FT5005: Assignment 1 (10 Marks in total)

**Due: 11:59pm of March 3rd 11:59pm Sunday**

1. **Overall Instructions:**

* Please submit one Jupyter file for all of your works. Do not send it separate files one-by-one. If you have to use multiple files, zip into one file for submission.
* We will not have extension of the deadline since you have about one month to work on A1.
* Do not include un-necessary codes into your submissions. Please add comments to explain your codes to help TA understand better your codes. **TA has the right to deduct up to 2 out of 10 total marks if you include many un-necessary codes or your code is not understandable.**
* Using GPT to help you is allowed in this class. I am more open-minded about this. This is like in the future, if you are assigned a programming task, you are allowed to use GPT to help you. I believe most of the assignment questions cannot be solved just by sending the questions to GPT.
* Copying from your classmates is strictly NOT allowed. Both students will receive 0 out of 10 if any kind of cheating is caught.

1. **Problem and Dataset DESCRIPTION**

This data is a subset of actual CRI datasets. I help you cleaned the data and all are in input.csv. The dataset information is provided below

1. Sample firms.
   1. USA Public firms in the Energy industry. There are three sub-industries.
   2. Industry classification code is BIGS (Bloomberg Classification Code)
      1. 131010: Oil & Gas production.
      2. 131011: Oil & Gas exploration.
      3. 131110: Renewable energy.
2. Sample period: from 1988 to 2023.
   1. You need to drop data <=2000 by yourself. It is your job to do this data preprocessing.
3. Y variable. Y is binary, Y=1 means default in the next 12 months. The default events are defined as (1) Bankruptcy filing (2) miss the payment of loan interest payment or principal payment or (3) any kind of debt restructuring.
4. X variables
   1. There are 13 numerical variables and one categorical input variable (industry). Do not use company name, company ID, or date in your model. You will be penalized heavily.
5. Hold-out test dataset: Please keep 2021, 2022, 2023 as hold-out test dataset. Do not use those rows in any training or validation. Only use those rows for final evaluation of performance.
6. **Questions**

**Question 1**. (3 Marks) Data pre-processing. Several requirements

* 1. Drop all rows that is before 2000/1/1. We keep the 2000 January data.
  2. (1 Mark) Create Onehot Encoding for the industry classification code variable. Self-study the sample code uploaded to A1 folder.
  3. (1 Mark) Winsorizing numerical variables at two tails by 5% each. Self-study the sample code uploaded to A1 folder.
  4. After the last two steps, use MinMaxScaler to normalize all 14 features to 0 to 1.
  5. Be careful that handling categorical variable and winsorization should be BEFORE c, so that the categorical variable is also scaled (especially when you use standardization, using min-max does not matter).
  6. **Output the summary statistics of your X and Y for TA to check.**
  7. (1 Mark) In general, after this step, you should carefully check the min. and max. of each of X variable. Please do so in A1. The definition of X is in the whitepaper’s page 4. List at least one issue that the min. or max of one variable looks suspicious and may have data error. This assignment question is about finance, less about data mining. But I want you to practice and understand the importance of this step. I will grade this in your final project.

**Question 2**. (3 Marks) End-to-End LightGBM

Now please use the data <=2020 to build a forecasting model by LightGBM that meets the following requirements

* You must do GridSearchCV with 2-fold cross validation. 2-fold is to save you time in this step.
* Tune the LightGBM by specifying the Grid properly. You decide the values and number of cases. If you are not tuning the correct parameters, you may be penalized.

1. (1 Mark) Report the best validation AUC (from two-fold cross validation) and the best model’s test AUC.
2. (1 Mark) Note that this approach of cross validation is WRONG for time series forecasting. The correct way is to use “time series split”. We do a primitive version of time series split in this question. You use the data <=2015 to train. You tune/validation by the performance of 2016 to 2020 to decide the best parameters. If you want to learn using the time series split more properly, sample code is also provided for your reference.
   1. You can compare the prediction performance of these two cases on the test dataset. On average, using grid search will overfit. So the validation performance will be much better than the test performance. In one case like in our A1, results may not be that obvious. But you can still compare 4 AUCs.
3. (1 Mark) Now redo your analysis by using PR-AUC as the performance ONLY in the grid search CV case (although it is wrong). The learning objective is to learn how to code PR-AUC for tuning. Sample code is also provided for your reference.

**Question 3**. (4 Marks) Yahoo Finance and Metrics Importance Score

1. Use the sample code to download the oil price from Yahoo Finance. It is fine that you do not save the file and directly merge with the input.csv data frame for further data analysis.
2. (2 Marks) Now is the challenging part. Add oil price (daily closing price) into your input features in Q2. Several more instructions below.
   1. Merge by date directly. Be careful in date formatting. Do not blindly match by string’s exact match.
   2. When on a date of input.csv there is no transaction price, you need to use the closest last trading day’s closing price. If you do not have the data in the very first row, feel free to drop those few rows.
   3. Create two columns as two input features, one is raw oil price. The other is a daily return by [Oil\_t/Oil\_(t-1)]-1
3. (2 Marks) With oil price, try to build a prediction model as in Q2. You are freely allowed to use any tuning method with LightGBM to build a prediction model. Report your validation PR-AUC and test PR-AUC. TA will grade this part also by your prediction performance. For unreasonable validation and test-performance (too high or too low), we will deduct up to 2 Marks for this part. In other words, the grading of this part will based on you correctly code your hypermeter tuning and your prediction performance is not too bad.
   1. This is not a data competition. We only penalized very poor prediction performance. The rule is if your prediction performance is top 60% percentile, then full marks in terms of performance as an example. Then for each bottom 10%, we deduct 0.5 out of 2 marks.
4. You can also check the input features importance score. Is oil price useful in your prediction model? This part is not graded. But check the importance score of oil should be helpful for your learning.