horizontal line

Project Summary

|  |  |
| --- | --- |
| Batch details | **PGP-DSE FT Apr’23 Gurgaon Batch** |
| Team members | **Mr. Samarth**  **Mr. Nishchay Chauhan**  **Mr. Pranshu Agarwal**  **Mr. Rahul Rana**  **Mr. Vikas Nehra**  **Mr. Shivam Singh** |
| Domain of Project | **Finance** |
| Proposed project title | **Home Credit Default Risk** |
| Group Number | **3** |
| Team Leader | **Mr. Samarth** |
| Mentor Name | **Mr. Mohit Sahu** |

Project Details

# OVERVIEW:

The purpose of this model is to develop and design an effective and efficient model for Home Credit Default Risk prediction in finance industry.

In order to deliver a positive loan experience, Home Credit makes use of a variety of alternative data like transactional information--to predict their clients' repayment abilities.

We will be using various statistical and machine learning methods to make these predictions.

Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their business to be successful. This necessitates the use of techniques like vintage analysis to establish these definitions based on historical data.

# Business problem statement (GOALS):

## Business Problem Understanding:

Credit Default Risk is a critical aspect of the financial industry, affecting lending decisions and risk management. A successful machine learning model can improve credit assessments, reduce defaults, and optimize credit allocation. However, it also raises the challenge of maintaining transparency and regulatory compliance, given the increasing use of complex machine learning models in credit risk assessment.

## Business Objective:

The task involves developing a machine learning model to assess the creditworthiness of Home Loan applicants. This is crucial for banks to decide whether to approve or reject Home loan applications. The task involves developing a machine learning model to assess the credit worthiness of the applicant. This is crucial for banks to decide whether to approve or reject the applications.

## Data Description:

* It includes all the previous records of credits provided by other financial institutions reported to credit bureau in the **Bureau table** consisting of **Columns :**

**SK\_ID\_CURR, SK\_ID\_BUREAU, CREDIT\_ACTIVE , CREDIT CURRENCY , DAYS\_CREDIT, CREDIT\_DAY\_OVERDUE [17 Columns]**

* Monthly Balances of previous credits in credit bureau in the **Bureau Balance table** consisting of **Columns :**

**SK\_ID\_BUREAU, MONTHS\_BALANCE, STATUS [3 Columns]**

* **Point of Sale Balances (POS Cash Balance)** table which provides insights about Monthly Balance snapshots of POS (Point Of Sales) and cash Loans that the applicant had with Home credit. It consists of **Columns:**

**SK\_ID\_PREV, SK\_ID\_CURR, MONTHS\_BALANCE, CNT\_INSTALLMENT, CNT\_INSTALLMENT\_FUTURE, NAME\_CONTRACT\_STATUS, SK\_DPD, SK\_DPD\_DEF**

**[8 Columns]**

* **Credit Card Balance** table which provides insights about the Monthly Balance of the previous credit cards that the applicant has with home credit. It consists of **Columns:**

**AMT\_CREDIT\_LIMIT\_ACTUAL, AMT\_DRAWINGS\_ATM\_CURRENT, AMT\_DRAWINGS\_\_CURRENT, AMT\_INST\_MIN\_REGULARITY, AMT\_PAYMENT\_CURRENT, AMT\_PAYMENT\_TOTAL\_CURRENT, AMT\_RECIVABLE\_PRINCIPAL, AMT\_RECIVABLE [23 Columns]**

* **Previous Application** table which provides insights about all previous applications for home credit Loans of clients who had applied for loan in the past. It consists of **Columns:**

**SK\_ID\_PREV, SK\_ID\_CURR, NAME\_CONTRACT\_TYPE, AMT\_ANNUITY, AMT\_APPLICATION, AMT\_CREDIT, AMT\_DOWN\_PAYMENT, AMT\_GOODS\_PRICE,**

**WEEKDAY\_APPR\_PROCESS\_START, HOURS\_APPR\_PROCESS\_START, RATE\_DOWN\_PAYMENT [37 Columns]**

* **Instalments Payments** table which provides insights about Repayment History for the previously observed distributed Loans from the organization with details of payments made and missed. It consists of **Columns:**

**SK\_ID\_PREV, SK\_ID\_CURR, NUM\_INSTALMENT\_VERSION, NUM\_INSTALMENT\_NUMBER, DAYS\_INSTALMENT, DAYS\_ENTRY\_PAYMENT,**

**AMT\_INSTALMENT, AMT\_PAYMENT [8 Columns]**

**Data Shape:**

|  |  |
| --- | --- |
| Number of tables | 9 |
| Number of combined columns | 120 |
| Number of rows | 307,512 |

1. **Data Dictionary of tables:**

* **Bureau :**

All client's previous credits provided by other financial institutions that were reported to Credit Bureau.

* **Bureau Balance :**

Monthly balances of previous credits in Credit Bureau.

