Lead Scoring Case Study Samary

1- Problem Statement

An education company named X Education sells online courses to industry professionals. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

2- Problem Solving Methodology

- Understanding Dataset & Data Cleaning & Preparation
- Applying Recursive feature elimination to identify the best performing subset of features for building the model.
- Building the model with features selected by RFE.
- Perform model evaluation with various metrics.
- Decide on the probability threshold value based on Optimal cutoff point.
- predict the dependent variable for the training data.
- Use the model for prediction on the test dataset.

3- Data Cleaning

- Remove columns which has only one unique value
- Handling 'Select' values in some columns
- Categorical Attributes Analysis (Null Values and duplication)
- Categorical Attributes Analysis (Drop Attributes)
- Numerical Attributes Analysis
- Assigning a Unique Category to NULL/SELECT values
- Binary Encoding

4- Data Preparation

- Dummy Encoding: create a dummy variables for all category variables
- TestTrain Split: Split The original dataframe was split into train and test dataset.
- Feature Scaling: Use standard scaler to scale the numerical variables

5- Build the Module

The module building go throw the following steps:

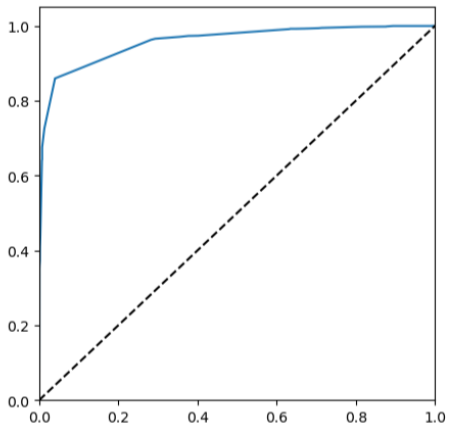
- Feature Selection using RFE.
- Build the Module.
- VIF check.
- Predicted values on the train set.
- Optimal Cutoff Point.

The Predicted values on the Train set.

- Overall accuracy 92.11 %
- Sensitivity 85.68 %
- Specificity 96.05 %

ROC curve

Since we got a value of 0.9582, our model seems to be doing well on the test dataset



6- Predictions on test set

The prediction result is:

- Accuracy: 93.17%
- Sensitivity: 87.93%
- Specificity: 96.26%
- F1 Scoure: 90.52%

7- Recommendations

Train Data Result:

- Accuracy : 92.11%
- Sensitivity : 85.68%
- Specificity : 96.05%
- F1 Scoure : 89.20%

Test Data Result:

- Accuracy : 93.17%
- Sensitivity : 87.93%
- Specificity : 96.26%
- F1 Scoure : 90.52%

Based on our model, some features are identified which contribute most to a Lead getting converted successfully.

The conversion probability of a lead increases with increase in values of the following features in descending order:

- Tags_Closed by Horizzon

- Tags_Lost to EINS
- Tags_Will revert after reading the email
- etc

The conversion probability of a lead increases with decrease in values of the following features in descending order:

- Last Notable Activity_Olark Chat Conversation
- Tags_Interested in other courses
- Last Notable Activity_Modified
- etc