# Loan Repayment Detection

Presented by : **Group 4**

Meghana Mendu (1002112896)

Lakshmi Chandrasekhar Aravapalli (1002052925)

Mrunmai Magar(1002092125)

# Agenda

01

Problem statement and Introduction

02

Implementation

03

Conclusion

04

Demonstration and QnA

# Problem Statement

Individuals with insufficient or without credit histories struggle to access traditional financial services like loans and credit cards.

Without credit, it's hard to build a positive financial profile, hindering access to affordable loans, housing, and employment.

Turning to alternative lenders like payday loans, including payday lenders, pawnshops, and other high-interest loan providers exacerbates the problem by charging exorbitant interest rates and imposing unfair terms and conditions trapping borrowers in high-interest debt cycles.

# Introduction

This project is employing a range of statistical and machine learning techniques to enhance their ability to predict creditworthiness and it aims to minimize the risk of rejecting clients who have the capacity to repay their loans.

Expanding upon this, the integration of statistical and machine learning approaches enables this project to analyze a diverse set of data points and variables, allowing for a more comprehensive assessment of each applicant's creditworthiness. This includes not only traditional financial indicators such as income and credit history but also alternative data sources such as payment behavior, employment history, and demographic information.

By considering a broader range of factors, Home Credit can make more accurate predictions about an applicant's ability and willingness to repay loans.

This approach not only increases the likelihood of successful loan repayment but also fosters a positive and mutually beneficial relationship between loan lenders and their clients.

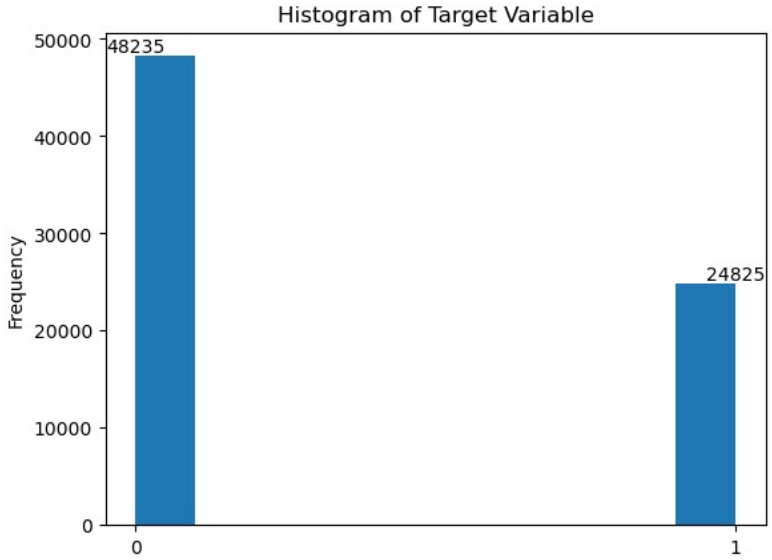
# Implementation

- Data Collection
- Data Pre-processing
- Exploratory data analysis
- Feature Engineering
- Model Development
- Model Evaluation

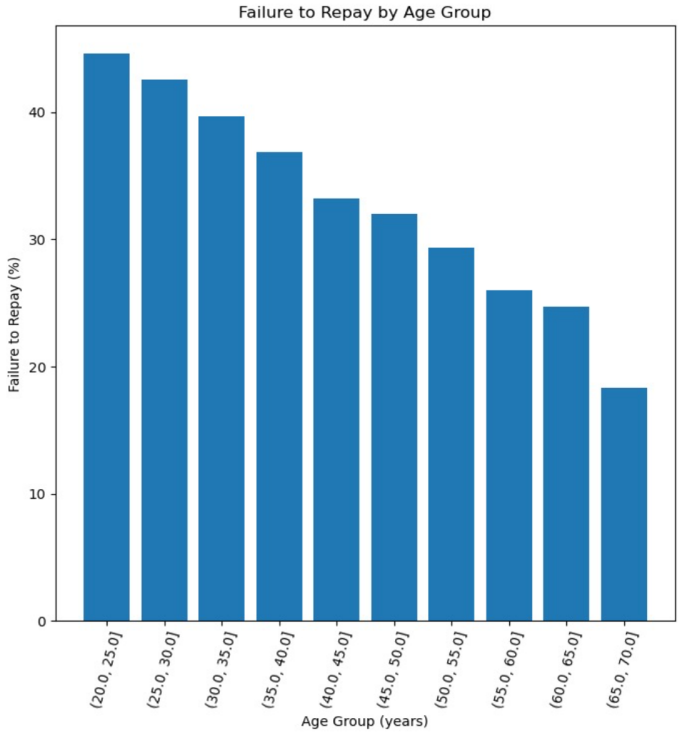
# Data Preprocessing

- The dataset has 42 columns in total. The columns which have more than 60% of values as NULL, are dropped.
- To handle remaining null values, columns with integers and float values are imputed with the mean of the column.
- In categorical columns the null values are replaced with Mode.

