Problem Statement

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a huge target lead conversion rate of around 80%

Solution Summary

1. Import and Inspect Data. Read and handle missing data
2. EDA. Start to get a feel of how the data is oriented. We compare all the columns with respect to 'Converted' column to analyse which attributes play a part in deciding the outcome of a Lead Conversion
3. Data preparation. Creating Dummy variables
4. Train Test Split. We divide the data set into test and train sections with a proportion of 70-30% values.
5. Feature Scaling. We use SKLearn's MinMaxScaler function
6. Model Building. Building the model Using the GLM method we build a regression model in the train dataset.
7. Feature Selection using RFE. We use RFE function to identify the top 15 features which we can use in building the model. We drop the rest of the columns from the train dataset and build another model using statsmodel.
8. ROC curve. Evaluating the model We determine the Confusion Matrix and the parameters like Sensitivity, Specificity, etc. We plot the ROC curve ,the Accuracy, Sensitivity and Specificity plot
9. Optimal Cutoff Point. we determine the optimal cut-off at 0.39 and got an accuracy of 81%.
10. Making prediction on the test set. We scale the test dataset with only transform and then predict the probabilities using the final model. On the test dataset we use the optimal cut-off of 0.39 and get an accuracy of 91%.
11. Assigning a Lead Score based on the Model.