**Project Proposal**

Credit risk analysis and financial risk prediction

**Project Title:**

Analyze and Predict Borrower Default Risk Based on Financial and Demographic Factors

**Purpose and Outcome:**

* ***Purpose:*** Building a model to predict the probability of borrower default based on financial and demographic factors. The project focuses on analyzing the factors that influence the probability of default, thereby developing a machine learning model that can help financial institutions assess credit risk more accurately.
* ***Outcome:*** Explore and analyze data to find out key factors that influence borrower default. Using python to build models to support making appropriate lending policy decisions.

**Dataset:**

* ***Source:*** Public dataset from Kaggle (Loan Default Prediction Dataset: <https://www.kaggle.com/datasets/nikhil1e9/loan-default>)
* ***Description:*** The data is collected from a loan information set that includes many variables related to demographics such as age, education level, marital status and finances such as income, loan amount, credit score, etc.
* ***Structure:***
* **LoanID:** A unique identifier for each loan
* **Age:** The age of the borrower
* **Income:** The annual income of the borrower
* **LoanAmount:** The amount of money being borrowed
* **CreditScore:** The credit score of the borrower, indicating their creditworthiness
* **MonthsEmployed:** The number of months the borrower has been employed
* **NumCreditLines:** The number of credit lines the borrower has open
* **InterestRate:** The interest rate for the loan
* **LoanTerm:** The term length of the loan in months
* **DTIRatio:** The Debt-to-Income ratio, indicating the borrower’s debt compared to their income
* **Education:** The highest level of education attained by the borrower (PhD, Master’s, Bachelor’s, High School)
* **EmploymentType:** The type employment status of the borrower (Full-time, Part-time, Self-employed, Unemployed)
* **MaritalStatus:** The marital status of the borrower (Single, Married, Divorced)
* **HasMortgage:** Whether the borrower has a mortgage (Yes, No)
* **HasDependents:** Whether the borrower has dependents (Yes, No)
* **LoanPurpose:** The purpose of the loan (Home, Auto, Education, Business, Other)
* **HasCoSigner:** Whether the loan has a co-signer (Yes, No)
* **Default:** The binary target variable indicating whether the loan defaulted

**Initial Analysis Plan:**

* ***Data cleaning:*** Check and handle missing values, handling outliers, encoding categorical variables, correct data types
* ***EDA:*** Generate summary statistics, visualize distributions, and analyze correlations between variables that influence individual defaults
* ***Analysis:*** Perform correlation analysis of the relationship between financial and demographic factors on default probability, data visualization, Random Forest model building and Logistic regression to predict borrower default probability.
* ***Visualization:*** Create visualizations like pie charts, bar charts, scatter charts, heatmaps, histograms to analyze factors leading to default.
* ***Data storytelling:*** Presents factors that cause borrower defaults, thereby providing recommendations to support companies with appropriate lending policies.