Summary

We need to identify the customer that are most probable to be converted for enrollment in the course.

1) Analysis of columns to see profiling of variables for checking variable types, distribution

Findings : We found columns with the value SELECT as per our understanding stands missing and needs to be treated.

Columns with more than 60% missing values are been dropped.

2) Data Treatment : Missing value and Outlier treatment

Missing are replaced with mean if no outlier else they are replaced with Median in case of numerical values.

In case of categorical we used mode to replace the missing value.

Outlier treatment is done base on business and variable understanding as Outlier are treated at 90, 95, 99^{th} percentile.

Summary

3) Data Preparation : We created dummy variables for categorical variables. Those variables which have 2 values only like Yes/No are been converted to 1/0 values as numerical variables. Values which have less than 5% frequency are clubbed to add stability to the data and hence model.

4) Test and Train Split : We have split the data into 70:30 and scaling of variables is done for variables which have values beyond the range of 0 to 1.

5) Variable Reduction : We have done the variable reduction using correlation and RFE(Recursive Feature Elimination).

6) VIF : Checking for variable elimination for values which are greater than 5. We have to recursively rebuilding the model until the stable stage is reached.

7) Checking the performance of model in terms of Sensitivity, Specificity, Accuracy, FPR, TPR, ROC curve and deciding the optimal cut-off which is 0.4 in our case.

Summary

8) Test dataset : Predicting the model on the test dataset and calculating the various performance metrices. We found all metrices are within the 5% range.

9) Below are the final model variables

	coef	Absolute
const	-4.3972	4.3972
Tags_Will revert after reading the email	4.3024	4.3024
Lead Origin_Lead Add Form	4.2262	4.2262
Tags_other	3.7069	3.7069
What is your current occupation_Working Professional	2.9169	2.9169
Last Activity_Olark Chat Conversation	-2.1973	2.1973
Last Activity_Page Visited on Website	-1.4706	1.4706
Last Activity_other	-1.4646	1.4646
Total Time Spent on Website	1.2784	1.2784
Do Not Email	-1.2691	1.2691
What is your current occupation_other	1.2638	1.2638
Lead Source_Olark Chat	0.9581	0.9581
Last Notable Activity_other	0.6534	0.6534
Lead Source_other	0.3445	0.3445

Summary

- Overall we see 13 variables coming the final model equation. Few top variables of these are :-

    a) Tags assigned to the customer indicating the current status of the lead

    b) Lead Origin – The origin identifier with which the customer was identified. So, one which is added by add form is is contributing

    c) Current Occupation Working Profession is adding a lot of value