MH81112 Python Quiz Question

Given current COVID 19 serious situation, it is important for a bank to build a prediction model to perform its client credit card default prediction, so that they can contact those clients who may default in advance to avoid high interest charge. Before starting the project, its CEO believes it is beneficial to do a POC project to evaluate the feasibility and effectiveness of the idea.

As an intern in the bank, you are assigned to do the POC project and present all the findings to the CEO. You are excited about this opportunity and hope you can do an impact and interesting work to impress the bank management, so that you can start building your own career in financial industry at this bank. To prepare for the presentation, you can download the data set Client_default.xlsx at NTULearn.

You have learned some classification methods in MH8112 using Python. You will explore the following suggested steps (feel free to add anything else, e.g. imbalance data, that you believe it is important to your work – useful to impress CEO):

1. Conduct **exploratory data analytics** to handle missing values and conduct visualizations to understand which features are related to the target variable.
2. Build **three classification models** (you may consider avoiding SVM as it could be very slow) and **an ensemble learning model** under 5-fold cross-validation (5FCV), and report their average Accuracy, MCC, Precision (for class Y=1), Recall (for class Y=1), F1-measure (for class Y=1).
3. Perform **feature engineering** (including both feature selection and new feature creation) to show if there are some performance differences using selected subset of features or newly added features. Note it is ok that performance does not improve given limited quiz time, but you should try feature engineering process.
4. Based on your findings, using **Markdown to summarize your major results and insights** that you will present to the bank CEO.

Your python program should be written in **Jupyter Notebook** (with proper comments) or **Google Colab** (with proper comments), and can be run properly (**you should keep your running results**), with **your name** **as the source code filename** (also write down your ***name*** and ***matriculation no*** in the beginning of your code).

You should send your ipynb code to xlli@ntu.edu.sg (no need to send the data) and cc'ed to yourself as well, to make sure you have sent it out successfully. Note I understand you may not be able to complete all the four questions. However, you just need to send whatever you have done and I will mark accordingly.

Good luck!

Data set Client_default.xlsx can be downloaded from NTULearn.

Variable Information:

1. **Customer demographic Variables**: CUSTOMER ID, AGE, GENDER, EDUCATION, MARRIAGE.
2. **Payment variables in last 6 months**: PAYMENT_1M, PAYMENT_2M, PAYMENT_3M, PAYMENT_4M, PAYMENT_5M, PAYMENT_6M. For example, PAYMENT_1M means its payment in last month (-1).

   Variable value: -2: no consumption, -1: pay duly in full, 0: The use of revolving credit, 1 = delay 1 month; 2 = delay 1 month, . . . , 8 = delay 8 month, 9 = delay 9 months and above.

   For the meaning of revolving credit. You can refer to

   https://www.investopedia.com/terms/r/revolvingcredit.asp

   or Chinese explanation

   https://wiki.mbalib.com/wiki/%E5%BE%AA%E7%8E%AF%E4%BF%A1%E7%94%A8#:~:text=%5B%E7%BC%96%E8%BE%91%5D-,%E4%BB%80%E4%B9%88%E6%98%AF%E5%BE%AA%E7%8E%AF%E4%BF%A1%E7%94%A8,%E9%87%91%E9%A2%9D%E5%B0%B1%E6%98%AF%E5%BE%AA%E7%8E%AF%E4%BF%A1%E7%94%A8%E4%BD%99%E9%A2%9D%E3%80%82
3. **Bill Amount variables in last 6 months**: BILLAMOUNT_1M, BILLAMOUNT_2M, BILLAMOUNT_3M, BILLAMOUNT_4M, BILLAMOUNT_5M, BILLAMOUNT_6M. For example, BILLAMOUNT_1M means its bill amount in last month.
4. **Amount Paid variables in last 6 months**: AMOUNTPAID_1M, AMOUNTPAID_2M, AMOUNTPAID_3M, AMOUNTPAID_4M, AMOUNTPAID_5M, AMOUNTPAID_6M. For example, AMOUNTPAID_1M means its amount paid in last month.
5. **Target Variable**: DEFAULT_PAYMENT_NEXT_MONTH**.**