Lead score case study summary report

This analysis is done for X Education to find ways to get more industry professionals to join their courses. The basic data provided us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used :

1. **Cleaning data:**
   The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to’ not provided’ so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to ‘India’, ‘outside India’ and ‘not provided’.

2. **EDA :**

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found.

3. Dummy variables:

The dummy variables were created and later on the dummies with ‘not provided’ elements were removed. For numeric values we used the MinMax Scaler.

4. Model Building :

Firstly, RFE was done to attain the top 20 irrelevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF <5 and p-value <0.05 were kept.

5. Determining feature importance

14 features have been used by our model to successfully predict if a lead will get converted or not.

- The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.
- Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.
- Similarly, features with high negative beta values contribute the least

6. Model Evaluation:

6a. Metrics Will be used for Evaluation - Accuracy ,Sensitivity, Specificity, Precision, Recall ,True Positive Rate , True Negative Rate,False Positive Rate ,False

Negative Rate, Positive Predictive Value Negative Predictive Value and F1 score.

6b.Plotting ROC Curve

Benefits of ROC Curve:-

- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

- 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.

6c. Optimal Cut-off Point

Optimal cut off probability is that probability where we get balanced sensitivity and specificity.

From Probability curve, **0.34** is the optimum point to take it as a cutoff probability. It shows accuracy , sensitivity and specificity for various probabilities.

7. Importance Features

There are some features which maximize the convergence rate such as ‘Lead origin_add form ‘ ,’Working professional’ etc and some features like ‘Do not email ‘ etc they decrease the convergence rate.