Lead Scoring Case study

SUMMARY

Problem Statement:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

The following steps are used:

1. **Understating the Data Set:**
   a. In The given data set there was total of **9247 records** with **37 attributes.**
   b. Data contains high number of missing values which we have handled by capped the null values to 40%, anything above 40% was dropped.
   c. After observing the columns we found the biasing in certain columns(i.e., one class is relatively higher than other). We need to drop these columns because they lack variation.
   d. Finally we have cut down to **9247 records** and **16 attributes.**
2. **Outlier Check:** We did some univariate analysis and then outlier treatment these were some potential outliers we did capping of 99%
3. **Visualizing The Data:** We did some bivariate analysis and observed
   a. Comparative to other lead origins categories 'Lead Add Form' category has the highest conversion ratio.
   b. 'Reference' category in 'Lead Source' column is doing good followed by 'Google' & 'Direct Traffic' in conversions.
   c. 'SMS Sent' category in 'Last Activity' column has highest conversion ratio followed by 'Email Opened'.
   d. 'Working Professional' seems to convert more as compared to 'Unemployed' Ones
4. **Dummy variables/Scaling:** Categorical variables were converted into dummy variables and scaling was done on both Train and Test dataset. Also, checks were done to reduce the dimensions of categories to make the model light-weight.
5. **Train-Test split:** Train test split was performed in the ratio of 70/30 % and logistic regression was initiated.
6. **Building Model:**
   a. For building a successful model, recursive feature elimination as well as variance inflation factor was used
   b. In the 7th Model Built, We observed the p-value are less than 0.05 and VIF values are less than 5. Therefore it seems that all the variables are significant and have low multicollinearity.
7. **Prediction**:
   a. After analysis using roc curve, After observing 'Accuracy Vs Sensitivity Vs Specificity' 0.3 Probability seems to be optimal cutoff

b. After Observing both 0.3 and 0.37 Cutoffs, 0.37 Gives a bit higher accuracy score.

Top three variables that the edtech company can focus on are:

1. Tags Will revert after reading the email
2. Lead Origin Lead Add Form
3. Lead Source_Welingak Website

Also if there is a scenario where company has attained its target before the quarter and now it needs to focus on new work, Then model can be tuned to attain high specificity, as specificity increases the model correctly predicts all the non conversions, so the cut-off value needs to be high to achieve this.