Summary

ML-MODEL: LEAD SCORING CASE STUDY

Team Member: Neha, Sunava Neogy & Naveen Kumar Upadhyay

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 customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. Cleaning of data:

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. If null values<10% impute them with median(numerical columns) and mode(if categorical columns). Few of the null values were changed to 'not provided' to not lose much data. Levelling of 'Country' column done as India, Outside India (as very few values other than India and 'Not Provided'(where values are null).

2. EDA:

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good, and no outliers were found.

3. Dummy Variables:

The dummy variables were created if the category in categorical variable>2. Remove the dummy column manually where it is suffix is "Not Provided". For numeric values we used the MinMaxScaler.

4. Train-Test split:

The split was done at 80% and 20% for train and test data respectively.

5. Model Building:

Firstly, RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

6. Model Evaluation:

A confusion matrix was made. Later on the optimum cut off value (using ROC curve cutoff .30) was used to find the accuracy, sensitivity, precision and specificity which came to be around 93% ,88% , 93% and 92% respectively.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.30 with accuracy-91%, sensitivity-92%,precision-93% and specificity of around 90 **%.**

8. **Precision – Recall:**

This method was also used to recheck and a cut off 0.39 was found with Precision around 91.5% and recall around 90% on the train data frame and on test data frame Precision -92 % and Recall around 90.5%.

It was found that the variables that mattered the most to decide the potential buyers are (In descending order):

1. Tags_Closed by Horizzon
2. Tags_Lost to EINS
3. Tags_invalid number
4. Tags_switched off
5. Tags_Already a student
6. Tags_Ringing
7. Tags_Not doing further education
8. Lead Source_Welingak Website
9. Tags_Will revert after reading the email
10. Total Time Spent on Website
11. Tags_Interested in full time MBA
12. Tags_Interested in other courses
13. Tags_opp hangup
14. Last Activity_SMS Sent
15. What matters most to you in choosing a course_Better Career Prospects
16. Tags_Graduation in progress
17. Last Notable Activity_Modified

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.