

Summary

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. **Cleaning data:**

The data was 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 'not provided' so as to not lose much data. Although they were later removed while making dummies.

2. **EDA:**

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no impactful outliers were found.

3. **Dummy Variables:**

The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

4. **Train-Test split:**

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

5. **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).

6. **Model Evaluation:**

A 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% approximately.

7. **Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.38 with accuracy, sensitivity and specificity of 80%.

8. **Precision – Recall:**

This method was also used to recheck and a cut off of 0.42 was found with Precision around 75% and recall around 76% on the test data frame.

Variables that mattered the most in identifying potential Customers:

1. Total visits.
2. The total time spends on Website.
3. When the lead origin was:
 - a. Lead Add form
4. When the last notable activity was:
 - a. Had a phone conversation
 - b. Unreachable (Though Customer was unreachable it has a positive coefficient)
5. When the last activity was:
 - a. SMS sent
6. When the Lead source is "olark chat".
7. When their current occupation is as a working professional.

Variables that negatively affected lead scoring:

1. When the lead origin was:
 - a. Landing Page submission
2. When the last activity was:
 - a. Olark Chat conversation
3. When the lead source is "olark chat".
4. When their current occupation is Not provided.

Conclusion:

X Education can leverage these insights to tailor their strategies, focusing on key variables to maximize conversions. By emphasizing positive factors like total visits and lead origin while mitigating negative influences, they can attract a significant number of potential customers and encourage them to enrol.
