# CASE STUDY SUMMARY

**Problem Description:**

An education company named X Education sells online courses to industry professionals.On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, 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%.

X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. A model is required to be built wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

**Approach:**

1. Data Cleaning , Outliers and duplicates :

- Clean the dataset we choose was to remove the redundant variables. After removing the redundant columns, we found that some columns are having label as ‘Select’ which means the customer has chosen not to answer this question.
- The ideal value to replace this label would be null value as the customer has not opted any option. Hence, we changed those labels from ‘Select’ to null values.
- Removed columns having more than 40% null values
- For remaining missing values, we have imputed values with maximum number of occurrences for a column.
- We found for one column is having two identical label names in different format. We fixed this issue by changes the labels names into one format.
- Next, we checked the outliers of the dataset. The visualization of those outliers we can see on the graph attached on the right side.
- Checked percentage of missing values in columns and dropped the columns which have missing values more than 40%.
- Group the similar specializations into a category to gain more meaningful insights

- There are days columns which is having NAN values, so we are converting that to 'Others’.
- Handling duplicates by dropping them.
- Created dummy variables and dropping the first column
- We started our analysis with our cleaned dataset by converting all the binary variables to ‘0’ and ‘1’ and multiple categories into dummy variables.
- Next, we checked the outliers of the dataset. The visualization of those outliers we can see on the graph attached on the right side.
- Outliers in logistic model are very sensitive hence we need to deal with it without losing our valuable information. This can be achieved by creating bins. Hence, we did it.

2. Model Building

- We build a model with all the features included and found there were many insignificant variables present in our model.
- We need to drop them, but we can’t do it one by one as it is time consuming and not an efficient way to do so.
- We did two RFE count because we want to find out our final model stability.
- We started creating our model with RFE count 19 and went dropping variables one by one until we reach the point where the model is having all significant variables and low VIF values.
- We evaluated our model by first predicting it. We created new dataset with original converted values and the prediction values.
- Performed feature scaling using the standard scaler.
- We have created range of points for which we will find the accuracy, sensitivity and specificity for each point and analyze which point to choose for probability cutoff.
- We created a graph which will show us the tradeoff between Precision and recall.
- We found that there is a trade-off between Precision and Recall and the meeting point is approximately at 0.38.

**Final Observation**

Train Data

- Accuracy : 90.53%
- Precision : 80%
- Recall : 91%
- F1-score : 85%

Test Data

- Accuracy: 88.03%
- Precision : 91%
- Recall : 85%
- F1-score : 88%

**Conclusion:**

To achieve maximum conversion with less amount of time spent on cold calling company should invest time on Hot Leads i.e, Leads who has more than 80% conversion rate. All the Potential Leads can be approached when the sales person have enough time to invest as the conversion rate is comparatively less.