# Logistic Regression (Module -9)

**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: Logistic 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 Logistic Regression model.**
   3. **Train and Test the data and compare accuracies by Confusion Matrix, plot ROC AUC curve.**
3. **Briefly explain the model output in the documentation.** **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

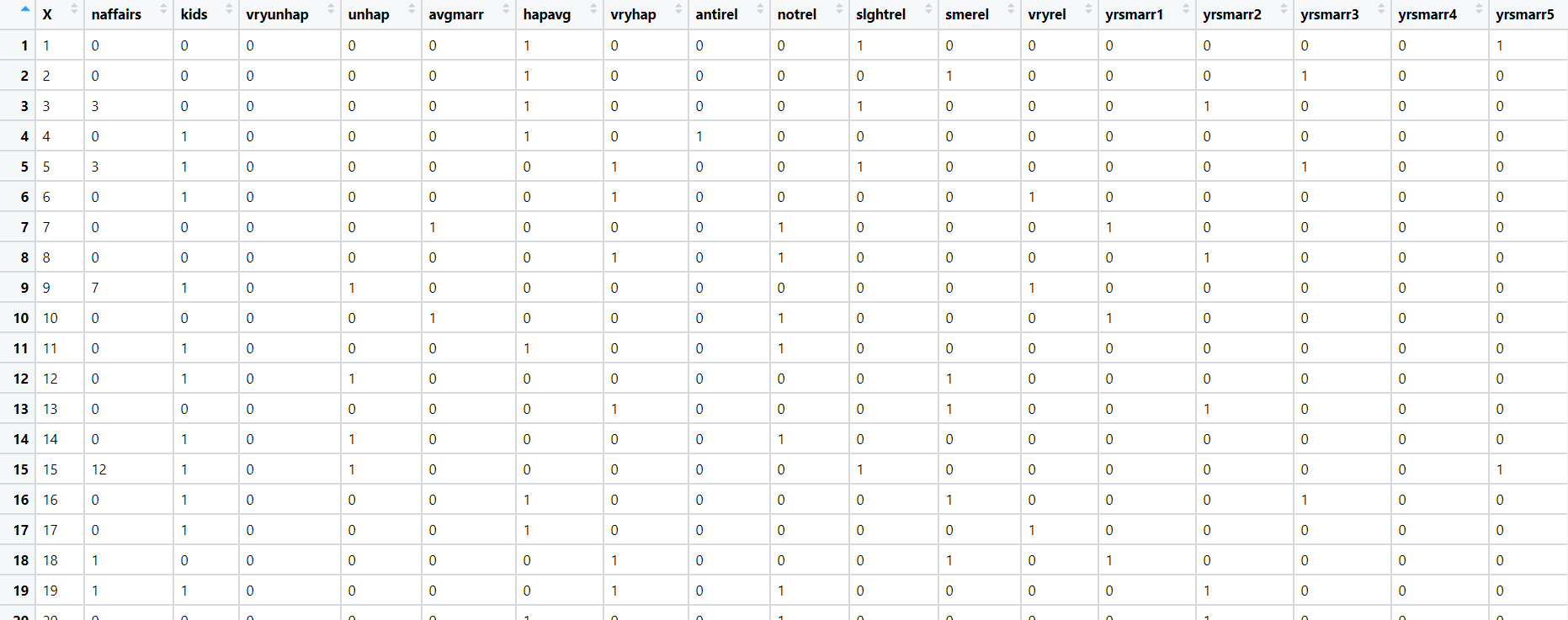
# Note:

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 psychological study has been conducted by a team of students at a University on married couples to determine the cause and effect on their married life and why they tend to have an extra marital affair, they have surveyed and collected a sample of data on which they would like to do further analysis to improve the relationship bond between couple, is it even possible to do so? Using your skills of Machine Learning apply Logistic Regression Model on the data and correctly classify whether a given person will have an affair or not given the set of attributes.

Convert naffairs column to Discreet Binary before proceeding with algorithm.

**Sol:**

**Business Objective:** To predict whether the person will have an external affair or not by using logistic regression model.

