Project Instructions

In the spreadsheet `loan_data.xlsx`, you are given data on 23,337 individual loan accounts. The `DPD_Data` tab summarizes the days past due (DPD) per account for the historical window October 1, 2019 to October 1, 2020. It indicates the number of months that the account was delinquent in paying the amount due for a given month. The `Accounts_Data` tab summarizes each account and its 13 variables, namely:

a. Account: account number
b. Sex: biological sex
c. Dependents: number of dependents
d. Civil_Status: civil status
e. House_Type: house type
f. Education: highest educational attainment
h. Yrs_Employed: number of years employed (from first to current employment)
i. Credit_Status: status whether the credit account has current (or ongoing) loan, non-earning (or non-paying) loan, or paid-off loan
j. Months_Loan: duration of the loan in months
k. Amortization: amortization amount per month
l. Purpose_Loan: purpose/reason for loaning
m. Gross_Salary: average gross salary per month
n. Credit_Ratio: calculated by dividing the amortization by the gross salary

As scorecard developers, your goal is to create a credit (behavior) scorecard based on the behaviors of the accounts for the past year. This project will be done by one group of three and two groups of two. The group with three members will be doing two models (Model 1 and Model 2), and three scorecards in total. The group with two members will just be doing Model 2, and two scorecards in total. Here are the guidelines:

1. Clean the data set if necessary. Check if there are outliers in the data set and deal with them properly if they exist.
2. State your definition of a "bad" account and confirm this "bad" definition by performing a roll-rate analysis on the historical delinquency performance of the accounts. Use also the current versus worst delinquency comparison method and compare the resulting "bad" definition with that of the roll-rate analysis.
3. Based on this definition, create a new variable Creditworthiness that indicates whether an account is "good" or "bad".
4. Split the data into training (80%) and test (20%) sets using `scorecard::split_df()`. Set the seed to 314.
5. Perform an exploratory data analysis on the 13 variables and on Creditworthiness.

Model 1: The General Credit Scorecard

1. Use `R` or `Excel` to obtain the optimal binning per variable. Calculate the weights of evidence (WOEs) per bin and the information value (IV) per variable. Do this for the training set and validate the binning using the test set by checking if the resulting WOEs still form a logical trend.
2. Replace all raw data values in the training and test sets with their corresponding WOEs which were calculated using the training set.

3. Perform a logistic regression on the train set in `R`. Experiment on which variables should be included in the model.
4. Evaluate the performance of the generated models on the test set.
5. Choose the best model and create the final scorecard.

Model 2: The Segmented Credit Scoring Model

In general, it is expected that model performance can improve if the population is divided into homogeneous subgroups, creating a credit scorecard for each subgroup. This is because these homogeneous subgroups may behave differently from others.

1. Segment the training set into two subgroups based on the purpose of loan: one group whose loans are "essential" (HOME REPAIR, TUITION, MEDICAL, and PAY DEBTS) and another whose loans are "non-essential" (LEISURE, AUTO, APPLIANCE, TRAVEL).
2. For each subgroup, create a scorecard and validate it using the same methods as in Model 1: The General Credit Scorecard.
3. Select the best model for each subgroup, and create the final scorecard for each subgroup.
4. (for the group with three members) Compare your results with the general credit scoring model.

Submission Guidelines

1. The required output is a model documentation report. Assume that you are reporting the results to the senior management.
2. The written output should contain all pertinent figures, tables, and graphs.
3. The report must be printed on A4 paper, preferably typset using LaTeX. Use 12 pt font size and standard font styles.
4. `R` codes and their corresponding outputs must be saved in an `R` notebook. Submit the `R` notebook together with the `Excel` files (such as the training and test sets, your WOE binning, and the final scorecards) used and generated in the project.
5. Compress all pertinent files (report, `R` notebook, `Excel` files) in one zip file.
6. The filename of the report **and** the zip file should be of the following format:

   MATH 271.1_Project2_[SurnamesofMembersInAlphabeticalOrder]

   For the e-mail submission, use the filename as the subject line of the e-mail.
7. Submit the zip file to jdumbrique@ateneo.edu **and** rfadri@ateneo.edu on or before **November 22, 2021 11:59 PM**.