# Logistic 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:RAKESH SETHU NP Batch ID: 11052022\_7.30PM**

**Topic: Logistic Regression**

**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:**



**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. And if not relevant, provide reasons and a description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

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

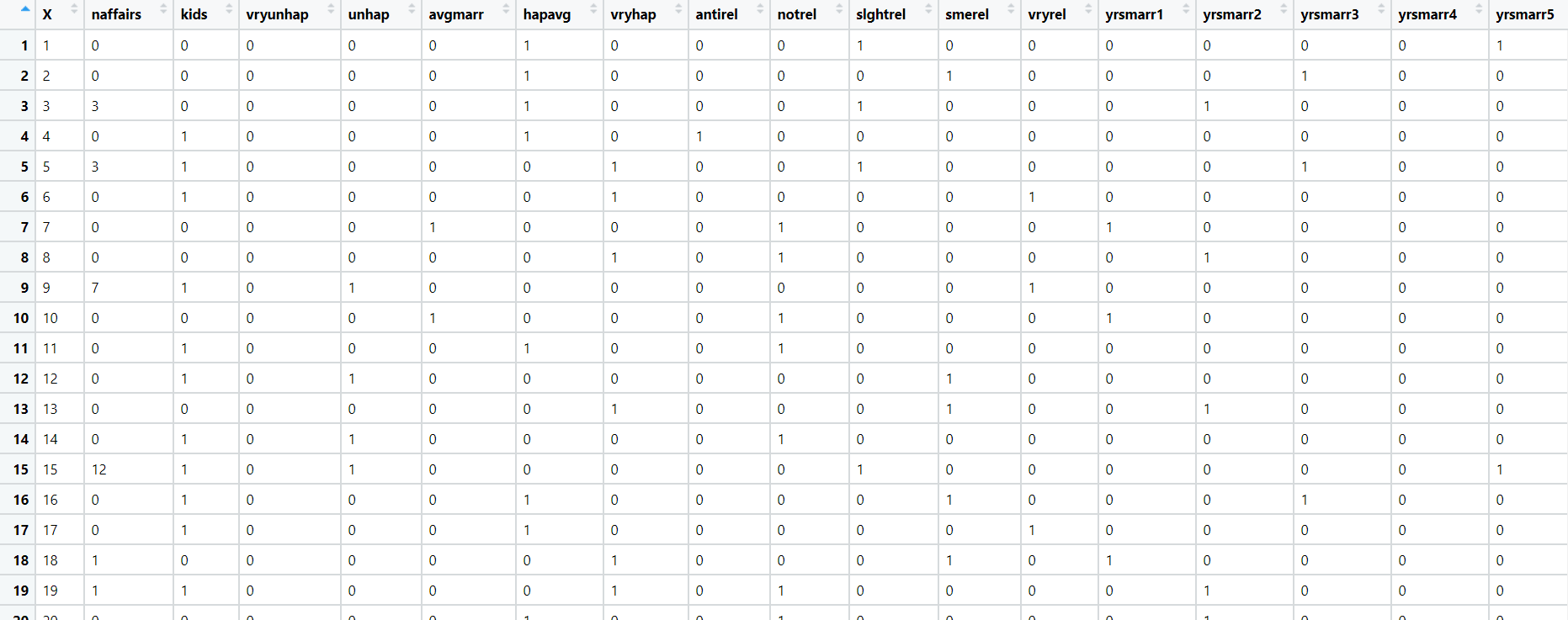
**3.2 Outlier Treatment.**

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 Logistic Regression model.**
   3. **Train and test the model and compare accuracies by building a confusion matrix, plotting ROC and 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?**

A screenshot of a cell phone

Description automatically generatedProblem Statement: -

1. A psychological study has been conducted by a team of students at a university on married couples to determine the cause of having an extra marital affair. They have surveyed and collected a sample of data on which they would like to do further analysis. Apply Logistic Regression on the data to correctly classify whether a given person will have an affair or not given the set of attributes. Convert the naffairs column to discrete binary type before proceeding with the algorithm.



**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 |

1. In this time and age of widespread internet usage, effective and targeted marketing plays a vital role. A marketing company would like to develop a strategy by analyzing their customer data. For this, data like age, location, time of activity, etc. has been collected to determine whether a user will click on an ad or not. Perform Logistic Regression on the given data to predict whether a user will click on an ad or not.

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

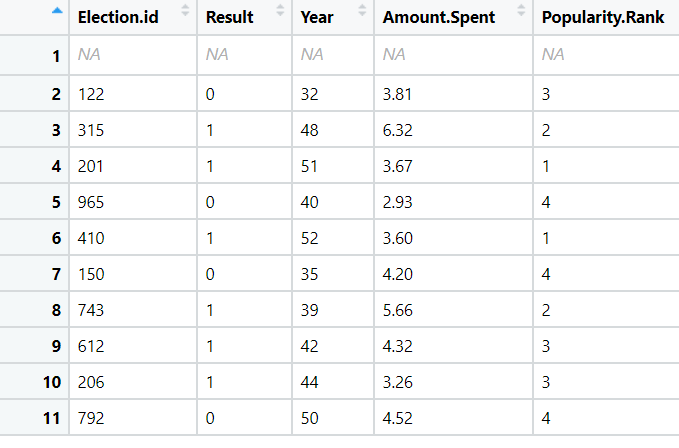
**Constraint:** Can only be used to predict discrete functions also non-linear problems can’t be solved because of linear decision surface.

|  |  |  |  |
| --- | --- | --- | --- |
| **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) |

A screenshot of a cell phone

Description automatically generated

1. Perform Logistic Regression on the dataset to predict whether a candidate will win or lose the election based on factors like amount of money spent and popularity rank.



**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** |
| 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) |

1. It is vital for banks that customers put in long term fixed deposits as they use it to pay interest to customers and it is not viable to ask every customer if they will put in a long-term deposit or not. So, build a Logistic Regression model to predict whether a customer will put in a long-term fixed deposit or not based on the different variables given in the data. The output variable in the dataset is Y which is binary. Snapshot of the dataset is given below.

**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. |

**A picture containing large

Description automatically generated**