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 websites 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, fill up a course form, or watch some videos. When these people fill out a form providing their email address or phone number, they are classified as 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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor.

Overall Approach to Solution:

Step 1: Reading and Understanding data

Imported the dataset and check the basics of the dataset to get an understanding of the data initially. [shape, describe, info]

Step 2: Data Cleaning and Preparation

- Checked for NULL values
- Converted the 'Select' values in certain variables to NaN
- Treated all the NULL values
    - If the NULL values in the column exceeded 40%, dropped the column
    - imputing – Mean and Median [Numerical data], Mode [Categorical data]
    - Dropped the rows if they were insignificant
- Combined some categories in certain variables to a single entity as they did not show much difference alone.
- Removed variables where only 1 value dominated
- Visualized the Numerical and Categorical data and made observations.

Step 3: Preparing the data for Modelling

Created dummy variables for the categorical data.

Step 4: Train-Test Split

Then it was time to divide the dataset to build the model, took a 70%-30% ratio for model building.

- Performed scaling of the 70% train dataset. Used StandardScaler() on the data which has mean as 0 and standard deviation as 1.
- Checked for correlation using a heatmap.

Step 5: Model Building

Used the RFE (Recursive Feature Elimination) to reduce the pool of leads to 15 features. Then, went ahead with manual model selection (statsmodel) and dropped features with high p-value (> 0.005) and high VIF (>= 5) until they were neutralized.

- Checked for correlation with a heatmap.

Step 6: Model Evaluation

Now, prediction on the train set. Initially used 0.5 as the threshold and obtained the overall accuracy, confusion matrix, sensitivity, and specificity.

Accuracy – 93.32 %, Sensitivity – 88.67 %, Specificity – 96.23 %
Precision – 93.65 %, Recall – 93.65 %

- Checked for the area under the ROC curve as a metric to evaluate the model – 0.98
- Found the optimal threshold - to be 0.3

Predicted the conversion this time with 0.3 as the threshold and calculated the overall accuracy, confusion matrix, sensitivity, and specificity [which turned out to be better than 0.5 as the threshold].

Accuracy – 92.64 %, Sensitivity – 91.94 %, Specificity – 93.08%
Precision – 89.28 %, Recall – 91.94 %

Step 7: Prediction on the Test set

- Scaled the test dataset using the StandardScaler () as in the train set [expect this time only transform was done and not fit]

Finally, prediction on the test set using 0.3 as the threshold cut-off. Calculated the overall accuracy, confusion matrix, sensitivity, and specificity. Which was very close to the percentages on the training dataset. Achieving more than 90% sensitivity.

Accuracy – 91.73 %, Sensitivity – 90.29 %, Specificity – 92.61 %
Precision – 87.46 %, Recall – 90.29 %