MUDRA LOANS RISK PREDICTION

GROUP : 2

TEAM LEADER : AKSHARA GUPTA

STUDENT ID : BC2023577

TEAM MEMBERS

S.NO.	MEMBER NAME	STUDENT ID	PARTICIPATION
1.	KABIR MOHSIN	BC2023064	ACTIVE
2.	RUDRANSH DWIVEDI	BC2023425	ACTIVE
3.	ANAMTA YUSUF	BC2023498	ACTIVE
4.	NISHANT GANGWAR	BC2023493	ACTIVE
5.	SAUMYA SINGH	BC2023479	ACTIVE
6.	LAVI SINGH	BC2023428	ACTIVE
7.	SHRUTYANSH MOHAN PATHAK	BC2023414	ACTIVE
8.	CHAHNA CHAND	BC2023210	ACTIVE
9.	HARSHITA CHAND	BC2023204	ACTIVE
10.	MANAS GANGWAR	BC2023342	ACTIVE
11.	NIMRA HIFZAN	BC2023512	ACTIVE
12.	RONIT MAURYA	BC2023038	ACTIVE
13.	KAUSHAL KUMAR	BC2023327	ACTIVE
14.	SHASHANK RAJPUT	BC2023657	ACTIVE

Table of Content

1.Objective

Developing a predictive model to assess the risk of default for Mudra loan applicants. Using historical loan data, the model should classify loan applications as low-risk or high-risk based on various applicant and loan attributes.

2.Libraries and Tools Used

Pandas: For data manipulation and cleaning.

NumPy: For numerical operations.

Matplotlib and Seaborn: For data visualization.

Scikit-learn: For data preprocessing, modeling, and evaluation.

3.Data Preprocessing

Data preprocessing in Python for machine learning involves preparing raw data for model training, including cleaning, transforming, and scaling data to improve model accuracy and performance. Key techniques include handling missing values, encoding categorical features, scaling numerical features, and splitting data into training and testing sets.

Data Cleaning

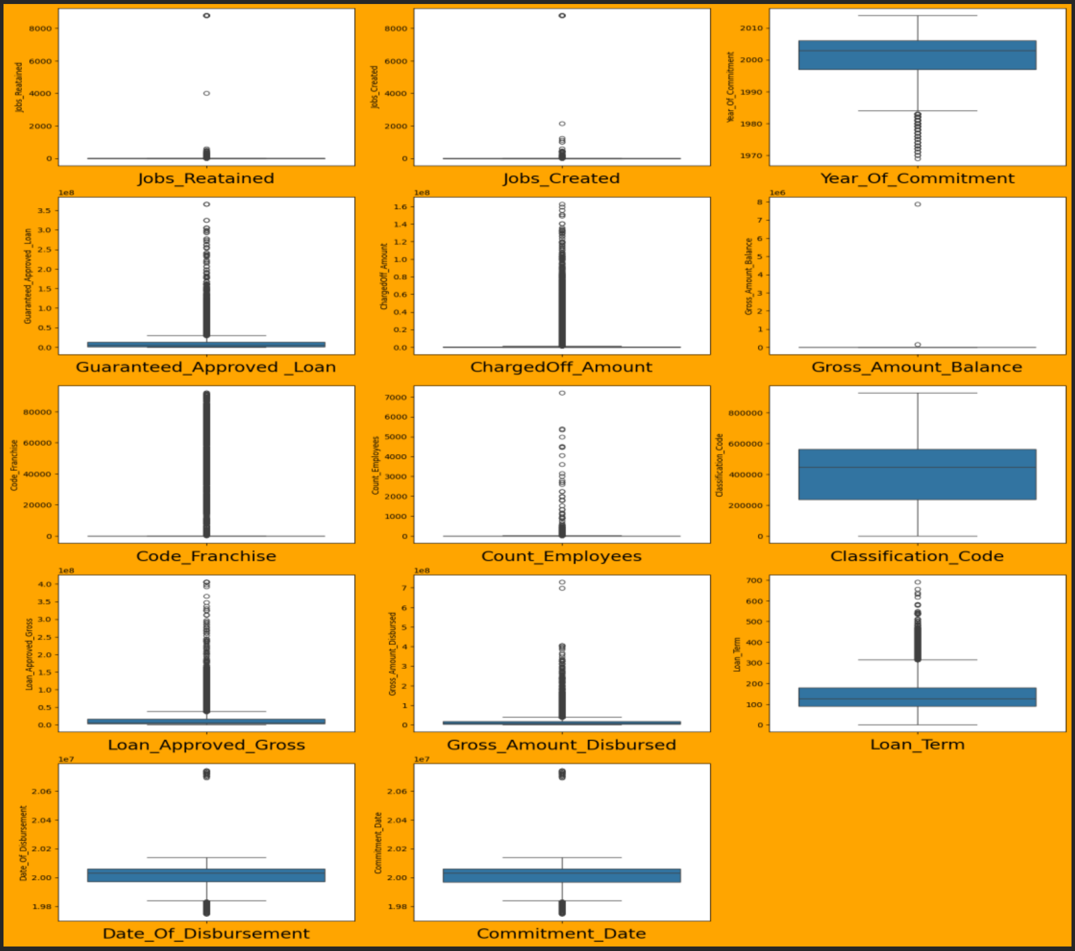
Handling Missing Values:
Deletion: Remove rows or columns with missing values (use with caution, as it can lead to data loss).
Imputation: Replace missing values with a suitable value (e.g., mean, median, mode, or a constant).

