Summary

We have performed an intensive analysis for an online education providing company named as **'X Education,** who use to sell their online courses to the industry professionals. Our goal was to analyse the lead data provided to us and to make an efficient model to predict the hot leads so that the sales team can focus on those leads to convert into their aspirant.

The following are the steps used:

1. **Cleaning data:**
The data was cleaned by deleting columns having null values greater than 45%. Missing values were imputed like in the city column, it had higher percentage of Mumbai city and value 'other' was imputed in place of NaN value in specialization column and in the column 'What is your current occupation' the value 'Unemployed' is placed in place of null values

2. Univariate Analysis and Bi-variate Analysis:
Univariate and bivariate analysis was done and based on the analysis we can say that the highest conversion rate is of 'Lead Add Form', at 94% and the 'Landing Page Submission' and 'API' have 36% and 31% conversion rate, respectively. 'Google' and 'Direct Traffic' is generating the maximum number of leads but it has a conversion rate of 40% and 32%. Based on 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.

3. **Dummy Variables:**
The dummy variables were created and then combined with original data set by concatenating both the dataset.

4. **Train-Test split:**
The split was done at 70% and 30% for train and test data respectively.

5. **Model Building:**
Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 3 and p-value < 0.05 were kept). Finally model 5 was considered as our final model.

6. **Model Evaluation:**
A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 81%, 70% and 89% respectively.

7. **Prediction:**
Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

8. **Precision – Recall:**
Precision and recall values were calculated and was found that precision more than 70% and recall value more 75% on the test data frame.

The X-Education should focus on the leads having

- lead origin - lead add form
- occupation - Working Professional
- Lead source - Wellingak website

Sales Team of the company should first focus on the 'Hot Leads' which are identified with having Lead Score above 35.

Once the Sales Team is done with the 'Hot Leads' only after that they should focus on the 'Cold Leads'(Customer having lead score <= 35).

Team can make some important variables mandatory to enter, such as city, specialization, occupation which can potentially explain Conversion better. It could be used in our model and build important decisions for the business.