### **Objective**

The goal of this project is to build a loan default prediction system for a consumer finance company to minimize financial risk. The system will analyze loan applicants’ data and predict the likelihood of default using machine learning algorithms. The project will also incorporate psychological data to provide deeper insights into an applicant's financial behavior, enhancing the predictive power beyond traditional financial attributes.

### **Dataset Description**

The dataset contains 255,347 records and 18 features, with both traditional financial data and psychological data.

#### **Financial Features:**

* **LoanID**: Unique identifier for each loan.
* **Age**: Applicant’s age.
* **Income**: Applicant’s income in dollars.
* **LoanAmount**: Loan amount requested by the applicant.
* **CreditScore**: Applicant’s credit score.
* **MonthsEmployed**: Number of months employed.
* **NumCreditLines**: Number of credit lines held by the applicant.
* **InterestRate**: Interest rate applicable to the loan.
* **LoanTerm**: Duration of the loan in months.
* **DTIRatio**: Debt-to-income ratio.
* **Education**: Applicant's education level.
* **EmploymentType**: Employment type (Full-time, Part-time, etc.).
* **MaritalStatus**: Marital status of the applicant.
* **HasMortgage**: Indicates if the applicant has an existing mortgage.
* **HasDependents**: Indicates if the applicant has dependents.
* **LoanPurpose**: Purpose of the loan (Business, Auto, etc.).
* **HasCoSigner**: Indicates if the applicant has a co-signer.
* **Default**: Target variable, 1 for default and 0 for non-default.

#### **Psychological Features:**

* **Risk Appetite**: Willingness to take financial risks.
* **Decision-Making Speed**: How quickly decisions are made.
* **Financial Stress Levels**: Applicant’s reported stress about finances.
* **Spending Behavior**: Historical spending patterns or behavioral tendencies.

### **Key Steps in the Project**

#### **1. Data Preprocessing**

* **Label Encoding**: Convert categorical columns (e.g., Education, EmploymentType, etc.) to numerical values.
* **Feature Engineering**:
  + Income-to-Loan Ratio: Income / LoanAmount
  + CreditScore-to-Age Ratio: CreditScore / Age
  + LoanTerm in years: Convert LoanTerm from months to years.
  + Interaction terms between important features like age, income, and loan amount.
* **Handling Imbalance**: Use **SMOTE** (Synthetic Minority Oversampling Technique) to address class imbalance (more non-defaults than defaults).

#### **2. Exploratory Data Analysis (EDA)**

* Perform **correlation analysis** between financial and psychological features and the target variable.
* Use visualizations like **scatter plots**, **histograms**, and **correlation heatmaps** to identify patterns and relationships.

#### **3. Machine Learning Models**

* **Random Forest Classifier**: Ideal for tabular data with robust classification performance.
* **K-Nearest Neighbors (KNN)**: Simple but effective classification algorithm based on proximity.
* **Neural Networks**: Use deep learning to capture complex patterns in the data.
* **Logistic Regression**: As a baseline for binary classification.

#### **4. Model Evaluation**

* **Accuracy, Precision, Recall, F1-score**: Standard metrics for evaluating the model's performance.
* **Confusion Matrix**: Visualization of true positives, false positives, etc.
* **ROC-AUC Curve**: For assessing model performance on distinguishing between default and non-default cases.

#### **5. Incorporating Psychological Data**

* **Exploratory Data Analysis** to assess how psychological variables correlate with loan default.
* **Merge psychological and financial data** to improve model performance and validate the significance of psychological factors in loan default prediction.

#### **6. Model Tuning and Evaluation**

* **Hyperparameter tuning** using **GridSearchCV** to optimize the performance of models like Random Forest and KNN.
* **Model validation** with cross-validation techniques.

### **Technology Stack**

#### **Backend:**

* **Flask**: Used to create a REST API to serve the machine learning model and handle prediction requests.
* **Python**: For machine learning, data preprocessing, and model training.

#### **Frontend:**

* **HTML/CSS/JavaScript**: A simple interface where users can input applicant data and receive predictions.

#### **Machine Learning Libraries:**

* **scikit-learn**: For preprocessing, model building, and evaluation.
* **imblearn**: For handling imbalanced data using SMOTE.
* **Keras**: For building deep learning models (if neural networks are part of the solution).

#### **Deployment:**

* **Flask** backend can be deployed on **Heroku**, **AWS**, or **Google Cloud** to serve the loan default prediction model as an API.

### **Psychological Data Integration**

To gather psychological data, applicants will complete a survey at the time of loan application. These features will then be merged with traditional financial data and used to train the machine learning model. By integrating both financial and psychological insights, the system will provide more accurate and holistic predictions.

### **Conclusion**

This project combines financial data with psychological analysis to predict loan defaults. The machine learning system not only uses standard financial indicators but also factors in behavioral insights for improved risk assessment. The system is built to minimize loan defaults and financial risk, offering a robust solution for consumer finance companies.