Build a Credit Default Risk Prediction Model with LightGBM
Project Overview

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

Modular code overview:

```
input
|_credit_risk_data.csv

documents
|_project_document.pdf
|_lightgbm_explanation.pdf

lib
|_model.ipynb
|_utils.py
|_hyperopt_results.csv

ml_pipeline
|_utils.py
|_processing.py
|_training.py

output

engine.py

requirements.txt

readme.md
```

Once you unzip the modular_code.zip file, you can find the following folders within it.

1. input
2. documents
3. lib
4. ml_pipeline
5. output
6. engine.py
7. requirements.txt
8. readme.md

1. The input folder contains the raw data that we have for analysis. In our case, it contains credit_risk_data.csv

2. The documents folder contains the supporting learning material.

3. The lib folder is a reference folder, and it contains the original ipython notebook as in the lectures.

4. The ml_pipeline is a folder that contains all the functions put into different python files, which are appropriately named. The engine.py script then calls these python functions to run the steps in one go to train the model and saves it in the output folder.

5. The requirements.txt file has all the required libraries with respective versions. Kindly install the file by using the command **pip install -r requirements.txt**

6. **All the instructions for running the code are present in readme.md file**

Project Takeaways

1. What is credit risk assessment?
2. Understand “default” using Roll Rate Analysis
3. What is the importance of “days past due” or “dpd” in credit risk?
4. How to perform exploratory data analysis for categorical target variable?
5. What is Target Encoding and why is it preferred over other encoding techniques?
6. How to make Feature Selection with Random Forest and Deep Decision Tree?
7. How to model LightGBM and optimize its parameters?
8. What is Hyperopt?
9. How does Hyperopt optimize model parameters?
10. How to do a model selection on the basis of metrics such as ROC AUC, and PR AUC?
11. What is Split and Gain and how is it used for Feature Importance?
12. What is SHAP?
13. How to visualize Feature Importance with SHAP?
14. What is the Class Rate Curve?
15. How to choose the right threshold for minimum credit loss?