1. Dataset **main\_loan\_base.csv**

The main\_loan\_base data set contains information about 20,000 loan accounts with a fictitious bank. The dataset also contains other relevant information about the borrowers. The features of the data set are as follows:

1. **loan\_acc\_num**: This is the unique loan account number associated with each loan account.
2. **customer\_name**: This is the name of the customer who has taken the loan.
3. **customer\_address**: This is the customer’s address.
4. **loan\_type**: This is the type of loan that the customer has taken, such as a car loan, two-wheeler loan or personal loan.
5. **loan\_amount**: This is the amount of the loan that the customer has taken.
6. **collateral\_value**: This is the value of the collateral provided by the customer to secure the loan.
7. **cheque\_bounces**: This is the number of cheques that have bounced for the customer in the past few months.
8. **number\_of\_loans**: This is the total number of loans that the customer has taken.
9. **missed\_repayments**: This is the total number of the customer’s missed loan repayments for any loan taken from the bank. This would also include any other loans that the customer has had with the bank.
10. **vintage\_in\_months**: the number of months since the customer has joined the bank.
11. **tenure\_years**: This is the total tenure of the loan in years. (Also referred to as loan tenor)
12. **interest**: This is the interest rate at which the loan is provided.
13. **monthly\_emi**: This is the monthly installment that the customer needs to pay to repay the loan.
14. **disbursal\_date**: This is the date on which the loan was disbursed to the customer.
15. **default\_date**: This is the date on which the loan has defaulted on the loan.

Dataset **repayment\_base.csv**

The repayment\_base dataset contains information about the repayments made by the customers for the loan accounts in the first data set, main\_loan\_base. It should be expected that accounts without any repayments would not be present in this dataset. It includes the following features:

1. **loan\_acc\_num**: This is the unique loan account number associated with each loan account.
2. **repayment\_date**: This is the date on which the repayment was made.
3. **repayment\_amount**: This is the amount of money that was repaid on the repayment date.

Dataset **monthly\_balance\_base.csv**

The monthly\_balance\_base dataset contains the details of the customer’s monthly bank balance for the duration of the loan. The features in this dataset include:

1. **loan\_acc\_num**: This is the unique loan account number associated with each loan account.
2. **date**: This is the date on which the monthly bank statement was generated.
3. **balance\_amount**: This is the amount of money that was present in the customer’s account on the specified date.

The three datasets can be merged using the loan\_acc\_num as the common key. The information from these datasets can be combined to get a more complete picture of the loan accounts, including the repayment history for each account.

**Caution**: When you merge the main\_loan\_base and repayment\_base datasets, it should be noted that not all accounts present in the main\_loan\_base will have corresponding entries in the repayment\_base. This results in the presence of null values in the merged data set, which must be properly handled before you proceed with the calculation of the loss given default (LGD) metric. Failure to address these null values may result in errors during the LGD calculation process.

