Lead Scoring Case Study

upGrad & IIITB | Data Science Program

- Gopi Jagini & Ashik Mohammed S

The main objective of this study is to help the organization named "**X Education**", which sells online courses to industry professionals, build a model to help them design a good marketing strategy to obtain improved lead conversion rate. Currently the organization is getting leads through different sources like Online Search, Reference, Ads, etc., and from professionals of various specializations. The current conversion rate is 30%, which is very low. T**wo main goals** of this study are, one to build a **logistic regression model** to assign a lead score between 0 and 100 to each lead and second is to **provide adjustments** in the model based on the company's requirements in the future.

So, here with the available data, we followed the below approach in building a suitable Logistic Regression model.

1. Imported required modules and created a data frame with the input data.
2. Performed initial data checks to check data type and missing values of each feature/variable.
3. Next, we cleansed the data. Dropped the columns with more than 40% null values. Also, those columns with one unique value have been dropped.
4. For columns with null percentage less than 40%, imputed empty cells with 'NA'. Eg: variable, 'Country' imputed NA with "India" (mode).
5. Performed Exploratory Data Analysis. First, Numerical variables to understand its distribution and outliers. Eg: Removed Outliers from 'Page Views per Visit'. Checked the influence of categorical variables against target variables.
6. Then we split the data into Train and Test sets in 70:30 ratio. Performed scaling of numerical variables to bring them onto a same scale of comparison.
7. Next, we started building 'Logistic Regression' model using 'sklearn' library. Used "RFE" methodology for feature selection based on p-value and decent VIF socres.
8. Created Prediction with Train data to check the performance of built model. Evaluated the same against certain parameters like Accuracy, Sensitivity and Specificity with the help of the confusion matrix.
9. We have drawn ROC curve to optimize the probability cut-off value.
10. Performed evaluation of the model on Test Data and used 0.42 as cut-off based on the trade-off between Precision and Recall.
11. Converted the Probability of conversion to Lead Score with a multiplication factor of 100.

Below are the top recommendations from the above analysis:

- Top three variables which have impact on the conversion percentage are:
  Total_Visits, Total_Time_Spent_on_Website, Lead_Add_Form.
- Top 3 categorical/dummy variables in the model which should be focused the most, for high lead conversion, are:
  Total Time Spent on Website, Lead 'Origin_Lead Add Form' and 'What is your current occupation_Working Professional'.
- The emphasis should be more on the Lead Sources.
  For Eg: Through Reference, the conversion rate/percentage is very high.
  Through Google Search, quantity of conversions is very high.
- Management related professionals should be targeted more for good conversion.
- New age social media platforms like Instagram, Twitter should be utilized for promotions (rather than Facebook).