Capstone Project Overview

**Project Name** –Loan Default Prediction

Financial loan services are leveraged by companies across many industries, from big banks to financial institutions to government loans. One of the primary objectives of companies with financial loan services is to decrease payment defaults and ensure that individuals are paying back their loans as expected. In order to do this efficiently and systematically, many companies employ machine learning to predict which individuals are at the highest risk of defaulting on their loans, so that proper interventions can be effectively deployed to the right audience.

**Expected Data Sources and Structure** –

Sample data set would be taken from Kaggle website. Loan Default Prediction dataset contains following important fields

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Column Name | Column Type | Data Type | Description |
| 1 | LoanID | Identifier | String | Unique identifier for each loan |
| 2 | Age | Feature | Integer | Age of the borrower |
| 3 | Income | Feature | Integer | Annual income of the borrower |
| 4 | LoanAmount | Feature | Integer | Amount of money being borrowed |
| 5 | CreditScore | Feature | Integer | Credit Score of the borrower |
| 6 | MonthsEmployed | Feature | Integer | Number of months the borrower has been employed |
| 7 | NumCreditLines | Feature | Integer | Number of credit lines the borrower has open |
| 8 | InterestRate | Feature | Float | Interest rate of the loan |
| 9 | LoanTerm | Feature | Integer | Term length of the loan in months |
| 10 | DTIRatio | Feature | String | Debt-to-income ratio indicating borrower’s debt compared to their income |
| 11 | Education | Feature | String | Highest level of education attained by the borrower |
| 12 | EmploymentType | Feature | String | Type of employment status of the borrower |
| 13 | MaritalStatus | Feature | String | Marital status of the borrower |
| 14 | HasMortgage | Feature | String | Whether the borrower has a mortgage |
| 15 | HasDependents | Feature | String | Whether the borrower has any dependents |
| 16 | LoanPurpose | Feature | String | Purpose of the loan (Home, Auto etc.) |
| 17 | HasCoSigner | Feature | String | Whether the loan has a co-signer (Yes or No) |
| 18 | Default | Target | Integer | Binary target variable indicating whether the loan has defaulted (1) or not (0) |

**Findings**

* Dataset is pretty clean.
* Income, Loan Amount, Interest Rate and Age are the most important features that influence loan default status.
* Load defaults are high for the age group between 18 to 40.
* Load defaults are high for the income group 15000 to 35000.
* Decision Tree Classifier, Random Forest Classifier, k-NN, Gradient Boosting Classifier and AdaBoost Classifier all of them show accuracy of around 88%

**Summary**

Based on our evaluation using ROC AUC, accuracy, precision, recall, and F1-score, the models achieve high overall accuracy (around 88%), but their precision, recall, and F1-scores remain suboptimal. This suggests that while the models reliably predict the majority class (non-defaulters), they struggle to correctly identify defaulters. Although the high ROC AUC values indicate that the models have reasonable discriminative power, the low recall and F1-scores reveal that the detection of the minority class (defaulters) is inadequate.   
  
Therefore, further optimization—such as applying rebalancing techniques, fine-tuning model parameters, or improving calibration—is necessary to enhance the detection of loan defaulters.