

Summary

The model is being built for Company X Education to identify strategies for converting potential users. The goal is to understand and validate the data to draw conclusions that help target the correct audience and improve the conversion rate. Let's discuss the steps followed in this process:

1. EDA:

- Checked for missing values in the data set and then Dropped all the columns which is having missing values greater than 3000.
- Here we can see Prospect ID and Lead Number are having 0 null values and having unique values, but these will not be helpful to us. So, we dropped these 2 columns.
- Also noticed that when we got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points. These include Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, 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. Since practically all of the values for these variables are No, it's best that we drop these columns as they won't help with our analysis.
- Dropped the null value rows present in the variable 'What matters most to you in choosing a course', 'Total Visits', 'Lead Source', 'What is your current occupation', 'Specialization'.
- We also worked on numerical variable and dummy variables.

2. Train-Test split & Scaling:

- The data was split into 70% for training and 30% for testing.
- Min-max scaling was applied to the variables: ['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'].

3. Model Building

- RFE was used for feature selection.
- Then 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.

4. Model Evaluation

Sensitivity – Specificity Evaluation:

For the Sensitivity-Specificity evaluation, we utilized the ROC curve to determine the optimal cutoff point.

Training Data:

- The area under the ROC curve (AUC) was 0.86, indicating strong model performance.
- After plotting the ROC curve, the optimal cutoff value was found to be 0.42, resulting in the following metrics:
 - **Accuracy: 79.5%**
 - **Sensitivity (True Positive Rate): 80.9%**
 - **Specificity (True Negative Rate): 78.2%**

Test Data:

- When evaluating the model on the test data, we observed:
 - **Accuracy: 78.4%**
 - **Sensitivity: 79.6%**
 - **Specificity: 77.3%**

Precision – Recall Evaluation:

For Precision-Recall evaluation, we focused on optimizing the balance between precision and recall.

Training Data:

- **Initially, the precision and recall values were:**
 - **Precision: 80.9%**
 - **Recall: 73.1%**
- To improve these values, we adjusted the cutoff value. After plotting, the optimal cutoff was determined to be 0.44, yielding:

- **Accuracy: 79.1%**
- **Precision: 78.09%**
- **Recall: 78.9%**

Test Data:

- On the test data, we obtained:
 - **Accuracy: 78.6%**
 - **Precision: 77.5%**
 - **Recall: 77.8%**

-
- When evaluating with Sensitivity-Specificity, the optimal cutoff value is 0.42.
 - When evaluating with Precision-Recall, the optimal cutoff value is 0.44.

The choice of evaluation metric depends on the business objective—whether maximizing sensitivity (recall) or precision is more important for conversion rates at Company X Education.

CONCLUSION

TOP 3 VARIABLE CONTRIBUTING TO CONVERSION:

1. **Total Time Spent on Website:** This feature shows the most significant impact on lead conversion. A higher time spent on the website indicates greater interest and engagement, which increases the likelihood of conversion.
2. **Lead Origin_Lead Add Form:** Leads generated through this origin tend to have a much higher conversion rate. This reflects a clear intent to enroll or inquire further.
3. **TotalVisits:** Although not as significant as the other two, this feature still plays a crucial role, as more visits often correlate with higher engagement, leading to conversion.

The model appears to predict the conversion rate effectively, providing the company with confidence to make informed decisions based on its outcomes.