Lead Score Summary

Started the case study by reading the dataset, performed the data cleaning activity, dropped the features which had null values more than 38% to 40%, for other features imputed with mode or median accordingly. Dropped the features which had imbalanced data

There were few column with value 'Select' , imputed those values with null. Also grouped the categorical value into one category where % of values where 1 or less than 1

Created dummy variables for categorical columns where the column had more than 2 categories. Divided the data set into train and test set in 70:30 ratio. Scaled the numeric features using minmax scaler

Built the model by considering 20 features, the features were selected using Recursive feature elimination method

Checked the optimal probability cutoff by finding the accuracy, sensitivity and specificity. The cut off considered is 0.35

By plotting the ROC curve, the area covered under curve was 87%

Metric on train data is Accuracy = 80.68%, Sensitivity = 80.34%

Specificity = 80.89%

Metric on test data is Accuracy = 81.80%, Sensitivity = 81.52%

Specificity = 81.98%

There is only 0.5% to 1% difference on train and test data's performance metrics. This implies that final model didn't overfit training data and is performing well.

We can conclude following points

- The customer/leads who fills the form are the potential leads.
- We must majorly focus on working professionals.
- We must majorly focus on leads whose last activity is SMS sent or Email opened.
- It's always good to focus on customers, who have spent significant time on our website.
- It's better to focus least on customers to whom the sent mail is bounced back. Working professional
- If the lead didn't fill specialization, he/she may not know what to study and are not right people to target. So, it's better to focus less on such cases.
- It's better to focus least on customers to whom the sent mail is bounced back.