

Machine Learning using Python

Exam Questions – Paper 1

[Time: 3 hrs]
[Total Marks: 40]

Part I: Supervised Learning

[Total Marks - 40]

Given is the 'Portugal Bank Marketing' dataset:

Bank client data:

- 1) **age** (numeric)
- 2) **job**: type of job (categorical: "admin.", "bluecollar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed", "unknown")
- 3) **marital**: marital status (categorical: "divorced", "married", "single", "unknown"; note: "divorced" means divorced or widowed)
- 4) **education**: education of individual (categorical: "basic.4y", "basic.6y", "basic.9y", "high.school", "illiterate", "professional.course", "university.degree", "unknown")
- 5) **default**: has credit in default? (categorical: "no", "yes", "unknown")
- 6) **housing**: has housing loan? (categorical: "no", "yes", "unknown")
- 7) **loan**: has personal loan? (categorical: "no", "yes", "unknown")

Related with the last contact of the current campaign:

- 8) **contact**: contact communication type (categorical: "cellular", "telephone")
- 9) **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
- 10) **dayofweek**: last contact day of the week (categorical: "mon", "tue", "wed", "thu", "fri")
- 11) **duration**: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

12) **campaign:** number of contacts performed during this campaign and for this client (numeric, includes last contact)

13) **pdays:** number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14) **previous:** number of contacts performed before this campaign and for this client (numeric)

15) **poutcome:** outcome of the previous marketing campaign (categorical: "failure", "nonexistent", "success")

Social and economic context attributes

16) **emp.var.rate:** employment variation rate - quarterly indicator (numeric)

17) **cons.price.idx:** consumer price index - monthly indicator (numeric)

18) **cons.conf.idx:** consumer confidence index - monthly indicator (numeric)

19) **concavepoints_se:** standard error for number of concave portions of the contour

20) **euribor3m:** euribor 3 month rate - daily indicator (numeric)

21) **nr.employed:** number of employees - quarterly indicator (numeric)

Output variable (desired target):

22) **y:** has the client subscribed a term deposit? (binary: "yes", "no")

Perform the following tasks:		Marks
Q1.	What does the primary analysis of several categorical features reveal?	[5]
Q2.	Perform the following Exploratory Data Analysis tasks: a. Missing Value Analysis b. Label Encoding wherever required c. Selecting important features based on Random Forest d. Standardize the data using the any one of the scalers provided by sklearn	[10]
Q3.	Build the following Supervised Learning models: a. Logistic Regression	[15]

b. AdaBoost

c. KNN

d. SVM

- Q4. Tabulate the performance metrics of all the above models **[10]**
and tell which model performs better in predicting if the
client will subscribe to term deposit or not