

ML Projects

Part 1- Supervised Learning

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) **concave points_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:

Q1. Perform EDA on the given data. What does the primary analysis of several categorical features reveal? □5 marks

Q.2. Perform the following pre-processing tasks: □10 marks

- a. Missing Value Analysis
- b. Label Encoding wherever required
- c. Selecting important features based on Random Forest
- d. Handling unbalanced data using SMOTE
- e. Standardize the data using any one of the scalers provided by sklearn

Q.3. Build the following Supervised Learning models: □5 marks

- a. Logistic Regression
- b. Decision Trees
- c. Random Forest

Q.4. Tabulate the performance metrics of all the above models, perform tuning of models and tell which model performs better in predicting if the client will subscribe to term deposit or not □10 marks

Part 2 – Unsupervised Learning

Given the 'credit_card' dataset, below is the data definition:

- 1) **CUSTID:** Identification of Credit Card holder (Categorical)
- 2) **BALANCE:** Balance amount left in their account to make purchases
- 3) **BALANCEFREQUENCY:** How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
- 4) **PURCHASES:** Amount of purchases made from account
- 5) **ONEOFFPURCHASES:** Maximum purchase amount done in one-go
- 6) **INSTALLMENTSPURCHASES:** Amount of purchase done in installment
- 7) **CASHADVANCE:** Cash in advance given by the user
- 8) **PURCHASESFREQUENCY:** How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
- 9) **ONEOFFPURCHASESFREQUENCY:** How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
- 10) **PURCHASESINSTALLMENTSFREQUENCY:** How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
- 11) **CASHADVANCEFREQUENCY:** How frequently the cash in advance being paid
- 12) **CASHADVANCETRX:** Number of Transactions made with "Cash in Advanced"
- 13) **PURCHASESTRX:** Number of purchase transactions made
- 14) **CREDITLIMIT:** Limit of Credit Card for user
- 15) **PAYMENTS:** Amount of Payment done by user
- 16) **MINIMUM_PAYMENTS:** Minimum amount of payments made by user
- 17) **PRCFULLPAYMENT:** Percent of full payment paid by user
- 18) **TENURE:** Tenure of credit card service for user

Perform the following tasks:

Q1. Perform EDA on the given data. What does the primary analysis of several categorical features reveal? □5 marks

Q2. Perform the following Exploratory Data Analysis tasks:□6 marks

- a. Missing Value Analysis
- b. Outlier Treatment using the Z-score method
- c. Deal with correlated variables

Q3. Perform dimensionality reduction using PCA such that the 95% of the variance is explained□2 marks

Q4. Find the optimum value of k for k-means clustering using the elbow method. Plot the elbow curve□2 marks

Q5. Find the optimum value of k for k-means clustering using the silhouette score method. Build a K-means clustering model and specify the number of observations in each cluster using a bar plot□5 marks