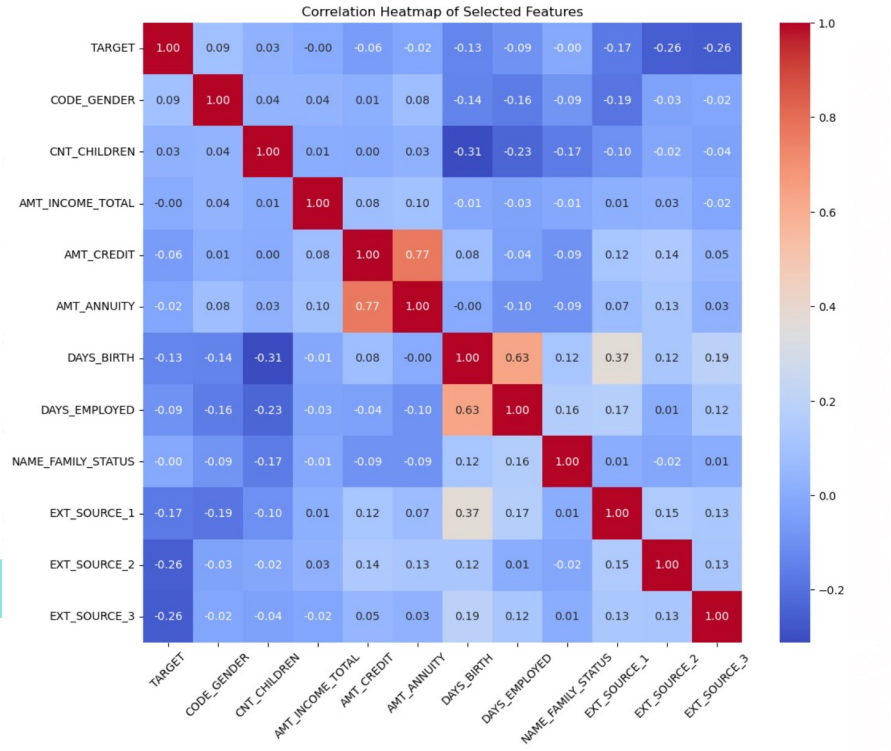
# Exploratory data analysis



# Exploratory data analysis



# Exploratory data analysis



# Feature Engineering

- In many datasets related to finance or demographics, age-related features are often recorded as the number of days relative to a reference point, typically the date of the loan application or another significant event.
- By recording age in terms of days relative to the current loan application date, it becomes easier to calculate the age of the individuals at the time of the loan application without needing to know their exact birth dates
- By multiplying the negative values by -1, the negative sign is removed, effectively making the values positive. Then, dividing by the number of days in a year converts the age from days to years, yielding the age of the individuals at the time of the loan application in years.

# Feature Engineering

- There are outliers in the feature ‘DAYS_EMPLOYED’. The outliers are first filled with ‘NULL’ values, and after preprocessing, these values are replaced with the mean.
- By using label encoder, we are converting all categorical values into numerical values.

Encoded values for column 'CODE_GENDER':
Original value: F, Encoded value: 0
Original value: M, Encoded value: 1
Original value: XNA, Encoded value: 2
Encoded values for column 'NAME_FAMILY_STATUS':
Original value: Civil marriage, Encoded value: 0
Original value: Married, Encoded value: 1
Original value: Separated, Encoded value: 2
Original value: Single / not married, Encoded value: 3
Original value: Unknown, Encoded value: 4
Original value: Widow, Encoded value: 5

# Model Development

- ## Logistic Regression:

  Logistic regression is particularly well-suited for binary classification tasks, where the goal is to predict one of two possible outcomes (e.g., ”yes” or ”no”, ”fraud” or ”not fraud”). Logistic regression assumes a linear relationship between the features and the target variable. While this assumption may not always hold true, logistic regression can still perform well if the relationship is approximately linear.

