**Credit Risk Modeling Case Study**

**Problem Statement:**

ABC Finance Company is finding it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for ABC finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants are capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

* If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
* If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

**The bank wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default**

**Now, what should you do?**

Identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

**Datasets:**

1. **'application\_data.csv'** contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

2. **'previous\_application.csv'** contains information about the client’s previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

**Datasets uploaded here**: [**https://drive.google.com/drive/folders/1IWQBgaO8e\_VDi6GBnluLJTlOQFSTPtcL?usp=sharing**](https://drive.google.com/drive/folders/1IWQBgaO8e_VDi6GBnluLJTlOQFSTPtcL?usp=sharing)

Note: You can refer to **Columns description** file to learn about the columns available in the datasets.

**How to begin?**

1. Start by importing the 'application\_train.csv'.
2. Check the structure of the data (shape, info, describe).

3. Data Quality Check and Missing values

● Find the percentage of missing values for all the columns.

● Remove columns with high missing percentage.

● For columns which has less percentage (around 13% or so), you need to check what will be the best metric to impute the missing values? Like if the column you are checking is a categorica column check, which category you can use to fill the nulls. For others check does mean or median can be imputed or not. Other cases may be imputing with 0. You need to do this task for some variables and not for all, say 5.

1. Check the datatypes of all the columns and change the datatype if required.
2. For numerical columns check for outliers and report them for at-least 3 variables. Treat them and analyze it.
3. Binning of continuous variables. Check if you need to bin any variable in different

categories. Do this for one or two columns.

**Analysis Steps**

1. Check the Imbalance percentage. No balancing technique required.
2. Divide the data into two sets, i.e. Target=1 and Target=0.
3. Perform univariate analysis for categorical variables for both 0 and 1. Compare the target variable across categories of categorical variables.
4. Find correlation for numerical columns for both the cases, i.e. 0 and 1.
5. Check the variables with highest correlation are the same in both the files or not?
6. Perform univariate for numerical variables for both 0 and 1. Compared the target variable across categories of continuous variables.
7. Perform bivariate analysis for numerical variables for both 0 and 1.
8. **Read “Previous Application” data.**
9. Merge the files
10. Perform univariate and bivariate analysis to find some pattern.