This table has one row for each month of history of every previous credit reported to Credit Bureau.

* **POS Cash Balance :**

This table has one row for each month of history of cash balance for each client.

* **Credit Card Balance :**

This table has one row for credit limit for each loan applicant.

* **Previous Application :**

All previous applications for Home Credit loans of clients who had applied for loan in the past with application amount, down payment and credit details.

* **Instalments Payments :**

Repayment history for the previously disbursed loans from the organisation with details of payments made and missed.

From these tables we will select the relevant features using SQL and apply the below approach on it.

1. **Approach:**
2. **Data Analysis, Cleaning/ Preprocessing :** The pre-processing of the dataset before performing ML functions involves the following:
   1. **Descriptive Analysis:** Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. Measures of variability help communicate the spread of distribution by describing the shape and spread of the data set.
   2. **Inferential Analysis:** Validating the inferences which are found with the help of descriptive analysis (Graphs) with the help of respective statistical tests if needed.
   3. **Treating Outliers:** Checking and analyzing for presence of Outliers in the numerical columns and treating those outliers using the IQR method or any relevant method.
   4. **Treating Missing Values:** Null values in the variables must be treated with suitable methods.
   5. **Encoding Categorical Variables:** Since, machine learning models are based on Mathematical equations and we can intuitively understand that it would cause some problem if we can either keep the Categorical data by encoding the categorical variable or we can drop by checking whether we need the variable for further modelling process because we would only want numbers in the equations.
   6. **Dropping Unnecessary Columns:** We are removing the columns which do not contribute to the model building or the columns which are of less, or of no importance .
3. **Exploratory Data Analysis:**

Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modeling the data in order to apply Machine learning. Plotting in EDA consists of Bar-plot, Box plot, Scatter plot and many more using Univariate, Bivariate and Multivariate Analysis.

# Data Preparation:

* 1. **Scaling:** It helps to normalize the data within a particular range and as well as in speeding up the calculations in an algorithm.
  2. **Train and Test Split of the Data:** The data is split into train and test in required ratio.

# Model Building:

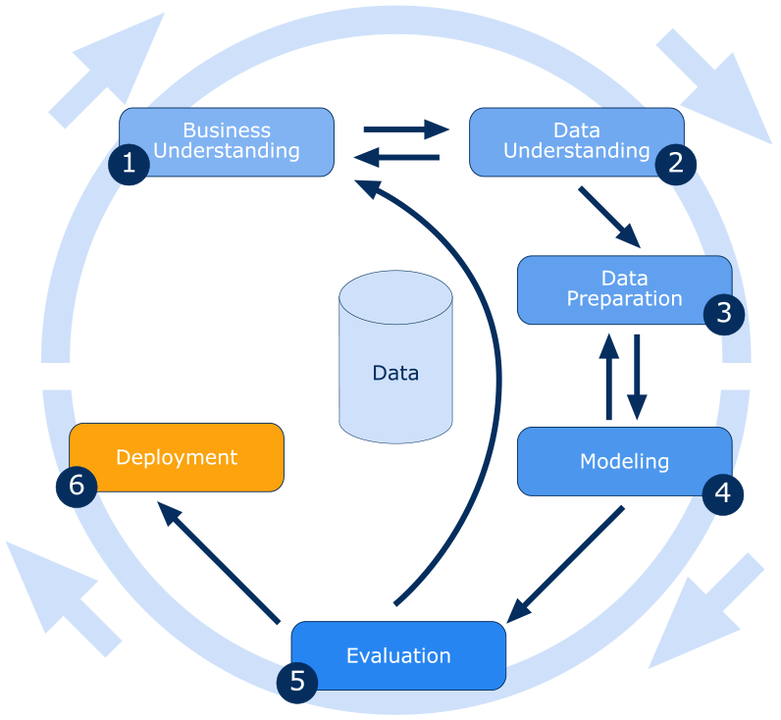
We will try to fit/train and test with below ML models and compare the performances.

1. Logistic Regression
2. Naive Bayes
3. KNN Classifier
4. Decision Tree
5. Random Forest
6. Bagging Classifier
7. Boosting Classifier
8. **Model Evaluation:**

Model Evaluation: Below metrics are used to evaluate the multi classification models performance.

1. Accuracy
2. Precision
3. Recall
4. F1-score
5. Confusion Matrix
6. RoC/AuC Score
7. **Model Deployment:**

In this step, we will save the best model(pickle) and come up with a method or function which takes patient data as input and re-admission status as output. We can try to productionize the deployment using flask.



1. **Conclusions:**

We will create a precise model which predicts the client repayment abilities. Status of whether the client will be able to repay the installments of the loan.

# REFERENCES:-

1. <https://www.kaggle.com/competitions/home-credit-default-risk/overview>
2. <https://www.homecredit.net/>
3. <https://www.linkedin.com/company/home-credit-international/>