

**ANL 252 (Online)**

**PYTHON FOR DATA ANALYTICS**

# **Tutor-Marked Assignment**

**July 2022 Presentation**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **CHANG GUO RUI, JOEL** | **Q1910355** |

**Tutorial Group: ­­­­­­­­­­­­ T 09**

**Instructor’s Name: Mr. MUNISH KUMAR**

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

**Question 1**

|  |  |
| --- | --- |
| **CATEGORICAL** | **NUMERIC** |
| GENDER | LIMIT |
| EDUCATION | BALANCE |
| MARITAL | INCOME |
| RATING | AGE |
| S(n) | B(n) |
|  | R(n) |

**Question 2**

1. **Replacing NaN Values**

Replacing NaN values with the relevant data for the variables ‘EDUCATION’ and ‘MARITAL’ as these values have blank values in them.

- EDUCATION is replaced with a ‘mode’ function which replaces the ‘blanks’ with the highest value in the variable. It is more logical to use the mode function as compared to using mean or median.

- MARITAL blanks are replaced with the value of ‘0’ which signify as ‘Others’ under the Appendix.

1. **Removing inaccurate data (Outliers)**

There are some outlier data under the ‘AGE’ column. Data like -1 and 199 does not make any logic, a negative value in age does not make any sense and currently, there is no human that has a life span of up to 199 years of age. Using ‘min’ and ‘max’ functions for that specific column, it was determined that there was a total of 10 outliers within the data, 5 for -1 and 5 for 199. These outliers were removed from the data as there was no additional info to use that can use to replace these data. Removing them was the most recommended choice.

The data frame has been dropped from 18769 rows to 18759 rows

1. **Data cleaning for Column ‘R3’**

Inconsistencies are found is ‘R3’. Values like ‘$’ and ‘,’ are found. A loop was created to remove the inconsistencies and was replaced with the data frame.

1. **Converting data types**

The data types have been converted to ensure the correct type is used for each variable.

**Question 3**

1. **Education vs Income**

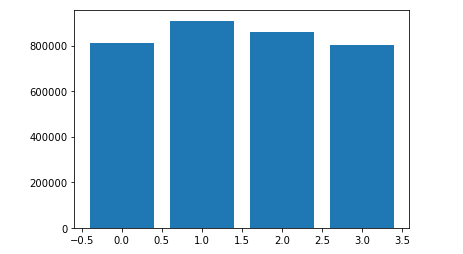


Figure 1 – Education Against Income

Based on appendix 1 under the variable of Education, 0 = Other, 1 = Postgraduate, 2 = Tertiary, 3 = High School. On figure 1, among the descriptions stated, the person with a postgraduate education has the highest paid income as compared to the rest. This is followed by tertiary, others, and then high school. Due to the lack of information, it is hard to determine ‘Others’ belong in which category.

1. **Gender vs Income / Gender vs Balance**

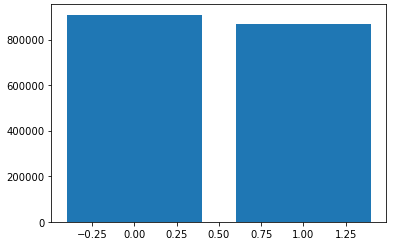


Figure 2 – Gender Against Income

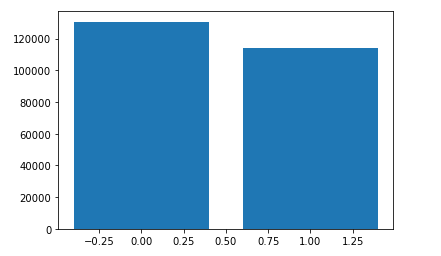


Figure 3 – Gender vs Balance

Using the data from the highest income of male and female, as seen from Figure 2, it shows that the male is earning slightly more than female. This is contributed to Figure 3 showing that male have higher credit balance compared to female.

