**Summary**

This analysis is done for X Education to find ways to get more industry professionals to join their courses. We as a data analyst needs to improve the lead conversion rate by finding out the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The data provided has features related to how the potential customers visit the site, the time they spend there, how they reached the site, their occupation, email preferences, etc

The following are the steps were used to reach our conclusion:

1. **Cleaning data:**

Following data cleansing activities were performed on the dataset.

* High Null values columns were dropped, since imputing the values would not be possible for the high percentage of missing data.
* Option “Select” was replaced by null value since it signifies the value was not chosen by the applicant for that field.
* Few single valued columns were dropped, since they didn’t constitute to our analysis.
* Null values in few important columns were treated as “Not Provided” to not lose data in those important columns.
* In Country column, since there were many data points from India and few from outside, we introduced new category the “Outside India” to club the date points together.

1. **EDA:**

A quick EDA was performed to analyse the dataset.

Many categorical variables were found irrelevant and were dropped later.

The numerical variables seem good. Outlier handling was done for numerical variables by capping the outliers.

1. **Data Preparation:**

Following data preparation activities were performed –

* Conversion of binary variables (Yes/No) to 1/0 was performed.
* The dummy variables were created for categorical variables.

1. **Train-Test split:**

The train-test split was done at 70% and 30% for train and test data respectively.

1. **Feature Scaling**

* Standard Scaler was used to scale the numeric variables before model training.

1. **Model Building:**

Firstly, RFE was performed to attain the top 15 relevant variables.

Later, using stats model some more reduction in variables were done by analysing the VIF values and p-value from the model statistics.

(The variables with VIF < 5 and p-value < 0.05 were kept).

1. **Model Evaluation:**

Following steps were taken in model evaluation phase –

* A confusion matrix was made.
* Various metrics were calculated:
  + Accuracy
  + Sensitivity
  + Specificity
  + Positive Predicted value
  + False Positive rate
* The optimum cut off value was found using ROC curve and the above metrics were again calculated.
* Accuracy, sensitivity and specificity all 3 were found to be close to 80% respectively.

1. **Precision – Recall:**

This method was also used to recheck and a cut-off of 0.33 was found with Precision around 79% and recall around 71% on the test data frame.

1. **Prediction:**

Prediction was done on the test set and with an optimum cut off as 0.33

**The accuracy, sensitivity and specificity, all 3 were found to be close to 80% respectively on test set.**

The Marketing team must focus on -

* Leads from **Lead Add Form**
* **Lead Source** being **Welingak website**
* Leads with **Current Occupation** as **Working professional**
* Leads with good **total time spent on website.**

Keeping these features into consideration, X Education can get high conversion ratio and focus on the prime leads and gain maximum buyers for their courses.