The problem statement was to predict whether the borrower will complete the payment during the next month or not.

My approach was to first understand the variable definitions and perform basic exploratory data analysis. I have checked the summary of the data to identify the characteristics of the different variables like mean and range for numerical variables and the different levels for categorical variables. I have also changed the values for m1 to m12 to binary class 0 and 1, based on whether the customer has defaulted or not.

I have checked whether there were any missing values or not. Also, I checked for outliers though I have not performed any outlier treatment.

I have also visualized the data for any underlying patterns.

I have checked the correlation between different variables and also normalized the data.

As this is an imbalanced class problem, I have used SMOTE to sample the data and create a balanced data.

For data modeling, I have split the data into 80:20 ratio. I have built the model in the training set and tested the data in the validation set. I have used different machine learning techniques. Then, on the basis of the precision value, I have selected the final model.

I have selected random forest as my final model.

Then, I have used the random forest model to predict the class in the test set.

library(data.table) 1.12.2

library(ggplot2) 3.2.1

library(plyr) 1.8.4

library(dplyr) 0.8.3

library(corrplot) 0.84

library(glmnet)

library(caret) 6.0-84

library(dummies) 1.5.6

library(randomForest) 4.6-14