Problem Statement:

X Education has appointed you to help them select the most promising leads. 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%.

Summary Report:

Here is the high level steps we did for the entire exercise

Exploratory Data Analysi:

1. Perform EDA on dataset imported like checking the shape, info and null percentage in the data.
2. Cleaned the data by dropping few columns for which the business significance is not seen based on Data Dictionary.
3. Imputed data on few columns by applying Statistical Imputations techniques (E.g: mode) on the column data.
4. Merged few Values of columns into a single column as few columns doesn't have good amount of data
5. Performed the Univariate Analysis based on the Target Variable 'Converted'.
6. Presented are few of the columns and their distribution with respect to Converted variable
7. Outlier Treatment:
    a. Identified the outliers in the numerical data columns.
    b. Addressed the outliers for the columns by capping the outlier value in the columns, and they are: "Total Visits", "Page Views Per Visit", and "Total Time Spent on Website"

Scaling of data:

    c. Created Dummies for the Categorical variables in the data.
    d. Performed the Train-test Data Split on the data.
    e. Performed the Feature Scaling using Standard Scaler.

Model Building:

8. Automatic Feature selection done by applying RFE on the Train data.
9. Dropped few features based on stats model's outcome which have high p-value.
10. Calculated VIF for the dataset to check if any features have high VIF values/multicollinearity.
11. Plotted ROC Curve to find the AUC (Area Under Curve) and identified the value to be : 0.95 (95%), which is a good value.
12. Optimal Cut-off point (probability) can be identified from the adjacent plot drawn where the intersection(balance) between Sensitivity and Specificity is noticed. Optimal Cut-Off probability here is: 0.45
13. Made predictions on the Test dataset to obtain the final model. And calculated the metrics of the model.

Inference and Coclusion

14. As per the motto we need to predict as many Actual "1's" as 1 and Actual "0's" as 0 and the target from CEO of X Education is to get 80 % lead conversion. So this means we need to have "Sensitivity" of our model to be at least 80 %, here we have got the 92.49% (i.e greater than 80%) and as per the first ask we need to predict correct 1's and 0's i.e accuracy is also good here with 86.94%. And the other metrics are also well in place:
    a. Sensitivity stands at 92.49
    b. Specificity stands at 82.76
    c. Accuracy stands at 86.94
    d. Precision stands at 80.14