LEADS SCORING CASE STUDY

Summary

Problem Description:

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 website and browse for courses.The company markets its courses on several websites and search engines like Google. Once these peopleland on the website, they might browse the courses or fill up a form for the course or watch some videos.When these people fill up a form providing their email address or phone number, they are classified to bea 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%. X Education needs help to select the most promising leads, i.e., the leads that are most likely to convert intopaying customers. A model is required to be built wherein a lead score is assigned 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 and the CEO has given a ballpark of the target lead conversion rate to be around 80%.

The following are the steps used:

1. **Cleaning data**: We set the cut off for null values more 3000rows for each column. The data was partially clean except for a few null values few columns and the option 'Select' had to be replaced with a null value since it did not give us much information in or removed as it would skew the data imputing such huge amount. Few of the null values were changed to 'Not Selected' so as to not lose much data. Although they were later removed while making dummies.
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. Univariate analysis of all categorical and numerical columns was done. Bivariate analysis was performed for categorical variables against target variable 'Converted'.
3. **Dummy Variables**: The dummy variables were created and later on the dummies with 'Not Selected' elements were removed. For numeric values we used the MinMaxScaler.
4. **Train-Test split**: The split was done at 70% and 30% for train and test data respectively.
5. **Model Building**: Firstly, 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 (The variables with VIF < 5 and p-value < 0.05 were kept).
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% each for train set. And for test set we are getting accuracy, sensitivity and specificity is coming around 81%.

7. **Prediction**: Prediction was done on the test data frame and with an optimum cut off as 0.38 with accuracy, sensitivity and specificity of 81%.
8. **Precision** – Recall: This method was also used to recheck and a cut off of 0.42 was found with Precision around 76% and recall around 75% on the test data frame.

Final Observations

Hot Variables: Based on the coefficient values of the final model, the 3 variables with the highest probability of converting a lead are:

i. Lead Source_Welingak Website (6.3286)
ii. TotalVisits (5.4923)
iii. Total Time Spent on Website (4.6118)

Test Data Values:

Accuracy - 80.7
Sensitivity - 81
Specificity - 80.5
Precision - 74
Recall - 76