Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals, due to which they require help in selecting the most promising leads i.e. the leads that are most likely to convert into paying customers.

The basic data which was provided to us, gave a lot of insight regarding the amount of time people spent upon visiting the website, how they reached the site and the conversion rate.

The company needs a model wherein a lead score is assigned to each of the leads, such that customers with a higher lead score have a higher conversion rate and those with a low lead score have a low conversion rate.

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

The following steps that were used to achieve this were →

1) **Cleaning data:** The first step was to drop the variables having unique values. A few null values were changed to 'not provided' so as to not lose data. Since, many were from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'
2) **EDA:** a quick EDA was done to check the condition of our data. It was seen that a lot of elements in the categorical variables were of no use. The numeric variables however seemed good and no outliers were found
3) **Dummy Variables:** Firstly, we created dummy variables for the categorical variables. Secondly, we removed all the repeated and redundant variables. For numeric values, we used the Min Max Scaler
4) **Train-Test split:** The split was done for 70% and 30% for train and test data.
5) **Feature Rescaling:**
   - We used the Min Max scaling to scale the original numeric variables
   - A heat map was then plotted to check the correlation amongst the variables
   - Highly correlated dummy variables were dropped
6) **Model Building:** 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(variables with VIF<5 and p-value<0.05 were kept)

7) **Model Evaluation:** A confusion-matrix was made. Optimum cut-off value(using ROC Curve) was used to find the accuracy, sensitivity & specificity which was around 80% each
8) **Based on the precision & Recall trade-off,** we got a cut-off value of 0.35
9) Then, the learnings were implemented to the test model, Accuracy was approximately = 89%, Sensitivity = 79% and Specificity = 96%

Conclusion:

- The lead score calculated in the test dataset shows the conversion rate of 83% which was clearly more than the ballpark of the target rate which was around 80%. Thus, this clearly meets the expectation of the CEO.
- It was found that the variables that mattered the most were (in descending order)→
  - Total time spent on the website
  - Total number of visits (when lead source was – Google, direct traffic, organic search, Welingak website)
- When the last activity was →
  - SMS
  - Olark chat conversation
- When the lead origin is Lead add Format
- When they are working professionals

Keeping the above points in mind, X Education can flourish as they have a good opportunity to convert almost all the potential buyers to full-time customers.