1. **Count of people vs Binned Income**

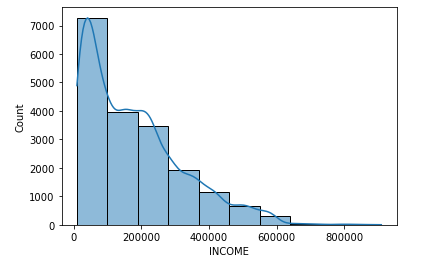


Figure 4 – Count vs Binned Income

Figure 4 shows the distribution of the number of customers belonging in a certain range of income. Majority of the customers have an income of 10,000 or more. Based on the data, the minimum income is at 10,000, which would mean that the data of anybody having an income of 9,999 and below is not captured. Which creates an insight of the credit facility only accepting customers with income of 10,000 and above.

1. **Heat map to determine correlation of variables**

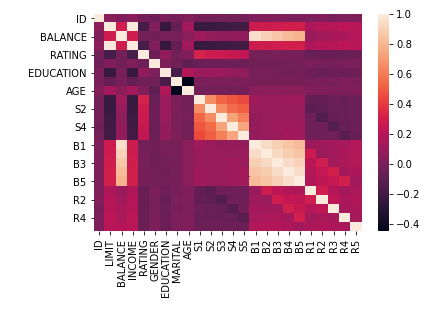


Figure 5 – Heat map to determine correlation of variables

Figure 5 shows a heat map to see how the variables are correlated. The lighter the box, the more correlation it has. S(n) repayment and B(n) billables are correlated. Like the B(n) billables and the customer’s credit balance. The correlation table shows how close together each variable is compared to other variables. This risk of multi-collinearity can happen and further data processing is needed.

1. **Education vs Income**

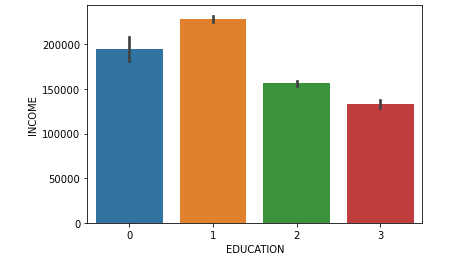


Figure 6 – Education vs Income

Figure 6 shows that the education of postgraduate earns the most while the education of high school earns the least. This shows us how education influences the income. This insight shows that the customers are living a modern society where the highly educated generally earns higher than the rest. Due to the lack of information given for the education of ‘others’ is it assumed that these customers might be entrepreneurs with education lower than high school. Or they might be customers who are unwilling to divulge their education information.

**Question 4**

As mentioned in figure 5, the risk of multi-collinearity can happen as there are independent variables with high correlation of more than 80%. Hence, we will be dropping the data of B2-B5 due to their high correlation with each other.

Data like ‘EDUCATION’, ‘GENDER’, ‘MARITAL’ will also be removed from the data frame as these data is not important towards predicting the customer’s billables from a credit facility perspective. Hence, there will be a total of 14 variables used to predict the result of B1.

The last 20 rows of the data were used as test data, training set is used to train the model, test set was used to validate model’s accuracy. Training set data was created and the result for training data was produced. Coefficient of determination was printed to explain the variability of the model. Intercept of the model was printed to determine the ‘y-intercept’ if all the variable data is 0. For each change in B1, the 14 variables will shift accordingly and the change in dependent variable varies at the co-efficient.

The trained model was used to predict the results of B1. A model of the predicted and actual results was produced to validate the accuracy of the last 20 rows of data form the data sheet.

The model is able to explain the 93% of the data variability, this is safe to assume that the model is doing a good job in prediction.

