Summary

- Purvi Padliya
- Arjun Mehtani

The case study aims to build a logistic model for an education company named X Education which sells online courses to industry professionals, so that it can assign a lead score to each of the leads which can be used by the company to target potential leads and get more people to join their course.

The model is developed using the following approach –

1. Data Understanding and Data Exploration – The dataset has 37 columns and 9240 rows. There are no duplicate rows present, it is validated using the unique identifier of dataset

2. Data Cleaning – In the leads dataset, 17 columns have null values which is cleaned by eliminating columns having null values more than 40%, other null values are handled by imputation

There are outliers present in the dataset which are handled using imputation of upper bound value

3. Data Visualization and Analysis (EDA)

There are outliers present in the dataset which are handled using imputation of upper bound value and heat map is plotted for numerical variables and count plot for categorical analysis (univariate and bivariate)

4. Data Preparation for model building

- Dummy variables are created for categorical variables
- The dataset is split in 70:30 ratio for training and test data
- For scaling numerical variables, MinMaxScaler method is used

5. Model Building and Model Evaluation

- RFE was used to identify 15 prominent features in model building
- Multiple models were created taking under consideration – p value should be less than 0.05 and VIF should be less than 5

- Accuracy score is calculated for the model which came out to be 86.42%
- A confusion matrix is created and using that other evaluation metrics like sensitivity, specificity, precision, recall was calculated
- A ROC graph is plotted and along with that classification report function was used to identify precision, recall, F1 score value
- Optimal cut off is calculated by plotting accuracy, sensitivity and specificity by plotting against various probabilities and optimal cut off is 0.27
- Using precision recall graph, cut off came at 0.35 and the accuracy and other evaluation matrix was formed
- The model was tested on test data and the accuracy and other metrics are similar to training data and thus can help to convert and identify potential leads

6. Final Model Variables –

The features which are used in the final model are as follows –

1. 'What is your current occupation_Unemployed'
2. 'TotalVisits'
3. 'What is your current occupation_Other' i.e. for which occupation detail is missing
4. 'Total Time Spent on Website'
5. 'Last Activity_Email Opened'
6. 'Last Activity_SMS Sent'
7. 'Lead Source_Olark Chat'
8. 'Tags_Will revert after reading the mail'
9. 'Lead Origin_Lead Add Form'
10. 'Last Activity_Email Bounced'
11. 'What is your current occupation_Student'

The X education company can use this model to target potential leads and get a higher conversion rates to make all the potential buyers to buy their course.