Question--Explain your intuition behind the features used for modeling. Are you creating new derived features?

Answer--The features used in the model aim to capture various dimensions of a borrower's financial behavior and profile. Key features like Principal_loan_amount, EMI_rate_in_percentage_of_disposable_income, and Savings_account_balance provide insights into the borrower's financial capacity. Has_coapplicant, Has_guarantor, and Loan_history variables indicate additional risk factors and credit history. Primary_applicant_age_in_years, Gender, and Number_of_dependents contribute demographic information. Derived features, such as Property_encoded and Purpose_encoded, transform categorical data into numerical form, enhancing model interpretability and performance.

Yes,I created new derive features using one hot encoding.

Question2--Are there values missing? If yes, how did you handle them.

Answer--Yes, there were missing values in the dataset. For columns with missing values greater than 50%, I dropped that column, while numerical columns with missing values were imputed using the median to ensure robust handling of outliers. Categorical columns with missing values were replaced with 'Unknown' to preserve the integrity of the data and avoid losing information.

Question3--How have you handled the categorical features?

Answer -Categorical features were handled through encoding methods to convert them into numerical format suitable for modeling. For nominal categorical features, like Property and Purpose, I used **one-hot encoding** to create binary columns for each category. For ordinal features, such as Loan_history, I applied **label encoding** to assign a unique integer to each category based on its order. This ensures that the model can effectively interpret and utilize the categorical data.

Question4--- . Describe the features correlation using correlation matrix. Tell us about few correlated feature & share your understanding on why they are correlated

Answer -Housing_rent and Marital_status are highly correlated, it implies a significant relationship between these features. Certain marital statuses might correlate with housing preferences. For example, married individuals or those with families may be more likely to rent larger homes compared to single individuals.

Question5— Do you plan to drop the correlated feature?

Answer -Yes, I planed to drop the highly correlated features Housing_rent and Marital_status. To address their redundancy, I used Principal Component Analysis (PCA) to combine them into a single feature, PCA_housing_marital, which captures their shared variance. This approach simplifies the dataset while retaining essential information.

Question6--Reason for choosing the ML algorithms used by you.

Answer -I used **Logistic Regression** for its simplicity and interpretability in binary classification, providing clear insights into feature impacts. **Random Forest** was chosen for its ability to handle complex interactions and imbalanced data by aggregating multiple decision trees, improving overall accuracy and robustness. Both algorithms complement each other, offering a balance of clarity and performance.

Question7--Which other ML algorithms did you consider but did not move forward with and why?

Answer-I considered **Gradient Boosting** but did not move forward with it due to unsatisfactory accuracy.

Question 8- Train two (at least) ML models to predict the credit risk & provide the confusion matrix for each model.

Answer-its done in notebook

Question 9---How you will select the hyperparameters for models trained in above step.

Answer- Random Forest--Using GridSearchCV from scikit-learn to exhaustively search over this grid to find the best combination.

Logistic Regression-- Adjusting the regularization parameter (C) to control the trade-off between fitting the training data well and keeping the model weights small. Using GridSearchCV to test various values for C.

Question10- Which metric(s) you will choose to select between the set of models.

Answer-To select between models, I would focus on **F1-score** and **Recall**. The F1-score balances precision and recall, making it suitable for imbalanced datasets, while recall is crucial for identifying high-risk applicants who are the primary concern in credit risk prediction.

Question11-Explain how you will export the trained models & deploy it for prediction in production.

Answer-Using joblib or pickle to save the trained model to a file.And to deploy we can use cloud service,Heroku ,docker etc