Case Study - Banking Domain

Business/Domain Understanding

Context

Financial markets are fundamental institutions in any developing economy. They play a crucial role in promoting economic growth by facilitating the channeling of saving decisions into productive investment. A major concern for financial institutions is credit risk, because if not managed properly, it can lead to a banking collapse.

In our banking system, banks have many products to sell but the main source of income of any bank is on its credit line. Loan is the core business part of banks. A bank's profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting.

Though a lot of people are applying for loans. However, they may have some difficulty in repaying the loan, due to their own capability to repay loan, their personal monetary terms, etc.. It’s hard to select the genuine applicant, who will repay the loan. While doing the process manually, a lot of misconception may happen to select the genuine applicant. The banks hold the risk of losing the amount loaned to the borrowers, which is basically regarded as "Credit risk".

Credit risk is the potential that a bank’s borrower or counterparty fails to meet its obligations in repaying the loan borrowed from the financial institutions ("banks").

By predicting the loan defaulters, the bank can reduce its Non-Performing Assets. This makes the study of this phenomenon very important.

Thus, the banks need to manage the credit risks in their portfolio both at the individual borrower and transactional level, as well as to consider the linkage between credit risks and other types of risk. This is because these are criteria to assess the success/failure of any banking lending activities.

Many research confirmed that machine learning technology is highly efficient to predict this situation. This technique is applied through learning from previous data.

What is Credit Risk?

Credit risk refers to the likelihood that a borrower will fail to meet their obligations in repaying a debt. In other words, it's the risk that a borrower might default on a loan or be unable to meet their contractual obligations, resulting in financial loss for the lender.

Managing and assessing credit risk is a crucial aspect of financial management for both lenders and borrowers. Lenders employ various tools and models, like credit scoring systems, to evaluate and mitigate credit risk. Borrowers, on the other hand, seek to maintain good creditworthiness to secure favorable lending terms and opportunities.

Credit Risk Dataset

Domain - Banking
Dataset - Click here to download the dataset: **credit_risk_dataset.csv**

Detailed **data description** of Credit Risk dataset:

Feature Name	Description
person_age	Age
person_income	Annual Income
person_home_ownership	Home ownership
person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default
cb_preson_cred_hist_length	Credit history length

SPRINT 1

Task - Exploratory Data Analysis

Assume that **you are working as a Data Scientist** with one of the world's leading financial institutions (like HSBC).

This is an open ended question. Kindly apply all your knowledge to perform an exploratory data analysis on the given dataset. It is known that the target variable is **Loan Status**.

However, you are mandatorily supposed to solve the below mentioned EDA Task for your presentation:

1. Which variables are most significant with respect to the target variable?
2. Explore the data distribution of each column. Identify some important patterns.
3. Insights and Recommendations (i.e. Data Driven Business Decision)

Write proper conclusions and provide recommendations to the bank based on the insights.

SPRINT 2

Task - Data Preparation and Model Building (Credit Risk Scoring)

What is Credit Risk Scoring?
Credit risk scoring, also known as credit scoring, is a statistical method used by financial institutions to evaluate the creditworthiness of individuals or businesses applying for credit. It is a way to assess the likelihood that a borrower will default on their financial obligations, such as repaying a loan or credit card debt.

Problem Statement - Given various features about a customer like Age, Income, Loan Amount, Loan Intent, Home Ownership etc.. , predict if in case the loan is given, will the customer default or not on the Loan payments.

Task - Prepare the data and build a model to predict if a customer is going to default or not.

Step - 1: Load the data
Step - 2: Document the below mentioned points properly:
- Identify the input and output/target variables.
- Identify the type of ML Task.
- Identify the Evaluation Metric.
    - For regression task - Mean Absolute Error
    - For classification task - Accuracy

Step - 3: Split the dataset into Training and Testing (recommended 75:25 split).
Step - 4: Data preparation on train data:
- For Numerical Variables - Standardization or Normalization (Fit and Transform)
- For Categorical - LabelEncoding or OneHotEncoding (Choose wisely)

Step - 5: Data preparation on test data:
- For Numerical Variables - Standardization (Transform)
- For Categorical - LabelEncoding or OneHotEncoding (Choose wisely)

Step - 6: Model Training Phase - Use all the algorithms mentioned below to train separate models:

- KNN
- Logistic Regression
- Support Vector Machines
- Decision Trees
- Random Forest

Step - 7: Predict and evaluate each model separately using the correct evaluation metric. Use metrics.accuracy_score(actual, predict).

Step - 8: Display a plot which shows all the algorithms applied along with the accuracies achieved. **Write your conclusion on the best algorithm for Credit Risk Scoring**.