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: Nukala Ayyappa Bharthwaj

Batch Id: DSWDMCOS 21012022

Topic: Multinomial Regression.

1. Business Problem

1.1. Objective

1.2. Constraints (if any)

2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

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:

3. Data Pre-processing

3.1 Data Cleaning, Feature Engineering, etc.

3.2 Outlier Imputation

4. Exploratory Data Analysis (EDA):

4.1. Summary

4.2. Univariate analysis

4.3. Bivariate analysis

5. Model Building

5.1 Build the model on the scaled data (try multiple options)

© 2013 - 2020 360DigiTMG. All Rights Reserved.

5.2 Perform Multinomial Regression Modl.

5.3 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.

6. 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.

© 2013 - 2020 360DigiTMG. All Rights Reserved.

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

Name of feature	Description	Data Type	Relevance
Id	I'd of the students	Ordinal	Irrelevant
Female	Gender of the student	Nominal	Relevant
Ses	Social economic status of the student	Ordinal	Relevant
Schtype	School type of the student	Nominal	Relevant
Prog	Program type of the student	Nominal	Relevant
read	Score of the student in read	Ratio	Relevant
Write	Score of the student in write	Ratio	Relevant
Math	Score of the student in math	Ratio	Relevant
Science	Score of the student in science	Ratio	Relevant
honors	Whether student honored or not	Nominal	Relevant

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.

	X	id	female	ses	schtyp	prog	read	write	math	science	honors
1	1	45	female	low	public	vocation	34	35	41	29	not enrolled
2	2	108	male	middle	public	general	34	33	41	36	not enrolled
3	3	15	male	high	public	vocation	39	39	44	26	not enrolled
4	4	67	male	low	public	vocation	37	37	42	33	not enrolled
5	5	153	male	middle	public	vocation	39	31	40	39	not enrolled
6	6	51	female	high	public	general	42	36	42	31	not enrolled
7	7	164	male	middle	public	vocation	31	36	46	39	not enrolled
8	8	133	male	middle	public	vocation	50	31	40	34	not enrolled
9	9	2	female	middle	public	vocation	39	41	33	42	not enrolled
10	10	53	male	middle	public	vocation	34	37	46	39	not enrolled
11	11	1	female	low	public	vocation	34	44	40	39	not enrolled
12	12	128	male	high	public	academic	39	33	38	47	not enrolled

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**.**

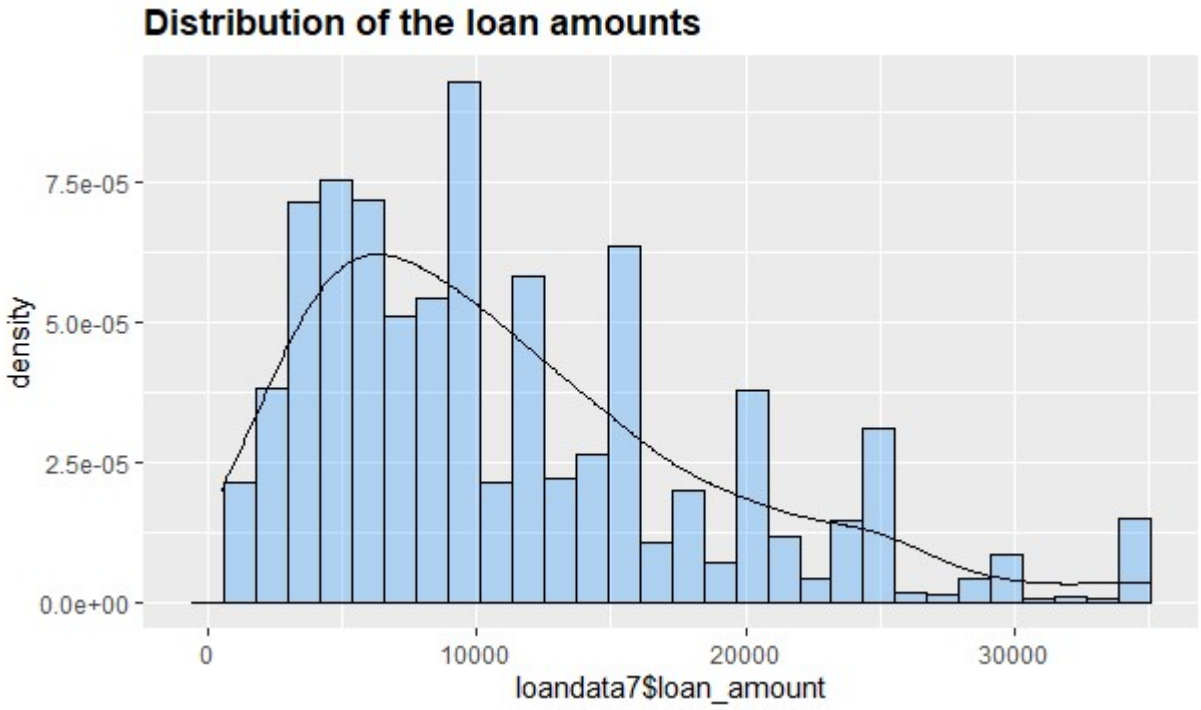
Data Types: There are 111 variables in the given data, all the variables are not used for doing the analysis on the data. The following are the selected variables from the given data to do the analysis.

