

**ANL252 (Online)**

**Python for Data Analytics**

# **End-of-Course Assessment**

**July 2022**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **Mako Wang Jun** | **Y2172686** |

**Tutorial Group: ­­­­­­­­­­­­ T 09**

**Instructor’s Name: Prof. Munish Kumar**

**Submission Date: 04/09/2022**

**1)**

The categorical variables are Rating, Gender, Educational and Marital.

The numeric variables are ID, Limit, Balance, Income, Age, S(n), B(n) and R(n).

**2)**

Firstly, we will replace blanks in the dataset to NaN so that the csv file is readable in python. We can do so by using df.replace(r'^\s+$', np.nan, regex=True).

Secondly, there are irrelevant data under the age column. An example is the age of -1 and 199 just from looking at the excel filter because it is not physically possible, and this may be due to system error during the input. Out of date or outlier entries may not be relevant for data analytics and could skew the results. Hence, we use interquartile range to filter out outliers in the dataset. The result removed the outliers from 18,769 rows to 18,596 rows to provide a more accurate representation of the majority

Thirdly, there are duplicate data in the dataset under the column ID. Therefore, we will proceed with using df.drop\_duplicates(subset=["ID"], keep="last", inplace=True). As a result, we have cleaned up the data set that are duplicated so as to provide a more accurate analysis.

Lastly, for the balance column of the dataset, there are inconsistent data in terms of decimal place, so we have converted the data to be consistent to ensure that it is analysed accurately. By using df = df.round({'BALANCE': 0}), we can standardised the data in a consistent manner and this makes the data more readable.

|  |  |
| --- | --- |
| **Customer total limit** | |
| Count | 18,593 |
| Mean | 167,901 |
| Standard Deviation | 129,168 |
| Minimum | 10,000 |
| 25% | 50,000 |
| 50% | 140,000 |
| 75% | 240,000 |
| Maximum | 800,000 |

**3)**

Chart, histogram

Description automatically generated

Firstly, from the chart created, we can see that the distribution is skewing towards the left, which shows that majority of the customers limit are around $100,000. Amount higher than $500,000 seemed to be on the lower side.

The mean customer total limit is $167,901. The mean value represents the average limit that are given to customers, and it is on the lower side of the scale. We can infer that it is a reasonable amount to provide customers with this total limit.

Chart, bar chart

Description automatically generated

Secondly, for gender of customers, 0 represents Male and 1 represents Female. From the chart, the customer’s demographics seems to be higher for females compared to males. Therefore, the majority of the customer base are females.

|  |  |
| --- | --- |
| **Education** | |
| Count | 18,580 |
| Mean | 2 |
| Standard Deviation | 1 |
| Minimum | 0 |
| 25% | 1 |
| 50% | 2 |
| 75% | 2 |
| Maximum | 3 |

Chart

Description automatically generated with low confidence

Thirdly, for highest education attained, (0 represents Others, 1 represents Postgraduate, 2 represents Tertiary and 3 represents High School), from the chart, we can see that majority of the customer have their highest education in Tertiary. Hence, we can say that majority tend to have a lower education and seems to take on credit. The minority are in Others education.

|  |  |
| --- | --- |
| **Marital** | |
| Count | 18,555 |
| Mean | 2 |
| Standard Deviation | 1 |
| Minimum | 0 |
| 25% | 1 |
| 50% | 2 |
| 75% | 2 |
| Maximum | 2 |

A picture containing chart

Description automatically generated

Fourthly, for Marital status, (0 represents Others, 1 represents Single and 2 represents Married). Based on the graph, we can see that majority of the customer base are Married and Single as they are around the same.

|  |  |
| --- | --- |
| **Age** | |
| Count | 18,593 |
| Mean | 35 |
| Standard Deviation | 9 |
| Minimum | 21 |
| 25% | 28 |
| 50% | 34 |
| 75% | 34 |
| Maximum | 60 |

Chart, histogram

Description automatically generated

Lastly, for customers age, the age ranges from 21 to 60 after removing the outliers based on interquartile range. From the mean of 35, the customers demographics lean towards middle aged people. The charts show a sharp increase which then follows by a slow drop in customers according to their age.

**4)**

I applied linear regression on the dataset so as to predict the variable B1. The mean squared error was used as the loss function and r2 against it. From there, I derived 0.79 from r2. The model was able to provide a 79% variance for the variable B1. The model applied was able to correctly represent and predict B1.

**5)**

B1 = 0.0379 \* LIMIT + 0.0488 \* BALANCE + 0.1879 \* INCOME + 0.0117 \* AGE + 0.0064 \* S(n) + 0.0023 \* R(n).

From there, we can see that credit limit and credit balance are positively correlated with billable amount. The factors in the equation represents a positive and upslope graph. Hence, we can ay that the limit, balance and income contributes to a positive relationship whereas age and repayment records do not play a significant role in the equation.