Summary

- The shape of leads dataset is 9240 rows and 37 columns.
- There are 7 numerical columns and 30 categorical columns
- There are many 'Select' values present in various columns in the dataset. These values correspond to the user having not made any selection.
- There are missing/null values in many columns.
- There are no duplicate values in the dataset
- The conversion rate of leads is 38.54%

Univariate Analysis and Bi-variate Analysis

Lead origin: ***From the above plot and conversion summary, we can infer that:***

- Lead Add Form has the highest conversion rate at 94%
- API and Landing Page Submission have 31% and 36% conversion rate but they generate maximum leads counts.
- Lead Import has the least amount of conversions and leads count.
- To improve overall lead conversion rate, focus should be on improving lead conversion rate of API and Landing Page Submission. Also,generate more leads from Lead Add form since they have a very good conversion rate

Lead Source: ***From the plot and conversion summary, we can infer that:***

- Google and direct traffic generates maximum number of leads but has conversion rate of 40% and 32%.
- Welingak website and References has highest conversion rates around 98% and 93% but generates less number of leads.
- Olark chat and organic search generates significant number of leads but their conversion rate is around 26% and 38%.
- Lead source in 'others' category such as Click2call', 'Live Chat', 'NC_EDM', 'Pay per Click Ads', 'Press_Release','Social Media', 'WeLearn', 'bing', 'blog', 'testone', 'welearnblog_Home', 'youtubechannel' generates very less leads.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic and google lead source. Generate more leads from reference and welingak website since they have a very good conversion rate

Do Not Email & Do Not Call: ***From the plot and conversion summary, we can infer that:***

- Around 99% of customers do not like to be called or receive emails about the course

From the above curve we can see that the optimal cutoff is at 0.35. This is the point where all the parameters - Accuracy,Sensitivity,Specificity are equally balanced

- As per our business objective, the recall percentage is more significant since we don't want to left out any hot leads which are willing to get converted.
- Hence Recall- 81% suggest a good model.

Recommendation

- The sales team of the X-Education should focus on the leads having lead origin - lead add form , occupation - Working Professional , Lead source - Wellingak website.
- Hot Leads are identified as 'Customers having lead score above 35. Sales Team of the company should first focus on the 'Hot Leads'
- The 'Cold Leads'(Customer having lead score <= 35) should be focused after the Sales Team is done with the 'Hot Leads'.
- There are many important variables like city, specialization , occupation which can potentially explain Conversion better.It is important for the management to make few of these information mandatory to fill , so that we can use in our model and build important decisions for the business.
- We have high recall score than precision score. Hence this model has an ability to adjust with the company’s requirements in coming future.
- High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.