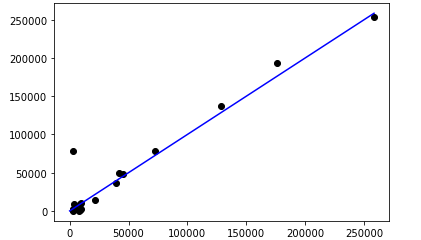


Figure 7 – Plot of linear regression model

**Question 5**

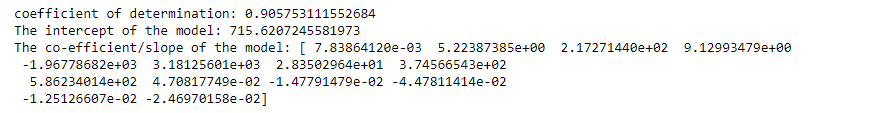


Figure 8 – Co-efficient/slope of model

Linear Regression Equation =

y = [(7.83864120e-03)(0.905753111552684)] + 715.6207245581973 ... ... [(-2.46970158e-02)(0.905753111552684)] + 715.6207245581973

Figure 7 shows the accuracy of the linear regression plot. Although there are some outlier data found, the model is still able to represent a straight line between the data points. This shows the model did well in terms of predicting the strong correlation between the prediction and testing data from Question 4.

**Jupyter Notebook Codes in Text Format**

import pandas as pd

import numpy as np

from sklearn.linear\_model import LinearRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn import preprocessing

from matplotlib import pyplot as plt

import seaborn as sns

from sklearn.metrics import mean\_squared\_error, r2\_score

df = pd.read\_csv('ECA\_data.csv')

df

# Question 2

# Data Pre-processing (1) Replacing NAN values

# We are able to see the NAN values in edu and martial therefore it's in our interest

# to fill the NA data with relevant data, we chose 0 because in the data dictionary

# it stated that 0 is for the Others type, which is what we will be using.

print(df.isnull().sum())

updated\_df = df

updated\_df['EDUCATION']=updated\_df['EDUCATION'].fillna(updated\_df['EDUCATION'].mode()[0])

updated\_df['MARITAL']=updated\_df['MARITAL'].fillna(0) # df.fillna(0)

updated\_df.info()

# Date Pre-processing (2) Removing inaccurate data

# There are inaccurate data (outliers) inputs in certain columns and

# therefore it is necessary to remove those.

min\_age = updated\_df['AGE'].min()

max\_age = updated\_df['AGE'].max()

print(updated\_df.loc[df['AGE'] == min\_age])

print(updated\_df.loc[df['AGE'] == max\_age])

# Drop rows with AGE = -1

updated\_df.drop([664], inplace=True)

updated\_df.drop([4242], inplace=True)

updated\_df.drop([15358], inplace=True)

updated\_df.drop([18165], inplace=True)

updated\_df.drop([18166], inplace=True)

# Drop rows with AGE = 199

updated\_df.drop([535], inplace=True)

updated\_df.drop([1617], inplace=True)

updated\_df.drop([7833], inplace=True)

updated\_df.drop([18082], inplace=True)

updated\_df.drop([18282], inplace=True)

# Check if the rows with inaccurate data have been dropped

updated\_df

# Data Pre-processing (3) Data Cleaning for Column R3

# In Column R3, some of the values have '$' and ',', which is inconsistent with the rest of the data

# Loop through the column and perform string manipulation to remove the inconsistencies

# Replace these new values in the df column

R3 = updated\_df['R3']

for row in R3:

if row[0] == '$':

new\_row = row[1:]

new\_row = new\_row.split(",")

new\_row = ''.join(new\_row)

R3.replace(row, new\_row, inplace=True)

# Data Pre-processing (4) Converting data types

df['R3'] = df['R3'].astype(int)

df['LIMIT'] = df['LIMIT'].astype(float)

df['INCOME'] = df['INCOME'].astype(float)

df['EDUCATION'] = df['EDUCATION'].astype(int)

df['MARITAL'] = df['MARITAL'].astype(int)

df.info()

# Question 3

updated\_df.head()

# Graph 1

# We are able to see the less correlation of the education against the income.

x = updated\_df['EDUCATION']

y = updated\_df['INCOME']

plt.bar(x, y)

plt.show()

# Graph 2

#We are able to see that males are earning slightly more than the females

# this can be contributed the higher credit balance that the males have compared to females

print("Gender vs Income")

x = updated\_df['GENDER']

y = updated\_df['INCOME']

plt.bar(x, y)

plt.show()

print("Gender vs Balance")

x1 = updated\_df['GENDER']

y1 = updated\_df['BALANCE']

plt.bar(x1, y1)

plt.show()

# Graph 3

#We are able to see the distribution of income.

