**Assessment Instructions**

The following assessment should be completed in either Python or R. For each numbered section below, write code and comments to solve the problem and to show your reasoning. If using R, please provide a completed .rmd or .pdf file. If using Python, please provide a completed Jupyter Notebook (.html or .pdf).

For sections that ask you to give outputs, you can paste the (commented) output directly below the code that runs it or you can provide outputs in separate files. How you present your results is entirely up to you.

Please return your script along with any supporting files within 48 hours of receiving the assessment.

Following the assessment, you will have 20 minutes to present your findings. More details can be found below in Question 5.

**Question 1**

Load in the dataset from the accompanying file "account-defaults.csv". Perform a simple exploratory data analysis, including summary statistics and visualizations of the distributions and relationships.

Overview of the "account-defaults.csv" dataset:

This dataset contains information about loan accounts that either went delinquent or stayed current on payments within the loan's first year. FirstYearDelinquency is the target variable, all others are predictors.

ID - a unique identifier corresponding to a loan

FirstYearDelinquency - indicates whether the loan went delinquent within the first year of the loan's life (delinquencies indicated by values of 1)

AgeOldestIdentityRecord - number of months since the first record was reported by a national credit source

AgeOldestAccount - number of months since the oldest account was opened

AgeNewestAutoAccount - number of months since the most recent auto loan or lease account was opened

TotalInquiries - total number of credit inquiries on record

AvgAgeAutoAccounts - average number of months since auto loan or lease accounts were opened

TotalAutoAccountsNeverDelinquent - total number of auto loan or lease accounts that were never delinquent

WorstDelinquency - worst status of days-delinquent on an account in the first 12 months of an account's life; values of '400' indicate '400 or greater'

HasInquiryTelecomm - indicates whether one or more telecommunications credit inquires are on record within the last 12 months (inquiries indicated by values of 1)

**Question 2**

Build one (or more) model(s) using the "account-defaults.csv" dataset. Please explain your choice of model(s).

The objective of modeling with this dataset is to be able to predict the probability that new accounts will become delinquent in the first year. Our primary concern is to differentiate lower-risk accounts from higher-risk accounts in a rank-order fashion (and not just to classify accounts as 'yes' or 'no' for delinquency).

Make sure to accomplish the following:

* Identify the strongest predictor variables and provide interpretations
* Identify and explain any issues or caveats with the data and the model(s)
* Calculate predictions and show model performance on out-of-sample data
* Summarize results in a fashion that differentiates lower-risk accounts from higher-risk accounts

**Question 3**

Split up the dataset by the WorstDelinquency variable. Run a regression of FirstYearDelinquency ~ TotalInquiries.

Extract the predictor's coefficient and p-value from each model. Combine the results in a list where the names of the list correspond to the values of WorstDelinquency.

**Question 4**

Prepare a scoring function to use your predictive model(s) in a production environment. To do this, you will need to accomplish the following:

* Ingest new loans for scoring. You may assume that new records are received in batches in the same .csv format as the "account-defaults.csv" dataset.
* Perform any required data preprocessing steps such as missing imputation.
* Score the preprocessed input data using your model(s).
* Output the scores so your business partners can match the scores to the new loans.

**Question 5**

Prepare a Powerpoint presentation that summarizes your findings and model(s). You will have 20 minutes to present followed by a question period. Your audience for this presentation is a team of analysts who are looking to use your model’s predictions to reduce their loan portfolio’s risk of default. Consider what information your business partners will need to know in order to use your model(s).