Summary of the project

Problem Statement:

X Education sells online courses to industry professionals. X Education requires assistance in identifying the most promising leads or those who are most likely to become paying customers.

The organisation requires a model in which each lead is allocated a lead score with the higher lead score customers having a higher conversion chance and the lower lead score customers having a lower conversion chance.

Solution steps:

1. Import libraries:
    - Import the required libraries
2. Load data
    - Load the data and convert the data columns and the text values to 'snake case' by replacing the spaces with '_'
3. Preliminary investigation of the data
    - Checked the shape, size of the data. Then we checked the missing values column wise and row wise as well. Check outlier in the data
4. Data preparation
    a. Replace select with NaN
        i. Replaced all "Select" with NaN and checked the missing percentage again
    b. Check unique columns in data and drop them
        i. Check the columns that have constant value in all rows and delete these rows
    c. Check no variance columns and drop them
        i. Check the columns which have more than one value but the distribution of class in one class is more than 90%. We will not observe any variance in data because of such columns so we will have to drop them
    d. Delete columns with more than 40% missing values
        i. Check the missing percentage per column and delete the rows with more than 40% missing values
        ii. Check the missing counts per row and delete the rows with most missing numbers
    e. Subset data based on occupation
        i. Subset the data based on the occupation and check the missing values again and subset based on the data
    f. Subset data based on total visits
        i. Subset data based on the total visit and check the missing values again
    g. Map country and City and create new country column
        i. Check the country and City columns and clean the data accordingly. Also create a new country variable based on the mapping
    h. Create new variables from numerical columns (feature engineering)
        i. Create new variables from the numerical columns

i. Check row wise missing counts and delete rows
    i. Check row wise missing counts and delete rows
j. Group categorical variables and reduce the classes
    i. Group the low frequency classes of the categorical variables to Others

5. EDA
    a. Check the distribution of categorical variable wrt the target variable
    b. Check the box plot for numerical variables
    c. Check correlation between numerical variables
6. Feature selection using RFE
    a. Out of the total features generated after doing feature engineering we need to find out the relevant features. We have used wrapper method to find the relevant features. For that we have used recursive feature elimination technique.
7. Split data to train test
    a. We have defined X and y and then split the data in 70:30 ratio as train and test.
8. One hot encoding (dummy variable creation)
    a. We have done the one hot encoding for all the categorical variables also we drop the first dummy column for each feature to avoid the dummy variable trap
9. Scaling numerical columns
    a. We did the scaling for the numerical variables; here we have used the normalization technique for scaling using min max scalar.
10. Built model on train data
    a. We built the base line model using the entire feature we have after one hot encoding and scaling. Also the we check the significance level for each variable with help of p-value
11. Check VIF
    a. Check the VIF for each features
12. Delete variables based on p value and VIF
    a. We can delete the most insignificant variable which has higher VIF value and run the model again. We will do this step till that time when we find all variables are significant and the VIF for them less than 5. In this model we found 10 such features
13. Check evaluation matrix for best model
    a. Check the accuracy, precision, recall (sensitivity), specificity and F1 score for the model. For train dataset we observed the Cross validated mean accuracy: 95.2%, the precision is : 96.12%, the recall is : 93.04%, the specificity is : 96.92% and F1 score is : 94.55%
    b. Check the AUC and ROC curve, in this model we have AUC is 98%
    c. Find the optimal cutoff values, in this model the optimum cutoff value is 0.28
14. Check model performance on test data
    a. Check the accuracy, precision, recall (sensitivity), specificity and F1 score for the model. For test dataset we observed the accuracy is : 94.73%, the precision is : 93.53%, the recall is : 94.84%, the specificity is : 94.63% and F1 score is : 94.18%
    b. Check the AUC and ROC curve, in this model we have AUC is 98%
    c. Find the optimal cutoff values, in this model the optimum cutoff value is 0.28
    d. Based on the cutoff value check the predicted values