Lead Score Prediction at X Education

X Education is an online course provider for industry professionals. Company markets its courses on multiple websites and search engines like google. The aim of the company is to increase its lead conversion for potential leads.

Below is a brief summary of steps followed for the analysis performed:

Data Understanding and Exploration:

1. Loading the data into data frames
2. Check basics of the data:
    a. Checked data for insights like data description, data information to find percentile values, mean, median values and shape of the data.
3. Data Exploration:
    a. Compiled a list of numerical columns.
    b. There are 7 numerical columns in the provided data.
    c. Plotted a heat map to check if there is any correlation between the columns.

Data Cleaning:

1. After analyzing the data, we found that there are columns that have a “Select” value which is set as default. Replaced the “Select” value by NAN values and treated the records as if they are null.
2. Calculated the percentage of null values present in the data. Found that there are many columns having more than 40% Null Values. Dropped these columns as they will not be providing any useful insights in model building.
3. Also there were columns such as Tags, Lead Quality those were given manually by the employee handling the cases. Since these columns are manually completed they will not be presenting any value for model building. Hence dropped these columns.
4. Country and City are providing geographical data and as this course is online, which can be accessed from anywhere, locations will not weigh as much for predictions. Hence opting to drop the columns.
5. Created the Dummy Variables for Categorical data.

Model Building and Evaluation:

After data was cleaned, proceeded further to create the model. Since data contains categorical variables, performed the logistic regression.

1. Scaling:

    Performed scaling on numeric columns so that data points are relatively
dispersed. 2. RFE:

    Used RFE to reduce the number of features form 29 to 12.
3. Assessing the model with Stats Model:

    Used stats model for checking the summary of the model.
4. Calculated the sensitivity of the model on train data which turned out to be 70%.

Plotting ROC Curve:

Earlier for predicting the values on train data, we used 0.5 as cut off value. Plotted ROC curve to get exact cut off value.

Finding Optimal Cut Off Value:

Calculated accuracy, sensitivity and specificity for various probability cut-offs and plotted them to get the above graph.

All the variables intersect each other at 0.30 , hence cut off value = 0.30

Making Predictions on Test Data:

When ran for Test Data, Model produced 83% Sensitivity and 80% Accuracy as per the desired goal.

Learnings from the Case Study:

1. Observe the categorical data carefully as it can contain a default value set.
   e.g.: “Select” value in this case.
2. Analyze all the columns to decide which to drop and which to use.
   a. Highly skewed Categorical columns need to be dropped as it won’t add much to analysis.
   b. Even we need to club the categories which are sparse in number in the categorical columns. Otherwise, it would lead to unnecessary numbers of dummy columns
   c. Also, columns containing redundant data as they will not help in model building. So, we see data cleaning is a very important phase of any regression modelling.
3. RFE:
   a. When a number of features are large, manual feature elimination is not possible.
   b. Use of RFE along with manual elimination is the optimal approach in most of the cases.
4. Finding Optimal Cut Off Frequency:
   a. Always find the cut off frequency using ROC Curve and plotting the Accuracy Sensitivity-Specificity graph.