# Summary

After understanding the business problem and requirement of X Education company, Logistic Regression modelling technique was used to address the requirements as per the specification.

Here is the approach followed.

**Data Sourcing**

1. Firstly, the Leads data was sourced and read to create the data frames to analyse and understand the data from business perspective using the data dictionary provided.

**Handling Data Quality Issues**

2. There were few anomalies like missing data or columns with values as ‘Select’ were identified, which are nothing but same as missing. Columns with missing values >35% were dropped for further analysis.
3. Columns with same values for all or most of the records were dropped like City, Country, Receive More Updates About Our Courses etc.
4. The columns ‘Total Visits’ and ‘Page Views Per Visit’ columns had outliers, however it was was also observed that rows with outliers had conversion potential, so outliers were imputed to median values so that it doesn’t impact the models and also important information is not removed.

**EDA**

5. Performed univariate and bi variate analysis to understand the data better for both categorical and numerical variables, so that based on this analysis we could chose the relevant information/variables for the model building.

**Data Preparation**

6. Prepared the data by taking only top 15 features using RFE method and then creating the dummy variables for the categorical columns.

**Model Building**

7. As part of model training, the p values and VIF for all the variables were checked, variables with p value > 0.05 and VIF >5 were manually eliminated.

**Model Evaluation**

8. Evaluated the model by creating Confusion Matrix and plotted the ROC Curve to find the area under the curve which was0.89.
9. Plotted accuracy, specificity and sensitivity to get Optimal Cutoff Point, which was 0.35, this gave the prediction Accuracy, Sensitivity and Specificity around 80%
10. Model was evaluated and predicted using Precision – Recall scores as well, which gave the optimum cut off 0.42 resulted in prediction of Accuracy of 81%, Precision 76% and Recall score of 75%.

From the above analysis it was observed that

- Following people are more likely to enquire about the courses or are potential conversion leads.
    - Unemployed
    - Looking for better career prospects
- The Source for most of the leads is ‘Google’, followed by ‘Direct Traffic’.
- Most leads chose not to put their ‘Specialization’.
- Leads with origin as ‘Landing Page Submission’ are most likely to get converted.
- Sending SMS was the last activity of most of the leads converted, followed by ones who had opened the emails sent by Sales/Marketing teams.
- They also chose ‘No’ for ‘Do Not Email’
- Key point to note is, unemployed people are looking for upskilling, so company should target Unemployed and Working Professionals, as they are more likely to get converted.