Lead Scoring

Machine Learning 1

Siva Gopavarapu

Manan Sharma

Saheli Paul

Problem Statement

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads
- Propose the changes which will help to increase the conversion rate
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be 80%.

Precursor

- The data provided was first cleaned to remove the null values and the outliers
- Then EDA analysis was performed rule out the variables that are less important

Model

For out use we use the Logistic regression model

For training we split the data into 2 parts train:70% and test:30%.

Feature Selection

We select top 15 features using RFE to reduce the model load. The top 15 features used for trainging are

- `Total Time Spent on Website`
- `Lead Origin_Lead Add Form`
- `Last Activity_Converted to Lead`
- `Last Activity_Email Bounced`
- `What is your current occupation_Working Professional`
- `Tags_Already a student`
- `Tags_Busy`
- `Tags_Closed by Horizzon`
- `Tags_Interested in other courses`
- `Tags_Lost to EINS`

- Tags_None
- Tags_Ringing
- Tags_Will revert after reading the email
- Tags_switched off
- Last Notable Activity_SMS Sent

It was found that following features have p value more than 0.5 and hence were removed.

- Tags_Interested in other courses
- Tags_Already a student
- Lead Origin_Lead Add Form

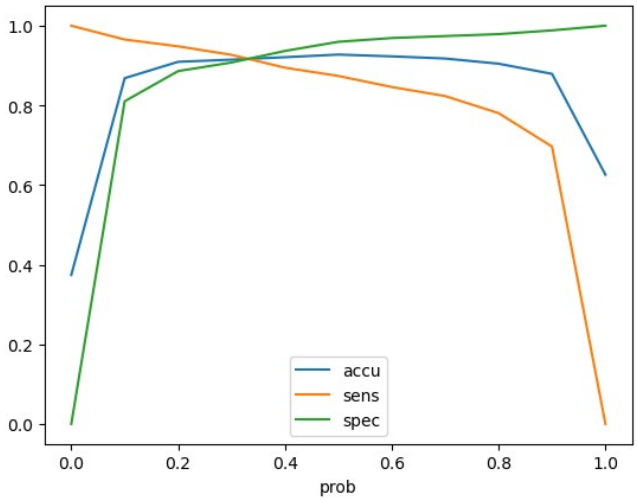
Multicollinearity Check

The remaining featured were tested for multicollinearity using VIF. The feature with highest score of 1.644171 was

Tags_Will revert after reading the email

Since the score is fairly low the feature was not removed.

Cutoff



For determining the cutoff we compared accuracy, specificity and sensitivity and found it to be 0.3.

Conclusion

The final results for the model are

Training:

- Accuracy: 91.48%
- Sensitivity: 92.68%
- Specificity: 90.76%

Testing:

- Accuracy: 91.57%
- Sensitivity: 92.11%
- Specificity: 91.24%