

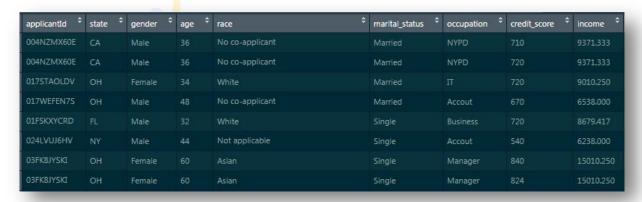
# **Project 4 - Loan Approval Prediction**

### **Problem Statement:**

You are the Senior Data Scientist at a major private bank. Since the last 6 months, the number of customers who are not able to repay their loan has increased. Keeping this in mind, you have to look at your customer data and analyze which customers should be given the loan approval and which customers should be denied.

## **Customer\_loan Dataset:**

The details regarding this 'customer\_loan' dataset are present in the data dictionary:



#### **Domain** – Banking

Lab Environment: R-Studio

## Tasks to be performed:

#### A) Data Preprocessing:

- a. Have a glance at the structure of the dataset and find if there are any missing values present
- b. Calculate the debt-to-income ratio and add it as a new column named 'dti'
- c. Create a new variable named 'loan\_decision\_status', where the value would be '0' if 'loan\_decision\_type' is equal to 'denied', else it would be '1'
  - i. Convert this variable into a factor
- d. Create a new data-set named 'customer\_loan\_refined', which would have these column numbers from the original dataframe (3,4,6,7,8,11,13,14)
- e. Encode 'gender', 'marital\_status', 'occupation', and 'loan\_type' as factors and then convert them into numeric



## **B)** Model Building:

- a. Divide the data into 'train' & 'test' sets and set the split-ratio to be 70%
- b. Apply feature scaling on all the columns of 'train' & 'test' set, except the 'loan\_decision\_status' column
- c. Apply principal component analysis on the first 7 columns of 'train' & 'test' set. The number of principal components obtained should be 2
- d. Build the naïve bayes model on the train set
- e. Predict the values on the test set
- f. Build a confusion matrix for actual values and predicted values

