Summary report on lead scoring case study

Problem Statement:-

An education company named X Education sells online courses to industry professionals. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor.

Objective:-

Our objective is to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers by building a model and try to get converted rate upto 80 % or more using few variables which are highly affected .

The steps followed are:-

1) Reading and understanding the data

i) Importing of libraries and dataset is done and then inspected the dataset.

2) Cleaning the data

ii) As we can observe that there are select values for many columns. So 'Select' as input converted into NA values.

iii) Removed those columns in which null values are high than 50 %.

iv) After that drop unnecessary columns with unique 1 value and columns created by sales department and highly skewed columns.

v) The rows which have null values more than 5% are dropped.

vi) The null values in Categorical columns and continuous columns imputed with mode, median respectively. After cleaning the data we have restored 99.99% of data.

3) Visualization of data

i) Checked the correlations between the variables.

ii) Univariate Analysis for numerical columns and categorical columns was performed.

iii) Bivariate analysis for categorical-categorical and numerical-categorical columns was performed.

4) Preparing the data for modelling

i) Columns which contains categories like yes and no, converted into 1and 0.

ii) Created dummy variables for some of the categorical variables and then dropped the unknown or other variables which contain Nan value or binning values. Dropped the repeated columns.

iii) Then performed outlier checking and treatment of outliers by IQR method.

iv) Now data is divided into two data set i.e train data set and test data set as 70% and 30% of prepared data. Scaled the train data and drop highly correlated column too

5) Modelling

i) Used logistic regression model with RFE technique with top 15 columns choose automatically by RFE. Later on done manual features elimination using P-value and VIF values, repeated till we got our desired features.

ii) Calculated Metrics beyond simply accuracy. Plotted the ROC curve and then on the basis of our business problem we have decided our Optimal Cutoff Point as 0.33

iii) Got Accuracy, Sensitivity and specificity, precision and recall on train data set are 80.44, 80.59,78.21,70.30,80.59

iv) Then done predictions on the test set and got Accuracy, Sensitivity and specificity, precision and recall on test data set are 80.81, 83.41,79.26,70.66,83,41

6) Final Analysis

i) Lead score was calculated for the data (train and test) and got the final sorted list (based on coeeficient values) of top 10 variables

7) Conclusion

i) In the end we get 6 features with positive coefficient(conversion probability increases) and 4 with negative coefficient (conversion probability decreases)

ii) For increasing lead conversion rate (more than 80%) target the potential leads whose leads score are in the range of 60-100(higher score would mean that the lead is hot, i.e. is most likely to convert) and also try to get leads from the top 10 features which are selected by the model as they contribute most to a Lead getting converted successfully.