Supervised Learning Project

Domain: Banking

Context:

This case is about a bank (Thera Bank) whose management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

Approach:

The file Bank.xls contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

We will implement Classification algorithms to differentiate people who will buy loans vs the who will not.

Attribute Information

- **ID** : Customer ID
- **Age** : Customer's age in completed years
- **Experience** : #years of professional experience
- **Income** : Annual income of the customer ($000)
- **ZIP Code** : Home Address ZIP code.
- **Family** : Family size of the customer
- **CCAvg** : Avg. spending on credit cards per month ($000)
- **Education** : Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
- **Mortgage** : Value of house mortgage if any. ($000)
- **Personal Loan** : Did this customer accept the personal loan offered in the last campaign?
- **Securities Account** : Does the customer have a securities account with the bank?
- **CD Account** : Does the customer have a certificate of deposit (CD) account with the bank?
- **Online** : Does the customer use internet banking facilities?
- **Credit card** : Does the customer use a credit card issued by UniversalBank?

Questions (Total 50 points)

1. Read the column description and ensure you understand each attribute well
2. Perform univariate analysis of each and every attribute - use an appropriate plot for a given attribute and mention your insights (5 points)
3. Perform correlation analysis among all the variables - you can use Pairplot and Correlation coefficients of every attribute with every other attribute (5 points)
4. One hot encode the Education variable (3 points)
5. Separate the data into dependant and independent variables and create training and test sets out of them (X_train, y_train, X_test, y_test) (2 points)
6. Use **StandardScaler( )** from sklearn, to transform the training and test data into scaled values (fit the StandardScaler object to the train data and transform train and test data using this object, making sure that the test set does not influence the values of the train set) (5 points)
7. Write a function which takes a model, X_train, X_test, y_train and y_test as input and returns the accuracy, recall, precision, specificity, f1_score of the model trained on the train set and evaluated on the test set (5 points)
8. Employ multiple Classification models (Logistic, K-NN, Naïve Bayes etc) and use the function from step 7 to train and get the metrics of the model (15 points)
9. Create a dataframe with the columns - “Model”, “accuracy”, “recall”, “precision”, “specificity”, “f1_score”. Populate the dataframe accordingly (5 points)
10. Give your reasoning on which is the best model in this case (5 points)