Lead Scoring Case Study

Summary

—

Team Members

Karishma Singh,
Ashish Pawar,
Shrikant Angre

Problem Statement

X Education sells online courses to industry professionals. X Education gets a lot of leads; however, its lead conversion rate is very poor.
X Education has appointed us to build a model such that:

Lead Score	**Higher lead score**	**Lower lead score**
Conversion chance	Higher ↑	Lower ↓

The CEO's expectation is → lead conversion rate to be around 80%

Step1: Reading and understanding the data:

1. Imported the csv file with the help of the library and got the number of rows and columns in the dataframe. (9204 rows and 37 columns)

Step2: Data Cleaning:

1. We dropped features which had
    a. more than **40%** of missing values,
    b. around 98% with single value,
    c. **70+% 'Select'** or **'Nan'** values
2. **'Select'** and '**Nan**' were converted to '**Not Selected**' for few variables.
3. Dropped '**Prospect ID**' and converted the '**Lead Number**' as an index column.

Step3: EDA:

1. Done **univariate**, **bivariate** and **multivariate** analysis with the help of boxplot, countplot and heatmap against target variable '**Converted**'.
2. Correlation between numerical and target variables was done with the help of heatmap.

Step4: Dummy Variables & Feature scaling:

1. Dummy variables were created for the categorical features
2. Scaled numeric values using MinMaxScaler
3. Checked correlation amongst variables using heatmap, as the number was high RFE was used later to select features

Step5: Train-Test Split: The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step6: Model Building:

1. RFE method was used to select top 15 features. Later removed features using VIF value (**>= 5**) and p-value (**>= 0.05**1).
2. After 3 model iterations we came to **12** most significant variables

Step7: Model Evaluation & Prediction:

1. Confusion matrix was made. Optimum cut off value (**0.36**) was used to find accuracy, sensitivity and specificity i.e., **81%, 79% and 82%** respectively.
2. Plotted the ROC curve for features, with an area coverage of **89%**.
3. Then, check if **80%** cases are correctly predicted based on the converted column.
4. Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be **80%**, Sensitivity= **80%**, Specificity= **81%**.

Step8: Precision & Recall:

This method was also used to recheck and a cut off of 0.42 was found with Precision → 72% and recall → 79% on the train data set & Precision → 72% and recall → 80% on the test data set

Step9: Conclusion:

The top Five variables in our model which are contributing most towards the probability of lead getting converted are

I. TotalVisits

II. Total Time Spent on Website

III. Lead Origin_Lead Add Form

IV. What is your current occupation_Working Professional

V. Lead Source_Welingak Website

These are selected basis their coefficient's value in our Logistic Regression model, which are 9.12, 4.51, 3.66, 2.52 and 1.84 respectively