Introduction

This is an individual assignment and will be a chance for you to perform an applied data science project on a real data set.

We will be working with the credit\_card\_df data frame in this project. This data set contains information on over 4,000 customers of a U.S. bank. The description of this data and the variables contained in it are provided below.

The objective of this project is to explore the factors that lead to customers canceling their credit card accounts and develop machine learning algorithms that will predict the likelihood of a customer canceling their account in the future.

To complete this assignment, students must download the R notebook template and open the file in their RStudio application. Please click the button below to download the template.

## **Credit Card Account Data**

The credit\_card\_df data frame contains information on the customers of a large U.S. bank which provides a number of financial services including multiple credit card offerings.

The bank is looking to see if it can determine the factors that lead to customers canceling their credit card account and whether it can predict if a customer will cancel their account in the future.

To maintain profits, banks must maximize the number of customers with credit lines. It is also in their best interests for customers to carry large credit card balances from month-to-month to maximize revenue from interest charges.

The bank has experienced record levels of customers closing their credit accounts in the past couple of years and this is leading to declining revenue.

The bank’s goal is to become better at identifying customers at risk of canceling their account to minimize financial losses.

Specifically, the broad questions that the company is trying to answer include:

* What are the factors that are associated with customers closing their credit card accounts?
* Is it possible to predict whether a customer will close their account? If so, how accurate are the predictions?
  + How many costly errors is the model expected to produce?
* Are there any actions or policies the company can implement to reduce the risk of losing their customers?

The data set contains a mixture of customer demographics and their financial behavior.

The outcome variable in this data is customer\_status. This variable records whether a customer eventually closed their account and indicates a financial loss to the company.

**Note**: The outcome variable has been coded as a factor with ‘closed\_account’ (**the positive class**) as the first level. This is the format that tidymodels expects for calculating model performance metrics. There is no need to recode this variable in your machine learning process.

### **Variable Information**

| **Variable** | **Definition** | **Data Type** |
| --- | --- | --- |
| customer\_status | Customer status (closed account or active) | Factor |
| age | Customer age | Numeric |
| dependents | Number of dependents in household | Numeric |
| education | Customer education level | Factor |
| marital\_status | Marital status | Factor |
| employment\_status | Employment status | Factor |
| income | Annual income (US Dollars) | Numeric |
| card\_type | Type of credit card | Factor |
| months\_since\_first\_account | Months since first credit card account activated | Numeric |
| total\_accounts | Total accounts (credit, checking, and savings) | Numeric |
| months\_inactive\_last\_year | Months without credit card activity last year | Numeric |
| contacted\_last\_year | Number of times contacted last year by sales representatives | Numeric |
| credit\_limit | Current credit limit | Numeric |
| utilization\_ratio | Average monthly balance to credit limit | Numeric |
| spend\_ratio\_q4\_q1 | Ratio of total Q4 to Q1 spending last year | Numeric |
| total\_spend\_last\_year | Total amount charged last year | Numeric |
| transactions\_last\_year | Number of transactions last year | Numeric |
| transaction\_ratio\_q4\_q1 | Ratio of total Q4 to Q1 transactions last year | Numeric |

### **Raw Data**

credit\_card\_df

# Data Analysis

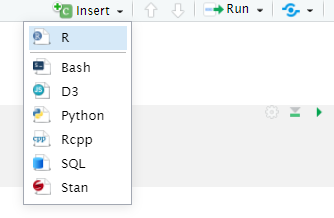
In this section, you must think of at least 5 relevant questions that explore the relationship between customer\_status and the other variables in the credit\_card\_df data set. The goal of your analysis should be discovering which variables drive the differences between customers who do and do not close their account.

You must answer each question and provide supporting data summaries with either a summary data frame (using dplyr/tidyr) or a plot (using ggplot) or both.

In total, you must have a minimum of 3 plots (created with ggplot) and 3 summary data frames (created with dplyr) for the exploratory data analysis section. Among the plots you produce, you must have at least 3 different types (ex. box plot, bar chart, histogram, scatter plot, etc…)

See the [Data Analysis Project](https://gmubusinessanalytics.netlify.app/data-analysis-project.html) for an example of a question answered with a summary table and plot.

**Note**: To add an R code chunk to any section of your project, you can use the keyboard shortcut Ctrl + Alt + i or the insert button at the top of your R project template notebook file.



Machine Learning

In this section of the project, you will fit **three classification algorithms** to predict the outcome variable,customer\_status.

You must follow the machine learning steps below.

The data splitting and feature engineering steps should only be done once so that your models are using the same data and feature engineering steps for training.

* Split the credit\_card\_df data into a training and test set (remember to set your seed)
* Specify a feature engineering pipeline with the recipes package
  + You can include steps such as skewness transformation, correlation filters, dummy variable encoding or any other steps you find appropriate
* Specify a parsnip model object
  + You may choose from the following classification algorithms:
    - Logistic Regression
    - LDA
    - QDA
    - KNN
    - Decision Tree
    - Random Forest
* Package your recipe and model into a workflow
* Fit your workflow to the training data
  + If your model has hyperparameters:
    - Split the training data into 5 folds for 5-fold cross validation using vfold\_cv (remember to set your seed)
    - Perform hyperparamter tuning with a random grid search using the grid\_random() function
    - Refer to the following tutorial for an example - [Random Grid Search](https://gmubusinessanalytics.netlify.app/lesson-08-r-tutorial.html#Hyperparameter_Tuning14)
    - Hyperparameter tuning can take a significant amount of computing time. Be careful not to set the size argument of grid\_random() too large. I recommend size = 10 or smaller.
    - Select the best model with select\_best() and finalize your workflow
* Evaluate model performance on the test set by plotting an ROC curve using autoplot() and calculating the area under the ROC curve on your test data

Summary of Results

Write a summary of your overall findings and recommendations to the executives at the bank. Think of this section as your closing remarks of a presentation, where you summarize your key findings, model performance, and make recommendations to improve customer retention and service at the bank.

Your executive summary must be written in a [professional tone](https://www.universalclass.com/articles/writing/business-writing/appropriate-tone-in-business-communications.htm), with minimal grammatical errors, and should include the following sections:

1. An introduction where you explain the business problem and goals of your data analysis
   * What problem(s) is this company trying to solve? Why are they important to their future success?
   * What was the goal of your analysis? What questions were you trying to answer and why do they matter?
2. Highlights and key findings from your Exploratory Data Analysis section
   * What were the interesting findings from your analysis and **why are they important for the business**?
   * This section is meant to **establish the need for your recommendations** in the following section
3. Your “best” classification model and an analysis of its performance
   * In this section you should talk about the expected error of your model on future data
     + To estimate future performance, you can use your model performance results on the **test data**
   * You should discuss at least one performance metric, such as an F1, sensitivity, specificity, or ROC AUC for your model. However, you must explain the results in an **intuitive, non-technical manner**. Your audience in this case are executives at a bank with limited knowledge of machine learning.
4. Your recommendations to the bank on how to reduce the number of customers closing their credit card accounts
   * Each recommendation must be supported by your data analysis results
   * You must clearly explain why you are making each recommendation and which results from your data analysis support this recommendation
   * You must also describe the potential business impact of your recommendation:
     + Why is this a good recommendation?
     + What benefits will the business achieve?