**Project Summary**

|  |  |
| --- | --- |
| Batch details | DSE-Gurgaon-Aug22 |
| Team members | 1.Lakesh Kumar Padhy  2. Samarth Behl  3. Ravi Joon  4. Upasana Sehrawat  5. Ramesh kumar |
| Domain of Project | Finance |
| Proposed project title | Bank Loan Defaulter |
| Group Number | 4 |
| Team Leader | Upasana Sehrawat |
| Mentor Name | Jatin Bedi |

Date:

Signature of the Mentor:

Signature of the team Leader:

# 

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Sl NO** | **Topic** | **Page No** |
| 1 | Overview | 3 |
| 2 | Business problem goals | 3,4,5 |
| 3 | Topic survey in brief | 5 |
| 4 | Critical assessment of topic survey | 6 |
| 5 | Methodology to be followed | 7,8,9,10,11 |
| 6 | References | 11,12 |

# Page | 3

**OVERVIEW**

**Dataset name –** **Bank-Loan-Defaulters**

**Introduction to the problem / domain/background details**

The domain chosen for the capstone project is of finance (Banking) sector. A loan default occurs when a borrower takes money from a bank and does not repay the loan. People often default on loans due to various reasons. Borrowers who default on loans not only damage their credit but also risk being sued and having their wages garnished. Let’s take a look at the types of defaults that happen and understand the various reasons why people take loans and learn how predicting loan default will work.

This case study aims to give us an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that we have learnt in the EDA module, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Project Details

# Business problem statement (GOALS)

1. **Business Problem Understanding:** Preserving the financial health of customers is of great relevance for companies involved with the financial system. But you may wonder how to preserve the financial health of each client? well, the solution for this problem consists of calculating the probability of payment of each client according to some variables and doing some strategies to anticipate customer needs.

# Page | 4

The dataset contains information about credit applicants. Banks, globally, use this kind of dataset and type of informative data to create models to help in deciding on who to accept/refuse for a loan.

We’re dealing with a supervised binary classification problem. The goal is to train the best machine learning model to maximize the predictive capability of deeply understanding the past customer’s profile minimizing the risk of future loan defaults.

After all the exploratory data analysis, cleansing and dealing with all the anomalies we might (will) find along the way, the patterns of a good/bad applicant will be exposed to be learned by machine learning models.

1. **Business Objective:** By using more data and analyzing customer default probability, the Bank Defaulter systems are able to predict behavior, thereby helping lenders come to a more conclusive decision based on data. ML allows innovative work on data analysis wherein a bespoke solution is being offered to consumers.

Financial sectors and social lending platforms are actively investing on lending. But financial institutions might face huge capital loss if they approved the loan without having any prior assessment of default risk of particular person. So, to avoid such risk we can use data science, Machine learning algorithms.

objective of this project is to predict the probability of default on a given obligation, in this case Bank Loan. This will allow the generation of strategies that minimize the risk of deterioration of the client's financial health. Additionally, to facilitate the development of collection strategies, it is proposed to use Bagging and Boosting algorithms to find homogeneous segments within the population and thus provide differential treatment to each customer.

1. **Approach:** With the development of big data and data mining technology, international scholars have formed rich research results on ML in Bank Loan prediction and evaluation. Because the goal of the Loan management of financial institutions is to optimize the business performance and minimize the risk, decision rules should be established to make credit decisions.

In this case we will be using Supervised learning classification concept. We are going to use hit and trial method using several algorithms (adaboost, Gradient Boost, xgboost, Random Forest technique, Naïve baye’s, KNN, Decision tree, Logistic model and etc.) to build models under supervised learning.

# Page | 5

1. **Conclusions:** Traditionally, default risk is gauged using standard measurement tools such as credit scores etc. In this project we are trying to provide models to identify defaulters and freeze their account.

# TOPIC SURVEY IN BRIEF

1. **Problem understanding:**

The bank risk every time it issue loan. Problem here is that can it reliably predict who is likely to default? If so, the bank may be able to prevent the loss by providing the customer with alternative options (such as forbearance or debt consolidation, etc.).

This dataset contains information on Bank Loan default payments

**Current solution to the problem:** At present, bank check for salary and job stability to decide if Bank Loan can be issued to a person or not.

1. **Proposed solution to the problem:** With proposed solution, we can predict with high efficiency whether a person can repay Loan amount based on their repayment history then the bank may be able to reduce the loses by freezing such person’s credit card.

