**Task:**

This report outlines the application of various data preprocessing and feature engineering techniques to develop a robust XGBoost classifier model for the given dataset, a loan default prediction data set from a finance company in the United States. The data set included information on previous loan applicants and whether they defaulted or not. The goal was to identify the patterns that indicate if a person is likely to default which may be used for taking actions such as denying the loan, reducing the amount of loan, lending at a higher interest rate for risky applicants and etc.

After going through the data dictionary provided, a significant amount of details about the data set could be obtained.

**Overview of Data:**

The data set had 145 columns including the label and 517788 rows.

**Data Preprocessing:**

Data preprocessing is a crucial step in the data analysis and machine learning pipeline that involves transforming raw data into a clean, organized format suitable for further analysis or modeling. This process typically includes several steps:

**Data cleaning:**

**Dropping Columns**

To clean data, the first step I took was to drop the columns (attributes). After testing several percentages such as 30%, 50% and 80%, I decided to drop the columns having more than 50% missing values from the training data set, the reason being that imputing values for those columns would not be much accurate due to the high percentage of non-accurate yet approximate values.

It resulted in reducing the number of attributes to 82 (including the label) from 145. (i.e. 63 attributes were dropped)

To clean the columns having categorical data columns, before imputing the mode, I observed the unique value counts of each categorical variable, which resulted in dropping few more columns as well.

['emp\_title','title','zip\_code','addr\_state'] these columns had significantly higher number of unique values, which made them unable to be encoded easily. In fact had I done so, I would have to deal with the curse of dimensionality in great numbers as well.

After observing the data dictionary, from the domain knowledge I have, I decided these columns cannot be crucial for the decision made here.

The unique values in the columns grade and sub grade gave the insights to drop grade column as well.

\*\*All the columns were dropped from train data, valid data and test data.

**Dropping Rows**

Then before handling the missing data, I listed out the null counts, the following numbers were observed. Hence dropped those rows along with the rows having emp length empty

A screenshot of a computer program

Description automatically generated

Before Imputing, the distributions were observed.

A screenshot of a graph

Description automatically generated

**Imputing**

According to the above, mean for slightly skewed, median for skewed and mode for categorical columns was imputed.

Before imputing the null values present in valid and test dimensionality was further reduced(i.e. no null values in train set)

**Handling outliers**

The model that was used here is XGBoost, which is said to be less sensitive to outliers due to the tree classification used. So outlier handling was not done here.