Null values

Null values are blank spaces in a dataset. They represent missing or unknown information. Some ways to handle null values are:
• Removing or deleting null values.
• Replacing null values with mean, median and mode.
• Using algorithm that can handle null values.

4.Outliers

In **machine learning**, an **outlier** is a data point that stands out a lot from the other data points in a set. The article explores the fundamentals of outlier and how it can be handled to solve machine learning problems.



Insight can be gained:- Outliers can provide valuable insights into exceptional cases.

Quality control:- It helps in detect quality control issues.

Understanding complexity:- It helps us to understand complex dataset.

Errors can be fixed:- Identify outliers can help detect errors in data collection.

5. Exploratory Data Analysis (EDA)

Univariate Analysis

Examining each variable individually to understand its distribution, using visualizations such as histograms and boxplots.

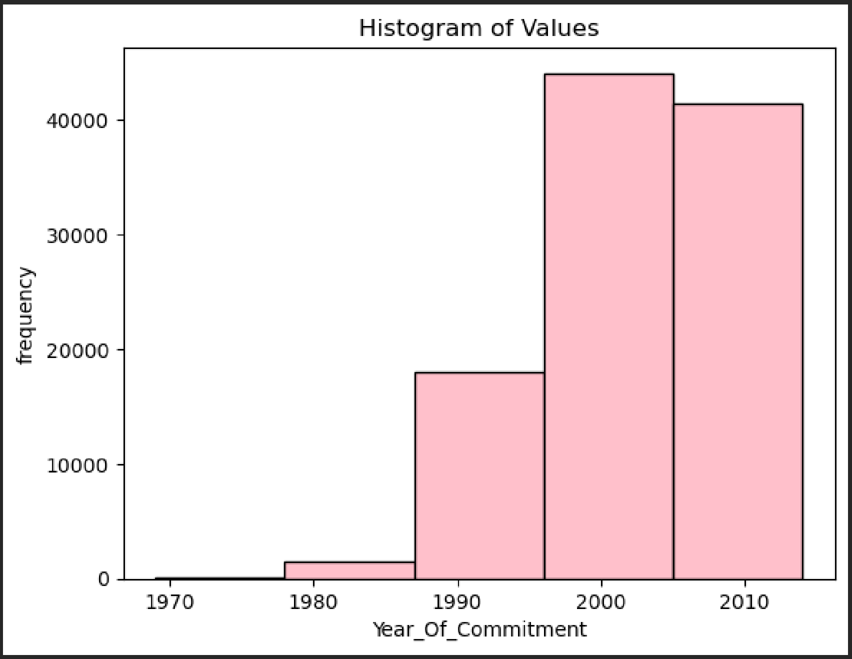
Bivariate Analysis

Exploring relationships between pairs of variables using scatterplots and correlation coefficients.

