# Lead Score Case Study Assignment Summary

## Problem Statement

* An education company named X Education sells online courses to industry professionals.
* The company markets its courses on several websites and search engines like Google which redirects to their website.
* People who fill up a form providing their email address or phone number, they are classified to be a lead. Leads can also be past referrals.
* Leads  --> Marketing Calls, Emails --> Converted
* Typical conversion rate – 30%
* Identify Hot leads
* Identify the model and assign a lead score according to the conversion rate

## Steps Involved

### Data Cleaning

* This involves dropping the columns with high percentage of missing values
* Replacing ‘Select ’ value in categorical columns with Nan
* Imputing missing values
* Dropping columns with no major significant value levels

### Exploratory Data Analysis

* API and Landing Page Submission have more **leads origins** but the conversion is 30-40%, however Lead Add Form has highest conversion rate
* **Lead Sources**: Google, Direct Traffic and Olark Chat are main contributors; however conversion Rate from reference leads and leads through welingak website is high.
* Leads spending more time on the website are more likely to be converted.
* **Last Activity**: Most of the leads have their **Email opened** as their last activity. Conversion rate for leads with last activity as **SMS Sent** is almost 60%.
* **Occupation:** Conversion rate is high for Working Professionals, however most of the leads are Unemployed with conversion rate around 30% - 35%
* **Tags (Current Status):**Tags with "will revert after reading the email have high leads as well as high conversion rates."
* **City-**Most leads are from Mumbai with around 30% conversion rate.
* **Last Notable activity**: More number of leads have modified as Last notable activity however with very low conversion rate, leads with 'SMS Sent' have high Conversion rate.

### Outliers handling

Capping the outliners in columns ‘Page Views Per Visit’

### Creating dummies

* Converting binary variables (Yes/No) to 1/0 for columns - 'Do Not Email', 'Do Not Call'
* Creating dummies for categorical columns for columns - 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Tags','Lead Quality', 'City', 'Last Notable Activity'

### Model Building

* Splitting the Data into Training and Testing Sets with train\_size = 0.7, test\_size = 0.3
* Rescaling the features for columns - 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'
* Logistic regression feature selection using RFE with 15 Features
* Drop down the features with high p values to avoid multicollinearity.
* Checking VIF and dropping column with high VIF

### Model Evaluation

* Getting the predicted values on the train set
* Creating a dataframe with the actual flag and the predicted probabilities.
* Plotting the ROC Curve to identify the optimal cut off point
* From the ROC curve plotted, 0.3 is the optimum point to take it as a cutoff probability.
* Calculated confusion matrix to check the overall accuracy
  1. The sensitivity of our logistic regression model -TP / float(TP+FN) = 0.86
  2. Specificity - TN / float(TN+FP) = 0.96
  3. False positive rate - FP/ float(TN+FP) = 0.41
  4. Positive predictive value -TP / float(TP+FP) = 0.93
  5. Negative predictive value - TN / float(TN+ FN) = 0.91
* Calculated Precision and Recall

Precision - TP / float (TP+FP) = 0.93

Recall - TP / float (TP+FN) = 0.85

### Making predictions on the test set

* Calculated confusion matrix to check the overall accuracy
  1. The sensitivity of our logistic regression model -TP / float(TP+FN) = 0.84
  2. Specificity - TN / float(TN+FP) = 0.96

## Conclusion

1. Top three most contributing variables in our model are:

* 1. ‘Last notable activity\_SMS sent’,
  2. ‘Lead quality\_not sure’
  3. ‘Tags\_will revert after reading the email’

1. Top 3 categorical/dummy variables in the model to be focused in order to increase probability of lead conversion are:
   1. Tags
   2. Last Notable Activity
   3. Lead Quality