Build a Credit Default Risk Prediction Model with LightGBM

**Project Description**

**Business Overview**

Credit Risk is the possibility of a loss resulting from a borrower's failure to repay a

loan or meet a contractual obligation. The primary goal of a credit risk assessment is to find out whether potential borrowers are creditworthy and have the means to repay their debts so that credit risk or loss can be minimized and the loan is granted to only creditworthy applicants.

 If the borrower shows an acceptable level of default risk, then their loan application can be approved upon agreed terms.

 This project involves understanding financial terminologies attached to credit risk and building a classification model for default prediction with LightGBM. Hyperparameter Optimization is done using the Hyperopt library and SHAP is used for model explainability.

**Aim**

To predict loan defaulters and minimize the risk of loss on the basis of credit history, employment, and demographic data.

**Data Description**

The dataset contains information about 143727 borrowers’ on various attributes such as employment type, work experience, income, dependents, total loans, total payment done, etc.

**Tech Stack**

* Language: Python
* Libraries:  pandas, numpy, matplotlib, seaborn, scikit\_learn, lightgbm, hyperopt, shap

**Approach**

* Data Reading
* Data Processing
  + Drop Columns
  + Split Data
* Define Label
  + Roll Rate Analysis
  + Window Roll Analysis
* Feature Engineering
  + Label
  + % Amount Paid as interest in past Loan Repayment
  + % of Loans defaulted in the last 2 years
* Exploratory Data Analysis (EDA)
  + Univariate Analysis
    - Numerical Summary: Min, Max, Mean, Median, etc
    - Categorical Summary: Top, Unique, Count, etc
  + Bivariate Analysis
    - Correlation Plot
    - Box Plots
* Target Encoding
* Feature Selection
  + Random Forest
  + Decision Tree
* ML Model Development
  + LightGBM
  + Hyperparameter Tuning using Hyperopt
* Model Selection
* Model Evaluation
  + ROC AUC
  + PR AUC
  + Score Distribution
* Feature Importance
  + Split and Gain
  + SHAP
* Class Rate Curve and Right Threshold