Loan Repayment Assessment in Banking

Problem Statement:

Welcome to KnowledgeHut AI hackathon – **Loan Repayment Assessment in Banking.** You are required to build and train a model that identifies a customer will repay or default from the loan dataset. This dataset is included in loan data, and provides a challenging classifier that will test what you have learnt in this course.

Task:

Your task is to build this model based on the details in this document and submit it. Please read the details carefully before attempting this hackathon.

You will need to decide the following:

1. Use the specific source or dataset for assess loan repayment shared with you
2. What is your intended data split ratio for training, validation, and test sets for the loan dataset? How do you plan to ensure randomness in this split?
3. Do you plan to explore the importance of these components further?
4. Do you anticipate class imbalance in the 'loan_status' feature, where
   Paid: Applicant has fully paid the loan (the principal and the interest rate)
   Defaulted: Applicant has not paid the installments in due time for a long period of time, i.e. Client has defaulted on the loan
   If so, how will you address this imbalance?
5. Will you normalize the features? If yes, what normalization techniques do you have in mind?
6. Do you intend to perform data preprocessing tasks such as outlier detection, missing value handling, or feature selection before training your model.

Your code should have the following:

- ✓ Statistics descriptive analysis
- ✓ EDA
- ✓ Data preprocessing
- ✓ Feature scaling
- ✓ Feature engineering
- ✓ Feature selection
- ✓ Build model
- ✓ Ensemble techniques - Bagging, Boosting
- ✓ Cross validation
- ✓ Grid search, Tuning Hyper parameters
- ✓ Evaluation metric: F1 - Score

Dataset:

Although the dataset is provided with limited columns you have some decisions to make around it. First, it's important to familiarize yourself with the dataset:

train_loan_data.csv:

You will note that in this dataset, there are 80,000 records and 28 features, and can be used to train the model

Data description of these features are given below

Column	Description
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_title	The job title supplied by the Borrower when applying for the loan.
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
Grade	LC assigned loan grade
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
initial_list_status	The initial listing status of the loan. Possible values are – W, F
num_actv_bc_tl	Number of currently active bankcard accounts.
mort_acc	Number of mortgage accounts.
tot_cur_bal	Total current balance of all accounts
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies.
Purpose	A category provided by the borrower for the loan request.
revol_bal	Total credit revolving balance
Title	The loan title provided by the borrower
total_acc	The total number of credit lines currently in the borrower's credit file
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
int_rate	Interest Rate on the loan
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
sub_grade	LC assigned loan subgrade
Term	The number of payments on the loan. Values are in months and can be either 36 or 60.

revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
Target	
loan_status	Status of the loan

test_loan_data.csv:
This dataset can be used to test your model, in this dataset there are 20,000 records and 27 features similar as in the train data (excluding the target 'loan-status')

test_results.csv:
This dataset contains the target ('loan_status' of the 20,000 test data) and can used to evaluate your model performance

Submission:
Please submit the Jupyter Notebook. This should clearly show:

1. The code you have written
2. The output of the test data
3. A comprehensive description of each code block, together with the decisions you've made and the rationale for those decisions.
4. A comprehensive description of what you also tried that did not work, and what lessons you have learnt from this hackathon.
5. Power point presentation on your work
6. Video Walkthrough
7. A GitHub link with all the above deliverables

Note:
• Use Python programming language
• Use a Google Colab, or a standard laptop/desktop to build the model
• This is a challenging dataset to predict accurately on, so iterate your approach over time making note of what works and what doesn't. In the long term, this is as useful to you as getting a high model accuracy.
• Avoid Plagiarism as the objective of this exercise is to give you a real-world project to build and we hope you will use this opportunity wisely to your benefit.