Summary

X Education sells online courses to industry professionals. The typical lead conversion rate at X education is around 30%, which is very poor the company wishes to identify ‘Hot Leads’.

The company requires to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead scores have a higher conversion chance and the customers with lower lead scores have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Based on the data provided the following procedure is followed to build the model:

Step 1: Analysis approach:

To increase the Lead conversion rate from 30% to around 80%, Machine-learning model has to be built based on the previous data available in the company, which helps predict the conversion of the customer with the accuracy required.

As this is a classification problem, a Logistic Regression model has to be built with the response variable as Conversion and others as feature variables.

Step 2: Solution Methodology:

The steps includes below:

Understanding of the data, which is as following:

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. Few of the null values were changed to ‘not provided’ so as to not lose much data. Although they were later removed while making dummies. Since there were many from India, the elements were changed to ‘India’

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 seems good and no outliers were found

Check and handle duplicate data, Drop columns that :

a) does not help much towards the analysis

b) If it contains large number of missing values.

We dropped features that are updated by the sales team in the data as the model building is based on the data collected from the student online.

- Check and handle NA and missing values, check and handle outliers in the data.
- Splitting the data to Test and Train

Model Building: logistic regression used for building the model and RFE technique is used for feature reduction.

- Prediction on train and test data
- Validation of the model- Confusion matrix , area under ROC curve
- Drawing conclusions and recommendations.

Step 3: Data Cleaning:

In data cleaning below is what was done:

Total Number of Rows was 37 and Total Number of Columns was 9240.

Column lead was dropped as it does not help in building the model and as the model building should be based on data collected from the customers online, we dropped features that were updated by the Sales team.

We also observed that the percentage of missing values in few features is above the threshold level of 40 percent which has been dropped, features with above 99% data imbalance have been dropped

Post which we Imputed missing values with median and mode however relevant and treated the outliers by dropping the outliers.

The Logistic regression Model results that came out were as below:

From the given 35 feature variables in the data, using Logistic Regression techniques, 8 feature variables have been identified to have a greater impact on predicting the conversion of the customer.

Then we moved in to the Model evaluation metrics and On model built we predicted the probability on train data and considered 0.3 as a cut off. Implies if the probability is greater than 0.3 it will be considered as 1 (one, lead) and less than 0.3 is considered as 0 (no lead).

Additionally, based on above calculation the confusion matrix was built and sensitivity of the model was checked, which is around than 0.77.

Step 4: Logistic regression Model results:

Feature Selection is done using RFE. Initially only 10 features are selected in the RFE model and based on these 10 features Logistic Regression model is build and is assessed.

Step 5 Model evaluation metrics:

Based on top 10 features , On model built we predicted the probability on train data and considered 0.3 as a cut off.

Additionally, based on above calculation the confusion matrix was built and sensitivity of the model was checked, which is greater than 0.8.

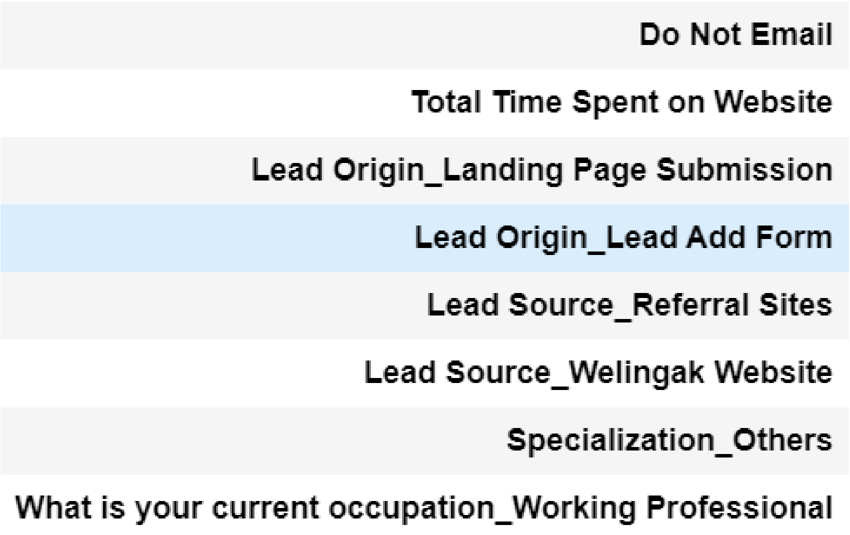
Step 6 Predictions on the test set:

Prediction was done on the test data frame and with an optimum cut off as 0.30 with

Accuracy of 79.5, sensitivity of 80.6%.

Conclusion:

The important features came from the Model are:



Do Not Email
Total Time Spent on Website
Lead Origin_Landing Page Submission
Lead Origin_Lead Add Form
Lead Source_Referral Sites
Lead Source_Welingak Website
Specialization_Others
What is your current occupation_Working Professional

Features namely Do Not Email and What is your current occupation_Unknown are negatively correlated and rest of the features are positively correlated to the target variable.

The lead conversion rate achieved is around 77% as the accuracy on the test data is 77%

---------------------------------- END ---
