SUMMARY

As a part of the Lead Scoring case study, we have been presented with the details how the company X Education pursues customer leads from various sources and tries to convert them to potential customers. The current conversion rate is quite low at 30%. So, we have been tasked to analyse the data and come up with a model which can make predictions to the order to 80% Lead conversion.

Data Cleaning:

- Columns with >30% nulls were dropped. Imputed few columns with missing data.
- Imputed data with mode or median according to the need.
- Other activities like outliers' treatment, fixing invalid data were carried out.
- About 98% of data was retained.

EDA:

- Conversion rate after data cleaning was only 38.5%.
- Performed univariate analysis for categorical and numerical variables. 'Lead Origin', 'Last Activity', 'Lead Source', etc. provide valuable insight on effect on target variable.

Data Preparation:

- We proceeded with encoding the categorical data with multiple levels into dummy variables.
- We also did binary mapping for Yes/No columns.
- Splitting Train & Test Sets: 70:30 ratio.
- Feature Scaling using Standardization.

Model Building:

- The training data is fed into a Generalized Linear Model (GLM).
- The ineffective variables are eliminated using RFE and VIF.
- Total 4 models were built before reaching final Model 5 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
- lrm5 was selected as final model which was used for making prediction on train and test set.

Model Evaluation:

- Confusion matrix was made and gave accuracy around 81%.
- ROC curve was plotted with 0.89 curve area value which was good.
- Prediction was done on the test data frame and with an optimum cut off of 0.37 and accuracy, specificity and sensitivity, all came around 81%.
- Precision – recall method was also used to recheck and a cut off of 0.42 was found with Precision around 75% and recall around 75% on the test data frame.
- At the 0.42 cut off, accuracy came as 81.37.
- After Observing both 0.37 and 0.42 Cut-offs, 0.37 gave a bit higher score for sensitivity and recall.
- All the metrics seemed good. Also, false positive rate was lower for 0.37 cut-off which will help in reduction of false predictions.
- So, used 0.37 Cut off to make final predictions.
- Final accuracy score at 0.37 cut-off was 81.03.

The top 3 Factors which can help in generating more successful leads are:

- Lead Origin_Lead Add Form
- Lead Source_Welingak Website
- What is your current occupation_Working Professional

Recommendations:

- Lead Add Format should be focussed on more to generate more leads.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay fees.