**Constraints:** Lack of analysis of affairs data of the people.

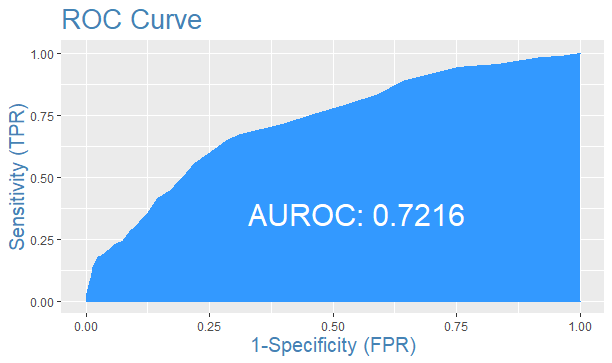
**Data Types:** the data given and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | description | Data Type | Relevance |
| X | study number | Ordinal | Irrelevant since it is case number |
| Naffairs | Number of affairs a person have | Nominal | Relevant |
| Kids | Weather they have kids or not | Nominal | Relevant |
| Vryunhap | Weather they are very un happy or not | Nominal | Relevant |
| Unhap | Weather they are unhappy or not | Nominal | Relevant |
| Avgmarr | Weather avgmarr or not | Nominal | Relevant |
| Hapavg | Weather Hapavg or not | Nominal | Relevant |
| Vryhap | Weather they are very happy or not | Nominal | Relevant |
| antirel | Weather they are antirel or not | Nominal | Relevant |
| Notrel | Weather they are notrel or not | Nominal | Relevant |
| Slghtrel | Weather they are Slghtrel or not | Nominal | Relevant |
| Smerel | Weather they are smerel or not | Nominal | Relevant |
| vryrel | Weather they are vryrel or not | Nominal | Relevant |
| yrsmarr1 | Numbers of years married | Nominal | Relevant |
| Yrsmarr2 | Numbers of years married | Nominal | Relevant |
| Yrsmarr3 | Numbers of years married | Nominal | Relevant |
| Yrsmarr4 | Numbers of years married | Nominal | Relevant |
| Yrsmarr5 | Numbers of years married | Nominal | Relevant |
| Yrsmarr6 | Numbers of years married | Nominal | Relevant |

**Data Pre-Processing:** The number of affairs column in the given data is in the range of 0-16 and the same is converted into binary data. If number of affairs is >0 then they are treated as 1 and the remaining are treated as 0.

Since there are no missing values in the given data the same data can be used further to do analysis.

**Logistic Regression:** Initially linear regression model is applied to see the predicted values of the output and then the Generalized Regression model is applied to get the correct predicted values. Accuracy of the model is 78.06% and the cut of value is 56% and the ROC curve the model is as follows:



A screenshot of a cell phone

Description automatically generated**ROC Curve for affairs data set**

**Problem Statement: -**

In this time and age of widespread internet usage, effective and targeted marketing play a vital role, a marketing company would like to develop their strategy by analyzing their customer data and how effectively they can do targeted marketing, for this historical data has been collected of users clicking on ad given different factors such as age, location, time of activity and more. Perform Logistic Regression on the given data and classify the user who click’s on ad’s and who does not click on ad.

A screenshot of a cell phone

Description automatically generated

**Sol:**

**Business Objective:** To predict whether who clicks on add or not by using logistic regression model.

**Constraints:** Lack of analysis of advertisement data of the users.

**Data Types:** the data given and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | description | Data Type | Relevance |
| Age | Age of the user | Ratio | Relevant |
| Area\_Income | Area income of the particular add | Ratio | Relevant |
| Daily Internet Usage | Daily internet usage of the user | Ratio | Relevant |
| Ad\_Topic\_Line | Topic line of the add | Nominal | Relevant |
| City | City of the user | Nominal | Relevant |
| Male | Whether the user is male or not | Nominal | Relevant |
| Country | Country of the user | Nominal | Relevant |
| Timestamp | Time stamp of the add | Ratio | Relevant |
| Clicked\_on\_Ad | Whether the custome seed the add or not | Nominal | Relevant |

**Data Pre-Processing:** Catagorical data is converted into numeric data so that it will be used for doing the analysis.

**Logistic Regression:** Initially linear regression model is applied to see the predicted values of the output and then the Generalized Regression model is applied to get the correct predicted values. Accuracy of the model is 86%. All the values are predicted corrected based on the given variables if the variables of the data changes then there will be change in the accuracy of the data.



**Problem Statement: -**

Prediction of election results has become trivial in these days, the outcome variable is (0/1) and the other factors that affect a candidate win or loss is amount of money spent, popularity and more. Perform Logistic Regression on the dataset and classify the candidates.

**Sol:**

**Business Objective:** To predict whether the candidate will win or lose the in election by using logistic regression model.

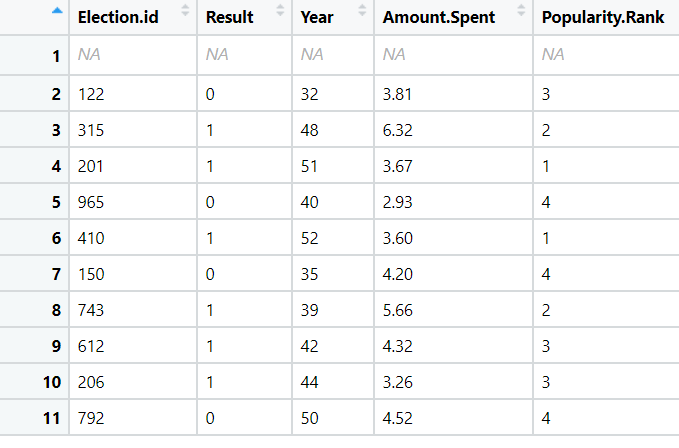
**Constraints:** Lack of analysis of election data of the members.

