# 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: Shahina Athar**

**Batch ID: 10122020**

**Topic: Logistic Regression**

**Grading 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 for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

**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.**

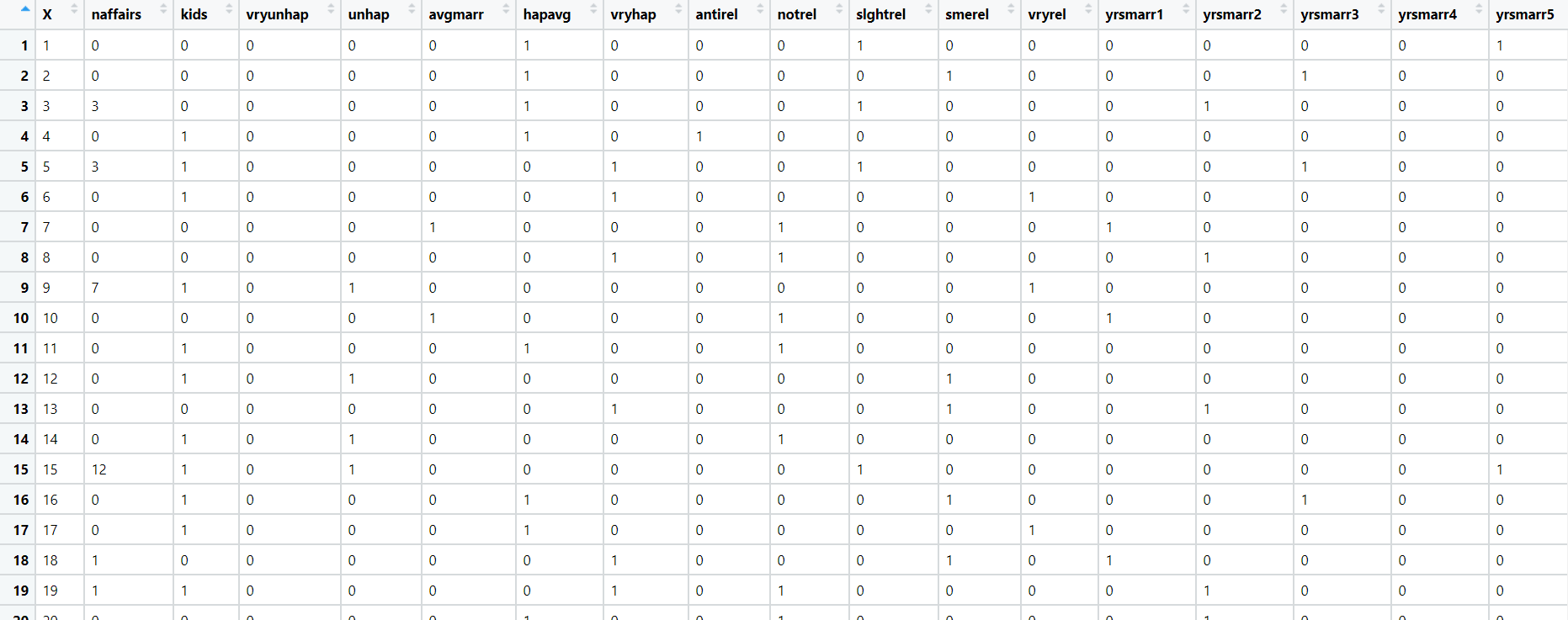
**3.2 Outlier Treatment.**

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.



**1. Business Problem:**

**a. What is the business objective?**

**Ans:** 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.

**2. Briefly explain the model output in the documentation.**

**Ans:** First of all we load the data then we remove unwanted columns, then check any null values, there is no null values in the data, Convert the naffairs column to discrete binary type before proceeding with the algorithm, Build model with logit function, then predict, then we get Optimal Threshold value in which TP Should be maximum as compare to FP, then we get fpr,tpr,1-fpr,tf,thresholds,then we plot ROC Curve, then we get Area under the ROC curve, add new column "pred" with all zeros, and then if this pred is greater than optimal threshold value value mark it as "1" otherwise "0", after that we get classification report, then split the data into train and test with (70-30)% , and then again build the model with logit function, get summary, Prediction on Test data set, Creating new column for storing predicted class of target variable, taking threshold value as 'optimal\_threshold' and above the threshold prob value will be treated as 1, and we get confusion matrix, with that confusion matrix we get accuracy, and get classification report, after that we get ROC & AUC CURVE, then we get roc\_auc\_test value but the value is is less than 0.8 so it is average model, now prediction on train data,Creating new column, filling all the cells with zeroes, taking threshold value and above the prob value will be treated as correct value, get confusion matrix, with that confusion matrix get accuracy, train and test accuracy is close enough so we can accept. We can say this model is works fine.

