Data Science Wizards Campus Hiring

Problem Statement: Classification problem for Loan Default Prediction

About NBFC:

A trusted Non-Banking Financial Company (NBFC) specializing in providing quick and accessible small loans tailored to the needs of individuals and small businesses. Our mission is to empower financial independence by offering flexible loan options, competitive interest rates, and a seamless approval process. With a strong commitment to customer satisfaction and financial inclusion, we aim to bridge the gap between traditional banking and underserved communities. Whether it's for personal emergencies, working capital, or small-scale projects, we are here to support your journey to success.

Task Summary:

The NBFC is developing a classification model to predict loan repayment behaviour, specifically identifying potential defaulters and non-defaulters. The primary goal is to enhance risk assessment and improve the loan approval process.

Data Overview (shared in separate files)

You have two data sets:

- Historic data: Loan disbursement application and their default and non-default status for past 2 years+ has been kept in the file. **File name**: train_data.xlsx
- Validation data: Loan disbursement application and their default and non-default status for past 3 months has been kept in the file. **File name**: test_data.xlsx

Columns:

column name	Description
customer_id	Unique identification for each customer application
transaction_date	transaction data
sub_grade	customer is classified into various grades based on geography, income and age
term	total loan tenure
home_ownership	status of home ownership of applicant
cibil_score	cibil score of applicants
total_no_of_acc	total number of bank accounts holded by applicant
annual_inc	annual income of the applicant
int_rate	interest rate charged by NBFC

purpose	purpose for taking loan, defined by applicant
loan_amnt	total loan amount
application_type	applicant type
installment	instalment amount
verification_status	applicant verification status
account_bal	total account balance as per previous month
emp_length	total years of employment experience
loan_status	loan status (1: default, 0: non-default)

Technical Requirements

- EDA:
  - Perform EDA and submit in a notebook named “eda.ipynb”.
  - For each analysis/chart code block add a comments markdown block.
- Modelling:
  - Prepare the training pipeline and submit it in a script named “model_.py”.
  - Create at least two different models.
  - Object-oriented, class-based approach to be followed. Model classes must consist of at least the following functions: load (to load data), preprocess (preprocessing steps), train (training steps), test (test steps, that also generate evaluation summary) and predict (for inference).
- Model Selection:
  - Prepare the model selection pipeline and submit it in a script named “model_selction.ipynb”.
  - Run different models prepared in the “Modelling” task and showcase evaluation metrics. You can perform hyperparameter tuning if required.
  - Finally, choose one model and provide a summary of “Why the model was chosen?” in the notebook at the end.
- The notebooks shared should be pre-runned and with the output cells.
- Use appropriate coding standards and provide in-line comments where required.
- Use Python 3.7 +
- The final submission should be a zip file with all the required files described above named “<Your Full Name>.zip”

Notes

- You will agree to only submit your own work! You are free to do research online or in the relevant literature.
- The given data sets are stochastically generated and are not real data of our company. Nevertheless, it is possible to make proper predictions based on the attributes.