SUMMARY

X Education is an education company that sells online courses to industry professionals. Many professionals land on their website and either browse through the courses, watch videos or fill up a form for a course. On filling the form with their email address or phone number, they are classified as a lead. However, only few of these leads get converted into paying customers. The conversion rate at X Education is around 30% which is poor as compared to the leads they get.

To increase the conversion rate, the company wishes to identify the most potential leads called 'Hot Leads'. They want us to build a model to identify these leads and help meet the expectations of the CEO of X Education of achieving a conversion rate of 80%.

Steps followed for the analysis:

1. Read/understand data and Data Preparation:

We first read the data and checked the features for null/missing values. There were many features with category as 'Select'. We replaced them with null values. We then handled the null values by either deleting that feature or imputing them with mode or creating a new category like 'Others'. Dropped columns not required for analysis and also handled outliers.

2. Exploratory Data Analysis (EDA):

We performed univariate and bivariate analysis on the data. The lead conversion rate was poor with only 37.85% converting. Leads spending more time on websites are more likely to convert.

3. Data Preparation:

i. Converted binary variables (Yes/No) to 1/0.
ii. We then created dummy variables for categorical features.
iii. We then split the data into train data (70%) and test data (30%)
iv. We scaled the features using StandardScaler().

4. Model Building:

We selected 20 features using RFE. We kept manually dropping features based on p-value and VIFs until we had features with p-value and VIF within the acceptable range. We made predictions on Model 7.

5. Model Evaluation:

We created a confusion matrix. We plotted the accuracy, sensitivity and specificity for various probabilities and from that curve found the optimal cutoff probability of 0.34. We then calculated the accuracy, sensitivity and specificity with the cutoff probability.

6. Make predictions on test data:

We scaled the features of the test data and made predictions on our final model. We then assigned lead scores.

Observations

Comparing the accuracy, sensitivity and specificity for train and test data.

Train Data

i. Accuracy – 81.0%
ii. Sensitivity – 81.9%
iii. Specificity – 80.5%

Test Data

i. Accuracy – 80.4%
ii. Sensitivity – 80.4%
iii. Specificity – 80.4%

Recommendations

The leads coming from ‘Welingak Websites’, ‘Reference’ and ‘Olark Chat’ are more likely to get converted. Working professionals are likely to have higher conversion rate. Leads who spend more time on the website are more likely to get converted. X Education should focus on these leads.