide recommendations or next steps based on your findings.

### 8. Prepare for Further Analysis

* **Objective**: Set the stage for in-depth statistical analysis or machine learning.
* **Actions**:
  + Summarize insights and hypotheses for further testing.
  + Export cleaned and transformed data.
  + Document your EDA process thoroughly for reproducibility.

### Tips

* Be curious and ask a lot of questions about your data.
* Be systematic; document your findings and the reasons behind the actions you take.
* Be critical; don't assume that the data you're working with is error-free or that it's representative.
* Remember that EDA is iterative; as you learn more about the data, you'll need to circle back and revise earlier steps.

### Output

* EDA typically results in a report or presentation accompanied by visualizations and summary tables. It should convey the insights gained from the data and any peculiarities that may need further investigation or consideration in future analyses.

Remember, EDA is more of an art than a science, and it’s often where data scientists spend most of their time. It requires a balance between statistical rigour and intuitive storytelling.