REPORT

Problem Statement –
Help X Education sell online courses to industry professionals; we need to help them sustain and grow their lead management. Main goal is to help them identify the hot leads, so that the sales team can concentrate on communicating with the most potential clients.

Solving Methodology –
We started with a basic EDA process to understand the data properly. The data included 9240 leads with 37 features. On further comprehension we found that data included null and missing (the value ‘Select’ which was as good as a null value) values at a good percentage. So, we needed to treat them before any further analysis. The columns with high percentage of missing values were removed, categorical columns were treated with mode values, numerical with median values and columns that were highly skewed were also removed.

After data cleaning, we did univariate and bivariate analysis. During univariate analysis, we dropped the outliers in various columns like Page Views Per Visit, Total Visits, etc. The conversion rate in the dataset was around 37%. For categorical variables where the distribution was highly skewed and more categorical levels, we grouped together lower count levels (for columns like What is your current occupation, Last Activity, Lead Source, Tags, Last Notable Activity, etc.) or dropped the features if the data was highly skewed with less categorical levels (for columns like Do Not Call, Search, Magazine, Newspaper Article, etc.).

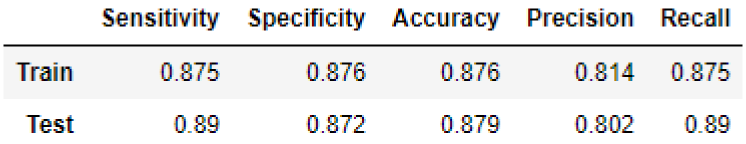
EDA also revealed that most of our revenue comes from the working professionals and unemployed, leads originating from Lead Add Form and leads which had opted for email communication.

Since the dataset had a lot of categorical features, we first visualized for any correlations among the existing features and dropped the required features to avoid multicollinearity. After creating dummy variables for categorical columns, we again dropped highly correlated features by visualizing the heatmap. The final count of the features in our dataset after creating dummy variables was 38.

After all the cleaning and data preparation we divided the data into train and test. We scaled the data using Min Max Scaler. In total we created 8 Models. First we started with all the available features, then we turned to RFE for selecting top 20 features and then dropped various columns on the basis of the p-values and VIFs. There were 14 features in the final model.

Finally, we made predictions using MODEL – 8. As we wanted our model to accurately predict the one’s, the model should have a high Recall or high Sensitivity. We previously decided for the threshold to be 0.5, but to define the optimal threshold we plotted Accuracy, Sensitivity and Specificity and the precision recall curve for various probability values. The optimal value turned out to be 0.332. The area under the ROC Curve also gave a high value of 0.95 which confirmed that our model was good. We created a Confusion Matrix and checked the sensitivity of the model on the train data which turned out to be 0.875, which was a decent turnout. Hence we made the predictions using the test data and evaluated the model on the test data set.

The Metrics of the Final Model are as follows.



	Sensitivity	Specificity	Accuracy	Precision	Recall
Train	0.875	0.876	0.876	0.814	0.875
Test	0.89	0.872	0.879	0.802	0.89