

Summary

This analysis for X education to boost the sales of their course we were provided with basic data about the potential customer.

The following are the steps used:

- i. **Evaluating & Cleaning Data:**
 - The data had few null values 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.
 - Few of the null values were changed to 'empty' so as to not lose much data. Many redundant column were dropped.
 - There were many values from India and few from outside, the elements were changed to 'india', 'not_india' and 'empty'.
- ii. **EDA:**
 - Various relevant graph were made.
 - Univariate and bivariate analysis.
 - Numeric values seems good and no outliers were found.
- iii. **Dummy Variables Creation:**
 - Dummy variables were create for categorical.
 - Scaling was over numerical variables.
- iv. **Train-Test split:**

Split was done at 75% and 25% for train and test data respectively.
- v. **Model Building:**
 - Firstly, RFE was done to attain the top 15 relevant variables.
 - Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with $VIF < 5$ and $p\text{-value} < 0.05$ were kept).
- vi. **Model Evaluation:**
 - Confusion matrix was made.
 - Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.
- vii. **Prediction:**

For a cut of 0.4 we have an accuracy of 0.8 and sensitivity and specificity of 0.753 and 0.836 respectively
- viii. **Precision - Recall:**

Method was also used to recheck and a cut off of 0.4 was found with precision around 74% and recall around 75% on the test data frame.

It was found that variables that mattered the most in the potential buyers are:

1. Total time spends on the Website.
2. Total number of visits.
3. When the lead source was
 - a. Welingak website
4. When the last activity was:
 - a. SMS
 - b. Olark chat conversation
5. When their current occupation is as a working professional.
6. When the lead origin is Lead add format.
7. When the last notable activity was:
 - a. Had a phone conversation
 - b. unreachable

By PRITHVI RAJ CHAUHAN
SARTHAK GUPTA
JEEVANJYOT SINGH