SCMA 648 Business Data Analytics

Assignment 1 - Importing Data and Reading

Dataset

BFSI is one of the sectors these days which generates the highest revenues and there are a lot of things which can be added or implemented in this domain and still there are challenges in this domain, which can be fixed by improving them one by one. The main aim is to reduce the issues faced by lenders and borrowers. Lending Club is one such organization that works on helping everyone on the same where they try to reduce the amount paid by the borrower and increase the amount for the lender.

This dataset has 150 features such as id, loan amount, member id, loan status, etc. This dataset has different data types like character, factor, numeric, and date. Different features associated with the finances of an individual have been considered. This can help a financial organization to improve its decision-making with respect to lending money.

Objective

To determine the frequency of candidates corresponding to their loan status.

Steps

1. Set the working directory to source file location using session > set working directory > to source file location. The first line of the file has been skipped using “skip = 1” as it is a text line. The features of the data can be found in the next line.

2. As mentioned earlier, the dataset has different data types. These have been declared concerning their data types. The variables have been named “character_columns” and “factor_columns” based on their data types. Based on “character_columns” and “factor_columns” the datatype of the dataset was created as the default is a character, a list of classes of was created with factor, and character was created. This was assigned with the respective datatype based on the features available in the variables created. Most of the attributes were numeric.

3. Now the actual data has to be read. The first line being text line has been skipped using “skip = 1”. The size of the dataset is large so “nrows” has been set to 122701. sep = “,” has been used as it is a csv file. “colClasses” is a default parameter which is a vector of classes. The variable “my_col_classes” has been assigned to it. The dataset consists of attribute names, so “header = TRUE”.

4. The feature “loan_status” has various categories under it. To determine the frequency of each category, a table has been created. This table can be saved as a text file using “write.table”. The datatypes for the table created are numeric, therefore “quote = FALSE” and there are no indices given, therefore “row.names = FALSE”.

5. Using the text file, the values have been extracted and copied in a word to make it aesthetic. The copied content was converted to a table using insert > convert text to table. Finally, a table design has been added.

   The frequency corresponding to each category can be seen below:

Loan Status	Frequency
Charged Off	18803
Current	40844
Default	9
Fully Paid	61673
In Grace Period	623
Late (16-30 days)	128
Late (31-120 days)	621

There are seven categories under the feature “loan_status”. Approximately 50% of the candidates (61673) have paid the loan in entirety. 33% of the candidates (40844) fall under the current category of loan status. Approximately 15% of the candidates (18803) couldn’t pay the loan amount and have been charged off. 623 candidates have been given the grace period to pay the loan amount. A total of 749 candidates are late for the payment either by 16 to 30 or 31 to 120 days. There are 9 loan defaulters as seen. The organization can focus on reducing the number of charged off and default candidates by rigorous background verification, regular follow-ups, strict supervision and policies.