Executive post graduation in Data science

Batch: DS C67

Assignment - Lead Scoring Case Study

Submitted by:

1) Tummalacheruvu Baji Babu (tummalacheruvu3120@gmail.com)
2) Ayushi Tyagi (tyagiayushi541998@gmail.com)
3) Kriti Tiwari (tiwarikriti54@gmail.com)

Summary

This analysis was carried out for X Education to determine methods for drawing more industry professionals to their courses. The initial data set revealed how prospective clients engage with the website, including their browsing behavior, duration of visits, sources of referrals, and rates of conversion.

I. Data Cleaning:

The dataset was generally in good shape, but a few null values needed to be addressed. We replaced "option select" with null since it lacked useful information. For categorical columns, some null values were updated using the mode to maintain data integrity, though these were ultimately discarded during the creation of dummy variables.

II. Dummy Variables:

We created dummy variables for the categorical columns. For the numerical data, we applied Standard Scaler for normalization.

III. Train-Test Split:

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

IV. Model Building & Training:

We employed Recursive Feature Elimination (RFE) to determine the top 15 relevant variables. The remaining variables were manually removed according to the criteria of VIF (< 3) and p-value (< 0.05).

V. Model Evaluation:

A confusion matrix was generated, and the ideal cutoff value was established by plotting accuracy, sensitivity, and specificity across different probabilities. The optimal cutoff probability we identified was 0.35, corresponding to accuracy, sensitivity, and specificity, each approximately between 70% and 80%.

Evaluation Metrics for the Training Dataset:

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

VI. Prediction:

Predictions were performed on the test dataset using an optimal cutoff of 0.35, yielding accuracy, sensitivity, and specificity that were similar, falling within the 70% to 80% range.

Evaluation Metrics for the Test Dataset:

- Accuracy: 0.80
- Sensitivity: ~0.81

- Specificity: 0.79
- Precision: 0.71
- Recall: 0.81

VII. Precision-Recall Analysis:

A precision-recall approach was also utilized to validate the results, with a cutoff of 0.35 resulting in a precision of approximately 71% and a recall of about 81% on the test dataset.

Final Conclusion

Key factors affecting 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 concentrating on these elements, X Education can greatly enhance the chances of turning potential buyers into course participants.