Few questions which can be answered by predicting defaulters are:

* + - How does the probability of default payment vary by categories of different demographic variables?
    - Which variables are the strongest predictors of default payment? Etc.

# Page | 6

# CRITICAL ASSESSMENT OF TOPIC SURVEY

1. **Find the key area, gaps identified in the topic survey where the project can add value to the customers and business.**

This project can add value to banking and finance sector. It will help bank and other finance sector which issue credit/loan/credit amounts without taking much risk of losing money or getting into loss. As we know, when banks goes under huge loss, it impact every common people and entire country economically. Therefore, this project aims to bridge this gap of uncertainty by utilizing a data-driven approach by using past data of Loan taken customers in conjunction with machine learning to predict whether or not a consumer will default on their Bank Loans**.**

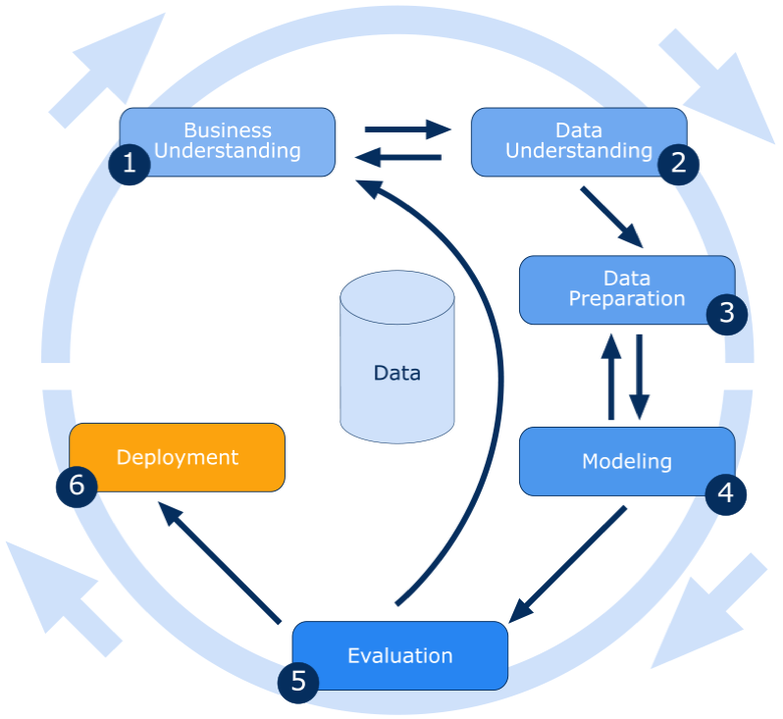
1. **What key gaps are you trying to solve?**

Currently the banks are considering only salary for issuing the Bank Loan. This increases the probability of occurrence of loss to the banks. This project aims to fill this gap, by considering other key variables and applying various machine Learning algorithms to predict, whether the consumer is going to be a defaulter or not.

Page | 7

# METHODOLOGY TO BE FOLLOWED

Machine Learning Modelling:



A machine learning model is an expression of an algorithm that combs through mountains of data to find patterns or make predictions. Fuelled by data, machine learning (ML) models are the mathematical engines of artificial intelligence.

**Machine learning modelling process includes the steps as shown in the figure above:**

**1.Business Understanding:** Machine learning (ML) models are built to solve complex business problems. In order to solve a problem we should know the right problem and the causes due to which such problem exists, which requires business understanding. Business Understanding

Page | **8**

refers to understanding the business and the problem which is to be solved and all the closely related attributes which affects the problem we are trying to solve.

**2.Data Understanding:** Data Understanding is the process of visualising, analysing and understanding the data. Various statistical methods like mean, standard deviation, correlation are used to find distributions and relations between the variables in the data, and how they effect our target variable or the problem we are trying to solve.

**3.Data Preparation:** Data preparation is a process which involves cleaning of data, treatment of outliers and missing values, scaling and normalising the data, this makes our data ready to be fed to various ML algorithms for training and testing.

**4.Modeling:** A machine learning model is an expression of an algorithm that combs through mountains of data to find patterns or make predictions. Fuelled by data, machine learning (ML) models are the mathematical engines of artificial intelligence.

**5.Evaluation:** Model Evaluation is the process through which we quantify the quality of a system's predictions. To do this, we measure the newly trained model performance on a new and independent dataset. This model will compare labeled data with it's own predictions.