Data visualization

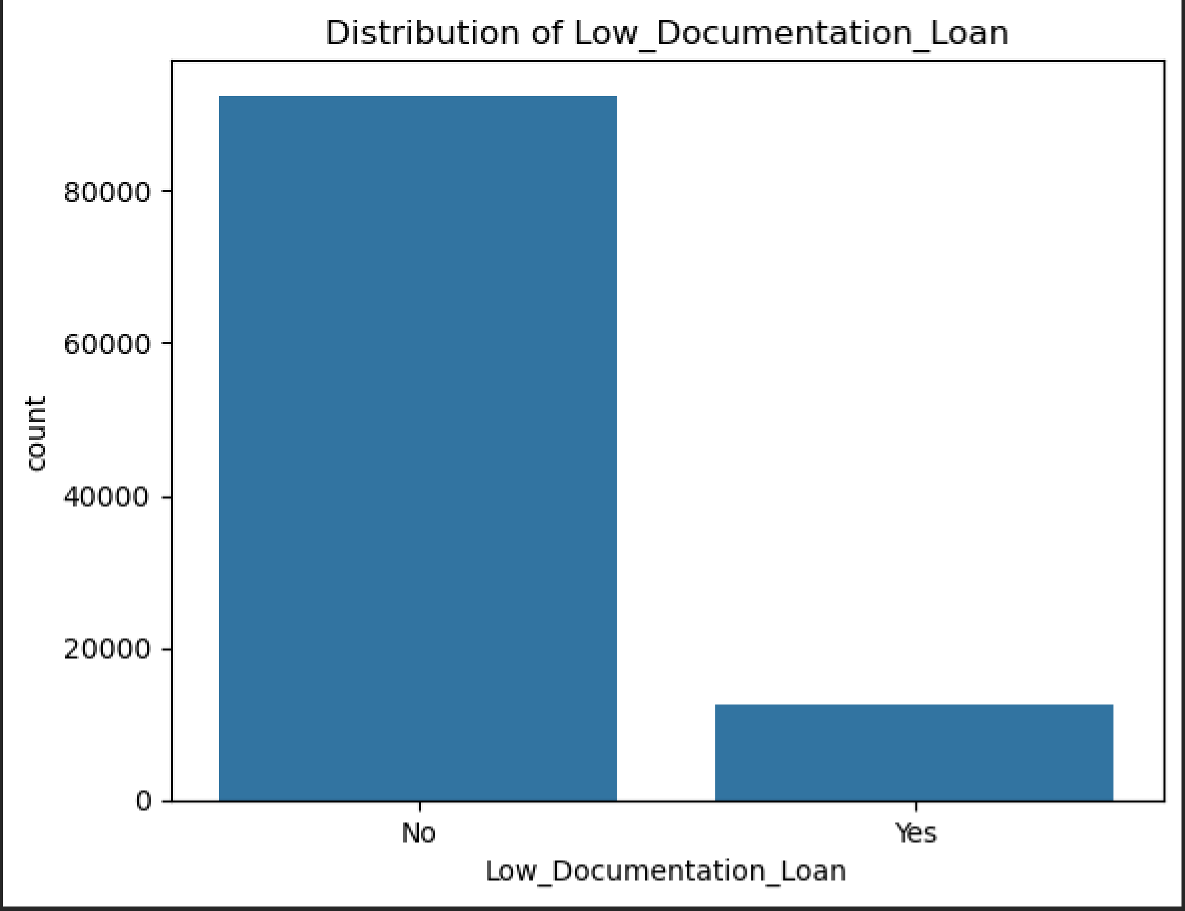
Histoplot

It is a combined visualization that displays the distribution of a single variable, typically showing a histogram alongside a kernel density estimation (KDE) or normal curve, and optionally a rug plot.



