IE7300
Statistical Learning

Topic:
Default of Credit Card Clients

Group 4:
Harsh Dalal
Krish Patel
Yesha Patel

Dataset Description:

The project is mainly focused on the capacity of payback, a common issue faced by consumers that overuse their credit cards and rack up substantial credit card debt. This issue can damage consumer confidence in finance, therefore presenting a significant problem for banks and credit cardholders.

The dataset which we are using is aimed at the case of customers and their default payments in Taiwan. This analysis can help customers to analyze their default payments, which will help determine their repayment credibility.

Dataset Attributes:

Following is the description of our dataset:

1) Number of instances: 30000
2) Number of Attributes: 24
3) Data set characteristics: Multivariate
4) Attribute Characteristics: Integer, Real

The project is about classifying one binary attribute, default payment (Y=1, N=0), as the response variable. There is a total of 23 diverse features and 1 target variable. The description of features is given below:

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 payments.

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:

-1 = pay duly; 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.

The goal of the project:
The main goal of the project is to predict whether the customer will make the default payment or not.

Steps to be followed:
After performing preprocessing and interpreting the dataset by implementing univariate and multivariate analysis, we will check for the distribution of the dataset. Our target variable is labeled as y, which tells us whether the customer will make the default payment or not.

Splitting the dataset:
The first step we do before proceeding is to split the dataset into a train and test set. This helps us get rid of any inherent bias that we would incorporate during the Feature Selection/EDA stage. We split our data in the ratios of 70, and 30. Training 70%, Testing 30%. We will further split our training data into training data and validation data on a 70-30 ratio.

We are planning to start our modeling by making use of the following models:

1) Naïve Bayes Classification
2) Classifications trees such as Random Forest, Decision Trees, etc.

The bigger challenge to finding the credibility of a customer's default payment is to find whether they can make a default payment. Therefore, we plan to also delve into determining the probability of a customer's default payment.

We will then see the performance of these models, and make changes to the model based on the performance.

This is our tentative plan and the models that we will be implementing might change based on our result.

References:
https://www.sciencedirect.com/science/article/pii/S0957417407006719
https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients