Project on '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 analyse 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:

applicantId	state	gender	age	race	marital_status	occupation	credit_score	income
004NZMX60E	CA	Male	36	No co-applicant	Married	NYPD	710	9371.333
004NZMX60E	CA	Male	36	No co-applicant	Married	NYPD	720	9371.333
017STAOLDV	OH	Female	34	White	Married	IT	720	9010.250
017WEFEN7S	OH	Male	48	No co-applicant	Married	Accout	670	6538.000
01FSKXYCRD	FL	Male	32	White	Single	Business	720	8679.417
024LVUJ6HV	NY	Male	44	Not applicable	Single	Accout	540	6238.000
03FK8JYSKI	OH	Female	60	Asian	Single	Manager	840	15010.250
03FK8JYSKI	OH	Female	60	Asian	Single	Manager	824	15010.250

Domain – Banking

Lab Environment: R-Studio

Tasks to be done:

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