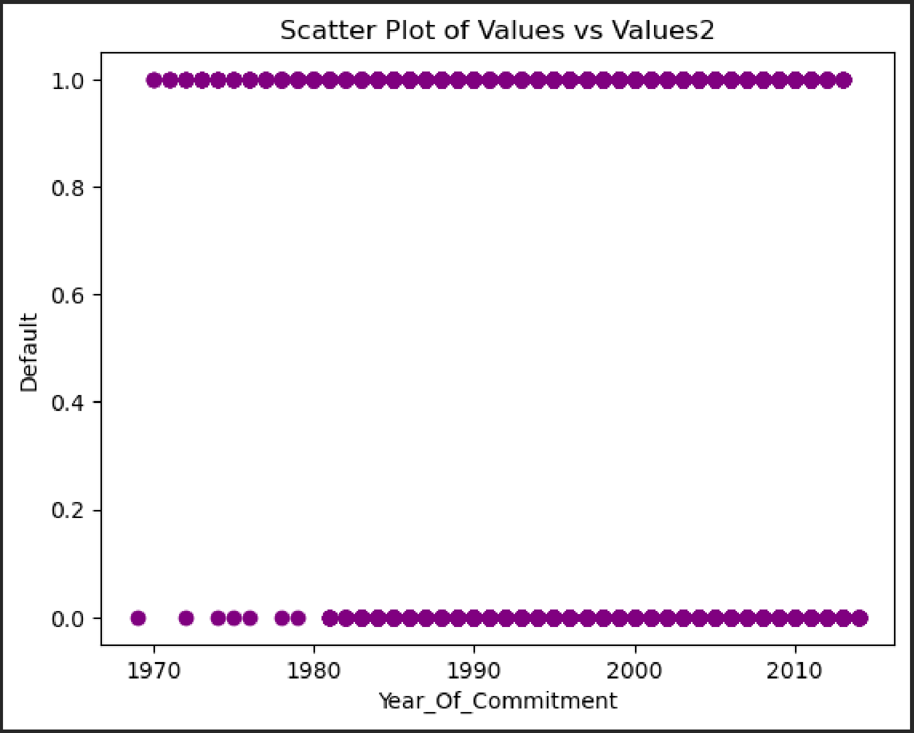
Count plot

A count plot resembles a histogram over a categorical variable as opposed to a quantitative one. You can compare counts across nested variables because the fundamental API and settings are the same as those for bar plot().



Scatterplot

A scatter plot is a diagram where each value in the data set is represented by a dot. The Matplotlib module has a method for drawing scatter plots, it needs two arrays of the same length, one for the values of the x-axis, and one for the values of the y-axis.



6.Encoders

Binary encoding

Binary encoding vs one-hot encoding are essential techniques for handling categorical data. Binary encoding efficiently converts categories into binary digits, making it suitable for high cardinality datasets. In contrast, one-hot encoding creates separate binary columns for each category.

7.Skewness Adjustment

In machine learning, skewness refers to the asymmetry or lack of symmetry in the distribution of data. A skewed distribution has a longer tail on one side than the other, meaning the data is not evenly spread around the mean. Skewness quantifies the degree and direction of asymmetry in a data distribution.

Types of Skewness:

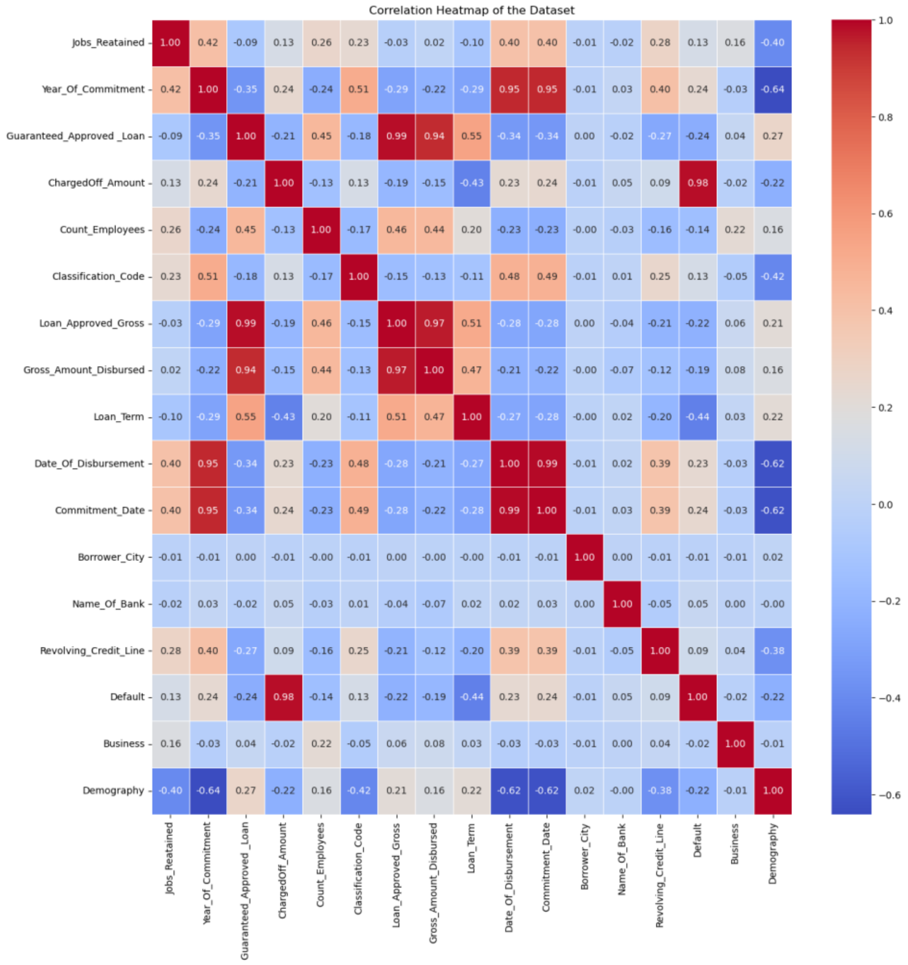
- Positive Skew (Right Skew): The tail of the distribution is longer on the right side, indicating that there are more data points on the lower end of the distribution.
- Negative Skew (Left Skew): The tail of the distribution is longer on the left side, indicating that there are more data points on the higher end of the distribution.

