Project: Capstone Project 1
Data Story

Approach 1

After we wrangled and cleaned the dataset, we started to explore the data in detail. The first step was to see the count and distributions of different variables from the dataset.

- How many cardholders are defaulters?
  We found that the 77.88% of the cardholders(23,364) did not default and 22.12% of the cardholders (6,636) default.
- How are the cardholders divided by gender?
  We have 60% female and 40% male clients.
- What are the education level, and Which education level does the most of the cardholders belong to?
  Most of our cardholder have University level education for their highest level of education. We have 35% with Graduate level education, 46% with University level, 16% with High School level, and 1.5% with Others as level of education.
- How many cardholders are married and how many are single?
  We have 53% married, 45% single, and rest as others.
- What age group is the majority of the cardholders?
  Most of our cardholders are of age group 20 to 40. This exploration can provide us with the demographic of different variables.

The main focus of this project is to create different Machine Learning Models to predict default, so let's find out some insights using different data visualization. We drew barplots to compare different variables with the default cardholders. From each plot, we learned which sub-variable effects default. We found that University level cardholders default more than other education level. From marital_status with default plot, we found that both married and single have very close number of default, and others have a very low default. Age group with default shows that age group 20 to 30 default more than any other age-group. We found that some cardholders have negative balance. This means that some cardholders are paying the bank more than their balance or some transaction of purchase may have been refunded to the credit card. When we compared balance column with paid column for the month of september, we saw that there is a

positive linear relationship between the balance and paid columns except for some balances. Looking at the plot, those balances may be paid by automatic payment every month.

From the analysis and visualizing the dataset, we found that the maximum number of defaulter are female, with University level education, and age between 20 and 30. When we compared male and female for each age group with default, we saw that more female of age group 20-30 default more, and both older male and older female after age group 50-80 default less. When we group gender and education column, we see that 2,2 or Female with University level education has the maximum default. When we compared gender, education and marital status with default, we found that a female, university level, and married cardholder has the maximum default.

Approach 2

After we wrangled and cleaned the dataset, we started to explore the data in detail. The first step was to see the count and distributions of different variables from the dataset.

- How many cardholders are defaulters?
  We found that about 72% of the cardholders(10,315) did not default and about 28% of the cardholders (4,042) default.
- How are the cardholders divided by gender?
  We have 60% female and 40% male clients.
- What are the education level, and Which education level does the most of the cardholders belong to?
  Most of our cardholder have University level education for their highest level of education. We have 30% with Graduate level education, 49% with University level, 19% with High School level, and 0.33% with Others as level of education.
- How many cardholders are married and how many are single?
  We have 53% married, 45% single, and rest as others.
- What age group is the majority of the cardholders?
  Most of our cardholders are of age group 20 to 40. This exploration can provide us with the demographic of different variables. Most of our clients are between age 20 and 30.

The main focus of this project is to create different Machine Learning Models to predict default, so let's find out some insights using different data visualization. We drew barplots to compare different variables with the default cardholders. From each plot, we learned which features or

variable effects default. When we compared different features with default, we found that the gender, age, education, and marital status has effect on cardholders default. We found that University level cardholders default more than other education level. From marital_status with default, we found that both married and single have very close number of default, and others have a very low default. Age group with default shows that age group 20 to 30 default more than any other age-group. We found that some cardholders have negative balance. This means that some cardholders are paying the bank more than their balance or some transaction of purchase may have been refunded to the credit card. When we compared balance column with paid column for the month of september, we saw that there is a positive linear relationship between the balance and paid columns except for some balances. Looking at the plot, those balances may be paid by automatic payment every month.

From the analysis and visualizing the dataset, we found that the maximum number of defaulter are female, and with University level education. When we compared male and female for each age group with default, we saw that female of age group 20-30 are the most default. When we group gender and education column, we see that 2,2 or Female with University level education default more than other group clients. When we compared gender, education and marital status with default, we found that group female, university level, and married cardholder had maximum default.

Approach 3

After we wrangled and cleaned the dataset, we started to explore the data in detail. The first step was to see the count and distributions of different variables from the dataset.

- How many cardholders are defaulters?
  We found that the 77.88% of the cardholders(23,364) did not default and 22.12% of the cardholders (6,636) default.
- How are the cardholders divided by gender?
  We have 60% female and 40% male clients.
- What are the education level, and Which education level does the most of the cardholders belong to?

Most of our cardholder have University level education for their highest level of education. We have 35% with Graduate level education, 48% with University level, 16% with High School level, and 0.4% with Others as level of education.

- How many cardholders are married and how many are single?
  We have 53% married, 45% single, and rest as others.
- What age group is the majority of the cardholders?
  Most of our cardholders are of age group 20 to 40. This exploration can provide us with the demographic of different variables.

The main focus of this project is to create different Machine Learning Models to predict default, so let's find out some insights using different data visualization.

We drew barplots to compare different variables with the default rate of total cardholders. From each plot, we learned which sub-variable has greater default. When we compared gender with default, we found that the female cardholders default more than male. We found that University level cardholders default more than other education level. From marital_status with default plot, we found that both married and single have very close rate of default, and others have a very low default rate. Age group with default shows that age group 20 to 30 default more than any other age-group. We found that some cardholders have negative balance. This means that some cardholders are paying the bank more than their balance or some transaction of purchase may have been refunded to the credit card. When we compared balance column with paid column for the month of september, we saw that there is a positive linear relationship between the balance and paid columns except for some balances. Looking at the plot, those balances may be paid by automatic payment every month.

From the analysis and visualizing the dataset, we found that the maximum number of defaulter are female, with University level education, and age 20-30. When we compared male and female for each age group with default, we saw that more female of age group 20-30 default more than male, and both older male and older female after age group 50-80 default less. When we group gender and education column, we see that 2,2 or Female with University level education default more than other group clients. When we compared gender, education and marital status with default, we found that group female, university level, and married cardholder had maximum default.