Capstone Project
Loan Default Prediction

Context

A major proportion of retail bank profit comes from interests in the form of home loans. These loans are borrowed by regular income/high-earning customers. Banks are most fearful of defaulters, as bad loans (NPA) usually eat up a major chunk of their profits. Therefore, it is important for banks to be judicious while approving loans for their customer base.

The approval process for the loans is multifaceted. Through this process, the bank tries to check the creditworthiness of the applicant on the basis of a manual study of various aspects of the application. The entire process is not only effort-intensive but also prone to wrong judgment/approval owing to human error and biases.

There have been attempts by many banks to automate this process by using heuristics. But with the advent of data science and machine learning, the focus has shifted to building machines that can learn this approval process and make it free of biases and more efficient. At the same time, one important thing to keep in mind is to make sure that the machine does not learn the biases that previously crept in because of the human approval process.

Problem Statement

A bank's consumer credit department aims to simplify the decision-making process for home equity lines of credit to be accepted. To do this, they will adopt the Equal Credit Opportunity Act's guidelines to establish an empirically derived and statistically sound model for credit scoring. The model will be based on the data obtained via the existing loan underwriting process from recent applicants who have been given credit. The model will be built from predictive modeling techniques, but the model created must be interpretable enough to provide a justification for any adverse behavior (rejections).

Objective

Build a classification model to predict clients who are likely to default on their loan and give recommendations to the bank on the important features to consider while approving a loan.

Data Dictionary

The Home Equity dataset (HMEQ) contains baseline and loan performance information for recent home equity loans. The target (BAD) is a binary variable that indicates whether an applicant has ultimately defaulted or has been severely delinquent. There are 12 input variables registered for each applicant.

- **BAD:** 1 = Client defaulted on loan, 0 = loan repaid
- **LOAN:** Amount of loan approved
- **MORTDUE:** Amount due on the existing mortgage
- **VALUE:** Current value of the property
- **REASON**: Reason for the loan request (HomeImp = home improvement, DebtCon= debt consolidation which means taking out a new loan to pay off other liabilities and consumer debts)
- **JOB:** The type of job that loan applicant has such as manager, self, etc.
- **YOJ:** Years at present job
- **DEROG:** Number of major derogatory reports (which indicates serious delinquency or late payments).
- **DELINQ:** Number of delinquent credit lines (a line of credit becomes delinquent when a borrower does not make the minimum required payments 30 to 60 days past the day on which the payments were due)
- **CLAGE:** Age of the oldest credit line in months
- **NINQ:** Number of recent credit inquiries
- **CLNO**: Number of existing credit lines
- **DEBTINC:** Debt-to-income ratio (all monthly debt payments divided by gross monthly income. This number is one of the ways lenders measure a borrower's ability to manage the monthly payments to repay the money they plan to borrow)