

Summary Report

This Logistic Regression analysis was done for converting more leads to join the courses offered.

The Data provided had lots of data that has to be removed, edited into meaningful/useful data. Some missing values have to be either dropped or assumed values before analysing.

We have to validate the conversion rate of each leads to identify the Hot Leads (Leads with potential conversion) and Cold Leads (Leads with lower conversion rates).

The following are the steps used:

1. Cleaning data:

The data had Null values/ Missing values (values which might not be filled by leads). Such fields had to be handled.

Fields with majority null values can be removed for further validation, Fields with minor misses can be handled by assigning data which constitute to the majority data of the field or a new field can be added like 'Not Specified' at suitable cases.

2. EDA:

An EDA was performed over the dataset once cleaning is done to visualise the relationship among various variables.

The outliers in few fields were removed. Some fields were dropped after EDA as they might not be helpful in model building (For example, country column was dropped as almost all the entries were from India except a few entries).

3. Dummy Variables:

The dummy variables were created and MinMaxScaler was used for numeric values.

4. Train-Test split:

The data split was done at 70% and 30% for train and test data respectively.

5. Model Building:

The top 15 variables were obtained using RFE. Then the variables which were irrelevant was removed one after another based on their VIF values and P-values. (The variables with $VIF < 5$ and $p\text{-value} < 0.05$ were kept).

6. Model Evaluation:

A confusion matrix was created. Optimum cut off value was found. The accuracy, sensitivity and specificity were also found using ROC curve.

7. Prediction:

Prediction was done on the test data frame and with an optimal cut off as 0.3.

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

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- current occupation_Working Professional
- Lead Source_Direct Traffic
- Lead Source_Google
- Lead Source_Organic Search

Considering this, the X Education can identify their Hot Leads and Cold Leads and focus on the hot leads for effective conversion of leads.