**Credit Risk Assessment**

Credit default risk is the chance that companies/individuals cannot make the required payments on their debt obligations, which can lead to a possibility of loss for a lender. Earlier credit analysts would perform risk management by analysing the borrower’s credentials and capabilities, but this was prone to errors. With the advent of Machine learning, ML algorithms can perform a credit risk assessment with better precision and much faster than any humans.

To start this [machine learning project](https://www.projectpro.io/article/machine-learning-projects-for-practice/459), download the[Credit Risk Dataset](https://www.kaggle.com/laotse/credit-risk-dataset). Load the dataset into a data frame and remove rows of data NaN values. Also, convert the categorical values into numerical values using Label encoding. Our data is imbalanced. Hence, we use the stratifiedKFold method to split the dataset into training and testing sets.

Machine learning algorithms that can be used are KNN, logistic regression, and XGBoost(Extreme Gradient Boosting) algorithms. You can use the[performance metrics](https://www.projectpro.io/data%20science-tutorial/performance-metrics-for-machine-learning-algorithm) like Accuracy, Precision, Recall, and F1 score to evaluate your model’s performance. However, since the training data was imbalanced, the Area Under the Curve for the ROC curve would be a better evaluation metric.

When analyzing a credit risk dataset with the columns provided, the goal is to understand and predict the factors influencing credit risk and loan default. Below are some key questions to guide the analysis:

**1. Demographic Analysis**

1. **Age Distribution:**
   * What is the distribution of the ages of the loan applicants?
   * Are there any age groups that are more likely to apply for loans?
2. **Income Distribution:**
   * How is the income distributed among the applicants?
   * Are there specific income brackets that are more prevalent?
3. **Employment Length:**
   * What is the average length of employment among applicants?
   * Does employment length vary significantly with age or income?

**2. Loan Characteristics**

1. **Loan Amount:**
   * What is the distribution of the loan amounts requested?
   * Are there trends in loan amounts based on income, age, or employment length?
2. **Loan Intent:**
   * What are the common purposes for which loans are taken (loan\_intent)?
   * How does the loan intent correlate with other factors like age, income, and loan amount?
3. **Loan Grade:**
   * What are the distributions of loan grades assigned to applicants?
   * How do loan grades relate to income, employment length, and credit history?
4. **Loan Interest Rate:**
   * What is the distribution of loan interest rates?
   * How do interest rates vary with loan grades and loan amounts?

**3. Credit Risk and Default Analysis**

1. **Default Status:**
   * What is the overall default rate in the dataset?
   * How does default status vary across different demographic groups (age, income, employment length)?
2. **Home Ownership:**
   * Is there a difference in default rates between different home ownership statuses (rent, own, mortgage)?
3. **Credit History:**
   * How does the length of credit history affect the likelihood of default?
   * Is there a relationship between having a default on file and new loan defaults?

**4. Correlation and Predictive Factors**

1. **Correlation Analysis:**
   * What are the correlations between loan default status and other variables (age, income, loan amount, etc.)?
2. **Predictive Modeling:**
   * Which factors are the most significant predictors of loan default?
   * Can we build a predictive model to assess the risk of default based on the available data?

**5. Financial Ratios and Metrics**

1. **Loan Percent Income:**
   * How does the loan amount as a percentage of income (loan\_percent\_income) affect default rates?
   * Is there an optimal threshold of loan\_percent\_income that minimizes default risk?

**6. Comparative Analysis**

1. **Loan Grades vs. Defaults:**
   * How do default rates compare across different loan grades?
   * Are certain loan grades more prone to default?
2. **Income vs. Default Rate:**
   * Is there a non-linear relationship between income levels and default rates?

**7. Exploratory Data Analysis (EDA)**

1. **Outliers:**
   * Are there any outliers or anomalies in the dataset (e.g., extremely high or low loan amounts, interest rates, etc.)?
   * How do these outliers affect the overall analysis?
2. **Trends Over Time:**
   * Are there any time-related trends in loan applications and defaults (if time data is available)?

**8. Socio-Economic Factors**

1. **Impact of Employment Length:**
   * Does the length of employment have a significant impact on the probability of default?
   * Is there a threshold employment length that is critical for loan approval or better loan terms?

By systematically answering these questions, you can gain a comprehensive understanding of the factors influencing credit risk and improve decision-making processes for loan approvals.