IE 582 Final Exam
Fall 21, due 23:59 of January 19th, 2022

Objective and Evaluation

Use the data provided with this exam:

- One file is the **training data** that you should use for model building.
- A second file contains **test data** to be used to evaluate your model. This file does not have the information of the labels. You are asked to provide your predictions to be evaluated for grading purposes. The details are provided below.

The training and testing data were selected randomly from the original data set.

Build a classification model for this data based on the methods described in the course. You are also free to use extended versions of the approaches covered in the lectures.

Note that this data set is created for credit scoring purposes. Due to privacy concerns, the name of the features are masked except the first four columns. These are:

- "loan_application_id": the id of the loan application (not a valid feature)
- "loan_amount": the amount of the requested loan
- "default": if the loan is paid pack, it is 0. Else it is 1. This is the target column
- “customer_age”: age of the customer (only feature that was not masked)
- rest: 5th column to 63th are features related to customer and loan application.

Performance measure:
Note that this problem is slightly different than what we have seen in the lectures in terms of the loss. The evaluation will be based on the cost of misclassification. Recall that we have discussed evaluation of classification models based on a misclassification cost matrix where we introduce a misclassification cost matrix and evaluate models based on the overall cost of the approach. However this approach assumes that the misclassification costs are the same for all instances (i.e. IE582 Fall21 Model Evaluation.pdf file slide no:11). Here the misclassification cost is not the same for the instances due to the varying loan amounts (provided in loan_amount column). For instance, correct classification of a default case for an amount of 100 units compared to a default case for an amount of 1000 units should not be considered as the same in the evaluation phase. Classification of the application for an amount 1000 units (which is from default class) is more valuable. This brings us how to compute the misclassification cost based on each instance. This depends on the type of the misclassification as:

- Classifying Default (Class 1) as Good (Class 0): You can assume that we will lose all the amount as the creditor since we will give the credit but the person does not pay the debt back.
- Classifying Good (Class 0) as Default (Class 1): You can assume that we will lose the profit which amounts to the 15% of the loan amount since will not approve the loan application although person pays the debt back.

The final performance will be evaluated based on **the total money lost based on your predictions for the test data.** Although loan amount is provided for evaluation purposes, you are free to use this information as a feature in your models. In real life cases, banks do not use this information in their credit scoring models since it is not realistic due to inflation and etc. This data covers a time period around a year therefore you are not expected to suffer from this fact in case you think that use of this information improves your model.

Note that you are expected to provide your predictions as labels (i.e. 1 or 0 - no probabilistic predictions). If you prefer to use an approach that provides probabilistic predictions, you may want to work on your threshold to find out the class labels. This threshold can be determined based on the defined cost.

Prediction and Report

Submit a written report with a brief description of your final method, how you evaluated your methods, and you choose the parameter settings. Probably fewer than five pages are sufficient. You can prepare your report in any format convenient to you (i.e. a document prepared in MS Word, Jupyter Notebook, Markdown, Latex and etc). Also attach your codes for your proposed model.

Your report should have the following format:

1. *Introduction:* Descriptive analysis of the given data (do not go into too much details, a brief summary of what you have as the data, type of the features and etc. will be fine).
2. *Approach:* Explain your approach to this problem.
3. *Results:* Provide your results and discussion.
4. *Conclusions and Future Work:* Summarize your findings and comments regarding your approach. What are possible extensions to have a better approach?

You are expected to submit predictions for the test data according to the following format in a csv file (name your file as “test_predictions.csv”):

loan_application_id,	**prediction**
1	1
2	0
…	…

Lastly, you are expected to upload a zip file named as “Surname_Name.zip” on Moodle containing three files. These are:

- Report: Preferably formated as pdf or html (word document is also fine)
- Predictions: predictions.csv file
- Code: Script (working code that have your final approach). Do not spend too much time in tidying your codes but they should be working for your final approach). This is in the format of your chosen scripting language (i.e. *.py or *.r, jupyter notebooks are also OK).