Lead Score Case Study Summary

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 requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution:

1. **Reading and Understanding the data:**
   Read the description of columns and loaded the data on jupyter platform.

2. **Data Cleaning:**
   Few columns had to be dropped because of high percentage of missing values (more than 35%). Also, few missing values were imputed with the median value for the numerical variables and created new classification variable in case of categorical variables. Furthermore, outliers for numerical variables were identified and removed.

3. **Exploratory Data Analysis:**
   A basic analysis of data helped to get better understanding of the dataset. While performing the EDA, few redundant columns were discovered and dropped. After this step, a basic understanding about the columns and data was comprehended.

4. **Creation of Dummy Variables:**
   Dummy variables for categorical variables were created. The original column and one dummy variable were dropped.

5. **Test Train Split:**
   The dataset was segmented in 70:30 proportion for training and testing.

6. **Feature Rescaling:**
   Using the MinMax Scaler all the numerical variables were scaled. Following this, an initial model was built from the stats model. This provided a complete statistical view of all the parameters of the model.

7. **Feature selection using RFE:**
   Initially, 20 top parameters were selected through the RFE method. Using the statistics generated, recursively we tried to eliminate the insignificant variables by looking at their respective p-values. Finally, we finalised with 14 significant parameters that had

significant effect on the lead. There was low multicollinearity for the selected variables as all of the parameters had VIF withing the accepted range.

After this, a dataset was created having converted probability values. We assumed that probability of more than 0.5 is 1 else 0.

Based on the above assumption, we derived the confusion matrix to calculate overall accuracy, sensitivity and specificity of the model.

8. **Plotted the ROC Curve:**

   The ROC curve was graphed to further strengthen the features. It was observed that the area covered by the curve was 89%.

9. **Finding the Optimal threshold Point:**

   In order to obtain the optimal cut-off, point a graph of Accuracy, Sensitivity and Specificity of different probability values were drawn. Fortunately, all the three parameters intersected at one point, at 0.38.

   Based on the new threshold value, nearly 80% of values were rightly predicted by the model. Furthermore, the accuracy was 80%, sensitivity was 79.5% and specificity was 82.81%.

10. **Computing the Precision and Recall values:**

    Since, business houses prefer precision and recall values, we calculated these values.

    The precision and recall values on the train dataset were 79.7% and 71% respectively.

11. **Prediction on the test dataset:**

    The learnings and the final model were implemented on test dataset and calculated the conversion probability based on the sensitivity and specificity metrics.

    The accuracy was found to be at 81%, sensitivity was observed to be 79% and specificity was noted at 82.4%.

    The precision and recall values on the test dataset were 73% and 79% respectively.