**Calculating credit worthiness for rural India**

**AIM:** The main of purpose of this project is to predict the loan amount could be given for rural customers based on other features and this can be done by descriptive analysis, EDA etc...

**WORKFLOWS:**

* **Understanding the Data**
* **Pre-processing and descriptive Analysis**
* **Model Building**
* **Parameter Tuning**

****

**UNDERSTANDING THE DATA:**

In this section, will look at the structure of the dataset, will be checking the features which are present in a given dataset and then will look at their data types.

An ideal dataset would be complete, with valid values for every observation. However, in reality, you will come across many “NULL” or “NaN” values and this will be handled in EDA (exploratory data analysis).

An outlier is a data point that is significantly different from other observations. Visualize the data using scatterplots, histograms and box and whisker plots and look for outliers.

**PRE-PROCESSING AND DESCRIPTIVE ANALYSIS:**

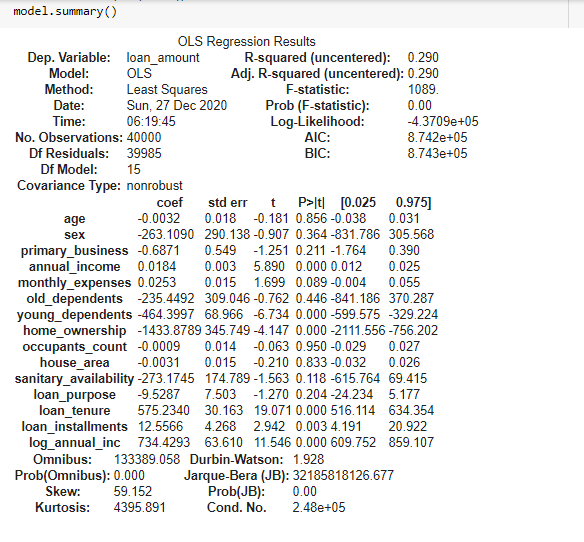
The data might contain redundancy features that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency, Drop the feature if it is not useful for predicting the loan amount.

Used descriptive analysis to summarize each features in dataset and visualize the data, in order to get an idea about each features. By using simple graphics analysis

**Major observation from the data:**

* Applicants who are female tends to have more applicant income whereas applicant who are male have least applicant income.
* Loan Amount is linearly dependent on Applicant income
* Most of the customer's taken loan installments between 12 and 50 months
* Most of the applicants( approx. 96%) taken loan tenure length as 12 months, and also this features is having constant values, will remove this feature in feature selection

Derived few feature based on target variable and modify the feature to make it useful for modeling, if necessary.



The output above shows that, when the other variables remain constant, if we compare two applicants whose 'Loan amount' differ by one unit, the applicant with higher 'loan amount' will, on average, have 0.75 units higher 'Income'.

Using the P>|t| result, we can infer that the variables 'loan amount' and **'Loan installments’, ‘log annual income’, ‘loan tenure’** are the **statistically significant variables,** as their p-value is less than 0.05.

DATA PRE-PROCESSING

BUILDING MODEL

MODEL TUNING

RAW DATA

Steps in knowledge extraction

**MODEL BUILDING:**

Several models have been used to predict loan amount such as Linear Regression, Decision tree, KNN, Random Forest, Lasso based on the model performance I have chosen the model and here I used RMSE to check the fitness of the model.

**PARAMETER TUNING:**

Here I used Grid Search to get best parameters, Grid Search passes all combinations of hyperparameters one by one into the model and check the result. Finally, it gives us the set of hyperparameters which gives the best result after passing in the model.