Summary

1. Lead Scoring case study is performed to understand the data set and identify the feature to improve the lead conversion rating.
2. Required libraries and data set are imported. Columns with more than 30% null values are removed. Columns which is not useful also dropped from the dataset.
3. Visualization of numerical and categorical variable is performed.
4. Dataset separated test train before model building. 70% train and 30% test used.
5. MinMax scaler used for rescaling the numeric variable.
6. Course fine tuning with RFE and 15 features are selected.
7. Statsmodel GLM method used to perform logistic regression.
8. 4 iteration done by checking VIF and P value (criteria – VIF less than 5 and P value less than 0.005).
9. Train data set prediction done and cut off 0.5 used randomly. Predicted probability >0.5 , converted and <0.5, not converted.
10. Confusion matric done, Accuracy, Sensitivity, Specificity = 79%, 74%, 84% respectively.
11. Cut off optimization is done using ROC curve , AUC = 0.87.
12. To find the best balance between sensitivity and specificity, confusion matrix created from 0 to 1 with multiple of 0.1 probability. Then plotted graphs for accuracy, sensitivity, specificity. Optimal point 0.42 selected from graph.
13. Accuracy, Sensitivity, Specificity are 79%, 79%, 79% respectively with new cutoff 0.42.
14. Test dataset observation are Accuracy, Sensitivity, Specificity are 76%, 92%, 68% respectively.