Lead Score Case Study – Summary

To analyse the given Lead score problem statement, we used the Logistic regression algorithm.

'select' has been replaced with np.nan in the EDA process, which then calculates the percentile of missing values in the data set. Missing values in more than 45% of the columns. As close to 97% of the data has 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 it is not present in the operator's drop down list or because the student has yet to select a #specialisation. As a result, NaN values will be replaced with 'Not Specified' in this case.

There appear to be three major groups of people: "unemployed/students and working professionals with disabilities." Working professionals appear to be taking the courses, so assigning Nan values with the mode "Unemployed" and combining Other/Housewife/Businessman as one category "other." Dropping the column 'what matters most to you in choosing a course' because most of them have a single value and are blank, and this variable does not carry enough information to be considered important. For the Tags, replace 'NaN' with 'Not Specified'. For the Lead source, NaN values were replaced and low frequency values were combined to 'others.'

The majority of leads are generated by Google and direct traffic. Lead conversion rates from referrals and website visits are both high. Focusing on olark chat, leads from organic search, direct traffic, and google leads could help improve overall lead conversion rate. A greater emphasis should be placed on providing appropriate incentives to references and improving the visiting website for future traffic.

API and Landing Page Submission appear to be generating a lot of leads and converting them as well. Although the Lead Add form appears to have a high conversion rate, it generates significantly fewer leads. Lead Import and Quick Add Form both generate a small number of leads. API and Landing Page Submission should be prioritised to increase conversion. To attract more leads, the Lead Add form should be targeted.

The vast majority of users engage in activities such as "Modified" or "Email Opened." Users who have received SMS appear to be more likely to convert, bringing the concept of personalization to the forefront. Dropping the rows with NaN values because the number of dropped rows is 2% and will have no effect on the analysis.

Because 'total visits' and 'page views per visit' have outliers and there was a sharp increase after the 90th percentile,' the top and bottom 1% of the column outlier values were removed. With the amount of time spent on the website, there appears to be a high likelihood of conversion. It is possible that an effort will be made to make the website more engaging and user friendly. Dummy variables on lead origin, specialisation, lead source, last activity, last notable activity, Tags were created and split into Train and Test datasets. Standard scaling was applied to the Train dataset.

15 features were chosen using RFE, with the exception of 'Lead Source Referral,' which has a p-value greater than 0.05 in Model 1 and 'Last Notable Activity SMS Sent,' which has a p-value greater than 0.05 in Model 2.

Model 3 is considered the stable model because it has p-values less than 0.05 and no multicollinearity, with 92.29% accuracy and 92.66% specificity when the ROC curve has a value of 0.97 and a cut-off of 0.3. Furthermore, the test set has a 92.78% accuracy and a 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