LEAD SCORE CASE STUDY BRIEF

PROBLEM STATEMENT

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The data provided gave 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 etc.

STEPS FOLLOWED

1. Cleaning data:

All the variables that wont be needed as a part of the analysis were dropped. The data was cleaned 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 and for other variables Nulls were replaced with either mean ()/median ()/ Mode ().

2. EDA:

EDA both Univariate and Bivariate was done to draw initial insights from the data The numeric values seems 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. The data was standardized using MinMaxScaler.

4. Train-Test split:

The split was done at 70% and 30% for train and test data respectively.

5. Model Building:

As first step RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value i.e. the variables with VIF < 5 and p-value < 0.05 were kept and the rest were dropped.

6. Model Evaluation:

A confusion matrix was made. Later on the optimum cut off value (using ROC curve)was used to find the accuracy, sensitivity and specificity which came to be around 80%.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off was around 0.4 with accuracy, sensitivity and specificity was around 80%.

8. Precision – Recall:

This is used was to recheck and a cut off of 0.5 was found with Precision around 82% and recall around 78.6% on the test data frame.

9. Conclusion:

The Accuracy, Precision and Recall/Sensitivity are showing promising scores in test set which is as expected after looking the same in train set evaluation steps.

Means the recall is having high score value than precision which is acceptable. In business terms, this model has an ability to adjust with the company's future requirements.

This concludes that the model is in stable state.

Important features responsible for good conversion rate or the ones which contributes more towards the probability of a lead getting converted are :

- Total Time Spent on Website

- Lead Origin_Lead Add Form

- What is your current occupation_Working Professional

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.