### **Project Title: Credit Risk Prediction using Machine Learning**

### **Project Objective:**

The primary goal of this project is to build a machine learning model to predict **loan default risk**, assisting lenders in evaluating whether a borrower is likely to default on their loan. By predicting this risk, the model helps improve decision-making processes, minimize financial losses, and better allocate resources within financial institutions.

### **Problem Definition:**

Lenders often face high financial risk when extending loans, particularly to borrowers who may not be able to repay them. **Credit risk analysis** is a crucial task for financial institutions, as it involves evaluating the likelihood that a borrower will default on a loan. A robust model can help flag high-risk individuals, ensuring that the right precautions are taken during the loan approval process.

However, this problem is complicated by:

* **Imbalanced classes**: There are usually far fewer defaults (class 1) than non-defaults (class 0).
* **High cost of false negatives**: Missing a potential defaulter is far costlier than rejecting a borrower who could have been a reliable repayer.

### **Approach:**

1. **Data Preprocessing:**
   1. Cleaned the data by handling missing values and applying **feature engineering** (e.g., converting categorical variables to numerical ones and scaling the features).
   2. Applied **SMOTE** (Synthetic Minority Over-sampling Technique) to handle the imbalanced data and ensure better model performance for the minority class (loan defaults).
2. **Model Selection:**
   1. Tried different machine learning models, including **LightGBM** (Gradient Boosting) and **Random Forest**, to predict loan default risks.
   2. Tuned the models and performed cross-validation for optimal performance.
3. **Thresholding:**
   1. Adjusted the decision threshold to **trade-off between precision and recall**, ensuring that the model’s focus was on **minimizing false negatives** (defaulters that go unnoticed) without being too lenient on false positives (declining trustworthy borrowers).
4. **Evaluation:**
   1. Used **AUC-ROC**, **accuracy**, **precision**, **recall**, and **f1-score** as metrics to evaluate the performance of the models.
   2. Focused on **recall** for the default class to minimize missing risky borrowers, which would otherwise have significant financial consequences.

### **Results:**

* The model achieved an **accuracy of 86.7%**, with a **recall of 0.86** for predicting loan defaults. This means the model successfully identifies 86% of the defaulters.
* The **precision for defaulters was 0.65**, which reflects an acceptable rate of false positives while maintaining strong recall.
* The **classification report** indicated a good balance between precision and recall, ensuring that both non-defaulters and defaulters were handled effectively.

### **Insights and Business Impact:**

* **Key Insight:** The model’s ability to detect defaulters is crucial in the financial sector, where missed risks can have significant financial consequences. The strong **recall** ensures that the model identifies most of the high-risk borrowers.
* **Business Impact:** This model can significantly reduce the risk of issuing loans to individuals who are likely to default, potentially saving financial institutions from large losses. The model could be deployed to flag high-risk applicants and assist loan officers in making more informed decisions.

### **Next Steps for Model Enhancement:**

* **Cost-Benefit Analysis:** Perform further evaluations to assess the cost of false positives (e.g., rejecting good borrowers) versus false negatives (e.g., approving risky borrowers).
* **Model Optimization:** Experiment with **Hyperparameter Tuning** and consider incorporating additional features like economic indicators or previous financial behavior to further improve predictions.
* **Deployment and Monitoring:** Integrate the model into a loan approval system and set up continuous monitoring for model performance, retraining it periodically as new data becomes available.

### **Conclusion:**

By implementing this credit risk prediction model, I was able to provide a solution that minimizes financial risks associated with loan defaults. The model focuses on **identifying defaulters**, allowing for more effective risk management and lending decisions. This project demonstrates my ability to handle complex real-world problems, apply machine learning techniques, and provide actionable insights to improve business outcomes.

### **Technologies Used:**

* **Python**: Pandas, NumPy, Scikit-learn, LightGBM, Matplotlib, Seaborn
* **Machine Learning Models**: LightGBM, Random Forest
* **Data Preprocessing**: SMOTE, Feature Scaling, One-Hot Encoding
* **Evaluation Metrics**: Accuracy, Precision, Recall, F1-Score, AUC-ROC