Summary Report

Problem Statement

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.

Solution:

We started by importing the required Libraries and Data Set, Reviewing the data set by using functions like Shape, info, and describe. Post understanding the data, we started with cleansing it.

Data Cleaning

Most of the values were not Selected during filling out the form, Therefore Values were reflected as “Select”, We replaced them with Null values throughout the set. Once completed we checked the total null value percentage in each column.

We decided to drop columns with a higher than 40% value as null and below-mentioned columns were dropped:

'Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Activity Score','Asymmetrique Profile Score','Lead Quality','Lead Profile','Tags','How did you hear about X Education'

Imputing missing values

We decided to replace below-mentioned null values in dataset:

'Country'='India'

['Specialization'] = 'Not Available'

['What is your current occupation'] = 'Not Available'

['What matters most to you in choosing a course'] = ('Not Available')

['City'] = ('Mumbai')

We decided to drop null value rows of 'TotalVisits', 'Last Activity', 'Lead Source' columns.

We dropped unwanted columns such as 'Lead Number', 'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque', Digital Advertisement, City, Last Notable Activity, What is your current occupation

Checking for Outliers

We used a box plot in columns 'TotalVisits' & 'Page Views Per Visit' to find Outliers in the data set.

Univariate Analysis

We used Countplot with below-mentioned columns to get better clarity of the data:

'Lead Origin', 'X Education Forums', 'Do Not Email', 'Do Not Call', 'Converted', 'Last Activity', 'Country', 'Newspaper', 'Through Recommendations', 'What matters most to you in choosing a course', Search, Specialization, Lead Source.

We used a histogram for Total Visits, Total Time Spent on Website & Page Views Per Visit.

Segmented Analysis

Bivariate Analysis

We used a heatmap to check bivariate analysis.

Data Preparation

- We found columns with values as objects and then created dummy variables out of the required Columns. Dropped columns which weren't required for model Building.

We split the data for train and test in 70-30 ratio, applied Min-max scaler, and took input as 'TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'.

Model Building

- Imported required Libraries like Logistic Regression and , RFE from Sklearn.
- Dropped one column at a time with a P-value higher than 0.05, and checked P-value and VIF score.
- Agreed with columns with P-value less than 0.05 and VIF score below 5.

Model Evaluation

- Calculated Sensitivity and specificity using metrics from SKlearn library.
- Created ROC curve on the train set.
- Checked accuracy, sensitivity & specificity at different values of probability cut-offs using confusion-matrix

Precision-Recall

So to increase the above percentage we need to change the cut-off value. After plotting we found the optimum cut-off value of **0.38** which gave an accuracy of 81%, precision at 73.40%, and recall of 77.89%.

Conclusion

- Train

  Accuracy : 0.81

  Sensitivity :0.77

  Specificity : 0.82

- Test

  Accuracy: 0.813

  Sensitivity : 0.79

  Specificity : 0.83

Important Parameters

- Total Visits
- Time spent on website
- What is your current occupation_Working Professional