Summary

Objective:

The objective of the case study is to increase the ratio of lead conversion for online education firm named X Education. X Education is currently struggling with low conversion for leads collected through various channel which is just 30 %.

In order to solve the problem, we need to build the logistic regression model that can analyze and predict the chance of people getting enrolled for the course. The model will also highlight the areas where X Education can focus for enhancing the lead conversion ratio to 80 % from currently 30 %.

Steps followed:

1. Data Cleaning:

From the original data set we found out that some variables has more than 30% missing values which needs to be dropped as it is not very useful. Also, we have observed that variables like City & Country has the class imbalance and not proven very useful in our analysis so we prefer to dropped them.

2. Exploratory Data Analysis:

We have found out that “Converted” is our target variable. We performed Uni-variate analysis on categorical variables to check which variables are significant for analysis. Then we check the relation between target variable and categorical variable to understand the effect of each independent variable.
We checked for the outliers in numerical variables by plotting BOX plot and found that few outliers in them but outliers treatment was not done as this may prove significant for model building.

3. Data Preparation:

The dummy variables have been added for categorical variables as a part of data pre-processing step. And scaling was performed on the numerical variables by using MinMax Scaler approach.

4. Train-Test Split:

The primary step for logistic model building process was done by doing the train & test split with ratio of 70% & 30% respectively.

5. Model building process:

We start building the model with 15 features selected through RFE method. Then basis upon the VIF & p-value for each variable we dropped some of the variables from the model (variables with VIF < 4 & p-value < 0.05 are considered). Our final model has 11 variables that helps in predicting the values for test / new data point.

6. Model Evaluation:

We have evaluated our model by creating the confusion matrix and plotting the ROC curve. And the optimal cut off was calculated by plotting Accuracy, Specificity & Sensitivity which comes around 0.42. With 0.42 as optimal cut-off Sensitivity & Specificity turns out around 79.34% & 78.85% which is quite significant.

7. Prediction:

We tested our model on the test set to check the accuracy of our model and found that it is correctly predicting the values as 1 & 0 for probability more than 0.42 & less than 0.42 respectively.

8. Precision - Recall

By using Precision- Recall approach we found that optimal cut-off is 0.44 which is quite closer to our model cut-off. Hence, our model is significant.

We can observe from our model feature which matter most in lead conversion are mentioned below (In descending order) :

- Total time spend on the Website
- Total number of visits
- When lead was sourced from:
    - Direct traffic on website
    - Olark Chat
    - Welingak website
- Basis upon the last activity:
    - SMS
    - Had phone Conversation
- When the lead is generated from the Landing Page Submission.
- Student are also likely to enroll for course but we should focus on which segment of student will join the course like those who have completed their Graduation or currently perusing their Post Graduation etc.
- We have not chosen the “What is your current occupation_Unemployed” as this segment of people might not have budget to enroll in the course.

- In order to reach the targeted conversion ratio which is 80%. X education must foucus on these features & segement for lead generation. As lead generated from these channel are most likely to get converted.