8.Heat map

A heatmap is a data visualisation technique that displays the value of a 2-D matrix or Data frame in a grid of coloured cells. The colour intensity represents the magnitude of the values in the matrix, making it easier to identify patterns, correlation or anomalies in the data.

Why we used it?

- It helped us in dealing with our dataset.
- Helped in visualisation of distribution plot.
- Helped in detecting outliers and patterns.
- Helped in spotting relationship between variables.



9.Spliting data to X and Y

X and Y split is a technique used in data preprocessing to split data into input features (X) and target variables (y).

- X (input features): The data that is used to predict the target variable. This is the data that is fed into the model.
- Y (target variable): The data that is being predicted. This is the output of the model.

Splitting data into X and Y is a crucial step in machine learning and modelling.

It allows you to:
Identify the relationships between the input features (X) and the target variable (y).
Train models to predict the target variable (y) based on the input features (X).
Evaluate the performance of models using metrics such as accuracy, precision, and recall.

10. Standard scaler

Standard Scaler is a technique used in data preprocessing to scale features to a standard range, typically between 0 and 1, or with a mean of 0 and a standard deviation of 1. This is also known as normalization or feature scaling.

The goal of Standard Scaler is to:
1. Prevent features with large ranges from dominating the model.
2. Improve model performance by reducing the effect of feature scale.
3. Enhance interpretability of model coefficients.

11. Train test split

Why Split the Data?

Splitting the data ensures the model is trained on one portion and evaluated on another, preventing overfitting.

Implementation Details

- **Scikit-learn Function: train test split.**
- **Typical split: 80% training and 20% testing.**

12. ML Algorithms

Machine learning (ML) algorithms are sets of instructions that allow computers to learn from data and make predictions or decisions without explicit programming, encompassing techniques like supervised, unsupervised, and reinforcement learning.

13. Classification matrix or regression matrix

A classification matrix sorts all cases from the model into categories, by determining whether the predicted value matched the actual value. All the cases in each category are then counted, and the totals are displayed in the matrix.

Regression metrics are quantitative measures used to evaluate the performance of regression models, which predict continuous outcomes, by quantifying the differences between predicted and actual values.

14.Logistic regression

Logistic regression aims to solve classification problems. It does this by predicting categorical outcomes, unlike linear regression that predicts a continuous outcome.

In the simplest case there are two outcomes, which is called binomial, an example of which is predicting if a tumor is malignant or benign. Other cases have more than two outcomes to classify, in this case it is called multinomial. A common example for multinomial logistic regression would be predicting the class of an iris flower between 3 different species.

Here we will be using basic logistic regression to predict a binomial variable. This means it has only two possible outcomes.

15.KNN

KNN is a simple, supervised machine learning (ML) algorithm that can be used for classification or regression tasks - and is also frequently used in missing value imputation. It is based on the idea that the observations closest to a given data point are the most "similar" observations in a data set, and we can therefore classify unforeseen points based on the values of the closest existing points. By choosing *K*, the user can select the number of nearby observations to use in the algorithm.

Here, we will show you how to implement the KNN algorithm for classification, and show how different values of *K* affect the results.

16.Random forest

A Random Forest Algorithm is a supervised machine learning algorithm that is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

17.Decision tree

Decision forest models are composed of decision trees. Decision forest learning algorithms (like random forests) rely, at least in part, on the learning of decision trees.

In this section of the course, you will study a small example dataset, and learn how a single decision tree is trained. In the next sections, you will learn how decision trees are combined to train decision forests.