**Milestone 1**

**Financial Institute data**

**Data description:**

1. **Find the attached data file in milestone 1 assignment folder.**
2. **Read the following detailed data description of Financial Institute. This is a part of real time current car loan data from one of the leading Texas financial institute.**

**Following are the variable details of the dataset.**

* Financial data collected between JUN-AUG 2022
* BranchID: Identification Number of Financial Institution’s branch office.
* Application Date: The date when an applicant applied for the Loan.
* Loan Sub-Type: This is a subcategory of the Loan type that are – “auto loan finance”, “new boat/ RV loan”, “new vehicle loan”, “other secured loan”, “used motorcycle loan”, “used vehicle loan”
* Loan Purpose: The main category of Loan that is “Vehicle” Loan category.
* Loan Type: This is a main Loan type categories – “bills”, “dealer purchase”, “lease buyout”, “private party purchase”, “purchase other”, “refinance”
* Requested Loan Amount: The amount in hundreds/thousands that has been requested by a Borrower.
* Approved Loan Amount: The amount in hundreds/thousands that has been approved by a financial institution.
* Funded Loan Amount: The amount in hundreds/thousands that has been funded (deposited in the borrower’s account) by a financial institution.
* Loan Status: This is the final status of loan applications in terms of- “approved”, “declined”.
* Declined Date: This is the day when the loan application is declined.
* Deciding Date: This is the day when the financial institute has taken decision on the application.
* Approved Date: This is the day when the financial institute has approved the application.
* Funded Date: This is the day when the financial institute has approved and deposited amount.
* Borrower Type: P and C type of borrower.
* Applicant State: This is the state where applicant have applied for the loan.
* Credit Score: The credit bureau score of a loan applicant.
* Type of Vehicle: ATV, BIGRIG, CAR, BOAT, MOTERCYCLE, OTHER, RV and blanks.
* Value of Asset: Exact assessed value of the vehicle in thousands.
* LTV Ratio: A loan-to-value ratio (LTV) is the total dollar value of your loan divided by the actual cash value (ACV) of your vehicle. It is 0 in the data for all records.
* Loan Class: 3 YR USED, 6-10 YR USED, 2 YR USED, 4 YR USED, 4 YR USED , 1 YR USED/REFI 60 MO, 60 MO NEW, 1 YR USED/REFI 72 MO, 72 MO New, TOKEN COLLETRAL.
* Loan Tier: 1.tov-vehicle T1, 1.tov-Vehicle T2, 1.tov-Vehicle T31.tov-Vehicle T4, 6.tov-Vehicle T1, 6.tov-Vehicle T2, 6.tov-Vehicle T3, 6.tov-Vehicle T4
* Loan Term: Duration of a vehicle loan in the months.
* Monthly Income: per month income of the borrower.
* Monthly debt: per month debt of the borrower.
* DTI Ratio: Your debt-to-income ratio (DTI) is all your monthly debt payments divided by your gross monthly income. This number is one-way lenders measure your ability to manage the monthly payments to repay the money you plan to borrow.
* Monthly Loan Payment: Loan payment installments towards the financial institute paid by borrower.
* Age of Employment in Months: borrower work experience in months.
* Age of Employment in Years: borrower work experience in years.
* Denial Reasons: Various denial reasons to disprove/approve the loans.
* Total you will have 1012 records (i.e., 1013 total excel rows) Where **Loan Status is the dependent variable for the Classification task**. All other are the independent variables.

1. **Run the Code for following tasks in Jupyter Notebook and produce the outputs wherever necessary. Answer all questions in your ipynb python jupyter file itself (you can use markdown cells as well wherever necessary) and submit the pdf file of your code run with results. (Failing to not submit the pdf file will result in 100% deduction in marks).**

***Step 1 :-*** *# importing all neccesory libraries*

import numpy as np

import pandas as pd

***Step 2*** *# import the dataset ->* ***here you have to write the code to import*** ('Finance\_data.csv')

***Step 3*** *#wrtie the code for viewing the first few rows of the dataset to see our imported data.*

**Step 4**# now lets **write the code** to eliminate some unwanted variables. So drop these listed variables- 'Loan\_Purpose','BranchID' ,'Application\_Date', 'Declined Date','Deciding Date',

'Approved Date', 'Funded Date','Denial Reasons'

***Step 5*** *# viewing statistical information about dataset*

**here you have to first write the code and then write a short summary that what have you understand about the generated results after running this code.**

**Ans:-**

***Step 6*** *# checking the number of missing data –* ***Write the code to see information related to missing values in data.***

*Now as we can see after running the code that we have missing values in ‘Loan Tier’, ‘Funded Loan Amount’ and ‘Credit Score’ variables. Here Loan Tier is categorical variable and other two has numeric values. So we need different treatment for both type of variables.*

***Step 7*** *#For the columns ‘Loan Tier’ we remove the rows with missing values.*

*# Dropping categorical data rows with missing values – write a code.*

***Step 8*** *#For the columns ‘Funded Loan Amount’ and 'Credit Score' we place 0 at null. –* ***here writes a code.***

***Step 9#*** *write a code to check again the number of missing data after running above code. Here you will not see any missing data.*

*#we have variety of variables (categorical and numerical) so we have to make the entire data in same format and that is numeric one. For that we have perform label encoding for all non-numeric or categorical variables.*

*We have 8 categorical variables- 'Loan\_SubType', 'Loan\_Type', 'Loan\_Status', Borrower\_Type', 'Applicant\_State', 'Type\_ofVehicle', 'Loan\_Class', 'Loan\_Tier'*

**Step 10 #Now run following code or any code of your choice to perform label encoding. Here I did for only one variable and you have to repeat it for other seven variables.**

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

dataset['Loan\_SubType\_Cat']=le.fit\_transform(dataset['Loan\_SubType'])

**Step 11***# Like above step 10 do the same for all other 7 categorical columns.*

*# to see the category name and assigned code run the following code. (****optional****)*

print(dataset[['Loan\_SubType\_Cat','Loan\_SubType']])

**Step 12**# Finally save your preprocessed data into new csv file as shown below. **Here you have to fill the blank for method name.**

dataset.------------- ('new\_finance.csv')

Now go where the 'new\_finance.csv' file has been stored in your system and open it in excel and list here all eight categorical variable’s values with assigned code. For example: the “1.tov-Vehicle T3” category of Loan\_Tier Variable get the value “2” as a code. Similarly, “1.tov-Vehicle T2” has get the “1” as a code.

Loan\_Tier\_Cat Loan\_Tier

2 1.tov-Vehicle T3

1 1.tov-Vehicle T2

**Now Tell me, in new csv file how many records do you have (e.g., 2000 or 1000 or 500 ..) and for next milestones we will use this new csv file.**

**Ans:**

**----------------\*\*\*\*\*\*\*\*\*\*\*----------------**