CIE 417 Project – Task 1

Names

**Ahmed Mahmoud – 201800683**

**Ahmed Elghamry – 201801254**

Problem definition

Nowadays, credit cards have become a crucial part of our lives due to the easy usage and pay-back options offered by credit cards companies. Also, discounts and offers on credit cards make them alluring to people. Although it may seem that using credit cards would ease human lives, it could be a debt trap if not used wisely. Hence, we chose this dataset to monitor the behavior of adults using credit cards and how they pay their debts and bills.

Dataset description

This dataset carries information about the default payments, credit data, history of payment, bill statements, and demographic factors of credit card clients in Taiwan from April to September 2005. The data set contains 25 features:

**ID**: ID of each client

**LIMIT\_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit

**SEX**: Gender (1=male, 2=female)

**EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

**MARRIAGE**: Marital status (1=married, 2=single, 3=others)

**AGE:** Age in years

**PAY\_0**: Repayment status in September, 2005 (-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)

**PAY\_2**: Repayment status in August, 2005 (scale same as above)

**PAY\_3**: Repayment status in July, 2005 (scale same as above)

**PAY\_4**: Repayment status in June, 2005 (scale same as above)

**PAY\_5**: Repayment status in May, 2005 (scale same as above)

**PAY\_6**: Repayment status in April, 2005 (scale same as above)

**BILL\_AMT1**: Amount of bill statement in September, 2005 (NT dollar)

**BILL\_AMT2**: Amount of bill statement in August, 2005 (NT dollar)

**BILL\_AMT3**: Amount of bill statement in July, 2005 (NT dollar)

**BILL\_AMT4**: Amount of bill statement in June, 2005 (NT dollar)

**BILL\_AMT5**: Amount of bill statement in May, 2005 (NT dollar)

**BILL\_AMT6**: Amount of bill statement in April, 2005 (NT dollar)

**PAY\_AMT1**: Amount of previous payment in September, 2005 (NT dollar)

**PAY\_AMT2**: Amount of previous payment in August, 2005 (NT dollar)

**PAY\_AMT3**: Amount of previous payment in July, 2005 (NT dollar)

**PAY\_AMT4**: Amount of previous payment in June, 2005 (NT dollar)

**PAY\_AMT5**: Amount of previous payment in May, 2005 (NT dollar)

**PAY\_AMT6**: Amount of previous payment in April, 2005 (NT dollar)

**default.payment.next.month**: Default payment (1=yes, 0=no)

The features used in this dataset seem to be sufficient because all the information needed to predict whether a client is going to pay the default payment or not is available.

Approach and methodology

# Data Preprocessing

The dataset included information about the clients (Age, Education, and Marital Status) and banking information regarding the payment and billing amounts in the previous six months. After visualizing this data as follows

# Model Selection

The target variable is a categorical binary variable to be classified. Hence, Logistic Regression is chosen to be our model to fit in our data.

# Model Development and Evaluation