Lead Score Case Study – Summary

For the given Lead score problem statement we used the Logistic regression algorithm to proceed with analysis.

In the EDA process, ‘select’ has been replaced with np.nan and then calculate the percentile of missing value in the data set. Columns with more than 45% missing values. Country column has been dropped as near to 97% data having country as “India” and ‘mumbai’ has been imputed for the missing values in ‘city’ as mumbai is the major city in the data set. Specialization may be blank because either its is not present in the drop down list of the operator or may be the student has still not chosen any #specialization yet. So NaN values will here be replaced with 'Not Specified'

It seems to be three major category of people -"Unemployed/Students and Working professionals with Working professionals seems to be going for the courses so imputing Nan values with mode "Unemployed" and combining Other /Housewife/Businessman as one category "other". Dropping the column ‘what matters most to you in choosing a course’ since most of them have a single value and blank,this variable is not carrying enough information to be an important variable. Replacing ‘NaN’ with ‘Not Specified’ with for the Tags. NaN values being replaced and low frequency values were combined to ‘others’ for Lead source.

Maximum number of leads are coming through Google and Direct traffic. Conversion Rate of leads through reference and through visiting website is high. Improvement of overall lead conversion rate could be achieved by focusing on olark chat, leads coming through organic search, direct traffic, and google leads .Larger focus should be done on giving proper incentives to references and improving visiting website for the coming traffic.

API and Landing Page Submission seems to be bringing lots of lead and converting too. Lead Add form seems to be having very good conversion rate but substantially less leads are coming with this medium. Lead Import and Quick Add Form get very few leads. API and Landing Page Submission should be focussed to improve the conversion .Lead Add form should be targeted to bring in more leads.

Majority of user are committing activities like "Modified" or "Email Opened" . Users which have been receiving SMS seems more likely to getting converted which being to the forefront concept of personalization. Dropping the rows with NaN values as rows that are being dropped are 2 % which will not impact the analysis.

As ‘total visits’ and ‘page views per visit’ have outliers and there was a sudden increase after 90th percentile, so removing top and bottom 1% of the column outlier values. There seems to be a strong possibility of conversion with the time spent on the website. May be a effort be made to make the website more engaging and user friendly. Created dummy variables on lead origin, specialization, lead source, last activity, Last notable activity, Tags and split into Train and Test, Standard scaling has been applied on train dataset.

15 features have been selected using RFE and dropping ‘Lead Source_Referral’ as it has p-value greater than 0.05 in Model 1 and dropping ‘Last Notable Activity_SMS Sent’ as it has greater p-value than 0.05 in Model 2.

Model 3, considered to be the stable model as it has p-values less than 0.05 and no multicollinearity has been observed with 92.29% accuracy and 92.66% specificity, when the ROC curve getting 0.97 value and 0.3 optimal cut-off. And test set having 92.78% accuracy and 93.26% specificity.

Below are few lead score generated on Test data with Prospect ID

Prospect ID	Lead Score
7681	0.024819
984	0.025692
8135	0.686054
6915	0.005880
2712	0.953208
244	0.002398