**6.Deployement:** Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data. However, in this project we are not going to perform model deployment.

**PROBLEM STATEMENT AND MODEL ANALYSIS**

**1.Business Understanding:** Bank usually helps in flow of money by lending money to people who needs (with interest) and giving interest to people who has kept money in their bank. Bank sector is always at risk while lending money to anyone (whether as loan or credit card) and bank needs to reduce their risk while doing the same.

**2.Data Understanding:**

The dataset contains information about credit applicants. Banks, globally, use this kind of dataset and type of informative data to create models to help in deciding on who to accept/refuse for a loan.

We’re dealing with a supervised binary classification problem. The goal is to train the best machine learning model to maximize the predictive capability of deeply understanding the past customer’s profile minimizing the risk of future loan defaults.

After all the exploratory data analysis, cleansing and dealing with all the anomalies we might (will) find along the way, the patterns of a good/bad applicant will be exposed to be learned by machine learning models

Page | **9**

* **Independent Variables**

|  |  |
| --- | --- |
| Column | Description |
| SK\_ID\_CURR | ID of loan in our sample |
| NAME\_CONTRACT\_TYPE | Identification if loan is cash or revolving |
| CODE\_GENDER | Gender of the client |
| FLAG\_OWN\_CAR | Flag if the client owns a car |
| FLAG\_OWN\_REALTY | Flag if client owns a house or flat |
| CNT\_CHILDREN | Number of children the client has |
| AMT\_INCOME\_TOTAL | Income of the client |
| AMT\_CREDIT | Credit amount of the loan |
| AMT\_ANNUITY | Loan annuity |
| AMT\_GOODS\_PRICE | For consumer loans it is the price of the goods for which the loan is given |
| NAME\_TYPE\_SUITE | Who was accompanying client when he was applying for the loan |
| NAME\_INCOME\_TYPE | Clients income type (businessman, working, maternity leave,…) |
| NAME\_EDUCATION\_TYPE | Level of highest education the client achieved |
| NAME\_FAMILY\_STATUS | Family status of the client |
| NAME\_HOUSING\_TYPE | What is the housing situation of the client (renting, living with parents, ...) |
| REGION\_POPULATION\_RELATIVE | Normalized population of region where client lives (higher number means the client lives in more populated region) |
| DAYS\_BIRTH | Client's age in days at the time of application |
| DAYS\_EMPLOYED | How many days before the application the person started current employment |
| DAYS\_REGISTRATION | How many days before the application did client change his registration |
| DAYS\_ID\_PUBLISH | How many days before the application did client change the identity document with which he applied for the loan |
| OWN\_CAR\_AGE | Age of client's car |
| FLAG\_MOBIL | Did client provide mobile phone (1=YES, 0=NO) |
| FLAG\_EMP\_PHONE | Did client provide work phone (1=YES, 0=NO) |
| FLAG\_WORK\_PHONE | Did client provide home phone (1=YES, 0=NO) |
| FLAG\_CONT\_MOBILE | Was mobile phone reachable (1=YES, 0=NO) |
| FLAG\_PHONE | Did client provide home phone (1=YES, 0=NO) |
| FLAG\_EMAIL | Did client provide email (1=YES, 0=NO) |
| OCCUPATION\_TYPE | What kind of occupation does the client have |
| CNT\_FAM\_MEMBERS | How many family members does client have |
| REGION\_RATING\_CLIENT | Our rating of the region where client lives (1,2,3) |
| REGION\_RATING\_CLIENT\_W\_CITY | Our rating of the region where client lives with taking city into account (1,2,3) |
| WEEKDAY\_APPR\_PROCESS\_START | On which day of the week did the client apply for the loan |
| HOUR\_APPR\_PROCESS\_START | Approximately at what hour did the client apply for the loan |
| REG\_REGION\_NOT\_LIVE\_REGION | Flag if client's permanent address does not match contact address (1=different, 0=same, at region level) |
| REG\_REGION\_NOT\_WORK\_REGION | Flag if client's permanent address does not match work address (1=different, 0=same, at region level) |
| LIVE\_REGION\_NOT\_WORK\_REGION | Flag if client's contact address does not match work address (1=different, 0=same, at region level) |
| REG\_CITY\_NOT\_LIVE\_CITY | Flag if client's permanent address does not match contact address (1=different, 0=same, at city level) |
| REG\_CITY\_NOT\_WORK\_CITY | Flag if client's permanent address does not match work address (1=different, 0=same, at city level) |
| LIVE\_CITY\_NOT\_WORK\_CITY | Flag if client's contact address does not match work address (1=different, 0=same, at city level) |
| ORGANIZATION\_TYPE | Type of organization where client works |
| EXT\_SOURCE\_1 | Normalized score from external data source |
| EXT\_SOURCE\_2 | Normalized score from external data source |
| EXT\_SOURCE\_3 | Normalized score from external data source |
| APARTMENTS\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| BASEMENTAREA\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BEGINEXPLUATATION\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BUILD\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| COMMONAREA\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ELEVATORS\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ENTRANCES\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMAX\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMIN\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LANDAREA\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAPARTMENTS\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAREA\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAPARTMENTS\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAREA\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| APARTMENTS\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| BASEMENTAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BEGINEXPLUATATION\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BUILD\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| COMMONAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ELEVATORS\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ENTRANCES\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMAX\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMIN\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LANDAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAPARTMENTS\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAPARTMENTS\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| APARTMENTS\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| BASEMENTAREA\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BEGINEXPLUATATION\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BUILD\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| COMMONAREA\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ELEVATORS\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ENTRANCES\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMAX\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMIN\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LANDAREA\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAPARTMENTS\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAREA\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAPARTMENTS\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAREA\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FONDKAPREMONT\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| HOUSETYPE\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| TOTALAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| WALLSMATERIAL\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| EMERGENCYSTATE\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings with observable 30 DPD (days past due) default |
