The written report, R script, and evaluations are due Thursday, May 7 at 8:00 AM. All three should be uploaded to the provided Dropbox link. The presentations will take place on Thursday, May 7 starting at 8:00 AM. This project should be completed in teams of 3 or 4 or your choosing. Should you need assistance finding a group, please let Dr. Weber know.

Overview:

Banks and credit card clients often have a high risk in that they don't know if an individual will default on their credit card payment or not. Specifically, banks want to reduce this risk by building a model that utilizes features they have in their database about each individual. The 25 features they have collected are:

- `limit_bal`: Amount of given credit in dollars, including individual and family/supplementary credit.
- `sex`: Gender (1=male, 2=female)
- `education`: (1=graduate school, 2=university, 3=high school, 4=others)
- `marriage`: Marital status (1=married, 2=single, 3=others)
- `age`: Age in years
- `bill_amt1`: Amount of bill statement in September, 2005 (dollars)
- `bill_amt2`: Amount of bill statement in August, 2005 (dollars)
- `bill_amt3`: Amount of bill statement in July, 2005 (dollars)
- `bill_amt4`: Amount of bill statement in June, 2005 (dollars)
- `bill_amt5`: Amount of bill statement in May, 2005 (dollars)
- `bill_amt6`: Amount of bill statement in April, 2005 (dollars)
- `pay_amt1`: Amount of previous payment in September, 2005 (dollars)
- `pay_amt2`: Amount of previous payment in August, 2005 (dollars)
- `pay_amt3`: Amount of previous payment in July, 2005 (dollars)
- `pay_amt4`: Amount of previous payment in June, 2005 (dollars)
- `pay_amt5`: Amount of previous payment in May, 2005 (dollars)
- `pay_amt6`: Amount of previous payment in April, 2005 (dollars)
- `default`: Default the payment of October 2005 (1=yes, 0=no)

They have hired you and your team to build a model that can help them predict whether a customer will default on their payment based on these features.

Project Requirements:

Using the skills you've acquired in class, build a machine learning model to accomplish this task. It should output a single probability for each individual which represents their likelihood of defaulting.

Coding

Your R script should contain the following four sections and be organized and well commented.:

1. **Loading and processing the data:** In this section, you will load both the training and competition datasets "final_train.csv" and "final_compete.csv" and process it appropriately. *Note:* Some of the categorical variables in the data are coded numerically. You should convert them in to the `factor` form before analysing and building a model.

The training dataset will be the set that you use to train, validate, and/or test your model to determine which model you would like to present. The competition dataset comprises of just the predictors. You'll use the final model you select to run on the competition dataset to output a vector of probabilities for each of the observations. You can submit these probabilities to Dr. Weber for the competition. More details below.

2. **Building the models:** In this section, you should try **at least four different methodologies**, e.g., logistic regression, LDA, decision trees, bagging, etc. *Note:* You will have to decide how to best use the training data to build, tune, and score your models. You may also want to consider adding new features (variables) to the dataset to improve your model.

3. **Testing the model and results:** In this section, you will compute the accuracy metric of your choice for all of your models.

4. **Deployment:** This section will be where you deploy your model on the competition dataset. Your chosen model should take in the competition predictors and return a vector of probabilities for defaulting. This vector should be combined with the IDs and be output as a CSV file in the same order as the competition dataset. Should you choose to enter the competition, you can then email Dr. Weber this CSV file. **Regardless of your participation in the competition, your code should create this CSV file.**

Report

Your group needs to turn in an executive report on your project. Your report should be a brief summary of what you have done to build the model. It should be well written and professional. Your report should include, but not be limited to, your answers to the following questions:

1. Which methodology have you used to build the model? Which one produces the "best" model in your opinion?

2. What are the results of your models? Report any scores or figures you feel necessary to explain your point. *Note:* The scores on a model might be the misclassification rate, the false positive rate, false negative rate, AUC, or the log loss:

$$\text{log-loss} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i))$$

where

- n is the number of individuals in the test set
- $\hat{y}_i$ is the predicted probability of an individual that default the payment.
- y_i is the individual's behavior (1 = default, 0 = not default)
- log() is the natural logarithmic function

The smaller the log loss score the better. In general, the log loss score is better than the misclassification rate as it also considers the uncertainty of the prediction.

Grading

- Coding submission – 20 pts
- Successful implementation of four or more methods – 20 pts (5 pt penalty for each missing method)
- Executive report – 25 pts
- Have a model that is better than the trivial model – 25 pts
- Have a model that is 10% better than the trivial model – 10 pts

The trivial model: Since the majority of individuals in the dataset will not default on their payment, the trivial model will have the predicted probabilities equal to the proportion of defaulting cases in the training set.

Bonus: Should your team choose to enter the competition, you will send the CSV file you created in the deployment stage to Dr. Weber. Dr. Weber will then compute your log loss score. The team(s) with the smallest log loss score will split a bonus **10 points** (if one team gets the smallest, then they receive 10 points. If two teams tie, each team gets 5 points, etc.). The log loss score from each submission will be posted on a Google Sheet, along with the anonymized team name. You may submit as many predictions as you want. When you email your submission to Dr. Weber, be sure to include your anonymized team name and the CSV file. The competition ends at 8:00 AM on Thursday, May 7.

Getting Started: Check out the "Getting Started" file on Blackboard for the calculation of the trivial model and the log loss score.

Peer Evaluation

An important part of any project in the workplace is self and team evaluation. You can't know how to improve unless you realize where your shortcomings are (both as a teammate and as a person). One way to do that is to reflect. The posted Peer Evaluation is designed to do that and is up on Blackboard along with this document. It is important to objectively and thoroughly evaluate your work and the work of your colleagues. This can be a very uncomfortable process, but it is a vital one.

Your Peer Evaluation is due at the same time as the rest of the documents and should be uploaded to the same Dropbox link. Everyone should do their own evaluation and upload it. These evaluations should be filled out individually and privately, and will remain **confidential**. So honestly and thoroughly evaluate yourself and your team. Your feedback will be considered in assigning final grades for this project. Please be as fair and honest as possible. Failure to submit a **genuine** evaluation on time will result in a severe reduction of your final grade.

Presentation

Your presentation should be styled as a presentation of your findings to the client bank who hired you. This means you'll have to explain your process, methodology, and the models to a non-data scientist (executive at the bank) as well as a data scientist/technical person (bank analyst). As a result, your explanation should include a broad description of the models you built, why you chose to build these particular ones, an overview of your process and how the models work, and your final model and its interpretation. Your presentation should be 10-15 minutes.

The following rubric will be used to assess your presentation.

Group: _____

Criteria	Points
Explanation of Idea: Presents information, findings, arguments, and supporting evidence clearly, concisely, and logically; audience can easily follow the line of reasoning and is convinced that the results are the best option.	/13
Organization: Meets all requirements for what should be included in the presentation. Has a clear and interesting introduction and conclusion. Organizes time well; no part of the presentation is too long or too short.	/7
Hosting Abilities: Professionally runs presentation and presentation software (i.e., Zoom). Switch off between team members speaking is seamless. Looks poised and confident. Speaks clearly; not too quickly or slowly. Changes tone and pace to maintain interest. Rarely uses filler words (e.g., "Um")	/12
Presentation Aids: Uses well-produced visual aids to enhance understanding of findings, reasoning, and evidence, and to add interest.	/12
Team Participation: All team members participate for about the same length of time. All team members are able to answer questions about the topic as a whole, not just their part of it.	/6
Total:	/50