**Housing Loan Repayment Risk Analysis**

**Bhavya Parmar NUID 002249226**

**Darshil Shah NUID 002814137**

**Team 21**

# Introduction:

**Background:**

The “Housing Loan Repayment Risk Analysis” dataset is a very essential source that will go a long way in advancing the cause of credit risk assessment in the financial sector. It is a compilation of very many pieces of loan application data, starting from financial status and demographics to credit history used in understanding loan repayment behaviors. In this perspective, the dataset paves the way for data-driven decision-making that allows detailed research on the borrower's risk factors. It comes as the instrument of developing predictive models for loan management and, in a way, mirrors the industry's march towards analytics-based more precise and effective risk mitigation strategies.

**Motivation:**

This initiative is propelled by the need to refine the precision of credit risk assessments within the finance industry, advancing beyond the confines of traditional assessment methods. By leveraging the “Housing Loan Repayment Risk Analysis” dataset, our goal is to harness sophisticated analytics and machine learning techniques to craft predictive models capable of accurately evaluating the risks associated with loan repayments. In essence, this project aspires to revolutionize risk management approaches, thus contributing to the stabilization of the financial sector and expanding the availability of credit, in harmony with the objectives of financial inclusion.

**Goal:**

The core goal of this initiative is to craft and deploy sophisticated predictive models capable of accurately predicting challenges in loan repayment. Utilizing the extensive data housed within the “Housing Loan Repayment Risk Analysis” dataset, our objectives include:

* Elevating the precision and efficiency of credit risk evaluations in the finance industry.
* Uncovering crucial indicators of loan repayment challenges through advanced analytics and machine learning.
* Diminishing loan default rates by equipping financial entities with more accurate risk assessment tools.
* Refining the loan approval decision-making process to be more knowledgeable and fairer.

This project is set to transform the practices of credit risk management fundamentally, bolstering the lending system's stability and trustworthiness.

# Methodology:

* **Data Preprocessing and Cleaning**: To ensure the “Housing Loan Repayment Risk Analysis” dataset's accuracy and relevance, the initial step will involve cleaning and preprocessing the data. This process will include imputation of missing values, standardization of formats and data types, detection and correction of anomalies or outliers, and ensuring the data is in a state ready for analysis.
* **Exploratory Data Analysis**: We will engage in exploratory data analysis (EDA) using tools such as correlation matrices, scatterplots, and histograms to uncover patterns and trends within the loan data. This phase will involve visual representation of data through various graphs and charts and the application of statistical techniques to identify correlations between different factors, such as financial background, employment status, and loan specifics (e.g., amount, type, and repayment terms).
* **Feature Engineering**: To enhance the precision of our predictive models, we will create new variables or feature selection based on domain knowledge and statistical methods to identify significant predictors of loan repayment challenges. Key factors likely to influence our feature selection include applicants' financial backgrounds, employment status, loan amount, annuity, type of loan, repayment terms, and credit history. This step is pivotal in refining our models to accurately predict loan repayment risks. We will use methods like one-hot encoding on the categorical columns like employment status, type of loan which will help in accuracy of the model.
* **Cross-validation:** Is a technique to ensure robust model performance. The dataset is divided into equal parts (folds), and the model is trained and validated across these folds. This method provides a comprehensive assessment of the model's ability to generalize. It further includes:
* Training Phase:In this phase, the model learns from the data by adjusting its parameters to predict outcomes based on input features. Typically, about 70-80% of the dataset is used for training, allowing the model to learn the relationship between inputs and outputs.
* Validation Phase: This phase helps in tuning the model's hyperparameters and preventing overfitting. By evaluating the model on a separate subset of the data (usually 10-15%), adjustments can be made to improve its ability to generalize to new data. This iterative process ensures the model performs well on data it hasn't seen during training.
* Testing Phase: In this phase assesses the model's performance on another set of unseen data—the test set. This final evaluation, using about 10-15% of the dataset, provides an unbiased measure of how the model will perform in real-world scenarios.
* **Model Selection and Evaluation:** We plan to utilize machine learning algorithms such as Logistic Regression, Decision Trees, Random Forests, and XGBoost Classifier to predict loan repayment challenges. These models will be trained on historical data from the “Housing Loan Repayment Risk Analysis” dataset to evaluate their accuracy.

# Description:

This project focuses on utilizing the “Housing Loan Repayment Risk Analysis” dataset to develop predictive models aimed at improving the precision of credit risk assessments within the finance industry. The dataset includes comprehensive loan application data, such as financial background, employment status, family circumstances, and loan specifics of thousands of applicants.

The Key columns from this dataset are:

* **Applicant Financial Background**: This column contains detailed information about the loan applicant's financial history, including income level, existing debts, and overall financial health. It's crucial for assessing the applicant's ability to repay the loan.
* **Employment Status:** Indicates whether the applicant is employed, unemployed, self-employed, or retired, providing insights into their stable income sources.
* **Family Circumstances**: Includes data on marital status, number of dependents, and living situation.
* **Loan Amount:** The total amount of money requested by the applicant. This figure is key to determining the scale of the loan and its repayment structure.
* **Annuity:** The regular payment amount the applicant is obligated to pay towards the loan. It helps in understanding the loan's impact on the applicant's monthly budget
* **Type of Loan:** Specifies the loan's purpose, such as mortgage, personal loan and auto loan.
* **Repayment Terms:** Details of the length of the loan and any specific repayment conditions. It's critical for understanding the loan's duration and the borrower's long-term financial commitment.
* **Credit History:** A record of the applicant's past borrowing and repayments. This column is vital for assessing the borrower's reliability and predicting their future loan repayment behavior.

**Data Source**: <https://www.kaggle.com/datasets/urstrulyvikas/house-loan-data-analysis>