Project 2

Machine Learning Capstone (Advanced Level)

📌 Project Title:

"Credit Risk Prediction System – End-to-End ML Model with Deployment"

🎯 Objective:

To build a complete Machine Learning solution that predicts whether a loan applicant is a high-risk or low-risk borrower, using real-world financial and demographic data.

This project simulates a real industry-level ML workflow from preprocessing to deployment.

📂 Dataset Example:

- Loan Approval or Credit Risk Dataset
- Contains features:
    - Income
    - Employment Type
    - Loan Amount
    - Loan Duration
    - Credit Score

- Number of Previous Defaults
- Collateral information
- Age, marital status, dependents
- Loan repayment history

🧠 Detailed Project Workflow:

1. Data Understanding & Requirement Analysis:

- Identify the business problem: reduce loan defaults.
- Understand the target variable: Risk (High/Low).
- Check data quality and identify necessary preprocessing steps.

---

2. Data Preprocessing & Feature Engineering:

- Handle missing values with advanced techniques (mean/median/group-based imputation).
- Encode categorical features using One-Hot Encoding.
- Create new features such as:
  - Debt-to-Income Ratio
  - Total Obligation Ratio
  - Age group categories
- Scale numerical features for uniformity.
- Remove outliers using IQR or z-score if required.

---

3. Exploratory Data Analysis (EDA):

Generate essential visual insights:

- Loan default patterns by age, income, employment, and credit score.
- Correlation heatmap to identify strong predictive features.
- Boxplots and distribution plots to analyze numeric features.
- Bar charts to understand class imbalance (if any).

---

4. Model Building:

Implement multiple models such as:

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier
- XGBoost
- Support Vector Machine (optional)

Compare them on performance and interpretability.

---

5. Model Optimization:

- Apply Hyperparameter Tuning using GridSearchCV or RandomizedSearchCV.
- Fine-tune important parameters such as:
    - Number of trees

- Learning rate
- Depth of trees
- Regularization strength
- Class weights for imbalanced data

---

6. Handling Imbalanced Data:

If dataset is imbalanced, apply advanced techniques:

- SMOTE Oversampling
- Class weights
- Undersampling of majority class
- Threshold tuning to improve sensitivity

---

7. Model Evaluation:

Evaluate using real-world metrics:

- Accuracy
- Precision (especially important in risk prediction)
- Recall (identify risky customers correctly)
- F1 Score
- ROC-AUC Curve
- Confusion Matrix

Pick the best model for deployment.

---

8. Model Deployment:

Deploy your final model using one of the following:

- **Streamlit App**
- **Flask API**
- **Gradio Interface**

Key UI elements:

- Input form for user details
- Clear prediction output (High Risk / Low Risk)
- Optional explanation section (feature importance)

---

9. Final Documentation:

Prepare a final project summary that includes:

- Problem statement
- Dataset explanation
- Preprocessing steps
- Visualizations
- Algorithms used
- Performance comparison
- Deployment details
- Learning outcomes and challenges faced

🎓 Learning Outcome:

By completing this major project, the intern will:

- Gain real-world experience in building a production-ready ML model
- Learn industry practices for preprocessing, EDA, modeling, and evaluation
- Understand deployment techniques and UI creation
- Strengthen their resume and portfolio with a professional ML project