Summary

The analysis is for the company ‘’X Education’, where they needed help in analyzing the root cause for the poor or low leads conversion rate i.e., 30%.The data provided helped us with a lot of insights on how the potential leads visit the site, the amount of time they spend etc.,

The company wants to target leads with high scores so that the conversion rate increases in a short span by targeting the ones which would be more potential leads than others.

Various steps have been followed in analyzing & then building the model which would have high metrics i.e., efficiency with respect to various factors based on the business problem. Here are the steps followed for building a logistic regression model:

1. **Data Understanding:**
   We looked at the data available to us, since understanding the data is the primary aspect of deciding what to do next & how to use the data available & whether the data available with us is sufficient or not etc.,

2. **Data Cleaning:**
   The data had to be cleaned since a lot of columns had null values and had to be handled. Hence, we removed those columns which had > 45% of null values. Later, the remaining columns with null values were imputed based on the column 7 its importance i.e., a few columns were replaced with either ‘Not Specified’ when a lead doesn’t specify the details or moved into ‘Others’ columns and later removed during dummy variable creation. A few columns were also imputed based on the mode of that column.

3. **EDA:**
   We had to drop all the unnecessary columns which would be of no use for the business & in our model building process. Univariate & Bivariate analysis were performed to check for outliers, duplicate values & other aspects.

4. **Dummy Variables Creation:**
   The dummy variables had to be created for binary & multi-level categorical columns so that they can be used for model building. Later, the unwanted dummy variables were dropped.

5. **Model Building:**
   We first split the train & test data to 70:30 ratio. Then we took help of Recursive feature Elimination (RFE) to decide the top 15 features to avoid multicollinearity & other factors. Now the model was tweaked i.e., features with high p-value were dropped followed by high Variance Inflation Factor (VIF) features. The final model features had very low p-values & VIF which is an ideal model i.e., a stable & efficient model to work upon.

6. **Model Prediction on Training Dataset:**
   At this stage, the model’s metrics were plotted using the confusion matrix. The optimum cutoff point/value i.e., ROC curve was used to find Accuracy, Sensitivity, Specificity, Precision & Recall which came out to be ‘0.97’ which was very close to 1.

   Accuracy : 92.27% **Precision** : 88.55% **Recall** : 91.48% **Sensitivity** : 91.48% **Specificity** : 92.76%

7. **Model Prediction on Test Dataset:**

   The model built was successful since the difference in the test dataset metrics is also similar to the train dataset i.e.,

   Accuracy : 92.82% **Precision** : 89.23% **Recall** : 92.32% **Sensitivity** : 92.32% **Specificity** : 93.13%

The most valuable features/columns for the company to consider to increase their lead conversion rate are:

- The total time spent on Website
- **When the lead source was:**

a. Direct traffic **b.** Welingak website

- Lead Origin_Lead Add Form
- Total visits to the website
- **Tags with:**

a. Will revert after reading the email **b.** Other_Tags **c.** Closed by Horizzon **d.** Ringing **e.** Interested in other courses **f.** Lost to EINS

- **When the last activity was:**

a. SMS **b.** Olark chat conversation