###### **House Loan Data Analysis**

**Description-:**

For safe and secure lending experience, it's important to analyse the past data. In this project, you have to build a deep learning model to predict the chance of default for future loans using the historical data. As you will see, this dataset is highly imbalanced and includes a lot of features that make this problem more challenging.

**Objective:** Create a model that predicts whether or not an applicant will be able to repay a loan using historical data.

**Domain:** Finance

**Analysis to be done**: Perform data preprocessing and build a deep learning prediction model.

**Steps to be done:**

⦁ Load the dataset that is given to you:-

* Data is loaded in below path:

df=pd.read\_csv("/content/drive/MyDrive/data\_project\_folder/loan\_data\_project.csv")

* data have row – 307511 and 122 -columns
* first drop SK\_ID\_CURR column due to having only serial number.

⦁ Check for null values in the dataset:

* Null value checked and found the maximum null value in OWN\_CAR\_AGE with 66.83 % null data.
* Containing More than 39% of null values in columns removed these are:
* 'Ext\_source\_1',
* 'Own\_car\_age',
* 'Commonarea\_avg',
* 'Floorsmin\_avg',
* 'Livingapartments\_avg',
* 'Commonarea\_mode',
* 'Nonlivingapartments\_avg',
* 'Floorsmin\_mode',
* 'Livingapartments\_mode',
* 'Nonlivingapartments\_mode',
* 'Commonarea\_medi',
* 'Floorsmin\_medi',
* 'Livingapartments\_medi',
* 'Nonlivingapartments\_medi'
* Rest of the columns containing null values is imputed by the mean and median of data.

⦁ Print percentage of default to payer of the dataset for the TARGET column:

* % of default to payer : 7.22%

⦁ Balance the dataset if the data is imbalanced:

* Imbalanced data for Target dependent columns containing:

Class no. of data

1. 131840
2. 9514

* After the SMOTE function use for balance data:

Class no. of data

1. 105472
2. 105472

⦁ Plot the balanced data or imbalanced data:

* Data plot by using pie chart

⦁ Encode the columns that is required for the model :-

* Many columns are string and it encode by using label encoding.
* No. of columns are :-
* 'NAME\_CONTRACT\_TYPE',
* 'CODE\_GENDER',
* 'FLAG\_OWN\_CAR',
* 'FLAG\_OWN\_REALTY',
* 'NAME\_TYPE\_SUITE',
* 'NAME\_INCOME\_TYPE',
* 'NAME\_EDUCATION\_TYPE',
* 'NAME\_FAMILY\_STATUS',
* 'NAME\_HOUSING\_TYPE',
* 'OCCUPATION\_TYPE',
* 'WEEKDAY\_APPR\_PROCESS\_START',
* 'ORGANIZATION\_TYPE',
* 'FONDKAPREMONT\_MODE',
* 'HOUSETYPE\_MODE',
* 'WALLSMATERIAL\_MODE',
* 'EMERGENCYSTATE\_MODE'

⦁ Calculate Sensitivity as a metrices :

Recall are:

No. of class –

1. 92.0
2. 20.0

⦁ Calculate area under receiver operating characteristics curve:

Roc cover calculated.

* build a deep learning model to predict the chance of default for future loans using the historical data.

Data build by deep learning model with by using activation function : Relu & sigmoid for dense layer ,optimizer – Adam, loss – “binary\_crossentropy”, epochs- 5