**Inference from Data Analysis of Loan Defaulter Dataset**

The dataset consists of **233153** rows and **23** columns with **Loan\_Default** as the target variable with values 0 or 1.

Columns taken as string datatype:

* Branch\_Id
* Supplier\_Id
* Manufacturer\_Id
* Employment\_Type
* State\_Id
* Unique\_Id

Columns taken as bool datatype:

* Mobileno\_Avl\_Flag (Redundant Value, Has the same value throughout the dataset)
* Aadhar\_Flag
* Pan\_Flag
* Voterid\_Flag
* Driving\_Flag
* Passport\_Flag
* Loan\_Default (Target Variable)

Columns taken as Integer datatype:

* Disbursed\_Amount
* Asset\_Cost
* Ltv
* Age (Derived from Date of Birth)
* New\_Accts\_In\_Last\_Six\_Months
* Delinquent\_Accts\_In\_Last\_Six\_Months
* Average\_Acct\_Age
* Credit\_History\_Length
* No\_Of\_Inquiries

**Distribution of Defaulters and Non-Defaulters (Target Variable):**

Non-Defaulter contributes to **78%** and Defaulter contributes to **22%**

1. **Branch Id**

There are **81** Unique Branches in the Dataset and out of those 81 branches**, the top 5 busiest branches in terms of Number of Customers are 2,67,3,5,36. But More than Count of Customers, Defaulter Percentage plays a major role and by that aspect 251,254,97,36,78** tops the list, andthese are the branches which will have the chances of defaulting customers more.

1. **Employment\_Type**

Has two Unique Values, ‘**Salaried’** and **‘Self Employed’** and null values was replaced by modal value. **Self Employed contributes to 22.7%** to **Defaulters** and **Salaried contributes to 20.3%**

1. **Manufacturer Id**

Manufacturer Id’s **86,45,51** tops the customer count but Manufacturer Id **48** tops the **defaulter percentage** and rest of the manufacturer id has more or less the mean defaulter percentage of the dataset.

1. **State Id**

State Id has 22 Unique values with State Id’s 4,3,6 tops the customer count and states such as 17,16,19 has minimum customers.

Highest defaulting % states are (which are above the mean default% (21.7) of the dataset)

State ID 13 - 30.7%

State ID 14 - 27.6%

State ID 2 - 27.1%

State ID 12 - 26.6%

State ID 17 - 24.6%

1. **Supplier Id**

Supplier Id has **2953** unique values and hence we have taken Supplier Id with at least 100 Customers for analysis, and this brings the unique value counts to **634.**

The top 5 Default Suppliers are **22994,23150,22127,17436,21242** are having more than double the **mean defaulter percentage** and these high defaulting suppliers needs to be analyzed by the management

1. **Bool Variables / Id Submitted:**

**Aadhar flag** is the ID submitted by **most customers** and **Passport** is the least submitted but passport has the **lowest chance of defaulting.**

This clearly shows Customer who submits passport as ID tend to be Non-Defaulters and Voter ID has the max chances of being a defaulter.

**Highest Defaulting Percentage of each ID:**

* Voter Id – 26%
* PAN Card – 22%
* Aadhar – 21%
* Driving – 20%
* Passport- 15%

We also find when a customer submits more than One ID’s, chances of defaulting decreases and it is minimum when a customer submits three Id’s.

Defaulter Percentage for One ID: 21.7%

Defaulter Percentage for Two ID: 21.5%

Defaulter Percentage for Three ID: 16.6%

**Numeric Variables:**

**Columns with Positive Pearson Co-Efficient Values**

1. **Disbursed Amount (Highly positively Skewed Data (4.5) and hence needs to be log transformed for model analysis)**

The **Pearson coefficient value is 0.07**, which means it has a positive co-relation with the target variable. More the Amount disbursed, the chances of defaulting is more and hence when a customer comes in with a request of loan with high amount to disburse, proper checks needs to be done.

1. **Asset\_Cost (Highly positively Skewed Data (6.1) and hence needs to be log transformed for model analysis)**

The **Pearson coefficient value is 0.01**, which means it has a positive co-relation with the target variable and this value is very negligible.

1. **LTV (Loan to Value):**

The **Pearson coefficient value is 0.09**, which means it has a positive co-relation with the target variable and this value is considerably high when compared to the other Pearson coefficient values in the dataset and hence will have a strong impact in the analysis.

The management can lower the LTV for all the customers to a certain extent to reduce losses or can scrutinize customers with higher LTV by making more checks.

1. **No\_Of\_Inquiries (0.04 corr coefficient):**

If a customer tends to inquire more about his/her loan, chances of defaulting is more and hence the management can keep count of the number of times a customer has inquired and can make more checks for such customers or can lower the disbursed amount.

**Columns with Negative Pearson Co-Efficient Values**

1. **Average\_Acct\_Age - corr coefficient ( -0.02 negative co relation):**
   1. Higher the loan tenure time, more is the chances of not defaulting for the customers. So, more the length of the loan, it is safe to say chances of defaulting is less. Customers who would want to default, will go for a lesser loan tenure time i.e., short loan and **hence management needs to be careful with customers who opt for lesser loan tenure and impose more scrutiny for the same.**
2. **Credit\_History\_Length - corr(-0.04 negative co relation) :**
   1. If a customer stays with the bank for **more time**, i.e. higher credit history length, the **chances of defaulting becomes less** and out of all the columns Credit\_History\_Length has the maximum negative Pearson co-efficient value.
3. **Age (-0.03 negative co-relation):**
   1. This relation tells us **youngsters tend to default more** with an age band of **26-34** and hence additional checks needs to be deployed by the management for younger customers seeking loan.
4. **New\_Accts\_In\_Last\_Six\_Months (-0.02** **negative co relation):**
   1. Very negligible value and can be ignored but can be combined with other columns to see the spread of defaulters.

**Linear Regression Model Explanation**

**Before proceeding with Linear Regression Model, following things are carried out:**

1. Highly Skewed Columns are transformed by using appropriate techniques

Columns such as:

* **No\_Of\_Inquiries**
* **Delinquent\_Accts\_In\_Last\_Six\_Months**
* **Asset\_Cost**
* **Disbursed\_Amount**

have very high skew value i.e., highly positively skewed, log transformation or square root transformation to reduce the skewness depending on the data.

1. All Categorical columns are encoded accordingly as shown:

* **Branch\_Id** – BinaryEncoding (due to high number of Unique Values)
* **Supplier\_Id** – BinaryEncoding (due to high number of Unique Values)
* **Manufacturer\_Id** - one hot encoding
* **Employment\_Type** - one hot encoding
* **State\_ID** - one hot encoding

**Model Building:**

* **Basemodel –** Consists of all the columns

**Cons:** Confusion matrix doesn’t predict values for Loan\_default = 1

* **model1 –** Consists of all the columns and **model is trained and tested by passing same ratio for Defaulters and Non-Defaulters in both train and test data by using the parameter stratify**

**Cons**: Still Confusion matrix doesn’t predict values for Loan\_default = 1 properly

* **model2– Undersampling the Dataset**

Result: Confusion matrix predict values for Loan\_default = 1 and hence can be taken as the final model