Lead Scoring Case Study

Report by

Sanyogita kharat, Rishi sahni and Ayush

Problem Statement:

An education company, X Education, sells online courses for working professionals. Many professionals interested in courses visit X Education's website, where they browse through available courses. We have data on the browsing behavior and limited customer data.

Once these visitors browse the website, watch course videos, and fill out forms, they become potential leads for the company. However, the conversion rate—the percentage of leads that ultimately make a purchase—is relatively low at 30%.

X Education aims to improve this conversion rate and make the lead conversion process more efficient. To achieve this, they need to identify the "hot leads"—those with a high probability of conversion.

The goal is to build a model that assigns a lead score to each lead, indicating the likelihood of conversion. A higher score implies a greater probability of the lead converting.

Data Processing:

1. Data Cleaning:

We began by examining the values and their proportions for each feature/column in the dataset.

During this process, we identified some columns that had "Select" as one of the categorical values. We initially replaced these with null values.Next, we calculated the proportion of null values in each column, and observed that data related to asymmetrical activities was missing for many customers. For categorical variables, when there was a clear mismatch in proportions, we filled the missing values with the mode. For example:

- **Occupation**: Unemployed

- **Country**: India
- **What matters most to you in choosing this course**: Better Career Prospects
- **Asymmetrique Activity Index**: Medium

After handling missing values, we removed all entries that still had null values from the dataset.

2. Feature Scaling:

We applied sklearn's MinMaxScaler to scale down three numerical features: 'time spent on website', 'TotalVisits', and 'number of pages viewed per visit'.

The data was then split into training and testing sets with an 80:20 ratio.

3. Feature Selection:

After one-hot encoding the categorical variables, we ended up with over 180 features. To reduce dimensionality, we used sklearn's Recursive Feature Elimination (RFE) module, and truncate the feature set down to 30 features.

We further removed features with high p-values, ultimately building a model using the following significant features:

- Email Permission
- Time spent on website
- Asymmetrique Activity Score
- Lead Source Reference
- Tags like 'Ringing' and 'Will revert after reading the email'
- Lead Quality indicators showing negativity are successfully predicting non converting leads
- Asymmetrique Activity Index
- Last Notable Activity (Modified)

One important insight was that the total time spent on the website significantly influenced lead conversion. Lead quality proved to be the best predictor of negative conversions. Leads with "Worst", "Maybe", or "Not Sure" in Lead Quality were highly unlikely to convert.

4. Logistic Regression Model:

From the logistic regression model we developed, we obtained the probabilities of lead conversion. These probabilities were mapped to binary outcomes using a threshold of 0.78.

Model Validation:

1. Training Set Evaluation:

The model was first evaluated on the training dataset. Using a threshold of 0.78, the model gave the following results:

- **Accuracy**: 91%
- **Sensitivity** (Recall): 97%
- **Specificity**: 93.7%

These results indicated that the model performed very well on the training data.

2. Test Set Evaluation:

We then tested the model on the unseen test dataset, and the results were similarly impressive:

- **Accuracy**: 90%
- **Sensitivity**: 87%
- **Specificity**: 100%

Conclusion:

The logistic regression model successfully predicts lead conversion with high accuracy, sensitivity, and specificity. By assigning a probability-based lead score, X Education can now efficiently prioritize hot leads, leading to a potentially higher conversion rate and more effective use of marketing and sales resources.