SUMMARY NOTE

LEAD SCORING CASE STUDY:

Problem Statement: An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goals:

Target is to help X Education increase the conversion rate. To acquire this we need to build a Logistic Regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted. Along with the above, we need to handle the problems faced by the company along with adjust to company's future requirements.

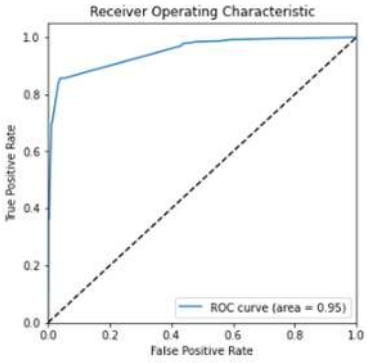
Approach:

The Steps we will be using in this assignment are:

1. Reading and understanding the data
2. Cleaning the data
3. Performing EDA
4. Creating Dummies
5. Splitting of Test and Train set
6. Building Model
7. Making Prediction
8. Model Evaluation
9. ROC Curve
10. Precision - Recall
11. Prediction on Test set

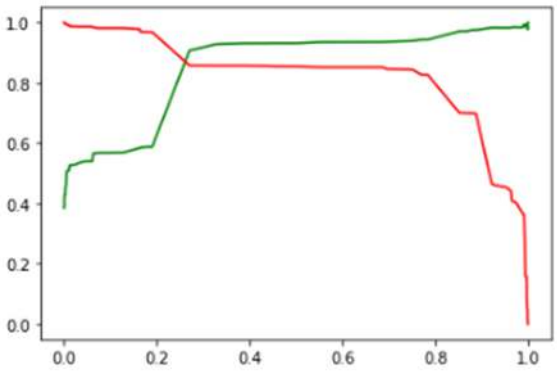
Step 1: All the libraries were imported and data is read.
Step 2: Columns having greater than 40% of null values are removed. However, since Lead Quality is an important parameter, so we imputed the blanks with Not Sure. All the columns which won't be contributing to the model has been removed. Remaining columns where the % of null values is 1% or less, the numbers of rows have been removed. Columns which are having 'Yes' and 'No' as values are mapped using binary digit 0/1.
Step 3: EDA is performed on the data set and parameters are checked which contribute to lead conversion. In few columns where data count is very less are mapped together under same category. For columns like 'Total Visit' and 'Total Time spent on Website' , the outliers are restricted to 95 percentile of the data.
Step 4: Dummy creation for all the categorical values. And then it is mapped with the original data set after removing the specific columns from the Original dataset.
Step 5: Data set is then bifurcated between Train and Test.
Step 6: RFE technique is used to remove attributes and build a model on those attributes that remain. RFE uses the model accuracy to identify which attribute contribute the most to predicting the target attribute. Using the p-values and vif, parameters contributing to model development is done.

Step 7: The predicted column is created which defines whether a lead will convert or not.
Step 8: Confusion matrix is created and the accuracy, sensitivity and specificity is checked.
Step 9: ROC curve is drawn, and accuracy is tested.



From the above it is observed that area under ROC curve is 0.95 which is high and it depicts that our final Model is a good one.

Step 10: Cut-Off is decided between 0.0 and 0.9. And from the cut-off plot 0.19 is taken as the optimum point. Then confusion matrix is calculated and accuracy, sensitivity and specificity is checked again.
Step 11: From Precision-Recall test, we finalize the cut-off to be 0.27. And the same is applied on Test data set.



After finalizing the optimum cut-off point, below are the observations:

Train Data:
Accuracy: 91.11%
Sensitivity: 85.73%
Specificity: 94.49%

Test Data:
Accuracy: 90.78%
Sensitivity: 84.12%
Specificity: 94.57%

The model predicts the conversion rate very well with a accuracy greater than 90%. Sales team should focus more on Lead scire which are high to increase the lead conversion rate.