# Multinomial Regression (Module -10)

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: CHINNI GANESH Batch ID:** 23/09/21

**Topic: Multinomial Regression.**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

**3.2 Outlier Imputation**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options)**
   2. **Perform Multinomial Regression Modl.**
   3. **Train and Test the data and compare accuracies by Confusion Matrix, plot ROC AUC curve.**

**5.4 Briefly explain the model output in the documentation.**

1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

# Note:

The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the model building (elaborating on steps mentioned above)

**Problem Statement:**

A University would like to effectively classify their student based on the Program they are enrolled into, perform multinomial regression on the given dataset and provide insights in the documentation.

**prog**: is a categorical variable indicating what type of program a student is in: “General” (1), “Academic” (2), or “Vocational” (3)

**Ses**: is a categorical variable indicating someone’s socioeconomic class: “Low” (1), “Middle” (2), and “High” (3)

**read, write, math, science** are their scores on different tests

**honors**: Whether they have enrolled or not

**Sol**:

**Business Objective**: classify the students based on the program they enrolled using multinomial regression.

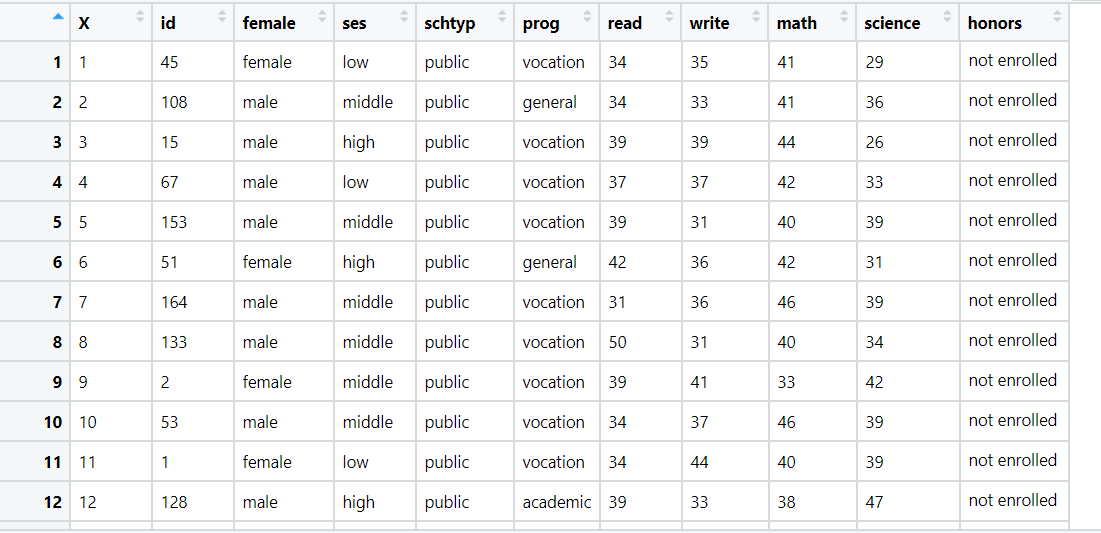
**Constraints**: Lack of analysis of the student’s data.

**Data Types:**  Given data and its data types are as follows

**Data Pre-Processing:**  all the variables of the given data is used for doing the analysis except the I’d of the student since it is not useful for the analyzing the data.

Some of the variables in the given data are Non numeric and the same data is converted into numeric data for the analysis of the data.

**Multinomial Regression:** for doing the analysis the output variable taken is program type of the student and the model is made for the students data which segregates the data with the predicted values on the data.





**Problem statement:**

You work for a consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

• If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

• If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

The data given below contains the information about past loan applicants and whether they ‘defaulted’ or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

When a person applies for a loan, there are two types of decisions that could be taken by the company:

1. Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

o Fully paid: Applicant has fully paid the loan (the principal and the interest rate)

o Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

o Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

2. Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through an online interface.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Perform Multinomial regression on the dataset in which loan\_status is the output (Y) variable and it has three levels in it.

**Sol:**

**Business Objective:** to predict whether the customer will default the loan or not using Multinomial regression model.

**Constraints:**  lack of analysis of the customer’s loan data**.**

**Data Types:** There are 111 variables in the given data, all the variables are not used for doing the analysis on the data. The following are the selected variables from the given data to do the analysis.

**Data Pre-Processing:** From the complete data given only 18 variables are selected for doing the analysis which are mentioned in the above data types table.

Target variable taken for doing the analysis is loan status which is renamed as default and the same is converted into numeric data.

Earliest\_credit\_line and issue\_d are the variable where the data is given in months and years the same is converted into date format for further analysis of the data.

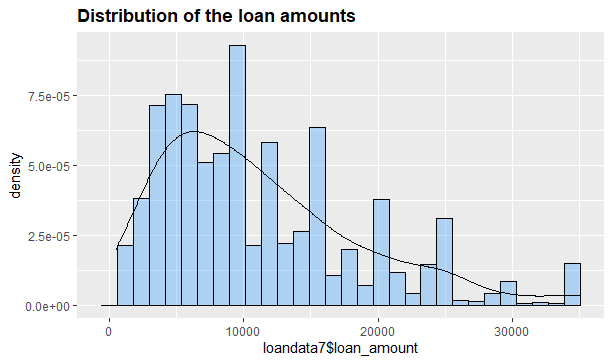
Since dates cannot be used directly for the analysis a new variable is created which is the difference of the earliest\_credit\_line and issue\_d variables known as time history.

Revol\_util id the column which is utilization percentage of the customer given in % format, this cannot be used directly for doing the analysis so the % is removed from the data so that it will be used for the analysis.

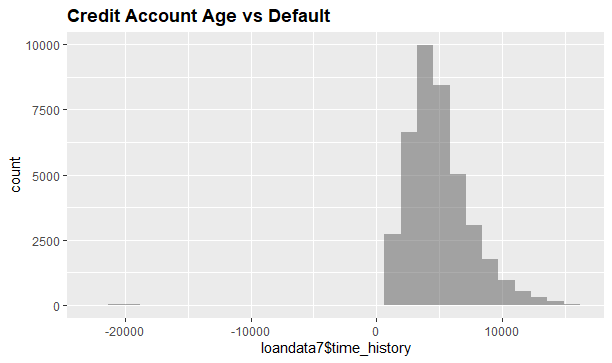
All the non numeric data is converted into numeric data so that it will be used for the analysis of the data.

**Exploratory Data Analysis:**

Histogram is plotted for the loan amount in the given data to see how it is distributed. The following histogram shows the distribution of loan amount of the data.



Histogram for the new variable created which is the difference of earliest\_credit\_line and issue\_d known as time history is shown as follows.



**Multinomial Regression:**

Cleaned data is spitted into training and testing data for doing the analysis. Cleaned data is then used for applying the multinomial regression model which predicts whether the customer defaults the loan or not. Accuracy of the model made is 99.9%.

A screenshot of a cell phone

Description automatically generated