Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. 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 or fill up a form for the course or watch some videos.
The people who fill these forms are classified as a lead.
As an analyst, it's our responsibility is to create a model which would assign a lead score to each of these leads.
This should help the Company to increase their lead conversion rate from 30% to 80%.

Method Summary:

I. Reading Data. Read and understand the structure of data given.

II. Data Quality check and corrections:

- Replacing Select Values with Nan.
    i. Select value is observed in certain columns which may be because the fields may not be mandatory within the form.
    ii. And so, the people willing to enroll in the course wouldn't have opted for a drop down option or missed to give a value while filling the form.
    iii. Hence, these values can be considered as Null or missing values.
- Dealing with Null / Missing values in df:
    i. Certain variables with an unbalanced spread can be dropped.
    ii. Other methods of filling the Nan with mode / median is conducted.
- Outlier Treatment:
    i. Soft capping is performed on upper tails.

III. Visualizing the Data:

- Check for imbalance in data.
- Univariate, Pair plot etc.

IV. Dummy Variables:

- Fields with Yes/No categories are converted to 1/0
- Other categorical variables are modified to dummy variables by dropping respective dummy column.

V. Test-Train Split:

- The given data set is divided into Train Data & Test Data at 7:3 ratio.

VI. Rescaling the numerical features:

- The numerical fields like: TotalVisits, Total Time Spent on Website are rescaled using Standardization.

VII. Feature selection using RFE:

- Recursive elimination method is used to select the best 30 variables contributing to the conversion rate.

VIII. Iteration of models:

- A pre-defined function is used to generate a model and VIF.
- Depending upon the p-values & VIF we can reduce down the variables.
- We have conducted 13 iterations to finalize with the best fitting model.

IX. Prediction and Model Evaluation:

- After optimizing the VIF and p-values we can go ahead and predict the new conversion status of each lead by taking the cut off as 50% probability.
- Verify with the model statistics, i.e. Accuracy, Sensitivity & Specificity.

X. Plotting the ROC Curve:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

XI. Optimizing the Cut-off value of Probability predicted:

- BY plotting the characteristics of Accuracy, Specificity & Sensitivity we will be able to find the best cut-off point.
- As per our approach we were able to see the intersect at **.35**!

XII. Predict the final conversion status of each lead using the optimized cut-off value of .35.

XIII. Model Evaluation on Train:

- **Accuracy**: 79.7%
- **Sensitivity**: 80.8%, **Conversion rate** meets the requirement.
- **Specificity**: 79.1%

XIV. Same model is applied on the test data set to check with the model performance.

XV. **Lead score is calculated = Predicted probability * 100**

XVI. **Model Evaluation on Test:**

- **Accuracy**: 80.4%
- **Sensitivity**: 81.1%
- **Specificity**: 80.0%

XVII. We can see that the final model is fetching almost similar model parameters which show us that the model is good to go!

XVIII. Also, the sensitivity is nothing but the conversion rate predicted using the model.

XIX. With the model we are able to predict a conversion rate of **81.1%** on test data.

End