Lead Scoring Case Study Summary

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

We have done the analysis to know which leads are mostly most likely to be converted into paying customers and expecting the target lead conversion rate to be around 80%.

We have used the following steps to prepare our analysis:

1. Loading and cleaning the data:

   Imported Leads.csv file and inspected the data frame. We dropped the columns with more than 35% of Null values in them. We also included imputing of missing values and checking for duplicates. Few of the Null values were changed to ‘not specified’ so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

2. EDA:

   We have used EDA in order to check the condition of our dataset and it is found that a lot of elements are irrelevant which belong to categorical variables. However numerical variables looks good. It is understandable from our EDA that there are many elements that have very little data and so will be of less relevance to our analysis.

3. Creation of Dummy Variables:

   The dummy variables were created and later the dummies with 'not specified’ elements were removed. For numeric values we used the MinMaxScaler.

4. Spitting dataset into train and test and Scaling:

   We have split the dataset into train and test at 70% and 30% respectively. Done Scaling for numerical columns.

5. Feature Selection and Model Building:

   Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

6. Predicting target variable using final model:

   Getting the Predicted values on the train set.

7. Model Evaluation:

   Confusion matrix was made. Determined the best cut-off value using ROC curve to find the Accuracy, Sensitivity and Specificity which came to be around 90% each. The area under **ROC curve is 0.97** indicating a very good predictive model.

8. Prediction on test data:

   Prediction was done on the test data frame accuracy, sensitivity and specificity all at almost 92% .

9. Precision – Recall:

   We got a Precision score of 89% and recall score of 91% on the test data.

10. Final Observation:

    Train Data:

    - Accuracy: 92.29%
    - Sensitivity: 91.70%
    - Specificity: 92.66%

    Test Data:

    - Accuracy: 92.78%
    - Sensitivity: 91.98%
    - Specificity: 93.26%