**BFSI Case Study**

**Problem Statement:**

A close-up of a graph and a person

Description automatically generated

**Data Understanding:**

Two data sets are presented

1. Demographic Data and
2. Credit Bureau data

Demographic Data: Which is also called Application data, This is the information obtained provided from the applicants at the time of the credit card application. It also contains customer information.

Credit Bureau Data: This is taken from the credit bureau and contains variables such as Outstanding balance etc.

**Data Preparation:**

• Checking all the columns for missing values

• Checking all the columns for Nas

• Checking the necessary columns for duplicate values

• Outliers detection using the quartiles and boxplots

**Outlier Detection:**

* There are few outliers in Age, having zero and negative values
* Other columns do not contain outliers, only the certain valid values lie outside the 3rdquartile range

**Data Cleaning:**

• Three duplicate application IDs were removed

• Negative and Zero values removed from Age, Income column

• Removed the rows where Gender, Marital Status, Profession was not mentioned

**Conclusion:**

Following seems to be the contributing factors (after basic logistic model):

• No of times 30 DPD or worse in last 6 months

• No of Inquiries in last 12 months excluding home auto loans.

• Presence of open home loan

• Total No of Trades

• No of dependents

Considering the classification problem of dividing the applicants in two categories based on the performance tag – Defaulters and Non Defaulters,

we can use two different models.

• Logistic Regression • Random Forest

• Not taking SVM into account as the amount of data is huge

• Segregating the data into test and train sets •

Will be using the drill down approach to remove the non significant variables on the basis of VIF and p-values.

• In random forest we need to vary the number of trees, min number of buckets and min number of leaves in a node.