Leads Scoring case study-Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. Although X Education gets a lot of leads, its lead conversion rate is very poor. The company wishes to identify the most potential leads, also known as 'Hot Leads'.

Solution Summary followed by us:

Step 1: Reading and Understanding the data.

a) Loading the data and understanding the data.
b) Data Cleaning which included data pre-processing, missing value imputation and outlier analysis.

Step 2 : Visualizing the data

Performed univaraiate analysis for both categorical and numerical data

Performed bi-variate and multi-variate analysis using pairplots and heatmaps.

To check the skewness of data, performed data imbalance analysis.

Step 3: Data Preparation for Modelling

Dummy variable creation for categorical data.

Step 4: Splitting the data into Test Train Split(we followed 70-30 split)

Feature Rescaling: We used the Min Max Scaling to scale the original numerical variables.

Step 5: Building a logistic regression model

a. Using the RFE, we went ahead and selected the 15 top important features.

b. From the summary which is generated each time we did rfe on the model, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.

Step 6: Making predictions using the final model(i.e model 6)

Predicting the probabilities (of "converted" value being 1) on the train set.

Step 7: Model Evaluation

We did predictions on train set first,

a.We plotted ROC curve for the features and we got it with an area coverage of 95% which further solidified the model.

b. Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.33.

c. Metrics for our final model, Accuracy : 88.54%, Sensitivity : 87.76% and Specificity : 89.02%

Then we made predictions on the test set,

The metrics we obtained are, Accuracy : 89.4%, Sensitivity : 87.89% and Specificity : 90.03%

Step 8: Conclusion:

After trying several models, we finally chose a model no 6 with the following characteristics:

-All variables have p-value < 0.05, showing significant features contributing towards Lead Conversion.

-All the features have very low VIF values, means hardly there is any multicollinearity among the features. This can be seen from the heat map.

-The ROC curve has a value of 1, which is very good!

-The overall accuracy of Around 89% at a probability threshold of 0.33 on the test dataset is also acceptable.