# Model Development

- **Random Forest:**

  Random forest is capable of capturing non-linear relationships and interaction effects between features, making it suitable for complex datasets with non-linear decision boundaries.

  Random forest is less prone to overfitting compared to some other algorithms, such as decision trees, due to the ensemble nature of the algorithm. By aggregating predictions from multiple decision trees, random forest reduces the variance of the model and improves generalization performance.

# Model Evaluation

- **Logistic Regression:**
  Accuracy for Logistic regression is around 67%

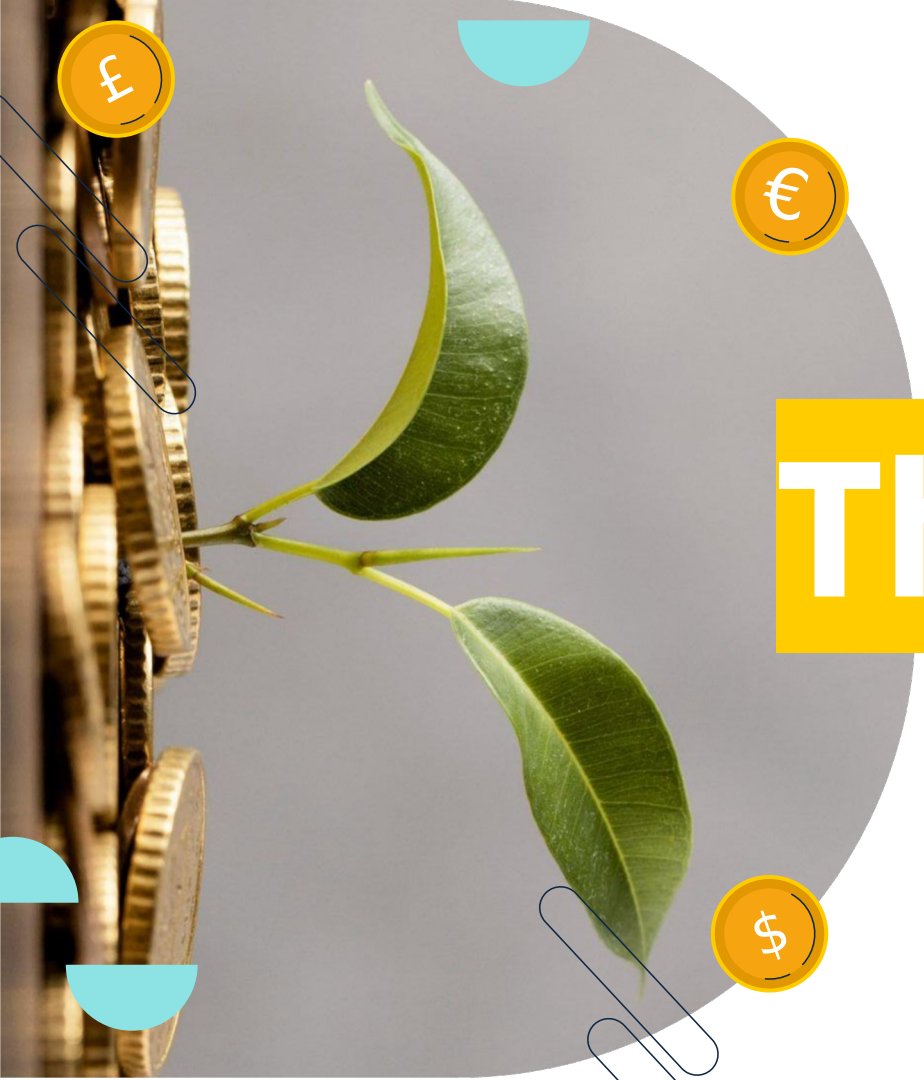
- **Random Forest:**
  Accuracy for Random forest is around 71%

- Logistic regression is a linear model, and random forest is a non-linear model, which makes it more suitable for complex datasets. Hence, we were able to get better accuracy in Random forest model.

# Conclusion

Overall, both models showed promise in loan default detection. Logistic regression offered transparency and interpretability, making it suitable for scenarios where understanding feature importance is essential. On the other hand, random forest excelled in predictive accuracy, especially in handling intricate datasets with non-linear relationships.

Future work may involve exploring ensemble methods or advanced algorithms to further improve model performance. Additionally, continuous monitoring and updating of the models with new data could ensure their effectiveness in detecting loan default over time.



# Thank You!