Hetal Khanapure (hetu.parmar@gmail.com)

Lead Scoring Case Study

Summary Report

Problem Statement:

- X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Solution Summary:

Step 1: Data Understanding

Read and understand data.

Step 2: Data Cleaning

- Convert ‘Select’ values to null value.
- Handling missing/null data: we dropped the variables having high missing value percentage.
- Dropped sales related variables to identify hot leads from the data generated directly from source.
- Dropping variable which have skewed data.
- Identify the outliers and treat them.

Step 3: Exploratory Data Analysis

Perform univariate, bivariate analysis for categorical variable and numerical variable to get more insight. Check the correlation metrics for continuous variables.

Step 4: Data Preparation

- Converted binary variables (yes/no) to 0 and 1.
- Created dummy features for categorical variables.

Step 5: Test/Train split

Divide the dataset into Test and Train dataset with proportion of 70%-30% rule.

Step 6: Rescaling the Features

We have used Standardization scaling to scale the numerical variables. After that we build basic model with all the features.

Step 7: Feature selection using RFE

- As we have more variables in model we used recursive feature elimination and selected top 15 features. Using the statistics generated by model we look at the p value and dropped insignificant variables until all the variable's p-value are significant.
- Then we checked the VIF of final variable to check multi-collinearity between variables.
- From confusion metrics we calculated accuracy, sensitivity, specificity, false positive rate, positive predictive value and negative predictive value.

Step 8: ROC Curve

We plotted ROC curve and we get area under the curve as 84% which is good.

Step 9: Finding optimal cutoff

We plotted a graph of accuracy, sensitivity and specificity of different probabilities to find out optimal cutoff value. The cutoff point found out to be 0.28.
Based on the cutoff we got

- Accuracy (79.3%)
- Sensitivity (78.7%)
- Specificity (79.6%)

Step 10: Precision and Recall metrics

We found out that Precision and recall values came out to be 70.6% and 78.9%. Based on precision and recall tradeoff Cutoff value came out to be 0.36.

Step 11: Prediction on Test Dataset

We implemented the learning to the test dataset and calculated the probability of lead conversion and also calculated the lead score. We found out on the test dataset

- Accuracy: 78.8%
- sensitivity: 78%
- specificity: 78.7%