**Project Title: Lending Club Loan Default Prediction**

**1. Objective**

The project’s primary goal was to create an end-to-end machine learning pipeline for predicting loan defaults using the Lending Club dataset. The pipeline integrated data preprocessing, analysis, model development, and deployment into a functional web application.

**2. Detailed Analysis**

**A. Data Preprocessing**

The data preprocessing phase ensured clean and relevant data for analysis and modeling:

* **Initial Cleaning:**
  + Removed irrelevant columns like id, member\_id, and policy\_code which did not contribute to predictive power​.
* **Handling Missing Data:**
  + Used median imputation for numerical fields and mode for categorical ones to handle missing values in features such as annual\_inc and loan\_term.
* **Outlier Detection:**
  + Identified extreme values in variables like loan\_amnt and dti. Applied transformations like log scaling to reduce the impact of skewed data distributions.
* **Feature Engineering:**
  + Derived meaningful metrics, e.g., income\_to\_loan\_ratio (ratio of annual\_inc to loan\_amnt) and categorized interest rates (low, medium, high).
  + Categorical data encoding (e.g., one-hot encoding) for variables like loan\_purpose.

**B. Exploratory Data Analysis**

A comprehensive exploratory analysis revealed critical insights:

1. **Loan Amount Distributions:**
   * The majority of loans were between $5,000 and $20,000.
   * Loans with higher amounts were more likely to default.
2. **Loan Purpose Analysis:**
   * Most loans were taken for debt consolidation, followed by credit card refinancing and small business investments.
   * Loans for business purposes had a significantly higher default rate.
3. **Interest Rates:**
   * Default likelihood increased with higher interest rates (>15%).
   * Visualization of rate trends confirmed its importance as a predictor.
4. **Debt-to-Income Ratio (dti):**
   * Borrowers with a dti greater than 40% showed higher risk of default.
   * This metric was among the most influential predictors.

**Visualization Summary:** Charts generated from the data included:

* Loan amount distribution.
* Loan purpose analysis.
* Feature importance rankings.

**C. Model Development**

Multiple machine learning models were evaluated for performance and robustness:

1. **Model Comparison:**
   * **Logistic Regression:** Baseline model, provided insights into linear relationships between features.
   * **Random Forest:** Introduced non-linear patterns and reduced overfitting using ensemble methods.
   * **Gradient Boosting (Final Model):** Achieved the best performance due to its ability to optimize weak learners and focus on misclassified instances.
2. **Performance Metrics:**
   * **Accuracy:** Final model achieved an accuracy of ~88%.
   * **Precision & Recall:** Balanced to minimize false negatives in default prediction.
   * **F1-Score:** High, confirming robustness in predicting both classes effectively.
3. **Feature Importance:**
   * Key predictors identified were:
     + int\_rate: Strongest predictor of loan default.
     + dti: Captured borrowers’ financial stress.
     + loan\_amnt: Higher amounts correlated with increased defaults.
     + term: Longer loan terms (60 months) indicated a higher risk of default.

**D. Web Application Deployment**

The project was deployed as a web application for real-time predictions:

1. **Frontend Development:**
   * Developed with **HTML**, **CSS**, and **JavaScript**​index​script​styles.
   * Key sections:
     + **Prediction Form:** Allowed users to input variables like loan\_amnt, dti, and annual\_inc.
     + **Prediction Results:** Displayed results dynamically with user-friendly messages (Defaulter in red, Non-Defaulter in green).
   * Styling ensured responsiveness and intuitive user interaction.
2. **Backend Simulation:**
   * A simulated backend predicted default status using a random probability generator (proof of concept)​script.
3. **Deployment on GitHub Pages:**
   * The application was successfully hosted on GitHub Pages, making it accessible via a shareable link.

**3. Results and Insights**

* The Gradient Boosting model demonstrated strong predictive capability, making it the model of choice for deployment.
* Feature engineering and EDA revealed important patterns in borrower behavior, lending insights into financial risk management.

**Key Insights:**

* Loans with higher interest rates and longer terms are at a higher risk of default.
* Borrowers with high debt-to-income ratios often struggle to repay loans.
* Feature selection significantly improved model performance by focusing on influential predictors.

**4. Impact and Applications**

This project highlights:

* Practical use of predictive analytics in financial risk assessment.
* Demonstrates end-to-end capabilities: data processing, model development, and real-world deployment.
* Provides an accessible interface for non-technical users to explore ML predictions.