Lead Scoring Assignment

Problem Statement: X education sells online courses to industry professionals and gets lots of leads wherein the lead conversation rate is poor. To make this process more efficient the company wishes to identify the most potential leads. If they identify this , conversion rate might increase and also sales team can focus on communicating with the potential leads.

Objective: X education wants to build a model to identify promising leads.

The following are the steps used for the analysis:

- Cleaning data:
    - The data was partially cleaned which has more than 35% of null values and the null values we used the option select.
    - There was not much of data loss when the null values were changed to ‘others’. With value counts the highest percentage category is replaced under the null values.
- EDA:
    - Data analysis is performed on both the category and numerical variables. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found.
    - Last activity , Last notable activity for SMS Sent is high with conversion rate. Lead origin: Lead Add form is also with high conversion, so people from these can be promising leads.
- Data Preparation:
    - The dummy variables were created for Lead origin, Source and what is your occupation and later redundant variables were removed . Standard scaler is used for scaling.
    - The split was done at 70% and 30% for train and test data respectively.
- Model Building:
    - RFE is used to reduce to top 15 relevant variables and recursively tried looking at the P-values in order to select most significant values and dropped insignificant values (The variables with VIF < 5 and p-value < 0.05 were kept).
    - Total 5 models were built and model 5 seems stable with VIF and p-values contains Do not email, Total time spent on website, Lead origin, Lead source, What is your current occupation variables.
- Model Evaluation:
    - A confusion matrix was made then the optimum cut off value using ROC curve was used to find the accuracy, sensitivity and specificity.
    - Prediction was done on the train data frame and with an optimum cut off as 0.3 with accuracy, sensitivity and specificity of around 72-85%.
    - Precision - Recall method is used to assess performance of a binary classification model, here we found Precision around ~67% and recall around ~83% on the train data set. Trade off is performed on Precision recall with thresholds to positive.

- Implemented the learnings on the test model with scaling and calculated conversion probability based on the metrics and found out the accuracy comes around 77%, specificity to 74% , sensitivity to 83% , Precision to 66% which is almost accurate.
- We have high recall score than precision score which is a sign of good model.

Important features that are responsible for good conversion rate:

Lead Origin_Lead Add Form, What is your current occupation_Working Professional, Total Time Spent on Website.