Batch DS C65

Assignment - Lead Scoring Case Study

Group Members

1. Swetha Ragavendra - 9845938104

2. Shibani Roy Choudhury - 9869568488

3. Suman Kumar- 9145242482

Summary

This analysis was conducted for X Education to identify strategies for attracting more industry professionals to enroll in their courses. The initial dataset provided insight into how potential customers interact with the site, including their visit patterns, time spent, referral sources, and conversion rates.

1. Data Cleaning:

The dataset was mostly clean, but some null values required attention. We replaced "option select" with null as it didn't provide useful information. Some null values were updated with mode in case of categorical columns, to retain data integrity, though they were later removed during dummy variable creation.

2. Dummy Variables:

We generated dummy variables, in case of categorical columns. For numerical data, we use Standard Scaler for scaling.

3. Train-Test Split:

The data was divided into 70% for training and 30% for testing.

4. Model Building & Training:

We used Recursive Feature Elimination (RFE) to identify the top 15 relevant variables. Remaining variables were manually excluded based on VIF (< 3) and p-value (< 0.05) criteria.

5. Model Evaluation:

A confusion matrix was created, and the optimal cutoff value was determined by plotting accuracy, sensitivity and specificity for various probabilities. The resulting optimal cut-off probability we got as .35, on that optimal accuracy, sensitivity, and specificity, which approximately between 70%-80% each.

Evaluation Metrics for the Training Dataset:

- Accuracy: 0.80
- Sensitivity: ~0.80
- Specificity: 0.79
- Precision: 0.71
- Recall: 0.81

6. Prediction:

Predictions were made on the test dataset using an optimal cutoff of 0.35, resulting in accuracy, sensitivity, and specificity was nearly same as in the range 70% - 80%.

Evaluation Metrics for the Test Dataset:

- Accuracy: 0.80
- Sensitivity: ~0.81
- Specificity: 0.79
- Precision: 0.71
- Recall: 0.81

7. Precision-Recall Analysis:

A precision-recall method was also used to confirm the findings, with a 0.35 cutoff yielding a precision of around 71% and recall of around 81% on the test data.

Conclusion

Key variables influencing potential leads include:

- Lead Origin: Add Form
- Lead Source: Welingak Website, Olark Chat
- Last Activity: SMS Sent, Email Opened
- Total Time Spent on Website

By focusing on these factors, X Education can significantly increase the likelihood of converting potential buyers into course participants.