SUMMARY

X Education sells online courses to industry professionals. This case study is done to analyze and find ways to get more industry professionals to join their courses.

The company needs a model to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO has given a ballpark of the target lead conversion rate to be around 80%.

The Steps used in the analysis are:

Step 1: Reading and Understanding the Data

We read, understand and inspect the data in this step.

Step 2: Data Cleaning

(a) Few columns have 'Select' value which means that the lead did not choose any option. So, we replaced it with Null values.
(b) Then, we drop the columns having 45% or more Null values.
(c) Then, we removed some of the variables because they are redundant and do not provide any useful information.

Step 3: EDA

We perform EDA on the data to check various categorical variables and the presence of outliers in numerical variables.

Step 4: Data Preparation

(a) We convert some binary variables to 0/1.
(b) We create Dummy Variables for categorical variables and dropped repeated variables.

Step 5: Train-Test Split

We split the data into train and test data at 70% and 30% respectively.

Step 6: Feature Scaling

Using Standard Scaler, we rescale the numerical features. Then, we plot Heatmap to check correlation between variables.

Step 7: Model Building

Firstly, we used RFE to select the features. Then, we remove the variables depending on their p-value and VIF value. Variables having p-value<0.05 and VIF<5 were significant.

Step 8: Predictions on the Train set

Here, we made a confusion matrix and checked the accuracy, sensitivity and specificity.

Step 9: Plotting the ROC Curve

We plot the ROC curve to check the tradeoff between sensitivity and specificity. The ROC curve is very decent one with 91% of area coverage.

Step 10: Finding Optimal Cutoff point

Optimum cutoff value was used to find accuracy, sensitivity and specificity, which came out decent. Then, we check for Precision and Recall and their tradeoff.

Step 11: Making Predictions on the Test Set

Predictions on the Test set was done. Precision and Recall was calculated as 89% and 66% respectively.

Conclusion:

(a) 714 leads have a high chance of getting converted.
(b) Precision of the model is good.
(c) The overall Accuracy of the model is good.
(d) The most relevant variables are Tags, Last Activity as Phone conversation, SMS sent, Lead Add Form.

X Education should focus on the above mentioned variables to get almost all their potential buyers to buy their course.