Lead Scoring Case Study - Summary

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customer visit the site, the time they spend there, how they reached the site and the conversion rate.

The Following are the steps followed in the case study:

1. Data Sourcing and Cleaning:

The Lead.csv dataset was read and the information have been clearly understood. While cleaning the data the value "Select" had to be replaced with a null value. Also there were so many null values present in few of the columns. Few of the null values were replaced with mode value so as to not lose much data. Once the Data cleaning was over there were nearly 27 columns present.

2. Data Visualization:

Visualization of both numerical and categorical variables was done against the target variable "Converted". The numeric values seem good and outliers were treated. It was found that a lot of elements in the data categorical variables were irrelevant and hence were dropped.

3. Data Preparation:

The following steps were followed in data preparation.

- Created Dummy variables for Categorical Data
- Splitting the data (Train – 70%/Test – 30%)
- Scaling the features (Standard Scalar was used)

4. Model Building:

Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables was removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were retained).

5. 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 approximately 80% each.

6. Prediction on Test Data:

Prediction was done on the test data set and with an optimum cut off as 0.42 with accuracy, sensitivity and specificity of approximately 80%.

7. Precision-Recall:

This method was also used to recheck and a cut off of 0.41 was found with Precision around 91% and recall around 69%.

8. Final Observation:

The Final Evaluation Metrics for the train Dataset:

- Accuracy: 0.84%
- Sensitivity: 0.88%
- Specificity: 0.81%

The Final Evaluation Metrics for the test Dataset:

- Accuracy: 0.82%
- Sensitivity: 0.86%
- Specificity: 0.80%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

9. Recommendations:

Following are the selected features which are more significant in predicting the conversion.

Features having positive impact on conversion probability:

Features with Positive Coefficient Values:

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Will revert after reading the email
- Tags_Busy
- Last Activity_Unsubscribed
- Last Notable Activity_SMS Sent
- Lead Source_Welingak Website
- Last Activity_Other Activities
- Lead Origin_Lead Add Form

Features having negative impact on conversion probability:

Features with Negative Coefficient Values:

- Specialization_Others
- Do Not Email

- Lead Origin_Landing Page Submission
- Tags_Ringing