# LEAD SCORING CASE STUDY SUMMARY

X Education sells online courses to industry professionals. The company needs a model wherein you a lead score is assigned to each of the leads.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

The following are the steps used:

1 **Cleaning data:**

a) Check percentage of null values in columns and drop the columns which have more than 45% missing values.
b) Some of the variables are created by the sales team once they contact the potential lead. We will drop these columns too.
c) Some of the columns have only 1 category These columns can be deleted.
d) Some of the columns have one of the value as "Select" These should be considered as null values. Data Value needs to be updated for these columns

2 **EDA:**

a. Univariate and bivariate analysis of categorical and numerical columns was performed.

b. Checked the correlations between the variables.

c. Detected Outliers and cap them using 99% - 1% quantile range.

3 **Dummy Variables and scaling :**

The dummy variables were created for categorical variables and scaling is done using StandardScaler.

4 **Train-Test split:**

The split was done at 70% and 30% for train and test data respectively.

5 **Model Building:**

RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

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 around ~81% each.

7. **Prediction:**

Prediction was done on the test data and with optimum cut off 0.34 with accuracy, sensitivity and specificity of 80.71%,81.79%,80.03% respectively. sensitivity - specificity Analysis has given more stability as compare to precision - recall Analysis

8 **Precision – Recall:**

This method was used to recheck and a cut off of 0.41 was found with Precision around 86% and recall around 85% on the test data frame.

**Conclusion :**

**Major indicators that a lead will get converted to a hot lead:**

1. **Lead Origin_Lead Add Form :** Chances of conversion of a lead are 2.8 times higher if a lead originate via lead add form than if a lead originate via API
2. **Occupation_Working Professional :**Chances of conversion of a lead are 2.39 Times higher if a lead is a Working Professional than if a lead is a businessman
3. **Lead_Source_Welingak website :**Chances of conversion of a lead are 2.46 times higher if a lead is a sourced via Welingak website than if a lead is a sourced via Direct traffic
4. **Lead Source_Olark Chat :**Chances of conversion of a lead are 1.09 times higher if a lead is a sourced via Olark Chat than if a lead is a sourced via Direct traffic

**Major indicators that a lead will NOT get converted to a hot lead:**

1. **Last_Activity_Olark chat conversation :**The chances of conversion of a lead are 0.61 times lower than if the last activity of the lead Olark chat conversation than if the activity shows converted to lead.
2. **Lead Origin_Landing Page Submission :** The chances of conversion of a lead are 1.02 times lower if the Lead Origin is landing page submission than if the lead origin of the lead is API.
3. **Do Not Email :** The chances of conversion of a lead are 1.18 times lower if the Lead Opted for do not email than if the lead didn't opt for the same.

**Recommendations :**

The company should use a leads score threshold of 34 to identify "Hot Leads" as at this threshold, Sensitivity Score of the model is around 81% which is as good as CEO's target of 80%.