Lead Scoring Case Study Summary

Objective

A model has to be built for X Education Company to help them select the most promising lead, i.e. the leads that are most likely to convert into paying customers .

Overview of dataset

- First step is to understand the dataset that we are working on.
- We look at the shape and the data types of various columns present in the dataset and get an understanding of what each column indicate.

Data Cleaning and analysis

Next we go ahead with cleaning the data.

- There are a few columns which contains 'select '. As they don't provide any information, those values are replaced with NAN.
- The data is checked for null values. Columns with null values above 40% are dropped.
- The values with the highest frequency in the column are used to replace the missing values in each column that are lower in numbers.
- We have analyzed the columns and got various insights after plotting them .
- There are few columns which do not offer any valuable information. Therefore it is preferable to drop these columns in order to avoid biasness in the model.

Dummy variables

- Dummy variables are created for 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'What matters most to you in choosing a course', 'City', 'Last Notable Activity' .
- The repeated values are dropped.

Splitting into Train and test dataset

- The dataset is divided into train and test dataset in the ratio of 70:30.
- Using Standard Scaler we have scaled the variables ‘TotalVisits’, 'Total Time Spent on Website' and 'Page Views Per Visit’.

Model Building

- Using Recursive Feature Elimination, top 20 features are selected for the model.
- Based on the P values are VIF values, we drop the insignificant variables.
- 15 variables are finalized which has P values <0.05 and VIF <5.
- We build a confusion matrix and find out various metrics such as accuracy, sensitivity and specificity.
- Using the ROC curve, we find out that the Area Under Curve is 0.88 which indicated that the model is good.
- An optimum threshold of 3.5 is chosen and the metrics are calculated again.
- The results found with new cutoff are **Accuracy = 0.81**, **Sensitivity = 0.81**, **Specificity = 0.80**.

Evaluation on Test Dataset

- The test dataset is first scaled.
- The model finalized in train dataset is applied on the test dataset
- The results found are **Accuracy = 0.79**, **Sensitivity = 0.78**, **Specificity = 0.80**.

Insights

- The company can contact customers whose lead source of knowing about the course is through 'Wellingkak website 'and through Reference.
- Customers who had their conversation over phone or had their last activity on SMS are most likely to be converted to a lead.
- Customers who have their current occupation as working professional are more likely to t to a lead.
- Company should contact the customers who spend more time on their website.