| DEF\_30\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings defaulted on 30 DPD (days past due) |
| OBS\_60\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings with observable 60 DPD (days past due) default |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings defaulted on 60 (days past due) DPD |
| DAYS\_LAST\_PHONE\_CHANGE | How many days before application did client change phone |
| FLAG\_DOCUMENT\_2 | Did client provide document 2 |
| FLAG\_DOCUMENT\_3 | Did client provide document 3 |
| FLAG\_DOCUMENT\_4 | Did client provide document 4 |
| FLAG\_DOCUMENT\_5 | Did client provide document 5 |
| FLAG\_DOCUMENT\_6 | Did client provide document 6 |
| FLAG\_DOCUMENT\_7 | Did client provide document 7 |
| FLAG\_DOCUMENT\_8 | Did client provide document 8 |
| FLAG\_DOCUMENT\_9 | Did client provide document 9 |
| FLAG\_DOCUMENT\_10 | Did client provide document 10 |
| FLAG\_DOCUMENT\_11 | Did client provide document 11 |
| FLAG\_DOCUMENT\_12 | Did client provide document 12 |
| FLAG\_DOCUMENT\_13 | Did client provide document 13 |
| FLAG\_DOCUMENT\_14 | Did client provide document 14 |
| FLAG\_DOCUMENT\_15 | Did client provide document 15 |
| FLAG\_DOCUMENT\_16 | Did client provide document 16 |
| FLAG\_DOCUMENT\_17 | Did client provide document 17 |
| FLAG\_DOCUMENT\_18 | Did client provide document 18 |
| FLAG\_DOCUMENT\_19 | Did client provide document 19 |
| FLAG\_DOCUMENT\_20 | Did client provide document 20 |
| FLAG\_DOCUMENT\_21 | Did client provide document 21 |
| AMT\_REQ\_CREDIT\_BUREAU\_HOUR | Number of enquiries to Credit Bureau about the client one hour before application |
| AMT\_REQ\_CREDIT\_BUREAU\_DAY | Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application) |
| AMT\_REQ\_CREDIT\_BUREAU\_WEEK | Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application) |
| AMT\_REQ\_CREDIT\_BUREAU\_MON | Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application) |
| AMT\_REQ\_CREDIT\_BUREAU\_QRT | Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application) |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application) |
| SK\_ID\_PREV | ID of previous credit in Home credit related to loan in our sample. (One loan in our sample can have 0,1,2 or more previous loan applications in Home Credit, previous application could, but not necessarily have to lead to credit) |
| SK\_ID\_CURR | ID of loan in our sample |
| NAME\_CONTRACT\_TYPE | Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application |
| AMT\_ANNUITY | Annuity of previous application |
| AMT\_APPLICATION | For how much credit did client ask on the previous application |
| AMT\_CREDIT | Final credit amount on the previous application. This differs from AMT\_APPLICATION in a way that the AMT\_APPLICATION is the amount for which the client initially applied for, but during our approval process he could have received different amount - AMT\_CREDIT |
| AMT\_DOWN\_PAYMENT | Down payment on the previous application |
| AMT\_GOODS\_PRICE | Goods price of good that client asked for (if applicable) on the previous application |
| WEEKDAY\_APPR\_PROCESS\_START | On which day of the week did the client apply for previous application |
| HOUR\_APPR\_PROCESS\_START | Approximately at what day hour did the client apply for the previous application |
| FLAG\_LAST\_APPL\_PER\_CONTRACT | Flag if it was last application for the previous contract. Sometimes by mistake of client or our clerk there could be more applications for one single contract |
| NFLAG\_LAST\_APPL\_IN\_DAY | Flag if the application was the last application per day of the client. Sometimes clients apply for more applications a day. Rarely it could also be error in our system that one application is in the database twice |
| NFLAG\_MICRO\_CASH | Flag Micro finance loan |
| RATE\_DOWN\_PAYMENT | Down payment rate normalized on previous credit |
| RATE\_INTEREST\_PRIMARY | Interest rate normalized on previous credit |
| RATE\_INTEREST\_PRIVILEGED | Interest rate normalized on previous credit |
| NAME\_CASH\_LOAN\_PURPOSE | Purpose of the cash loan |
| NAME\_CONTRACT\_STATUS | Contract status (approved, cancelled, ...) of previous application |
| DAYS\_DECISION | Relative to current application when was the decision about previous application made |
| NAME\_PAYMENT\_TYPE | Payment method that client chose to pay for the previous application |
| CODE\_REJECT\_REASON | Why was the previous application rejected |
| NAME\_TYPE\_SUITE | Who accompanied client when applying for the previous application |
| NAME\_CLIENT\_TYPE | Was the client old or new client when applying for the previous application |
| NAME\_GOODS\_CATEGORY | What kind of goods did the client apply for in the previous application |
| NAME\_PORTFOLIO | Was the previous application for CASH, POS, CAR, … |
| NAME\_PRODUCT\_TYPE | Was the previous application x-sell o walk-in |
| CHANNEL\_TYPE | Through which channel we acquired the client on the previous application |
| SELLERPLACE\_AREA | Selling area of seller place of the previous application |
| NAME\_SELLER\_INDUSTRY | The industry of the seller |
| CNT\_PAYMENT | Term of previous credit at application of the previous application |
| NAME\_YIELD\_GROUP | Grouped interest rate into small medium and high of the previous application |
| PRODUCT\_COMBINATION | Detailed product combination of the previous application |
| DAYS\_FIRST\_DRAWING | Relative to application date of current application when was the first disbursement of the previous application |
| DAYS\_FIRST\_DUE | Relative to application date of current application when was the first due supposed to be of the previous application |
| DAYS\_LAST\_DUE\_1ST\_VERSION | Relative to application date of current application when was the first due of the previous application |
| DAYS\_LAST\_DUE | Relative to application date of current application when was the last due date of the previous application |
| DAYS\_TERMINATION | Relative to application date of current application when was the expected termination of the previous application |
| NFLAG\_INSURED\_ON\_APPROVAL | Did the client requested insurance during the previous application |

