Machine Learning:

About the Data:

The dataset consists of all the basic bank details of the customers, and the dataset has categorical and numerical variables. It has missing values and more anomalies. It consists of the customer's bank and processional details over the years so we can explore more of it. You have given the two data sets namely Train and Test datasets. Kindly use the TRAIN dataset to build and test the model. And use the Test dataset to submit your results.

SUBMISSION INSTRUCTIONS:

1. Predict the labels for the test dataset. And combine the same with ID.
2. Save it as a data frame.
3. And Upload the same predicted file with the solution file in a zip folder.
4. Do not drop any rows, The number of rows should be matched with the sample output that has attached below.

Sample output:

Sample_Output.csv

Attributes:

1. ID - Represents a unique identification of an entry.
2. CUSTOMER ID - This represents the unique identification of a person.
3. MONTH - Represents the month of the year.
4. NAME - Represents the name of a person.
5. AGE - Represents the age of the person.
6. SSN - Represents the social security number of the person.
7. OCCUPATION - Represents the occupation of the person.
8. ANNUAL INCOME - Represents the yearly income of the person.
9. MONTHLY IN-HAND SALARY - Represents the monthly base salary of a person.
10. NUM BANK ACCOUNTS - This represents the number of bank accounts a person holds.

11. NUM CREDIT CARD - This represents the number of other credit cards held by the person.
12. INTEREST Rate - This represents the interest rate on a credit card.
13. NUM OF LOAN - Represents the number of loans taken from the bank.
14. TYPE OF LOAN - Represents the type of loan taken by the person.
15. DELAY FROM DUE DATE - Represents the average number of days delayed from the payment date.
16. NUM OF DELAYED PAYMENT - Represents the average number of payments delayed by a person.
17. CHANGED CREDIT LIMIT - This represents the percentage change in the credit card limit.
18. NUM CREDIT INQUIRIES - Represents the number of credit card inquiries.
19. CREDIT MIX - This represents the classification of the mix of credits.
20. OUTSTANDING DEBT - This represents the remaining debt to be paid(in USD).
21. CREDIT UTILIZATION RATIO - This represents the utilization ratio of credit cards.
22. CREDIT HISTORY AGE - This represents the age of the credit history of the person.
23. PAYMENT OF MIN AMOUNT - Represents whether only the minimum amount was paid by the person.
24. TOTAL EMI PER MONTH - Represents the monthly EMI payments(in USD).
25. AMOUNT INVESTED MONTHLY - Represents the monthly amount invested by the customer(in USD)
26. PAYMENT BEHAVIOUR - Represents the payment behavior of the customer (in USD)
27. MONTHLY BALANCE - Represents the monthly amount of the customer (in USD).

Section A: ***Title: Classification model***

Problem Statement:

You are working as a data scientist in a global finance company. Over the years, the company has collected basic bank details and gathered a lot of credit-related information. The management wants to build an intelligent system to segregate the people into credit score brackets to reduce manual efforts.

Objective:

The objective of this project is to explore the data to identify the pattern that causes the person to have a good or bad or standard type credit score and build a machine learning model that should be able to predict or classify the credit score type.

Steps that are to be followed:

Step 1: Understand the business problem.

Step 2: Import all the libraries and set up all the requirements that you will be needed(optional).

Step 3: Read the train and test data sets, and check for the datatypes.

Step 4: Fix the problem that the features have been wrongly identified.

Note: One way to clean the training and testing data is to combine both train and test datasets. Then do the cleaning.

i) Clean the anomalies in the categorical variables. A few anomalies have been mentioned here.

1. Occupation - _____
2. SSN - #F%$D@*&8
3. Payment Behaviour - !@9#%8

Replace the above anomalies by replacing them with the mode of each customer.

ii) Clean the anomalies for numerical variables.

Ex:

1. Age has above 8000 values therefore replace the values that are above 100 or 85 with median values.

Note: I encourage you to replace the abnormal values with customer ID-wise median replacement for the customers who have above and below abnormal values. (customer-wise median means the median value for each customer. Example: customer aaa Annual Income has anomaly value in one row so replace that with customer aaa Annual Income median value.)

2. Few columns have - negative values but they are not supposed to have negative values. So replace the negative values with the median by doing a customer-wise median.

3. Go through each variable and find out the other mistakes. And handle them.

Step 6: Convert the Credit_History_Age datatype variable into float data types by taking only year and month. Example. 22 years and 1 month → 22.1. And the Payment_of_Min_Amount column you might find some other weird values apart from Yes and No. And If you have combined the train and test datasets, then change the month's names into its number.

Step 7: Find out the missing values in the data frame and handle them in the best way possible. One way of solving this is by imputing the missing values with a customer-wise median.

Step 8: Perform Univariate, Bivariate, and Multivariate analyses to find the factors that affect the Target variables.

Step 9: Separate your Train dataset and test data set if you combined them in the initial steps. (In this step only segregate train and test datasets based on the length of the train and test dataset)

Note: Your given test dataset is only for validating and submitting the results. Only Use the Train dataset to perform the train test split in the coming steps. Do not use Test to build the model and test the model, Since there is no target variable in the test data set you can not test the model performance with the test data set. Thus we only consider the training dataset and split that into X_train and X_test.

Step 11: Perform the Statistical analysis to prove where the independent variables have an effect on the Target variables.

Example: Few statistical analyses:

1. Check whether the Annual income across all the target variables is significantly the same. Let's fix the alpha is 0.05. Make sure the data is normal and the variance is equal. If not use a Non-parametric statistical test.
2. Check if there is an independence of the Occupation and Credit Score. The significant level is 0.05.
3. Check if there is a relationship between the Payment Behaviour and Credit Score.
4. Check Statistically that the Credit_Utilization_Ratio median values are significantly not different across the target variable classes.

Step 12: Encode the categorical variables with related technologies. Change the target variable classes as (poor to 0, Standard to 1, Good to 2).

Step 13: Scale the numerical features (optional).

Step 14: Use a train test split on the dataset called a train.

Step 15: Build the base model. Observe how the model is performing.

Step 16: Build other models and choose the model which gives the best results.

Step 17: Perform Feature selection using different feature selection methods.

Step 19: Tune the final model using Grid search CV or Randomized CV or any other methods.

Step 20: Perform Cross-validation for the final model by setting the best parameters.

Step 21: Use the validation data set called test dataset to get predict the target variables once it is done create a data frame with ID and predicted values.

Step 22: Write down the overall business insight.

Step 23: Save the dataset that you created in step 22.

Step 24. Submit the solution file and output dataset.