Lead Score Case Study – Logistic Regression Assignment

Brief Summary

This analysis is carried out for X Education in an effort to attract more business professionals to their courses. We learned a lot from the fundamental data on how potential customers use the site, how long they stay there, how they got there, and the conversion rate.

There were 9240 rows and 37 columns present in the data set initially. Except for a few null values, the data was mostly clean. However, the option ‘Select’ had to be changed to a null value because it provided little useful information. To avoid losing too much data, a few of the null values were changed to "not known." Little data cleaning is done by dropping the columns with more than 40% missing values & grouping values with low count to remove bias.

To quickly assess the state of our data, an EDA was performed. It was discovered that several of the categorical variables' components were unnecessary, so handled accordingly. There is also seen some outliers in numerical variables & that is also taken care of.

Further moving to data preparation for logistic regression, first we have created dummy variables & scaled numerical variables using Standard scaler. Model building started after splitting the data into Train -Test set and feature selection is done using RFE & Automated approach.

First, the top 15 relevant variables were determined by RFE. Later, based on the VIF values and p-value, the remaining variables were manually deleted (the variables with VIF < 5 and p-value < 0.05 were retained).

For Model evaluation, a confusion matrix was created. Later, the accuracy, sensitivity, and specificity were determined using the ROC curve, and they all came to be around 85%-89% each.

Predictions were done on the test data set and with an optimum cut off as 0.3 with good accuracy, sensitivity and specificity. Precision & Recall also checked & to be found as 83% & 89% with ~0.35 optimal cut-off.

So, we have achieved an overall accuracy of about 0.87 on our Logistic Regression model. That is, there is 87% chance that our predicted leads

will be converted. This meets the CEO's target of at least 80% lead conversion.

Some Findings: -

Significant variables to predict the lead conversion are : -

- Tags_Closed by Horizon
- Tags Ringing
- Tags Will revert after reading the email
- Tags Interested in other courses
- Last Notable Activity SMS Sent
- Total Time Spent on Website
- Last Activity Olark Chat Conversation
- Lead Origin Landing Page Submission

With these in mind, X Education can succeed since they have a very good probability of persuading nearly all prospective customers to change their minds and purchase their courses.

Let us compare the values obtained for Train & Test:

Train Data:-

Accuracy : 89.19%

Sensitivity : 89.15%

Specificity : 89.22%

Test Data:-

Accuracy : 87.12%

Sensitivity : 89.36%

Specificity : 85.82%