SUMMARY REPORT

An education company named X Education sells online courses to industry professionals. Although X Education gets a lot of leads, its lead conversion rate is very poor. Their target lead conversion rate is 80%. To do the prediction, since this is a classification problem, we build a logistic regression model.

Steps for EDA and building the model:

Data Cleaning

- The rows having null values are dropped.
- Columns with more than 3000 missing values are dropped.
- Columns having only a single value does not help the model and is also dropped.
- Any other columns which does not help in model building and prediction is also dropped.

Data Preparation for Modelling

- Dummy variables are created for all the categorical columns having multiple levels.
- Min Max scaling is done for the columns with numeric variables.
- Test train split of 70:30 ratio is done.

Model Building & Evaluation on Train and Test Set:

- RFE is used to select 15 features.
- Initial model built and then features having a p-value greater than 0.05 are dropped and VIF greater than 5.
- Predictions are done on the training set and an arbitrary cut off value of 0.5 is chosen.
- Sensitivity and specificity of 73% and 83% is obtained and area under ROC curve is 0.86.
- Accuracy, specificity and sensitivity of model for different cutoffs is plotted and 0.42 is chosen as optimal cutoff.
- Transformations are done on the test set and using 0.42 as cutoff we get a new column with the Converted or Not flag.
- Overall accuracy, sensitivity and specificity found to be 78.4%, 77.9% and 78.9% respectively.
- Precision and recall found to be 80.5% and 73.9% respectively and based on precision recall tradeoff cutoff point is changed to 0.44.
- Predictions are done on the test set using 0.44 as cutoff and Final models accuracy, precision and recall on test set found to be 78.6%, 78.2% and 76.7% respectively.