Summary

The ask here was to build a logistic regression model on an education company's (X Education) leads data and to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

We started with data cleaning. We dropped the columns where the null value was more than 40%. We also removed unique identifiers like 'Prospect ID' and 'Lead Number' because they were of no use in the model building. We also removed the columns where one value was having a much higher share like 'Country', 'Do Not Call', 'Search', 'Magazine' etc. For Specialization and Occupation, we replaced Nulls/Select with *Not Specified* and *Not Provided* respectively because we thought these two columns are important and thus didn't drop these columns. In order to avoid class imbalance, we grouped values of smaller frequencies together and named them as *Others*. We then performed EDA on the cleaned dataset and found some useful insights:

- 'Working Professionals' had a very high conversion rate.
- Customers identified through 'Add Lead Form' are very much likely to convert.
- People who are spending more time on the website are more likely to convert.

We then created dummy variables for the required categorical variables. After splitting the dataset into train-test set and scaling the training dataset, we used RFE to get the top 15 most important features. Then, we build logistic regression model using those features. The p-value and VIF of all the features were in acceptable range.

We plotted the ROC curve and the area under the curve came out to be **0.97**. In order to find optimal cutoff, we calculated accuracy, sensitivity and specificity for various probability cutoffs. **0.26** came out to be the optimal cutoff point. The model accuracy on the training set was **92.27%** and sensitivity (92.72%) and specificity (93.17%) were also very close. We then predicted the conversions in test data and there the model accuracy came out to be **93.23%**. The sensitivity was 91.94% and specificity was 94.02%. This looks like a really good model.

We would like to suggest following recommendations:

- The company should focus more on 'Working Professionals' as their conversion rate is very high.
- Customers identified through 'Add Lead Form' are very much likely to convert.
- The company should also focus more on people who have specialization in a Management course
- People who are spending more time on the website also needs to be targeted aggressively
- Tags, LeadSource and LastActivity are the most important factors causing conversions.