Lead Score Case Study

Summary Report

A detailed analysis using Logistic Regression is done for an education company X Education that sells online courses to industry professionals. The raw data from the past of 9000 data points was provided, from which Hot Leads i.e. leads likely to convert into paying customers are identified and assigned lead score based on the most significant model.

Steps followed are:

1. Cleaning Data:
   The data was first cleaned by:
   a. Replacing "Select" by NaN values
   b. Converting all the values in the dataset into lowercase
   c. Removing variables with more than 35% missing values
   d. Replacing import columns NaN values with 'Not Provided'
   e. Identifying the country **India** with maximum customer views.

2. Exploratory Data Analysis (EDA):
   To check the distribution of our dataset and correlation among the variables we performed EDA. It was found that many categorical variables were highly correlated hence dropped before building the model, and there were no visible outliers in the data set.

3. Data Preparation:
   i. Dummy Variable: Dummy variables for 8 category variables were created and later dummies with not provided dummies and the first dummy of each variables were removed from the data.
   ii. The numeric values are appropriate and scaled using MinMax Scaler,
   iii. Train and Test split of data set.

4. Model Building:
   On the train data we use RFE method and identify 15 the most relevant features from the data set. After dropping the high VIF and insignificant variables or features from the data set we finally get a 12 feature model all significant at 5% level with VIF less than 5 for all.

   Final Model:

$$
\begin{aligned}
Converted = \; & -1.25 + 4.76 TotalVisits + 4.55 Total\ Time\ Spent\ on\ Website \\
& + 2.68 Lead\ Origin_{leadaddform} - 1.47 Lead\ Source_{DirectTraffic} \\
& - 1.16\ Lead\ Source_{Google} - 1.26\ Lead\ Source_{OrganicSearch} \\
& + 2.59\ Lead\ Source_{WelingakWebsite} - 1.42\ Do\ Not\ Mail_{yes} \\
& - 1.47\ Lead\ Activity_{OlakChatConversation} + 1.30\ Lead\ Activity_{SMSSent} \\
& + 2.79 What\ is\ your\ current\ occuaption_{WorkingProfessional} \\
& + 1.68\ Last\ Notable\ Activity_{Unreachable}
\end{aligned}
$$

5. Model Evaluation
   Using Confusion matrix and ROC – accuracy, sensitivity and specificity were calculated. At the cutoff of 35%, we achieved ROC to be 87% and:
   a. Accuracy 80%
   b. Sensitivity 80.28%
   c. Specificity 79.93% =~ 80%

   This show that the final model gives high consistency and accuracy in data as the results have minimum variance.

6. Prediction
   The data was then tested on the model by predicting the y-converted values and we identify the hot leads based on the lead score >=80%.

7. Precision- Recall
   Finally we check for precision at the cutoff rate of 41% and we achieve Precision at 73% and Recall to be 75%.

Thus the increase in the following most relevant factors that would increase the conversion rate are:

a. Total Visits
b. Total Time Spent on Website
c. Lead Origin-Lead Add Form
d. Lead Source_Welingak Website
e. Lead Activity SMS sent
f. Current occupation_ Working Professional