Summary Report of Lead Score Case Study

An X Education company sells online courses and markets its website on search engines like Google and other media platforms. Once the user lands on website and fills up personal details then, the user is classified as a lead. Once the lead enrolls for the course, it is classified as converted. The current conversion rate is 37.85%. In order to increase conversion rate we built a model which will assign lead score to each lead. This will help sales team to focus on hot leads.

The data has 9240 rows, 37 columns. During data preparation process, we dropped 6 columns having <70% missing values, dropped some unnecessary columns and rows with null values <2%.

According to the observations based on EDA , in order to increase the conversion rate, we can suggest following points:

We need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' Lead Origins and also increasing the number of leads from 'Lead Add Form'.

We need to focus on increasing the conversion rate through 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from 'Reference' and 'Welingak Website'.

Websites can be made more appealing so as to increase the time of the Users on website.

We should focus on leads having last activity as Email Opened and SMS sent by making a call to those leads.

We need to increase the number of Working Professionals and unemployed leads by reaching out to them through different social media sites such as LinkedIn etc.

After EDA we created dummy variables for certain columns and did train-test data split in ratio of 70:30. We dropped some of the features which were highly correlated. We used RFE for feature selection. We dropped some features with VIF>5 and p-value>0.05.

In order to evaluate the model we took help of multiple evaluation metrics. According to the ROC curve plot, we inferred that, the curve is inclined towards Y-axis and area under the curve is more. Hence, the model is accurate. After plotting, optimum probability cut-off plot and precision-recall plot, we inferred that, the optimal cut-off point is approximately 0.27.

After performing above tests for evaluating the model, we built final Logistic Regression model with 14 features. The model predicts probability of the target variable having certain value. The cut-off probability is used to obtain predicted value of target variable from 0 to 100 which is "Lead Score".

The optimum cut off is 0.27 i.e. any lead with probability of getting converted greater than 27% is predicted as Hot Lead (customer will convert) and any lead with 0.27 or less probability of getting converted is predicted as Cold Lead (customer will not convert).

The final model has Sensitivity of 0.928, this means the model is able to predict 92% customers out of all the converted customers, (Positive conversion) correctly. The final model has Precision of 0.68, this means 68% of predicted hot leads are True Hot Leads.

IIITB Data Science Batch C44 Group Partners

- Divya Shah.
- A. Bhargav Kumar.
- Siddhesh Parab.