# Lead Scoring Case Study Summary

## Problem Statement:

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.

On any given day, many professionals who are interested in the courses land on their website and browse for courses and they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

## Solution Summary:

**Step1: Reading and Understanding Data**
Reading the given file and analyzing the data to understand the various aspects of it.

**Step2: Data Cleaning**
We dropped the variables that had high percentage of NULL values in them. This step also included imputing the missing values as and where required with frequent values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and handled as and wherever required.

**Step3: Exploratory Data Analysis**
We started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, there were few variables that were identified to have only one value in all rows. These variables were dropped.

**Step4: Creating Dummy Variables**
We created dummies for the categorical variables. After analyzing, we dropped the first column and added the results to master data frame.

**Step5: Test Train Split**
The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

**Step6: Feature Scaling**
We used the Fit-Transform Scaling to scale the original numerical variables. Next, using the stats model we created our initial model, which gave us a complete statistical view of all the parameters of model.

**Step7: Feature selection using RFE**

Using the Recursive Feature Elimination, we selected the 15 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values. Finally, we arrived at the 10 most significant variables. The VIF's for these variables were also found to be good i.e. below 5.

We created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0. Based on this assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model. We also calculated the 'Sensitivity' and the 'Specificity' matrices to understand how reliable the model is.

**Step8: Finding the Optimal Cutoff Point**

We plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.35.

Based on the new value we could observe that close to 83% values were rightly predicted by the model.

We observed the new values of the 'accuracy=76%, 'sensitivity=83%', 'specificity=72%' approx. Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 83%.

**Step9: Computing the Precision and Recall metrics**

We found out the Precision value to be 65% and Recall value to be 83% on the train data set.

Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.35.

**Step10: Making Predictions on Test Set**

We implemented the learning to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 77%, Sensitivity=81%, Specificity=74%.

## Conclusion

While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

Accuracy, Sensitivity and Specificity values of test set are around 77%, 81% and 74% which are approximately closer to the respective values calculated using trained set.

Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%.

## Recommendation

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

1. Total Time Spent on Website
2. Lead Origin_Lead Add Form

3. What is your current occupation_Working Professional