* **Target Variable**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Application\_data- | TARGET- | Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases) |

Page | **10**

**3.Data Preparation:**

●Cleaning data- check the datatypes as per the data dictionary

● Setting Target (Defaulter) as target variable. Shrink the table by dropping irrelevant variables.

● Exploring data through visualizations- Diagram/ Each Value analysis (Univariate analysis

& Describe)

● Missing Values / Outliers treatment

● Scaling of numerical variables

● Encode categorical variables

● Transformation (if any)

**4.Modeling:** After preparing the dataset, we are going to build generalized models under different supervised learning classification algorithms. Such as adaboost, Gradient Boost, xgboost, Random Forest technique, Naïve baye’s, KNN, Decision tree, Logistic model etc. Then we are going to evaluate these various models and select the best model.

**5.Evaluation:**

For model evaluation, we are going to use multiple evaluation methods, such as:

### Classification accuracy

### Confusion matrix

### Precision and recall

### F1 score

### Sensitivity and specificity

### ROC curve and AUC

**Assumptions and Limitations**

Page |

REFERENCES

**kaggle datasets download -d gauravduttakiit/loan-defaulter**

**Declaration: This is to declare that the dataset that we are using for our capstone project does not have any relevant legality associated to it** **and can be used to showcase the work we do on it as a presentation in Great Learning.**

### 