Summary Report of Case Study

The problem statement was that an education company named X Education sells online courses to industry professionals. Many professionals who are interested in the courses land on their website and browse for courses. We are required to find the most important features and their influence on the target variable, which in this case is “Converted”.

As for this problem, we use Logical Regression method to analyse and find optimum model. The steps involved in the completion of this case study are as below:

1. Data understanding

   After uploading the necessary dataset, we started understanding the data set. Which is the first step in any analytical problem.

2. Data cleaning

   We came across different errors like missing values, data type error, outliers in numerical variables, etc. For missing values percentage greater than 45%, we removed the variables. For small percentage of missing values, we imputed it with mode or median, depending upon the data type of the variable. For data type error, we changed, Yes/No values in Boolean columns with 1 and 0. For outliers, we quantized based on 0-95%.

3. EDA and Visualization

   we got some insights about the natural trend of the data.

   - The variable "spent on website" have a positive correlation with the "Converted" variable.
   - Data traffic and Google are relatively better lead sources.
   - Almost all the converted users are from India. So, there is no valuable insight from that variable.
   - People specialized more in Finance management are more likely to convert.
   - Unemployed users are more interested in taking the course than working professionals.
   - Users who will revert after reading the Email are very likely to convert, but users who keeps avoiding the phone calls are very less likely to convert.
   - Users who sent SMS and the one's whose Email opened as their last activity are more likely to convert.

4. Data preparation

   We made dummy variables for the categorical variables and concatenated it with the original data set.

5. Train-Test splitting

   We split the entire data set into Train and Test set in the ratio, 70:30 and started the scaling.

6. Model building
   We selected the features using RFE and reduced the number of features to 20. Then eliminating the features one by one using Stats model approach. Here P-value >0.05 (ie.5%) are eliminated and VIF value > 5 are eliminated and finalized the 8th model.

7. Predictions on Train Set
   After that using ROC curve, we found out the area as 0.88. which is great. After that optimum cutoff point as 0.33 is got by plotting the Accuracy, Sensitivity and Specificity curve. Then modelled it again with that cutoff. Precision and recall is also found out and plotted.

8. Predictions on Test Set
   The Acc, Spec, Sens values for both the Train and Test sets are found to be comparable.

Final observations on the model are given by

	Train data	**Test data**
Accuracy	0.806	0.805
Sensitivity	0.804	0.797
Specificity	0.807	0.810

Conclusions and recommendations from the data is:

- Company should focus more on leads from form fillups.
- Company should focus more on Working professionals and unemployed users.
- Lead source from Wellingak website should be focused more.
- Users who spend more time on the website are more likely to convert. So keep an eye on them.
- Company should focus less on users, whose Email got bounced as the Email ID given are most probably wrong.
- Company should focus less on users with specialization of Hospital management and focus more on users of Financial management.