**Data Assessment Document**

**1. Project Title**

Debt Risk Analysis and Loan Default Prediction

**2. Data Sources**

* Source Name: Unpublished data from our internal analysis
* Data Collection Method: Database Extraction from Finacle software.
* Date of Collection: 20th September 2024

**3. Initial Data Overview**

* Size of the Dataset: There are 32 columns and 54681 rows in the dataset.
* Type of Data: The data comprises of categorical, numerical, text, datetime of datatype.
* File Format: The file is XLSX format and is opened in excel.

**4. Dataset Features**

The dependent variables and independent variables consist in the data are provided below:

|  |  |
| --- | --- |
| **Variable** | **Dependency** |
| BRANCHNAME | Independent |
| STATE | Independent |
| CYDESC | Independent |
| SCHM\_TYPE | Independent |
| ACTYPEDESC | Independent |
| SEGMENT | Independent |
| ACOPENDATE | Independent |
| INTRATE | Independent |
| OUTSTANDING | Independent |
| MID\_RATE | Independent |
| LCY\_OUTSTANDING | Independent |
| LIMIT | Independent |
| LIMITEXPIRYDATE | Independent |
| EI\_END\_DATE | Independent |
| NORMALACCRUED | Independent |
| PRINOVERDUE | Independent |
| PRIN\_OVERDUE\_DATE | Independent |
| PRINODDAYS | Independent |
| INTERESTOVERDUE | Independent |
| INTEREST\_OVERDUE\_DATE | Independent |
| INTOVDDAYS | Independent |
| PENAL\_INT\_ON\_INT | Dependent |
| COVID\_PRN\_EXT | Independent |
| COVID\_INT\_EXT | Independent |
| CLASSIFICATION | Dependent |
| NRB9\_3 | Independent |
| Final Classification | Dependent |
| LLP% | Dependent |
| LLP Amount | Dependent |
| DCGF & Hydro Moratorium | Independent |
| Additional LLP | Independent |
| Total LLP | Dependent |

**5. Ethical Assessment**

The process of data collection is intended to protect customer privacy by anonymizing data and implementing strong security measures to prevent unauthorized access. The data is used solely for the purposes stated during collection. Secondary uses of the data is not practiced and ensure practices are not harmful or discriminatory.

Through the project we intend to recognize and mitigate potential biases in datasets and algorithms to prevent reinforcing inequalities in marketing strategies.

**6. Next Steps**

* The next step is to do EDA involving summarizing and visualization of data to find the patterns, trends and relationship and any other anomalies.
* To assess the dependencies of the dependent variables with other independent variables.
* Also, assess the target variables for the model preparation and predictions.