## Summary

### Objective:

X Education has asked to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. It requires to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

### Steps to Analyze:

Data was cleaned and missing values were dropped, categorized and in some cases assigned a not known class.

Data visualization gave interesting insights into the module; however a detailed logistic analysis needs to be done to attain a higher accuracy and better analysis.

Before model building, data needs to be prepared. This involves converting the categorical variables to dummy variables and then numerical variables were scaled using min max scaler. Train test split of 70% and 30% respectively was done.

Later to decrease the features and prevent overfitting RFE technique for feature reduction was used. After dropping the features with high VIF value and high p values, we are left with 9 variables.

Model 1 was built wherein the leads having a probability of less than 0.45 were considered as not converted. These predicted, y target variable is tabulated alongside the y actual and a confusion matrix is constructed. The performance parameters were not satisfactory as accuracy was around 80% but sensitivity is low.

High sensitivity will lead to high lead score for customers not likely to convert, whereas highly specific will show customers with low lead score having no conversion. We can afford to have a high sensitivity so that we don’t loose on the customers who would have converted. So, a tradeoff between specificity and sensitivity is made. The decision for threshold cut off value for probability is made using ROC curve.

An optimum value of 0.3 probability was chosen. Later performance parameters obtained on train and test dataset are as follows:

Confusion matrix for train data set :

[[2994, 888],
[ 394, 1991]]

- Accuracy: 79.54%%
- Sensitivity: 72.45%
- Specificity: 86.60%
- FPR : 13.40%
- Positive predictive value: 76.87%
- Negative predictive value: 83.65%

Confusion matrix for test dataset

[[1267, 409],
[ 154, 856]]

- Accuracy: 79.04%
- Sensitivity: 81.09%
- Specificity: 78.82%
- Precision score: 67.67%
- Recall: 84.75%

**Outcomes:**

The top three variables are:

- Total time spent on website
- Lead Origin
- Lead source
- Customers filling out add forms are more likely
- Customers engaging in an Olark chat window are more likely to convert
- If mode of last communication is SMS more likely to convert

**Suggestions:**

- Drop SMS to the probable high lead score customers.
- Make groups and follow up based on the categories of customers like strong leads, not so strong and weak leads. Make email draft to be sent to these categories.
- Keep the draft emails short and engaging, catch the attention by keeping some free engaging sessions on the subject matter, this will increase the time spent on website.
- Never assume whether its worth deleting cold leads, let the customer tell that themselves.