**Learner Report for Exploratory Data Analysis**

### Overall Impression

In this classification task, I first used Jupyter Notebook to import numpy, pandas, matplotlib and seaborn libraries. Then, I loaded preprocessed dataset of ‘DefaultofCreditCardClients.csv’, in which I updated to the actual month info based on the PDF explanation in the plan of attack. Further, I used some basic Python functions to check for the data types, null info and summary of dataset (count, mean, standard deviation, etc.) to get a big picture of this dataset. Different from previous tasks, I do not need to build any models in this task. However, I need to apply the data visualization tools in-depth in order to perform the Exploratory Data Analysis (EDA).

### Data Visualization

Prior to data discretization, I used seaborn to generate a few histogram, distribution chart and bar chart, to get a general idea of how the data look like at the first glance. As I can see, the majority of the clients are age between 25-40, with the peak around 28 years old. The default rate is 22%.

### Data Discretization

#### Step 1 – Categorize the Age Column

When exploring the entire dataset, I found that the original AGE column ranges from 21 to 79 years old. I think It is better to bin them into several category with 14-15 years old apart. Thus, I created 4 bins (20-35, 36-50, 51-65, 66-80) to break them up. Here I used pd.cut () function and relabel to [1,2,3,4].

#### Step 2 – Reorganize the Education Column

Initially we have 7 EDUCATION classes (from 0 to 6), in which 1 stands for graduate school, 2 stands for university, 3 stands for high school, and the rest of the values 0,4,5,6 all stand for others. In this case, I believe, it is better to rename 4,5,6 to 0 as well, so it is straightforward and easy to manage. Here I applied .replace() function to consolidate all others to 0. Doing so makes it easy for future data analysis.

### Factors Affecting Default

#### Demographic Factors

* Based on my analysis, males have 24% of default rate Vs. females have 21% default rate. Marital factor is close to even split on default among married and single. There is a significant trend of not default among singles.
* For those have graduate education level tend to have lower default rate than university and high school graduate.
* High school graduates seem has a flat line which shows no obvious relationship between age and default rate, but its default probability is the greatest among all four education groups.

#### Credit Limit Factors

Clients with lower credit limit are more likely to default.

### Did you learn anything of potential business value from this analysis?

Yes, of course! EDA is a wonderful tool that allows us to uncover patterns and insights with visual methods within data. EDA is often the first step of data modelling process. EDA is also more acceptable for business audiences as I can use these plots and charts in the presentation to enhance the future modelling results.

### What are the main lessons you've learned from this experience?

Data discretization on age and education is a necessary step, as it can generate less categorical data and make things easier to handle and to explain.

### What recommendations would you give to the Guido regarding your findings?

As I discussed earlier, people with higher education or higher credit limit, especially females, are less likely to default on their future credit card payment. Based on this finding, I would recommend Guido to pre-select the target group of people by screening their credit limits before issuing loans. Therefore, it can help to ensure that customers will have a higher probability to pay back their loans in the future.

In conclusion, we can approve customers with high certainty, eg: single females with higher education level. As this target group is unlikely to default on their loans.