#it follows a skewed distribution where most people are earning lesser

#and a few with high income.

sns.histplot(x=updated\_df['INCOME'], bins=10, kde=True)

#Graph 4

#Heat map to see how the variables are correlated

#We are able to see that Billable and Repayment are correlated

#We can also see that balance has higher correlation with B(n) variables

#multi-collinearity can happen and this needs further data processing

sns.heatmap(round(updated\_df.corr(),2))

# Correlation Table - same as above, but with the values

corr=updated\_df.corr()

corr.style.background\_gradient(cmap='coolwarm')

#Graph 5

#We are able to see how education has an effect on the income

#this shows that postgraudate generally earn the most while

#high school education earn the least

sns.barplot(x="EDUCATION", y="INCOME", data=updated\_df);

# Question 4

#As stated in visualisation portion the risk of multi-collinearity can happen

# As there are indpendent variables with high correlation (more than 80%)

#Hence we will be dropping B2-B5 as they have high correlation between each other

#we also dropped income as it is highly correlated with limit

# we also dropped these columns: education, gender, marital -> explain from bank perspective

X = updated\_df[['LIMIT', 'BALANCE', 'RATING', 'AGE', 'S1', 'S2', 'S3', 'S4', 'S5', 'R1', 'R2', 'R3', 'R4', 'R5']]

y = updated\_df['B1']

corr=X.corr()

corr.style.background\_gradient(cmap='coolwarm')

# Split the data into training/testing sets

# Use the last 20 rows as test data

# Training set is use to train the model, test set is use to validate model's accuracy

X\_train = X[:-20]

X\_test = X[-20:]

# Split the targets into training/testing sets

y\_train = y[:-20]

y\_test = y[-20:]

# To train the data with training set

model = LinearRegression()

model.fit(X\_train, y\_train)

#R\*\*2 Result for Training Data

r\_sq = model.score(X\_train, y\_train)

print(f"coefficient of determination: {r\_sq}")

#We are able to see that this model can explain 90% of variability of the model

#To retrieve the intercept:

print(f"The intercept of the model: {model.intercept\_}")

#this means that if all variable are 0, the y is at 20564

#For retrieving the slope:

print(f"The co-efficient/slope of the model: {model.coef\_}")

# this means that for each unit of change B1

# the 14 variables used to predict will have to shift accordingly

# the change in dependent variable varies at {coef}%

# using the trained model to predict B1

y\_pred = model.predict(X\_test)

#result: predicted vs actual

# predicted values: generated by the model

# actual: the last 20 rows of the data - used to validate the accuracy

df = pd.DataFrame({'Actual': y\_test, 'Predicted': y\_pred})

df

#The model is able to explain 93% of the data variability

# this is safe to assume that model is doing a good job in prediction

print("Coefficient of determination: %.2f" % r2\_score(y\_test, y\_pred))

# # Plot outputs

# Predicted vs Actual

# We can see that the model is doing very well in terms of predicting

#As there is a very strong correlation between pred and test

plt.scatter(y\_pred, y\_test, color="black")

p1 = max(max(y\_pred), max(y\_test))

p2 = min(min(y\_pred), min(y\_test))

plt.plot([p1, p2], [p1, p2], 'b-')

plt.show()

#For residual plot shows how much variability between the actual data and the predicted

# as residual appears to be near 0 and this also proves that prediction are correct

residual = y\_test - y\_pred

sns.regplot(y = residual, x = y\_pred, data = None, scatter = True, color = 'red')