Name of feature	Description	Data Type	Relevance
Loan_status	Loan status of the customer	Nominal	Relevant
Issue_d	Issued date to the customer	Ordinal	Relevant
Loan_amount	Loan amount of the customer	Ratio	Relevant
emp_title	Position of the customer	Nominal	Relevant
emt_length	Job experience of the customer	Ordinal	Relevant
Verification_status	Whether the customer details are verified or not	Nominal	Relevant
home_ownership	Home ownership of the customer	Nominal	Relevant
Annual_inc	Annual income of the customer	Ratio	Relevant
Purpose	Loan purpose of the customer	Nominal	Relevant
Inq_last_6months	Income of customer in last 6 months	Ratio	Relevant
Desc	Description for the loan	Nominal	Relevant
Open_acc	Open account of the customer	Ratio	Relevant
Pub_res	Public records of the customer whether its there or not	Nominal	Relevant
Revol_util	Utilization percentage of the customer	Ratio	Relevant
Dti	debt-to-income percentage of the customer	Ratio	Relevant
Total_acc	Total account balance of the customer	Ratio	Relevant
Delinq_2yrs	Whether loan declined in past 2 years for customer or not	Nominal	Relevant
Earliest_cr_line	Earliest credit line of the customer	Ratio	Relevant
Mths_since_last_declinq	Months since last decline of the customer	Ratio	Relevant

© 2013 - 2020 360DigiTMG. All Rights Reserved.

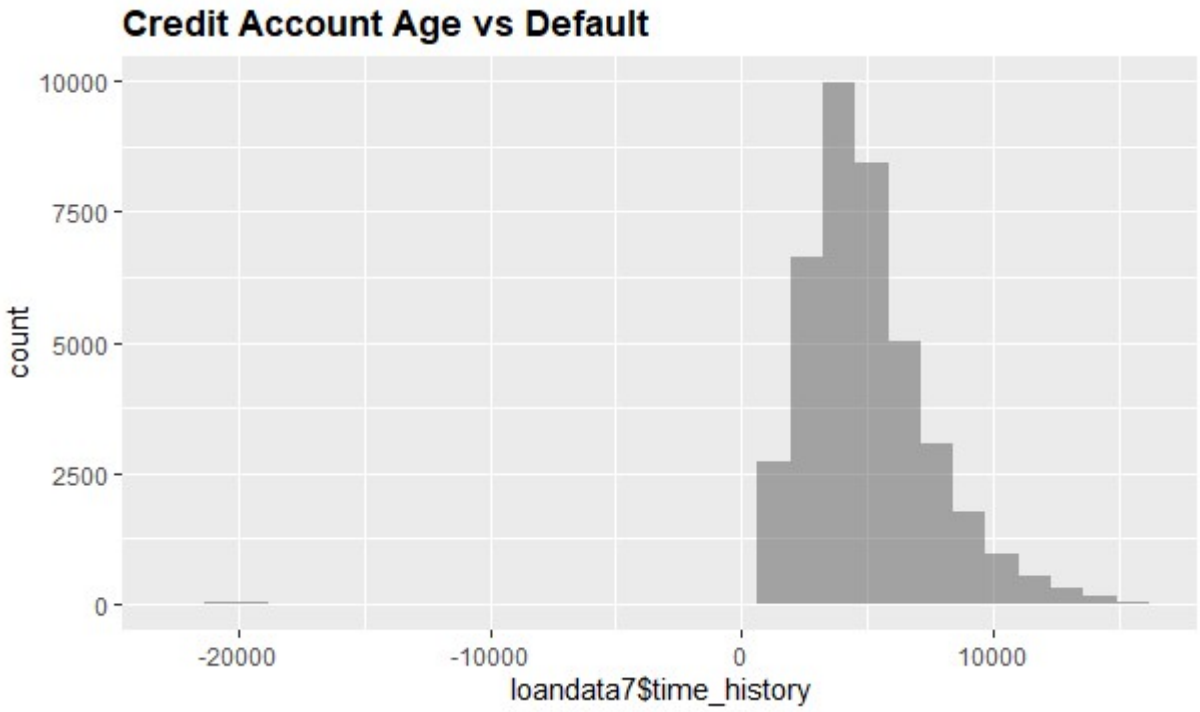
Data Pre-Processing: From the complete data given only 18 variables are selected for doing the analysis which are mentioned in the above data types table.
Target variable taken for doing the analysis is loan status which is renamed as default and the same is converted into numeric data.
Earliest_credit_line 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.



