Problem Statement:

The Bank wants to understand the driving factors (or driving variables) behind loan defaults, i.e. the variables which are strong indicators for loan defaults. The company can utilise this knowledge to scrutinise genuine customers to whom they can lend loans. It will be helpful for portfolio and risk assessment.

To Achieve the model following methods were followed:

1. Cleaning The Data:

Originally we had two csv to work with, which was application_data.csv which had all the current loan application data, and the next data we had previous_application.csv which had all the data of the loan applicants previously applied. There were around 122 columns in our Loan Application data and around 37 columns in our previous application data. Out of which we took 19 columns in our current loan applicant and took 10 columns out of previous loan applicant data, For column reduction we reduced 20 documents columns into one single merged document column. There were around <0 % Not Applicable and Not Available data which was not dropped as it was not contributing a significant amount in our data.

2. EDA:

Exploratory data analysis was done on the dataset, which enabled us to understand, prepare and draw conclusions on the data.

In the given problem there were two datasets. One described the customers' information who were applying for the loan and the other dataset gave the details from the bank's end about the same customers. We had to see the factors which affected the bank's decision to approve or reject a loan application. For this we first identified the outliers in the dataset that do affect our data significantly in certain columns. Next we identified data imbalance in different columns and visualised it with the help of pie charts and bar graphs.

Then we did bivariate analysis on different columns and found the correlation and observed our data using HeatMap. We also merged the two datasets to make more clear and detailed observations.We Analysed our data to check what kind of data got loan approval and what factors lead to rejection of a loan Approved.

4. Model Building:

Now for model building, we have all continuous data and categorical data so the best kind of model for such kind of prediction would be *Binary Logistic Regression.* Here our Response would be binary whether the loan should be Approved or Refused.

5. Conclusion:

Target Variable:

Loan Status :

Top-5 Major variables to consider for loan prediction:

1. Target
2. Documents Required
3. Loan Repayment
4. Amount Annuity
5. Amount Income Total