Logistic Regression Lab activity - Instructions

Agenda

1. Get the data

2. Data Pre-processing

3. Build a model

4. Predictions

5. Model validation

Dataset: "bank.txt"

**Problem statement**: Classifying whether the subject subscribes to a term deposit or not

Data Description:

The dataset is from a bank, using which we have to predict whether the subject subscribes to a

term deposit or not

The dataset has the following attributes:

1 - age (numeric)

2 - job : type of job (categorical:

"admin.","unknown","unemployed","management","housemaid","entrepreneur","student",

"blue-collar","self-employed","retired","technician","services")

3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means

divorced or widowed)

4 - education (categorical: "unknown","secondary","primary","tertiary")

5 - default: has credit in default? (binary: "yes","no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes","no")

8 - loan: has personal loan? (binary: "yes","no")

9 - contact: contact communication type (categorical: "unknown","telephone","cellular")

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

13 - campaign: number of contacts performed during this campaign and for this client (numeric,

includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous

campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical:

"unknown","other","failure","success")

Response Variable (desired target):

17 - y : has the client subscribed to a term deposit?(binary: "yes","no")

Get the data(and understand it):

1. Read the bank.txt

Data Pre-processing:

2. Use earlier activities to see if there are any null values in each of the columns.

Model Building

3. Apply Logistic regression on the data and see if the individual

coefficients are significant.

4. Below is the code for ROC curve.

5. FInd the AUC.

Model Validation

6. Build the confusion matrix by comparing actuals and predicted values.

7. Get the error metrics from confusion matrix.

A few questions to be investigated:

What do you think about the performance of the model?

What is the impact of choosing the cut-off on the performance? (try with different variables)

What is the best metric for this problem statement?

Will standardising the data help?