Lead Score Case Study – 7th Dec 2020

Name: Muruganand Kumaran & Karthik Veluchamy Sarguru

Solution Acceptance Criteria:

The output logistic regression model must able to generate Lead Scores between 1 – 100 to showcase the cutoff point that determines whether it is “Hot” lead or “Cold” lead. The Sensitivity is expected to be close to 80% so that we can recommend right variables that potentially the sales team can consider moving forward for leads to be converted positively.

Solution Approach:

Identify Data:
Data is exported into pandas dataframe and analyzed for nulls & missing values. Those columns which have higher percentage of missing or null values are removed. We removed columns like 'Tags','Lead Quality','Lead Profile','Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Activity Score','Asymmetrique Profile Score', and 'How did you hear about X Education' which got more than 30% of null values.
We see several categorical variables which got high level of data skewness. They are also removed. They are 'Magazine','Receive More Updates About Our Courses','I agree to pay the amount through cheque','Get updates on DM Content', and 'Update me on Supply Chain Content'.

EDA:
Bivariate analysis are performed against the converted column and noticed the categorical variables like Last Notable activity, Specialization, Occupation & Lead Source. We see even distribution and removed all biased categorical columns.

Model Building:

Create Dummies:
Dummies are created for categorical columns 'Lead Source','Lead Origin','Specialization','Last Notable Activity','City','Country','A free copy of Mastering The Interview', and 'What is your current occupation'.

Splitting Test & Train set:
Using train_test_split function split the dataset into test & training dataframes.

Scaling:
Scale the numerical columns like 'TotalVisits', 'Total Time Spent on Website', and 'Page Views P er Visit' using function fit_transform

Feature selection based on RFE:

Using recursive feature elimination method, we rank the variables to be considered for initial modeling and considered the output for first iteration.

Iteration 1:

Based on Iteration 1 output the P-Values are analyzed. The “Occupation_HouseWife” features is removed based on highest P-Values.

Feature selection based on VIF:

After Iteration-1 VIF analysis is performed to identify multicollinearity between features. We identified features where VIF > 5 and removed lead_origin_Landing Page Submission.

Iteration 2, 3 & 4:

After Iteration 2, 3 & 4 we removed columns related to city are removed based on highest P-Values. Final VIF is also performed to make sure the features selected doesn’t cause any multicollinearity by verifying values < 5.

Model Assessment:

Initially probability > 0.5 is considered as convertible lead and proceed with populating the confusion matrix. We obtained accuracy, specificity, sensitivity and plotted ROC. We obtained accuracy as close to ~80%

Finding optimal cutoff point:

We calculated accuracy, sensitivity and specificity for various probability cutoffs and plotted them to identify the ideal point to determine “Hot” Leads point. We get them as 0.33. After obtaining the cutoff point we calculated accuracy as 79% which is good.

Model Testing:

Model testing is performed on test dataset and obtained accuracy as 80% which shows the model is success and works as expected. Based on probability the lead score is generated between 0-100 and determined score > 33 is “Hot” lead.

Conclusion:

We recommend criteria to be considered for identifying “Hot” Leads:

1. Had a phone call.
2. Mumbai city
3. Other country.
4. SMS sent already.
5. Who spend more time on the website and filled in the form.
6. Landing from Welingak website.
7. Occupation – Working professional.