

## Scenario

You are working as a risk analyst with a bank. Apart from the other banking and loan services, the bank also provides credit card services that are a very important source of revenue for the bank. The bank wants to understand the demographics and other characteristics of its customers that accept a credit card offer and that do not accept a credit card.

Usually, the observational data for these kinds of problems is somewhat limited in that often the company sees only those who respond to an offer. To get around this, the bank designs a focused marketing study, with 18,000 current bank customers. This focused approach allows the bank to know who does and does not respond to the offer, and to use existing demographic data that is already available on each customer.

## Objective

Build a model that will provide insight into why some bank customers accept credit card offers. There are also other potential areas of opportunities that the bank wants to understand from the data.

Your senior management has also posted these other questions that will help them better understand their customers.

## Dataset

All the necessary files, as well as the dataset, can be found in the following repository: [**Mid-bootcamp project - Classification**](https://github.com/ironhack-edu/data_mid_bootcamp_project_classification).

The data set consists of information on 18,000 current bank customers in the study. These are the definitions of data points provided:

* **Customer Number:** A sequential number assigned to the customers (this column is hidden and excluded – this unique identifier will not be used directly).
* **Offer Accepted:** Did the customer accept (Yes) or reject (No) the offer. Reward: The type of reward program offered for the card.
* **Mailer Type:** Letter or postcard.
* **Income Level:** Low, Medium, or High.
* **Bank Accounts Open:** How many non-credit-card accounts are held by the customer.
* **Overdraft Protection:** Does the customer have overdraft protection on their checking account(s) (Yes or No).
* **Credit Rating:** Low, Medium, or High.
* **Credit Cards Held:** The number of credit cards held at the bank.
* **Homes Owned:** The number of homes owned by the customer.
* **Household Size:** The number of individuals in the family.
* **Own Your Home:** Does the customer own their home? (Yes or No).
* **Average Balance:** Average account balance (across all accounts over time). Q1, Q2, Q3, and Q4
* **Balance:** The average balance for each quarter in the last year

## Expected Outcomes

We encourage you to thoroughly understand your data and take the necessary steps to prepare your data for modeling before building exploratory or predictive models.

Since this is a classification model, you can use logistic regression for classification for building a model. You are also encouraged to use other models in your project, including KNN classifiers or decision trees.

### 1. Explore the data

To **explore the data**, you can use the techniques that have been discussed in class. Some of them include using the describe method, checking null values, using Matplotlib, and Seaborn for developing visualizations.

The data has many categorical and numerical variables. Explore the nature of data for these variables before you start with the **data cleaning** process and then **data pre-processing** (scaling numerical variables and encoding categorical variables).

For the target variable (Offer accepted – Yes/No), it is also an important to check the data imbalance ie., the number of people who responded with a yes vs. the number of people who responded with a no.

### 2. Build a Model

Use different models to compare the accuracies and find the model that best fits your data. You can use the measures of accuracies that have been discussed in class. Please note that while comparing different models, make sure you use the same measure of accuracy as a benchmark.

### 3. Visualize

You will also use **Tableau** to further visually explore the data. You will deep dive in the data for customers who accepted the offer vs the customers who did not and check their characteristics. E.g., we select the Yes level in Offer Accepted and then examine the distribution of accepted offers across the other variables in our data set and similarly for people who did not accept the offer.

## Deliverables

1. **A slides deck**: use [slides.com](https://slides.com/) for simple yet effective templates.
2. **A presentation**: Your public presentation should last between 5 and 7 minutes.
3. **SQL Queries**: Access the [SQL Questions file](https://github.com/ironhack-edu/data_mid_bootcamp_project_classification/blob/master/sql_questions_classification.md) and make sure you run these queries using SQL.
4. **A Tableau Dashboard**: Follow the instructions in [this file](https://github.com/ironhack-edu/data_mid_bootcamp_project_classification/blob/master/tableau_classification.md) to build in Tableau.
5. **Python code**: Your code will be reviewed. Make sure you follow the best practices explained in class so far.

## Evaluation criteria

* Critical Thinking
* Communication
* Transparency
* Thoroughness

### Rubrics

[These are the rubrics](https://education-team-2020.s3-eu-west-1.amazonaws.com/data-analytics/DATA2020-MidBootcamp+Project+Rubrics+-+Sheet1.pdf) we use to assess your work.

## Tips & Tricks

* Organize yourself (don’t get lost!). Respect deadlines.
* Ask for help but don’t forget that Google is your friend.
* Define a simple approach first. You never know how the data can betray you. 
* Document your work.
* Learn about the problem and what research has been done before you.
* Before making a graph, think about what you want to represent.