**Project Name - Credit Card Default Prediction**

##### **Project Type - Classification**

##### **Contribution - Team**

##### **Team Member 1 –Aditya**

##### **Team Member 2 –Swapnil**

# ****Project Summary****

# ****Objective****

# **To predict whether a customer in future will default in credit card payment or not. If a person does not default than we can increase his/her credit card limit and provide better offers.**

# ****Data Size** (30001, 25)**

# ****Data Preprocessing****

# ****Data Cleaning** We will go through each and every column and analyse each column**

# **Remove first row as it is not correct column names**

# ****1) Id** –**

# **Unique for each customer so no interpretation and is sequence hence drop the column**

# ****2) Preprocessing of**** LIMIT\_BAL ****column–****

* No Missing/Null Values.
* We have to change the column type from object to numeric.
* By analyzing the column graphically and analyzing its **Distribution Plot** we can clearly see that the column is **+ve skewed**, So we will transform the column into square root and create a new column named **LIMIT\_BAL\_sqrt**. Now, we will plot **boxplot** and **replace the outliers** with their nearest limit (i.e upper or lower).

**3) Preprocessing of SEX, MARRIAGE, EDUCATION & PAY-**

* By analyzing SEX, MARRIAGE, and EDUCATION & PAY column for now we can clearly see that is is a categorical column so we will do encoding of the gender column during transformation.
* We will do **Category Aggregation** in **MARRIAGE, EDUCATION & PAY** to **reduce noise**, **Handle sparsity** and **improving model generalization.**

### **4) Preprocessing for Billing and Payment Columns.**

* **Data Quality Checks**:
* All columns are verified to have no missing or null values.
* **Type Conversion**:
* Each column's data type will be converted from object to numeric for appropriate analysis.
* **Transformation**:
* To address the positive skewness and the presence of negative values in each column, a cube root transformation will be applied. This will create new columns (e.g., BILL\_AMT1\_cbrt, BILL\_AMT2\_cbrt, etc.).
* **Outlier Detection and Treatment**:
* A boxplot will be generated for each transformed column to identify outliers.
* Outliers will be replaced with their nearest limits (upper or lower) based on the boxplot analysis.
* **Binning**:
* To reduce the number of unique values in the continuous variables, each transformed column will be binned into discrete categories, resulting in new columns (e.g., BILL\_AMT1\_cbrt\_binned, BILL\_AMT2\_cbrt\_binned, etc.).

**5) Preprocessing of AGE Column-**

* All columns are verified to have no missing or null values.
* Each column's data type will be converted from object to numeric for appropriate analysis.
* To address the positive skewness values in each column, a square root transformation will be applied. This will create new columns (e.g., AGE\_sqrt.).

**Dropping columns:**

('LIMIT\_BAL','AGE','BILL\_AMT1', 'BILL\_AMT2','BILL\_AMT3', 'BILL\_AMT4', 'BILL\_AMT5', 'BILL\_AMT6', 'PAY\_AMT1', 'PAY\_AMT2', 'PAY\_AMT3', 'PAY\_AMT4', 'PAY\_AMT5', 'PAY\_AMT6', 'BILL\_AMT1\_cbrt','BILL\_AMT2\_cbrt','BILL\_AMT3\_cbrt','BILL\_AMT4\_cbrt', 'BILL\_AMT5\_cbrt','BILL\_AMT6\_cbrt','PAY\_AMT1\_cbrt','PAY\_AMT2\_cbrt', 'PAY\_AMT3\_cbrt','PAY\_AMT4\_cbrt','PAY\_AMT5\_cbrt','PAY\_AMT6\_cbrt'])

The above columns being dropped are either deemed unnecessary for the analysis or may introduce multicollinearity, which could negatively impact model performance. Specifically, columns like LIMIT\_BAL and AGE might have been transformed or derived into other features that are more suitable for modeling.

**Type Conversions:**

Here, specific columns are being converted to integer data types. This step is important for ensuring that the data is in a suitable format for machine learning algorithms, which often require numerical input. Columns related to payment status (PAY\_0 to PAY\_6) and the target variable (default payment next month) are explicitly converted to integers to facilitate model training.