Here’s your customized scenario and task instructions in the same structured format as the previous one.

**Scenario: Credit Risk Prediction at Bank of Kigali**

Imagine you are a **data scientist** at **Bank of Kigali**, tasked with developing a machine learning model to predict whether a loan applicant is likely to **default on a loan**. Your model will help the bank assess credit risk and make informed lending decisions based on historical loan data.

Your task is to:

1. **Load and preprocess the dataset**, ensuring it is clean and ready for modeling.
2. **Perform exploratory data analysis (EDA)** to understand key trends, correlations, and insights.
3. **Train multiple machine learning models**, evaluate their performance, and select the best one for deployment.
4. **Test the model** using appropriate evaluation metrics to ensure reliability in predicting loan default.

The dataset contains the following features:

* person\_age: Age of the applicant.
* person\_income: Monthly income of the applicant.
* person\_home\_ownership: Homeownership status (e.g., "Rent", "Mortgage", "Own").
* person\_emp\_length: Employment length in years.
* loan\_intent: Purpose of the loan (e.g., "Debt Consolidation", "Education", etc.).
* loan\_grade: Assigned loan grade (e.g., "A", "B", "C").
* loan\_amnt: Loan amount requested.
* loan\_int\_rate: Interest rate on the loan.
* loan\_status: Whether the loan was defaulted (Target Variable).
* loan\_percent\_income: Loan amount as a percentage of income.
* cb\_person\_default\_on\_file: Whether the applicant has a prior default ("Y" or "N").
* cb\_person\_cred\_hist\_length: Length of credit history in years.

Once you complete the analysis and model training, you will prepare a report with insights, model performance metrics, and a recommendation for deployment.

**Task Instructions and Evaluation Criteria**

| **Task** | **Marks** | **Done/Not Done** |
| --- | --- | --- |
| **Data Preprocessing** |  |  |
| 1. Load the dataset (credit\_data.csv) and check for missing values. | 5 |  |
| 2. Handle missing values and perform necessary data cleaning. | 5 |  |
| 3. Convert categorical variables (person\_home\_ownership, loan\_intent, loan\_grade, cb\_person\_default\_on\_file) into numerical format. | 10 |  |
| 4. Normalize or scale numerical features (person\_income, loan\_amnt, loan\_int\_rate, etc.). | 5 |  |
| 5. Split the dataset into training and testing sets. | 5 |  |
| **Exploratory Data Analysis (EDA)** |  |  |
| 6. Generate summary statistics and visualize feature distributions. | 5 |  |
| 7. Analyze correlations between features and loan\_status. | 5 |  |
| 8. Detect and handle outliers in numerical features. | 5 |  |
| **Model Training and Selection** |  |  |
| 9. Train at least three machine learning models (e.g., Logistic Regression, Random Forest, XGBoost). | 10 |  |
| 10. Tune hyperparameters using **GridSearchCV** or **RandomizedSearchCV** for better performance. | 5 |  |
| **Model Evaluation and Testing** |  |  |
| 11. Evaluate models using **accuracy, precision, recall, F1-score, and ROC-AUC**. | 10 |  |
| 12. Interpret feature importance to understand key predictors of loan default. | 5 |  |
| 13. Generate and interpret a confusion matrix for the best-performing model. | 5 |  |
| 14. Perform cross-validation to check model robustness. | 5 |  |
| **Final Submission** |  |  |
| 15. Prepare a report summarizing findings, key insights, and model recommendations. | 10 |  |
| 16. Provide all Python scripts and Jupyter Notebook files. | 5 |  |
| **Total Marks** | **100** |  |

This structured approach ensures that all steps—**preprocessing, EDA, model training, evaluation, and reporting**—are covered. Do you need any refinements? 🚀