Summary Report for Case Study

Problem Statement :

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.

Once these people land on the website, they might browse the courses or fill up a form.people fill up a form providing their email address or phone number.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.

There are a lot of leads generated in the initial stage (top) but only a few of them are Converted as a lead.In the middle stage.Help them to find hot leads and give education company to strategy to convert more leads and focus more some variable.

Below is the Step Which I used in Assignment to find the Optimal Result :

1) Importing libraries and Load the Data & Inspect the Data

2) Data Cleaning :

In this step, after inspecting data I decide to drop more than 30% missing value and less than 30% value I fill the data with highest Frequency..

3) Checking Outlier in the Dataset :

In the Dataset there is some outlier in the “Total Visit”,” Total time spent on website” ,” Page views per visit ”

Also I Remove the outlier in data set using quantile value 0.5 to 0.95

4) Exploratory Data Analysis (EDA) :

As per the the Dataset Converted is the target variable so I ll perform the univariate analysis with allthe column one by one.

Based on Analysis we can see some of the column is not much important for the Further Analysis. So we can Drop that Column.

5) Checking Correlation Matrix :

check the correlation matrix and I can say that correlation matrix is good because it is not highly correlated with each other.

6) Data Preparation :

Create the Dummy Variable and onverting some binary variables (Yes/No) to 0/1.

For categorical variables with multiple levels, create dummy features (one-hot encoded).

7) Split the Data into Train & Test.

8) Feature Scaling :

We perform the Feature Scaling on “ TotalVisits ” ,” Total Time Spent on Website ”
” Page Views Per Visit ”

9) Model Building :

9.1) Feature Selection Using RFE : using Automated technique we choose the 15 Variable as output.

9.2) Running First Model : If the model has the high P-value then Drop that Column one by one and Run model again unitl all the variable p-value is < 0.5 obtain.

9.3) Check the VIF value on feature Variabel : if the VIF value has > 0.5. Drop that column one by one until Vif value has < 0.5

9.4) Check the Model Accuracy : And also check the model sensitivity , calculate specificity, Calculate false postive rate , positive predictive value

9.5) Ploting the ROC curve : It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

**9.6) Choose the optimal Cutoff : **Optimal cutoff probability is that prob where we get balanced sensitivity and specificity

From the curve 0.37 is the optimum point to take it as a cutoff probability.

9.7) Perform Precision and Recall : also see confusion matrix and perofrm Precision and recall tradeoff From the curve above, 0.43 is the optimum point to take it as a cutoff probability.

Take the avarage of optimum cutoffs obtained by Accuracy, Sensitivity, specificity curve and Precision Recall Curve as the optimum cutoff for predictions i.e. (0.37 + 0.43)/2 = 0.4

10) **Make Prediction on Test DataSet also**

At the end of the Analysis and model building using logistic Regression I can say that the model Has the **81 % Accuracy**

Top three variables in model which contribute most probability of a lead getting converted.

Ans :

1) Lead Origin Lead Add Form

2) Last Activity had a Phone Conversation

3) Lead Source_Welingak Website

Also I can Give **Recommendation & Important variable** Company should focus on more to convert more lead

1) API & Landing page Submission

2) Lead Add Form

3) Google , Olark chat,Direct Traffic

4) Total Time spend on Website

5) City

6) E-mail Opend

7) SMS Alert

8) Lead Source Welingak Website

9) Had a Phone Convertion