Summary

X Education is an education company which sells online courses to industry professionals. The company gets leads from different sources both online and past referrals. When people fill up a form providing their email address or phone number, they are classified to be a lead on the website. The company gets lots of leads, but its lead conversion rate is poor. With the given data, the company expects us to build a model that helps in assigning a lead score such that, higher the lead score the higher the chance for the lead to get converted. The CEO of the company expects the lead conversion rate to improve to 80%.

Based on the data provided the following procedure is followed to build the model:

Step 1: Data Understanding and Data Cleaning:

Data cleaning is done by removing irrelevant features. The 'select' in the data is replaced with NA. Missing values in few features are replaced with median and mode. Few of the features which do not explain any variance are dropped.

Step 2: Creation of Binary form and dummy variables:

The features with "yes" and "no" are converted to 1 and 0. The categorical features with more levels are converted to dummy variables.

Step 3: Initial Model Building

Before Model Building the data is split into Train and Test. On numeric variable data scaling is done. Logistic Regression Model is built with all the relevant features.

Step 4: Feature Selection

Feature Selection is done using RFE. Initially only 10 features are selected in the RFE model and based on these 10 features Logistic Regression model is built and is assessed.

Step 5 Model Building

Based on the top 10 features, model is built and assessed and based on p-value few features are removed one after other until all the features in the model have p-value less than 0.05 and vif less than 5. We get final model with 7 variables having all the variables p-value less than 0.05 and vif less than 5.

Step 6 Model Evaluation:

Based on the model built, we predicted the probability values on train data and randomly considered 0.3 as cut off. If the probability is greater than 0.3 it is considered as 1 (lead) and

less than 0.3 it is considered as 0(not a lead). Based on it, we calculate confusion matrix and check the sensitivity of the model which is greater than 0.8.

We plot the ROC curve and find the area under the ROC curve to be 0.87 which is a good value.

Then we calculate the accuracy, sensitivity and specificity for various probability cutoffs and try to plot accuracy sensitivity and specificity for various probabilities and observe the curves meet between 0.2 and 0.4. This implies our initial assumption of 0.3 cutoff is right.

We try to predict the test values based on the above model and we get accuracy of 0.75 and sensitivity of 0.82.

Conclusion:

Final Logistic Regression Model built is-

Target Variable = -0.7262 - 1.3535(Do Not Email) + 1.1397(Total Time Spent on Website) + 2.4352(Lead Origin_Lead Add Form) + 2.2047(Lead Source_Welingak Website) + 1.0224 (Country_unknown) -1.4077(What is your current occupation_Unknown) + 2.3755(What is your current occupation_Working Professional)

The important features are:

1. Do not Email

2. Total Time Spent on Website

3. Lead Origin_LeadAdd Form

4. Lead Source_Welingak Website

5. Country Unknown

6. What is your current occupation_Unknown

7. What is your current occupation_Working Professional

In which the features “Do Not Email” and “What is your current occupation_Unknown” are negatively correlated, and rest of the features are positively correlated to the target variable.

---------- END ----------