© 2013 - 2020 360DigiTMG. All Rights Reserved.

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 61.26%.

id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership
1077501	1296599	5000	5000	4975	36 months	10.65%	162.87	B	B2		10+ years	RENT
1077430	1314167	2500	2500	2500	60 months	15.27%	59.83	C	C4	Ryder	< 1 year	RENT
1077175	1313524	2400	2400	2400	36 months	15.96%	84.33	C	C5		10+ years	RENT
1076863	1277178	10000	10000	10000	36 months	13.49%	339.31	C	C1	AIR RESOURCES B(	10+ years	RENT
1075358	1311748	3000	3000	3000	60 months	12.69%	67.79	B	B5	University Medical G	1 year	RENT
1075269	1311441	5000	5000	5000	36 months	7.90%	156.46	A	A4	Veolia Transportator	3 years	RENT
1069639	1304742	7000	7000	7000	60 months	15.96%	170.08	C	C5	Southern Star Photo	8 years	RENT
1072053	1288686	3000	3000	3000	36 months	18.64%	109.43	E	E1	MKC Accounting	9 years	RENT
1071795	1306957	5600	5600	5600	60 months	21.28%	152.39	F	F2		4 years	OWN
1071570	1306721	5375	5375	5350	60 months	12.69%	121.45	B	B5	Starbucks	< 1 year	RENT
1070078	1305201	6500	6500	6500	60 months	14.65%	153.45	C	C3	Southwest Rural met	5 years	OWN
1069908	1305008	12000	12000	12000	36 months	12.69%	402.54	B	B5	UCLA	10+ years	OWN
1064687	1298717	9000	9000	9000	36 months	13.49%	305.38	C	C1	Va. Dept of Conserva	< 1 year	RENT
1069866	1304956	3000	3000	3000	36 months	9.91%	96.68	B	B1	Target	3 years	RENT
1069057	1303503	10000	10000	10000	36 months	10.65%	325.74	B	B2	SFMTA	3 years	RENT
1069759	1304871	1000	1000	1000	36 months	16.29%	35.31	D	D1	Internal revenue Ser	< 1 year	RENT
1065775	1299699	10000	10000	10000	36 months	15.27%	347.98	C	C4	Chin's Restaurant	4 years	RENT
1069971	1304884	3600	3600	3600	36 months	6.03%	109.57	A	A1	Duracell	10+ years	MORTGAGE
1062474	1294539	6000	6000	6000	36 months	11.71%	198.46	B	B3	Connection Inspecti	1 year	MORTGAGE
1069742	1304855	9200	9200	9200	36 months	6.03%	280.01	A	A1	Network Interpreting	6 years	RENT
1069740	1284848	20250	20250	19142.16108	60 months	15.27%	484.63	C	C4	Archdiocese of Galv	3 years	RENT
1039153	1269083	21000	21000	21000	36 months	12.42%	701.73	B	B4	Osram Sylvania	10+ years	RENT
1069710	1304821	10000	10000	10000	36 months	11.71%	330.76	B	B3	Value Air	10+ years	OWN
1069700	1304810	10000	10000	10000	36 months	11.71%	330.76	B	B3	Wells Fargo Bank	5 years	RENT

© 2013 - 2020 360DigiTMG. All Rights Reserved.