Homework assignment 2

Instructions: *You should hand up a document report of **not more than** 10 pages long, inclusive of charts (if any) through Polymall's assignment submission portal. Negotiate with your lecturer's for any extension of this deadline if need be.*

Task description

The data you have was used to study credit card default behaviour in Taiwan. The variables in this dataset are described below:

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005.

The measurement scale for the repayment status is: -2 = balance paid in full and no transactions this period; -1 = Balance paid in full but account has a positive balance at the end of period due to recent transactions for which payment has not yet come due; 0 = customer paid the minimum due amount but not the entire balance. ; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

(*Source*: Taken from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)

You may use any software (KNIME, Minitab, Python, R etc…) to study and report on the following business questions:

MS9002: Data Mining Techniques

1. Evaluate and compare at least **two** predictive models for predicting customer default. You will be assessed on whether you have chosen a suitable evaluation metric and the persuasiveness of your recommendation. Take note that the class label distribution is **unbalanced.**

2. Study and report on the risk factors associated with credit default. Possible risk factors include the features/attributes of the data set and includes the variables relating to payment history and status (variables X6-X23). Note that you may also **derive** additional features from the underlying dataset and study their relation to default risk. Your report / study need not include a final predictive model, but rather to develop a profile of a typical credit card defaulter. We suggest using a **Naïve Bayes** and **Logistic Regression** models to aid you in developing this profile.

Remarks
It is not necessary to include screenshots of KNIME workflows in your submission report.