# Multinomial Regression

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: MD SABIULLAH Batch ID:**  PDS 07052024

**Topic: Multinomial Regression.**

**Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered as correct submission.**

**2. Ensure that you submit your assignments correctly and in full. Resubmission is not allowed.**

**3. Post the submission you can evaluate your work by referring to keys provided. (will be available only post the submission).**

**Hints:**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

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

**Data Dictionary :**

**Data Dictionary :**

**id: A unique LC assigned ID for the loan listing**

**member\_id: A unique LC assigned ID for the borrower member**

**loan\_amnt: The listed amount of the loan applied for by the borrower**

**funded\_amnt: The total amount committed to that loan**

**funded\_amnt\_inv: The total amount committed by investors for that loan**

**term: The number of payments on the loan. Values are in months and can be either 36 or 60**

**int\_rate: Interest rate on the loan**

**installment: The monthly payment owed by the borrower if the loan originates**

**grade: LC assigned loan grade**

**sub\_grade: LC assigned loan subgrade**

**emp\_title: The job title supplied by the borrower when applying for the loan**

**emp\_length: Employment length in years**

**home\_ownership: The home ownership status provided by the borrower during registration or obtained from the credit report**

**annual\_inc: The self-reported annual income provided by the borrower during registration**

**verification\_status: Indicates if income was verified by LC, not verified, or if the income source was verified**

**issue\_d: The month which the loan was funded**

**loan\_status: Current status of the loan**

**pymnt\_plan: Indicates if a payment plan has been put in place for the loan**

**url: URL for the LC page with listing data**

**desc: Loan description provided by the borrower**

**purpose: A category provided by the borrower for the loan request**

**title: The loan title provided by the borrower**

**zip\_code: The first 3 numbers of the zip code provided by the borrower in the loan application**

**addr\_state: The state provided by the borrower in the loan application**

**dti: A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income**

**delinq\_2yrs: The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years**

**earliest\_cr\_line: The month the borrower's earliest reported credit line was opened**

**inq\_last\_6mths: The number of inquiries in past 6 months (excluding auto and mortgage inquiries)**

**mths\_since\_last\_delinq: The number of months since the borrower's last delinquency**

**mths\_since\_last\_record: The number of months since the last public record**

**open\_acc: The number of open credit lines in the borrower's credit file**

**pub\_rec: Number of derogatory public records**

**revol\_bal: Total credit revolving balance**

**revol\_util: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit**

**total\_acc: The total number of credit lines currently in the borrower's credit file**

**initial\_list\_status: The initial listing status of the loan**

**out\_prncp: Remaining outstanding principal for total amount funded**

**out\_prncp\_inv: Remaining outstanding principal for portion of total amount funded by investors**

**total\_pymnt: Payments received to date for total amount funded**

**total\_pymnt\_inv: Payments received to date for portion of total amount funded by investors**

**total\_rec\_prncp: Principal received to date**

**total\_rec\_int: Interest received to date**

**total\_rec\_late\_fee: Late fees received to date**

**recoveries: Post charge off gross recovery**

**collection\_recovery\_fee: Post charge off collection fee**

**last\_pymnt\_d: Last month payment was received**

**last\_pymnt\_amnt: Last total payment amount received**

**next\_pymnt\_d: Next scheduled payment date**

**last\_credit\_pull\_d: The most recent month LC pulled credit for this loan**

**collections\_12\_mths\_ex\_med: Number of collections in 12 months excluding medical collections**

**mths\_since\_last\_major\_derog: Months since most recent 90-day or worse rating**

**policy\_code: Publicly available policy\_code=1, new products not publicly available policy\_code=2**

**"""**

**Using Python codes perform:**

1. **Data Pre-processing**

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

**3.2 Outlier treatment.**

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. **Build a Multinomial Regression model.**
   3. **Train and test the model and compare accuracies by confusion matrix, ROC & AUC curves.**
   4. **Briefly explain the model output in the documentation.**
3. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statement:**

1. You work for a consumer finance company that specializes in lending 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 information about past loan applicants and whether they ‘defaulted’4 or not. The aim is to identify patterns that indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of the loan, lending (for 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:

* Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
* Current: Applicant is in the process of paying the installments, i.e., the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.
* Charged-off: Applicant has not paid the installments 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)

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). 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 labeled as 'charged-off' are the 'defaulters'.

If one can 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.

A screenshot of a cell phone

Description automatically generated