Group Name: Tech Geeks
Name: Dhanashri Jagadale
Email: dhanashri.jagadale1998@gmail.com
College: Munster Technological University
Specialization: Data Science

Problem Description: The problem is to develop a predictive model that can assess the credit worthiness of potential future customers of a financial institution. The available data set consists of 807 past loan customer cases, each with 14 attributes including financial standing, reason for the loan, employment, demographic information, foreign national status, years of residence in the district, and the outcome/label variable Credit Standing, which classifies each case as either a good loan or bad loan. The objective is to build a model that accurately predicts the credit standing of new loan applications, using the available data as the training set. The model should be able to identify the key factors that determine creditworthiness and provide insights to help the financial institution make better lending decisions.

Data cleansing and transformation done on the data:

1. Handling missing values
   As the data contains missing values, we need to handle them as well. We can either remove the missing values or fill them with some values like the mean or median of the feature. The number of missing values is very small and there is no meaningful pattern to their occurrence therefore, Deleting missing values can be a reasonable approach. I have used dropna() function to remove any missing values before calculating the skewness. This is important because missing values can cause errors in the skewness calculation.
2. Handling skewness
   Months.since.Checking.Acct.opened:Since the skewness value is greater than 1, this means the distribution is highly skewed to the right. In this case, we can apply a log transformation to reduce the skewness. After applying the log transformation, we can check the skewness value again to see if it has reduced.
   Residence.Time.In.current.district: Since the skewness value is close to 0, this means the distribution is approximately symmetric. In this case, we don't need to apply any transformation.
   Age: Since the skewness value is close to 1, this means the distribution is slightly skewed to the right. In this case, we can apply a log transformation to reduce the skewness. The log transformation can be applied.

3. Handling outliers

   Outliers can have a significant impact on data analysis, so it's important to handle them appropriately. I have dropped the rows with outliers using the drop method with the index of the rows containing outliers
4. Next I have used the factorize function to factorize the categorical data to accurately analyze it further without any errors.
5. Changed the datatype of all numerical data to integer to avoid inconsistency in the data.
6. Renamed column name with appropriate naming conventions.
7. To build the best machine learning model for this dataset, it is important to first preprocess the data by encoding the categorical variables and scaling the numerical variables. I have done this using techniques such as one-hot encoding for categorical variables and standard scaling for numerical variables.