Lead Scoring Case Study Summary

Problem Statement :

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. 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%.

Solution Summary:

Step 1: Reading and Understanding Data.

- The Leads data provided by Education company has 9240 rows and 37 columns
- There are a few rows with missing values, that needs attention. Such as
    - Lead Source
    - Total Visits
    - Page Views Per Visit
    - Country
    - Last Activity
    - What is your current occupation
    - What matters most to you in choosing a course
- There are certain columns that have 'Select' as label possibly because these are dropdown.
- It looks like there are outliers present in 'Total Time Spent on Website', 'TotalVisits', 'Page Views Per Visit', since the mean is almost twice or more than 50th percentile

Step 2: Data Cleaning:

- The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information.
- We dropped the variables that had high percentage of NULL values in them.
- creation of new classification variables in case of categorical variables
- The outliers were identified and removed.

Step 3: Data Analysis

- A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant.
- In this step, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped.

Step 4: Creating Dummy Variables

- we went on with creating dummy data for the categorical variables.

Step 5: Test Train Split:

- The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Lead Scoring Case Study Summary

Step 6: Feature Rescaling

- We used the Slandered Scalar to scale the original numerical variables.
- Then using the stats model, we created our initial model, which would give us a complete statistical view of all the parameters of our model.

Step 7: Feature selection using RFE

- Using the Recursive Feature Elimination we went ahead and selected the 15 top important features.
- Using the statistics generated, we recursively variables removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

Step 8 Model Evaluation

- Prediction was done on the test data frame and with an optimum cut off as 0.35.
- A confusion matrix was made for train and test data sets.
- Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity.
- Evaluation Metrics on Train Data
    - Accuracy – 79.24%
    - sensitivity – 79.58%
    - specificity – 79.02%
    - false positive rate – 20.97%
    - positive predictive value - 70%
    - Negative predictive value – 86.28%
    - Overall conversion rate predicted – 80%
- Evaluation Metrics on Train Data
    - Accuracy – 80.13%
    - sensitivity – 80.81%
    - specificity – 79.72%
    - Precision – 70.69%
    - Recall – 80.81%
    - Overall conversion rate predicted – 81%