**SUMMARY**

This analysis is performed for X education and to find ways to get more industry professionals to join their courses. The dataset provided gave us information about how the customers visit the site, the time they spend over there. Then how they reached the site and the conversion rate and built the model to target the customers to reach the hot leads.

Here, the logistic regression model is used to predict the probability of conversion of a customer.

The following technical steps are performed:

1. **Data cleaning**:

* To clean the given dataset, we choose to remove the duplicate variables.
* The dataset was partially cleaned except for a few null values.
* Dropped the high percentage of Null value more than 45%
* Treated the missing values by imputing statistical method (Mean, Mode, Median)
* Detected and treated the outliers.

1. **EDA:**

* Eda was done to check the condition of the data, found that categorical variables were irrelevant, the numeric values with outliers.
* Performed univariate and bivariate analysis for both continuous and categorical variable w.r.t Target variable.

1. **Dummy Variables:**

* The dummy variables are created for the categorical columns like 'Lead Origin', 'LeadSource', 'Country', 'Specialization', 'What is your current occupation', 'Tags', 'City'.

1. **Scaling:**

* Used standard scalar to scale the data for continuous variables like TotalVisits', 'Total Time Spent on Website’.

1. **Train – Test Split:**

* The data was divided into 70-30 split between train and test.

1. **Model building:**

* Model Building involved:

Feature Selection Using RFE, Improvising the model further inspecting VIF and p-values.

1. **Model Evaluation:**

* Evaluated using the metrics like accuracy, sensitivity, specificity and precision and recall values

1. **Prediction:**

* Predictions were made on test variable using the final model obtained from the train dataset

1. **Conclusion:**

* The logistic regression model is used to predict the probability of conversion of a potential lead.
* Optimum cut off chosen is **0.35** i.e., any lead with greater than 0.35 the probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.35 or less probability of converting is predicted as Cold Lead (customer will not convert).
* Features used in Final Model are:

'Do Not Email', 'Total Time Spent on Website', 'Lead Origin\_Landing Page Submission', 'Lead Origin\_Lead Add Form', 'Lead Source\_Welingak Website', 'Specialization\_Unknown', 'What is your current occupation\_Working Professional', 'Tags\_Busy', 'Tags\_Closed by Horizzon', 'Tags\_Graduation in progress', 'Tags\_Lost to EINS', 'Tags\_Others', 'Tags\_Will revert after reading the email’.

* The final model has **Sensitivity (Recall)** value of 0.855, this means the model is able to predict **85.5%** leads out of all the converted leads, (Positive conversion) correctly.
* The final model has **Precision** of 0.782, this means **78.2%** of predicted hot leads are True Hot Leads.