Summary – Lead\_conversion\_case study

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 which type of customers are our potential customers, how they use the site, how long they stay there, how they got there, and the conversion rate and the lead score. The procedures are as follows:

1. Cleaning data: Except for null values in some columns, 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, only a small number of the null values were changed to 'not specified' or ‘others’. However, they were later taken out while creating dummies. Some columns such as country was dropped out as 70% people were from india and 27% values were missing.

2. EDA: To quickly assess the state of our data, an EDA was performed. It was discovered that several of

the categorical variables with “Yes” , “No” which was transformed to 1/0 were unnecessary , similarly for categorical variables make variables were present contributing to very low frequency so all of them were combined which follows similar pattern.

3. Dummy Variables: The fake variables were made, then later the fakes with the 'Not\_specified' or “others were dropped. We utilized the StandardScaler to scale numerical numbers.

4. Train-Test split: The split was done at 70% and 30% for train and test data respectively.

5. Model Building: First, correlation was seen and highly correlated variables wer removed then RFE was used to identify the top 15 important factors. Later, the remaining variables were manually deleted based on their VIF values and p-value (the variables with VIF <2 and p-value <0.05 were maintained).

6. Model Evaluation: A confusion matrix was created. Later, the ideal cut off value (using the ROC curve)

was utilized to determine the accuracy, sensitivity, and specificity, which came to be about 90% each.

7. Prediction: Prediction was performed on the test data frame with an ideal cut off of 0.35 and nearly 90% accuracy, sensitivity, and specificity.

8. Precision – Recall: This approach was also used to retest, and a cut off of 0.40 was discovered on the

test data frame, with precision around 93% and recall around 85%. The variables that mattered the most

to potential customers with their coefficient are mentioned below

|  |  |
| --- | --- |
| **Features** | **coef** |
| Tags\_Closed by Horizzon | 6.7735 |
| Tags\_Lost to EINS | 5.6049 |
| Tags\_Will revert after reading the email | 4.4724 |
| Lead Origin\_Lead Add Form | 3.6828 |
| Last Notable Activity\_SMS Sent | 3.0727 |
| Total Time Spent on Website | 1.0814 |
| Last Notable Activity\_Email Opened | 1.0366 |
| Lead Source\_Olark Chat | 0.9617 |
| Tags\_Busy | 0.7109 |
| Lead Source\_Direct Traffic | -0.5416 |
| const | -2.4789 |
| Tags\_Other\_tags | -2.9731 |

Keeping these in mind, X Education may use the model since it has good possibility of convincing practically all potential consumers to alter their minds and purchase their courses