Probability of Default Prediction Model

Problem Statement:

You just been appointed as Data Scientist and your first task is to create Expected Loss Model. Consisting of Probability of Default (PD) model, a loss given default (LGD) model, and an exposure-at-default (EAD) model for a portfolio of individual's credits. But for this time we can start from the PD Model first in order to predict loan defaults based on loan application.

You can create feature engineering by using any combination of the features in the dataset to make your predictions accuracy higher. Some features will be easier to use than others.

Variable	**Definition**
Loan_ID	A unique id for the loan.
Grade	The grade of the loan (e.g. A, B, C, ...).
Home_Ownership	The home ownership status provided by the borrower during registration. Values are: Rent, Own, Mortgage.
Purpose	A category provided by the borrower for the loan request.
Verification_Status	Indicates if income was verified, not verified, or if the income source was verified.
Term	Credit term period (in months).
Emp_Length_Int	Employment length in years.
Mths_Since_Issue_D	The number of months since the borrower's last loan issue date.
Int_Rate	Interest rate of the loan application.
Mths_Since_Earliest_Cr_Line	The number of months since the borrower's first loan issue date.
Acc_Now_Delinq	The total number of loan account that currently delinquent.
Inq_Last_6mths	The number of inquiries by borrowers during the past 6 months.
Annual_Inc	The annual income provided by the borrower during registration.
Dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's self-reported monthly income.
Good_Bad	Indicates if the loan was good or bad (0 = default, 1 = non-default)

You're required to:

- **Build a machine learning model in a Python language** using the data provided with the highest accuracy as good as you can
- **Deploy/ predict the testing dataset** using machine learning model that you already developed
- **Create Model Development Documentation (ipynb format) & Presentation (PPT Slides)**

File for Submissions:

1) **Final dataset** with selected features for training & testing dataset (give your final training and testing dataset that already predict)

2) **Documentation** describing the flow, packages used, functions created, etc. (in ipynb file format), with contents as follows:

   A) Data Preparation & Exploration (EDA)

   - What did you discover from the datasets given?
   - Include any charts that you created as well.
   - How strong is your predictive power of your independent variable?
   - What is the relationship between variables and your data distribution?
   - How clean your data? Is there any missing value?
   - Is there any outliers? How you handle it?
   - Should we add new independent variable to increase predictive power?

   B) Model Development

   - Develop at least 3 different algorithms to solve this problem.
   - How did you optimized the model?
   - How did you make sure the model is stable enough?
   - Did you use ensemble method to improve model performance?

   C) Model Evaluation

   - For model performance criteria at least you must include : *accuracy, recall, precision, F1 Score, Log Loss, AUC Score, and confusion matrix result*
   - What is the final model algorithm that you used? Please explain!

3) **Presentations (PPT Slides)**
   - Create max. 7 slides (excluding cover page/ dividers) and should include:
     - ✓ Business problem & how to solved it
     - ✓ Model development process i.e. missing data assumption, new variables creation, etc. (1 slide)
     - ✓ Model selection process i.e. model assessment, error rate, etc. (1-2 slide)
     - ✓ Final model, key features, and model performance (1-2 slides)