**Data Types:** the data given and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | description | Data Type | Relevance |
| Election-id | Id of the member | Ordinal | Irrelevant since it is I’d number |
| Result | Result for the member whether won or not | Nominal | Relevant |
| Year | Age of the member | Nominal | Relevant |
| Amount Spent | Amount spent by the member | Nominal | Relevant |
| Popularity Rank | Popularity rank of the member | Nominal | Relevant |

**Data Pre-Processing:** Since election id is not useful for the analysis that column is removed and the data is used for further analysis.

**Logistic Regression:** Initially linear regression model is applied to see the predicted values of the output and then the Generalized Regression model is applied to get the correct predicted values. Accuracy of the model is 100%. All the values are predicted corrected based on the given variables if the variables of the data changes then there will be change in the accuracy of the data.



**Problem Statement:**

In Financial Institutions getting their customers to do a fixed deposit in the banks is a vital and at most important for the bank as they bank uses it and pays an interest amount to those deposited customers. To ask every customer for a term deposit is not viable as well as time consuming process, can you come up with a Logistic Regression model to predict customers who will do a term deposit or not.

The output variable in the dataset is Y which is discreet and binary. Snapshot of the dataset is given below**.**

**A picture containing large

Description automatically generated**

**Sol:**

**Business Objective:** To predict the customer who will do term deposit or not by using logistic regression model.

**Constraints:** Lack of analysis of Customer data.

**Data Types:** the data given and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | description | Data Type | Relevance |
| Age | Age of the customer | Ratio | Relevant |
| Default | Whether customer is default or not | Nominal | Relevant |
| Balance | Balance of the customer account | Ratio | Relevant |
| Housing | Whether the customer is housing or not | Nominal | Relevant |
| Loan | Whether the customer is having loan or not | Nominal | Relevant |
| Duration | Duration of the loan | Ratio | Relevant |
| Campaign | Whether the customer is campaign or not | Nominal | Relevant |
| Pdays | Whether the customer is paid in days or not | Nominal | Relevant |
| Previous | Whether the customer is is previous or not | Nominal | Relevant |
| poutfailure | Whether the customer is poutfailure or not | Nominal | Relevant |
| poutother | Whether the customer is poutother or not | Nominal | Relevant |
| poutsuccess | Whether the customer is poutsuccess or not | Nominal | Relevant |
| poutunknown | Whether the customer is poutunknown or not | Nominal | Relevant |
| con\_cellular | Whether the customer is having coc\_cellur or not | Nominal | Relevant |
| con\_telephone | Whether the customer is having con\_telph or not | Nominal | Relevant |
| con\_unknown | Whether the customer is having unk\_cont or not | Nominal | Relevant |
| divorced | Whether the customer is divorced or not | Nominal | Relevant |
| married | Whether the customer is married or not | Nominal | Relevant |
| single | Whether the customer is single or not | Nominal | Relevant |
| joadmin. | Whether the customer job is admin or not | Nominal | Relevant |
| joblue.collar | Whether the customer job is blue.col or not | Nominal | Relevant |
| joentrepreneur | Whether the customer job is entrepreneur or not | Nominal | Relevant |
| johousemaid | Whether the customer job is house maid or not | Nominal | Relevant |
| jomanagement | Whether the customer job is mang. Or not | Nominal | Relevant |
| joretired | Whether the customer job is retired or not | Nominal | Relevant |
| joself.employed | Whether the customer job is employed or not | Nominal | Relevant |
| joservices | Whether the customer job is service or not | Nominal | Relevant |
| jostudent | Whether the customer student or not | Nominal | Relevant |
| jotechnician | Whether the customer job is technician or not | Nominal | Relevant |
| jounemployed | Whether the customer job is unemployed or not | Nominal | Relevant |
| jounknown | Whether the customer job is unknown or not | Nominal | Relevant |
| y | Output variable whether the customer paid or not | Nominal | Relevant |

**Data Pre-Processing:** All the given data is numeric some of them are continues and some are binary data, it is used for doing the further analysis.

**Logistic Regression:** Initially linear regression model is applied to see the predicted values of the output and then the Generalized Regression model is applied to get the correct predicted values. Accuracy of the model is 90.01% and the cut of value is 46.9% and the ROC curve the model is as follows:

