# 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:** Nithin D Souza

**Batch Id:** 05012021-10AM

**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.**
   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.**

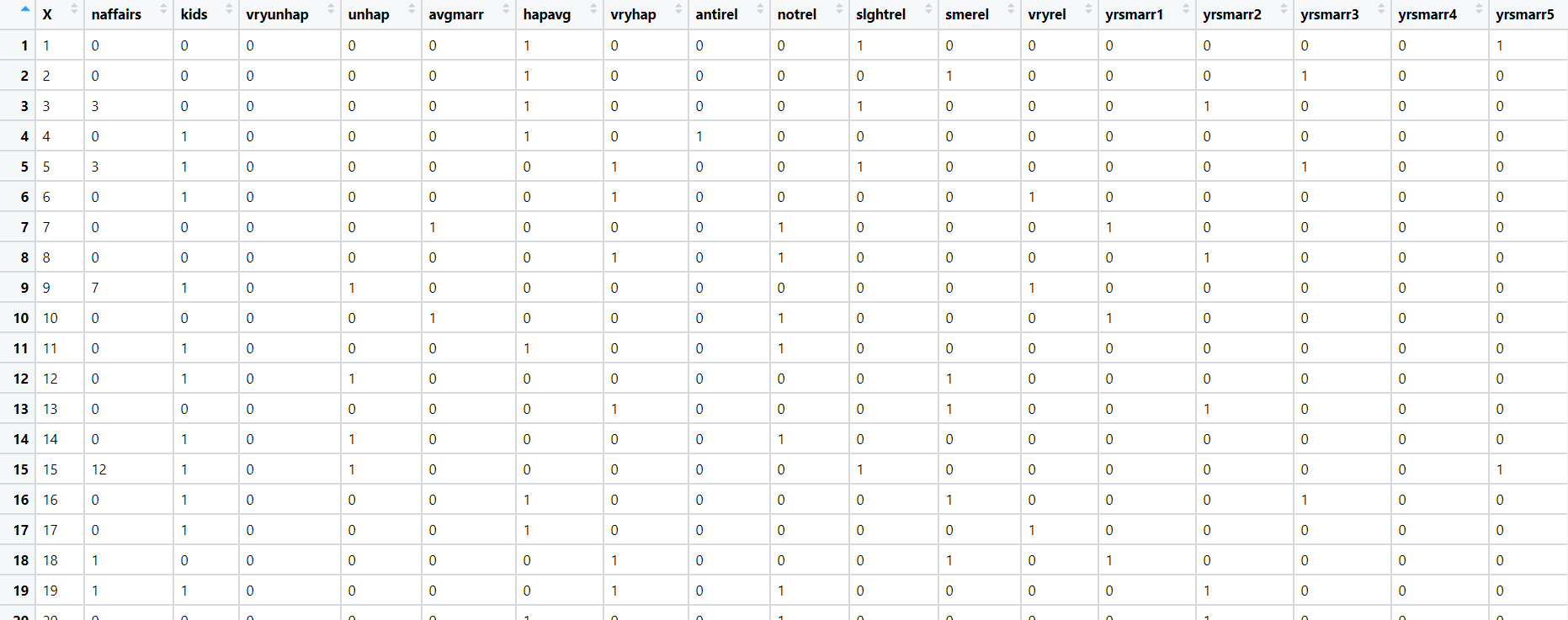
# Note:

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 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.

**A screenshot of a cell phone

Description automatically generated**

**Objective:** Maximize the accuracy in relating dependent variable and one or more independent variable.

**Constraint:** Can only be used to predict discrete functions also non-linear problems can’t be solved because of linear decision surface.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Discerption** | **Type** | **Relevance** |
| Naffairs | Number of affairs | Discrete, Count | It provides useful information |
| Kids | Number of kids | Discrete, Count | It provides useful information |
| Vryunhap | Mood of a person | Discrete,  Ordinal | It provides useful information |
| Unhap | Rate of marriage | Continuous, Ratio | It provides useful information |
| Avgmarr | Rate of marriage | Continuous, Ratio | It provides useful information |
| Hapavg | Rate of marriage | Continuous, Ratio | It provides useful information |
| Vryhap | Rate of marriage | Continuous, Ratio | It provides useful information |
| Antirel | Religious Rating | Continuous, Ratio | It provides useful information |
| Notrel | Religious Rating | Continuous, Ratio | It provides useful information |
| Slghtrel | Religious Rating | Continuous, Ratio | It provides useful information |
| Smerel | Religious Rating | Continuous, Ratio | It provides useful information |
| Vryrel | Religious Rating | Continuous, Ratio | It provides useful information |
| yrsmarr1 | Years married | Discrete, Count | It provides useful information |
| yrsmarr2 | Years married | Discrete, Count | It provides useful information |
| yrsmarr3 | Years married | Discrete, Count | It provides useful information |
| yrsmarr4 | Years married | Discrete, Count | Useful information |
| yrsmarr5 | Years married | Discrete, Count | Useful information |
| yrsmarr6 | Years married | Discrete, Count | Useful information |

**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



**Objective:** Maximize the accuracy in relating categorical dependent variable and one or more independent variable.

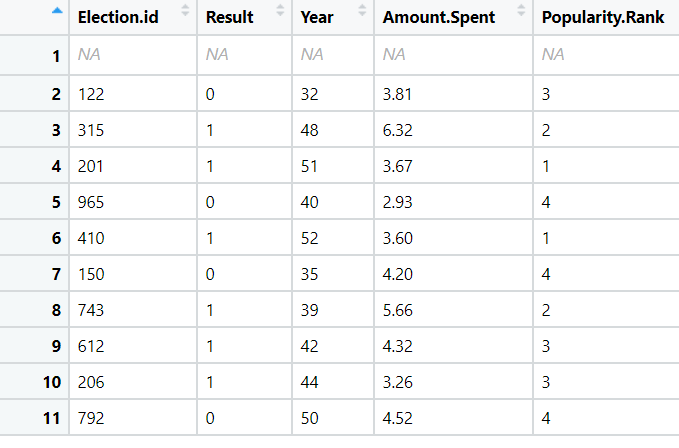
**Constraint:** Can only be used to predict discrete functions also non-linear problems can’t be solved because of linear decision surface.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Discerption** | **Type** | **Relevance** |
| Daily\_Time\_ Spent \_on\_Site | Time spent on site by user | continuous, Ratio | Relevant, it provides useful information.  (independent) |
| Age | Age of user | Discrete | Relevant, it provides useful information.  (independent) |
| Area\_Income | User income | continuous, Ratio | Relevant, it provides useful information.  (independent) |
| Daily Internet Usage | Internet usage of users | continuous, Ratio | Relevant, it provides useful information.  (independent) |
| Ad\_Topic\_Line | Topic of ad popped on site | Nominal | Irrelevant, does not provide useful information. |
| City | City of user | Nominal | Irrelevant, does not provide useful information. |
| Male | Gender of user | Ordinal | Relevant, it provides useful information.  (independent) |
| Country | Country of user | Nominal | Irrelevant, does not provide useful information. |
| Timestamp | Time at which event occurred | Nominal | Irrelevant, does not provide useful information. |
| Clicked\_on\_Ad | Action of user for the ad popped up. | Ordinal | Relevant, it provides useful information also a response variable (dependent) |



**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.





**Objective:** Maximize the accuracy in relating categorical dependent variable and one or more independent variable.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Discerption** | **Type** | **Relevance** |
| Election.id | ID of candidate | Nominal | Irrelevant, does not provide useful information. |
| Result | Result of election | Ordinal | Relevant, it provides useful information also a response variable (dependent) |
| Year | Age of candidate | discrete | Relevant, it provides useful information. (independent) |
| Amount.spent | Amount spent by candidate | Continues, Ratio | Relevant, it provides useful information. (independent) |
| Popularity.Rank | Rank for popularity of candidate | Discrete | Relevant, it provides useful information. (independent) |

**Constraint:** Can only be used to predict discrete functions also non-linear problems can’t be solved because of linear decision surface.

**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**

**Objective:** Maximize the accuracy in relating categorical dependent variable and one or more independent variable.

**Constraint:** Can only be used to predict discrete functions also non-linear problems can’t be solved because of linear decision surface.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Discerption** | **Type** | **Relevance** |
| age | age of each Pearson | discrete | Relevant, it provides useful information. (independent |
| default | default amount of each person | continuous | Irrelevant, does not provide useful information. |
| balance | balance amount of each person | continuous | Relevant, it provides useful information. (independent |
| housing | no of house | continuous | Relevant, it provides useful information. (independent |
| loan | no of loan taken | continuous | Relevant, it provides useful information. (independent |
| duration | duration taken to pay loan | discrete | Relevant, it provides useful information. (independent |
| campaign | no of campaign done | discrete | Irrelevant, does not provide useful information. |
| pdays | no of pay days | discrete | Irrelevant, does not provide useful information. |
| previous | previous data | ratio | Irrelevant, does not provide useful information. |
| poutfailure | poutfailure of each person | discrete | Irrelevant, does not provide useful information. |
| poutother | poutother of person | discrete | Irrelevant, does not provide useful information. |
| poutsuccess | pout success of each person | discrete | Irrelevant, does not provide useful information. |
| poutunknown | pout known of person | discrete | Irrelevant, does not provide useful information. |
| con\_cellular | cellular contact no given | discrete | Relevant, it provides useful information. (independent |
| con\_telephone | telephone contact no given | discrete | Relevant, it provides useful information. (independent |
| con\_unknown | unknown contact | discrete | Relevant, it provides useful information. (independent |
| divorced | person is divorced | discrete | Relevant, it provides useful information. (independent |
| married | person is married | discrete | Relevant, it provides useful information. (independent |
| single | person is single | discrete | Relevant, it provides useful information. (independent |
| joadmin. | person is admin or not | discrete | Irrelevant, does not provide useful information. |
| joblue.collar | person is collar or not | discrete | Irrelevant, does not provide useful information. |
| joentrepreneur | person is entrepreneur or not | discrete | Irrelevant, does not provide useful information. |
| johousemaid | person is housemaid or not | discrete | Irrelevant, does not provide useful information. |
| jomanagement | person is management or not | discrete | Irrelevant, does not provide useful information. |
| joretired | person is retired or not | discrete | Irrelevant, does not provide useful information. |
| joself.employed | person is self-employed or not | discrete | Irrelevant, does not provide useful information. |
| joservices | person is in service or not | discrete | Irrelevant, does not provide useful information. |
| jostudent | person is student or not | discrete | Irrelevant, does not provide useful information. |
| jotechnician | person is technician or not | discrete | Irrelevant, does not provide useful information. |
| jounemployed | person is employed or not | discrete | Irrelevant, does not provide useful information. |
| jounknown | about unknown person | discrete | Irrelevant, does not provide useful information. |
| y | output of details given | discrete | Irrelevant, does not provide useful information. |