Project Workflow

1. Data Ingestion

- Read accepted and rejected loan datasets from raw CSV files.
- Merge datasets into a single DataFrame for modeling.
- Ensure consistent feature names and data types.

2. Data Cleaning

- Handle missing values (drop or impute based on context).
- Drop irrelevant columns (e.g., unique IDs, URLs).
- **Loan purpose categorization:**
    - Original dataset had 30k+ loan purposes.
    - Applied **TF-IDF clustering** locally to group similar purposes.
    - Selected **top 10 examples per cluster** and used **OpenAI API** to assign meaningful labels.
    - Reduced 30k+ categories to **50 relevant categories**.

3. Feature Encoding & Normalization

- Encode categorical features using **one-hot encoding**:
    - Loan grade
    - Purpose
    - Home ownership
- Normalize numerical features using **MinMaxScaler** or **StandardScaler**:
    - Loan amount
    - Debt-to-income ratio (DTI)
    - Income

4. Feature Engineering

- Create flags and interaction features:
    - `high_dti_flag = 1 if dti > 30 else 0`
    - `long_term_employment = 1 if emp_length >= 10 else 0`
    - `loan_dti_interaction = loan_amount * dti`
- Optional: Derived ratios like loan-to-income.

5. Modeling (Risk & Profit Tradeoff)

- Train models to predict loan default probability:
    - **Logistic Regression**
    - **Random Forest**
    - **XGBoost**
- Tune hyperparameters using **GridSearchCV** or **RandomizedSearchCV**.
- Include **expected profit modeling**:
    - `profit = profit_per_good_loan` if repaid
    - `loss = loss_per_default` if defaulted

6. Cross-Validation & Evaluation

- **K-Fold Stratified CV** for robust performance.
- Evaluate metrics:
  - **ROC-AUC**
  - **F1-score**
  - **Expected Profit** (custom business metric)
- Visualizations:
  - **ROC Curves** for all models
  - **Profit Curves** to show cumulative profit vs. number of loans issued