**3.Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Ans:** 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. With the help of Logistic Regression we can correctly classify whether a given person will have an affair or not.

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.

A screenshot of a cell phone

Description automatically generated

**1. Business Problem:**

**a. What is the business objective?**

**Ans:** 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.

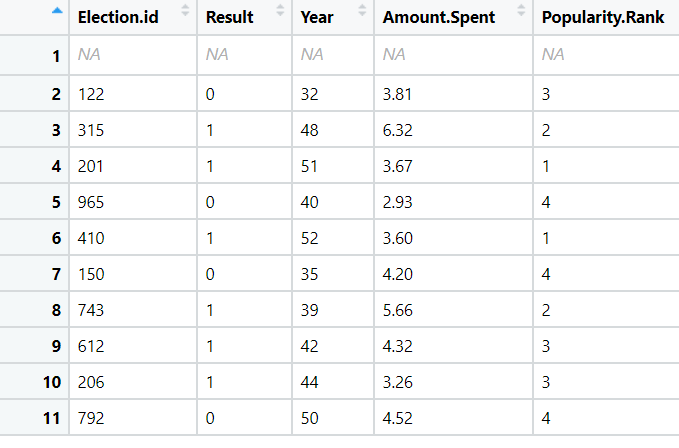
**2. Briefly explain the model output in the documentation.**

**Ans:** First of all we load the data then we remove unwanted columns, then check any null values, there is no null values in the data, Build model with logit function, then predict, then we get Optimal Threshold value in which TP Should be maximum as compare to FP, then we get fpr,tpr,1-fpr,tf,thresholds,then we plot ROC Curve, then we get Area under the ROC curve, add new column "pred" with all zeros, and then if this pred is greater than optimal threshold value value mark it as "1" otherwise "0", after that we get classification report, then split the data into train and test with (70-30)% , and then again build the model with logit function, get summary, Prediction on Test data set, Creating new column for storing predicted class of target variable, taking threshold value as 'optimal\_threshold' and above the threshold prob value will be treated as 1, and we get confusion matrix, with that confusion matrix we get accuracy, and get classification report, after that we get ROC & AUC CURVE, then we get roc\_auc\_test value but the value is is less than 0.8 so it is average model, now prediction on train data,Creating new column, filling all the cells with zeroes, taking threshold value and above the prob value will be treated as correct value, get confusion matrix, with that confusion matrix get accuracy, train and test accuracy are same so we can accept. We can say this model is good model.

**3.Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Ans:** 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. With the help of Logistic Regression Marketing Company can easily predict whether a user will click on an ad or not.

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.



**1. Business Problem:**

**a. What is the business objective?**

**Ans:** 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.

**2. Briefly explain the model output in the documentation.**

**Ans:** First of all we load the data, then check any null values, drop null values, then we divide our data into input and output. Build Logistic Regression model, then predict, get confusion matrix,accuracy, As accuracy = 0.90 , which is greater than 0.5;Threshold value>0.5=1 else ,Threshold value<0.5=0 then we get ROC Curve plotting and finding AUC value, auc accuracy: 0.875 - Good model, it is greater than 0.8, then split the data into train and test with (70-30)% , and then again build the Logistic Regression model. After we do Testing model in that we get confusion matrix, also get accuracy, same for Training Model in that we get confusion matrix, also we get accuracy.

**3.Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Ans:** with the help of Logistic Regression we can easily predict whether a candidate will win or lose the election based on factors like amount of money spent and popularity rank.

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.

**A picture containing large

Description automatically generated**

**1. Business Problem:**

**a. What is the business objective?**

**Ans:** 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, We as a Data Scientist 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.

**2. Briefly explain the model output in the documentation.**

**Ans:** First of all we load the data, then check any null values, no null values, then we divide our data into input and output. Build Logistic Regression model, then predict, get confusion matrix, accuracy, As accuracy = 89% , which is greater than 0.5;Threshold value>0.5=1 else ,Threshold value<0.5=0 then we get ROC Curve plotting and finding AUC value, auc(Area under the curve) accuracy: 60% , then split the data into train and test with (70-30)% , and then again build the Logistic Regression model. After we do Testing model in that we get confusion matrix, also get accuracy, same for Training Model in that we get confusion matrix, also we get accuracy, train and test accuracy is close enough so it is good model..

**3.Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Ans:** **:** 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, with the help of Logistic Regression model bank can easily predict whether a customer will put in a long-term fixed deposit or not.

