Lead Scoring Case Study

Summary

Logistic regression model was built for X Education Company which sells online courses to industry professionals. The company needed help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. Following are the steps used for analysis:

Step 1: Reading and Understanding the Data
Read and analyse the data.

Step 2: Data Preparation
Columns having more than 45% missing values were removed. The option 'Select' had to be replaced from categorical variables with a null value since it did not give us much information. Few of the null values were changed to 'Not Specified' so as to not lose much data. Single value features have been dropped as they don't serve any purpose for analysis. Data Imbalance observed in some columns was treated by removing those columns. Outlier treatment was done on numeric columns.

Step 3: Exploratory Data Analysis
Exploratory Data Analysis was done to check the condition of our data. Univariate and bivariate analysis was done on categorical and numerical variables. Relation of individual columns with target variable (Converted) was checked to get some insights on lead conversion.

Step 4: Data Modelling
The dummy variables were created for categorical variables.

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

For numeric variables we used the StandardScaler for feature scaling.

Step 5: Model Building
Firstly, automated feature selection was performed using RFE to get the top 15 relevant variables. Later the rest of the variables were removed by manual elimination process depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

Step 6: Model Evaluation
A confusion matrix was made and model accuracy was calculated which was 92.37% on training data. The optimum cut off value was calculated which came to be around 0.5. ROC curve was plot to visualize tradeoff between sensitivity and specificity.

Based on optimal cut-off value, predictions were made on train data and evaluation metrics were calculated.

Step 7: Making predictions on the test set
Prediction was done on the test data frame and with an optimum cut off as 0.5 and with accuracy, sensitivity and specificity of around 92%. Precision around 88.46% and recall around 91.67% on the test data frame was observed.

Step 8: Checking Conversion Rate

As per the lead score calculated on final data shows conversion rate of 88%. Hence, we can say the business requirement of achieving 80% conversion rate was achieved.

Step 9: Conclusion

The Accuracy, Precision and Recall score we got from the test data are in the acceptable region

In business terms, our model is having stability and accuracy with adaptive environment skills. Means it will adjust with the company's requirement changes made in coming future.
The top three variables contributing for lead conversion are as below:

- ✓ Tags_Closed by Horizzon
- ✓ Tags_Lost to EINS
- ✓ Tag_We will revert after reading the email