SUMMARY

Goals of the Case Study

There are few goals for this case study:

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

The following are the steps:

1 Cleaning the Data: -

Firstly all the columns which have select value were replaced by NaN as select means a person has not selected any option. The data was cleaned by dropping columns having more than 40 % of null values. Many columns were having 90-95% value either yes or no, since this means zero variance therefore we will drop these columns as well.

2. EDA: -

A quick EDA was done(both univariate and bivariate analysis) to check the data. Few outliers were found in numerical columns and they were capped.

Total Visits and page per view visit have outliers hence capping them.

3. **Dummy Variables: -** The dummy variables were created.
4. **Scaling: -** For numeric values scaling was done using MinMaxScaler.

5. Train-Test Split: -The split was done at 70% and 30% for train and test data respectively.

6. Model Building: -

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

7. Model Evaluation: -

A confusion matrix was made and **the optimum cut off value was 0.337** and the training set has the **Sensitivity of 80.37 %, Specificity of 80.93 % and Accuracy of 80.72 %**

8. Prediction: -

Prediction was done on the test data frame and with an **optimum cut off as 0.337** with **Sensitivity of 80.46 %, Specificity of 81.40 % and Accuracy of 81.02 %**

9. Precision-Recall: -

This model was also rechecked using precision and recall and a cut off **0.41**. With **Precision of 72.20 % , Recall of 80.37 % for training set** and **Precision of 73.84 % and Recall of 80.45 %**

Top three features are with their coefficients

a) Total Time Spent on Website - 4.545584

b) Lead Origin Lead Add Form - 4.268686

c) What is your current occupation_Working Professional - 2.808038

As these three columns are contributing more towards increasing the conversion rate.

Recommendations

1. We need to focus on employed or working professionals as they have higher chances of enrolling themselves.

2. We need to give briefing to leads as in what are the opportunities they will be getting after doing this particular course and it will lead to their professional growth.

3. Use broadcast messages, emails to reach out to the maximum audience.

4. Do not focus on students as they are already involved in some courses.

5. Focus on features with positive coefficients for targeted marketing strategies.

6. More budget can be spent on Olark chat as more leads are coming from them in terms of advertising.

7. Incentives or discounts for providing reference that convert to lead.

8. The company should make calls who spend more time on website.

9. They should not make more calls to the leads who chose the option of Do not email as yes.

10. More the number of visits to the website more are the chances of the lead conversi on.