Lead Scoring Case Study Summary

Business Understanding

1. 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.
2. The company markets its courses on several websites and search engines. The people land on the website and up a form providing their details, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Problem statement

1. 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%.
2. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objectives

1. The company requires to build a model to get most potential leads that are most likely to convert into paying customers.
2. Assign a lead score (between 0 and 100) 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.
3. The lead conversion rate of the model souled be greater than 80 % i.e. company wants at build the model with 80 % accuracy

Solution summary:

1. Importing lead Data:
   Import and understand the lead data, making lead number as index to retain the lead identity

2. Data Cleaning and Exploratory Data Analysis:
   a. There are no duplicate entries in the table.
   b. Deleted the Prospect ID.
   c. 'Select' value in the data is replaced by 'null' value or 'not specified' and then Deleted column with missing value greater than 40%.
   d. Imputing the remaining missing value with mean or median.
   e. For numerical variable, Outliers are imputed with threshold value of the data i.e. Q3+1.5 IQR.
   f. Deleting non relative column i.e. column having all unique value (categorical column) or only one unique value.
   g. Biased entry with data more than 95% biased are deleted.
   h. Column having only one unique vale are deleted.

3. Dummy Variables:
   Dummy variable is created deleted the dummy variable with "not Specified" and "others" element. Then merged the dummy variable with the lead data and deleted the primary variables.

4. Test Train Split:

   The data is split randomly with 70% train data and 30% test data, then for numeric variable of train data we have used Standard Scaler with fit_transform. And on test numeric data we have used Standard Scaler with fit.

   We also have split the data into Xtrain, ytrain, Xtest and ytest based on target variable

5. Feature selection using RFE:

   Using the Recursive Feature Elimination we have selected the 20 top important variables.

6. Model Building

   Using Generalized linear model, we have built the model with 20 top important variables. Then the variable which are having high p-value i.e. >0.05 and high VIF i.e. >5 is deleted one by one for every model until we left with p-value < 0.05 and VIF < 5.

7. Optimal Cutoff Point:

   We have plotted the 'Accuracy', 'Sensitivity', and 'Specificity' vs Lead score (Probability). And find optimum cutoff point at lead score = 30.

   Based on the lead score and confusion matrix we have calculated

   Accuracy=92.39%,
   Sensitivity=91.57%
   Specificity=92.90%

8. ROC Curve:

   ROC curve is plotted and fond that the area under ROC curve is 0.97 which is closed to 1. This implies the model is good.

9. Precision and recall tradeoff

   We have also visualized the Precision and recall tradeoff curve and the cutoff value we got is approximately 40.

   Precision = 88.83 %
   Recall = 91.57%

10. Predictions on Test Set.

    Based on the logistic regression model we have calculated the predicted lead score for test data and calculated the

    Accuracy=92.53%,
    Sensitivity=92.60%
    Specificity=92.49%
    Precision = 88.94 %
    Recall = 92.60 %

    Accuracy of the test and train data is nearly same and >80%