Lead Score Case Study

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, The company markets its courses on several websites and search engines

some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Expectation:

The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. the target lead conversion rate to be around 80%.

Approach:

We have solved this problem with the logistic regression model. Start with data analysis.

Checked the data type of all the variable whether they are as expected or not, then checked for null value and from the data we get that many columns have null value, then we find the percentage of null value present.

For few columns there was 'select' value which was nothing but the missing value so converted it to the null value

Dropped the column which has null value greater than 45% and dropped the row which have null value 1 or 2 %.

Checked the value count for each column and where there was skewed value for the column then also dropped the column.

Performed outlier treatment with capping and EDA

Now for the categorical variable we had so many categories so this was an issue as while creating dummy variable there was lots of dummy created so to solve this, we have done bucketing of categories whose value count was less.

Converted all the binary variable into 1 or 0. There was many binary variables for which around 99 or 100 % data was either only yes or only no, so dropped all such column as this was highly skewed and will not help in our analysis.

After data cleaning, EDA and data preparation, proceed for modelling by splitting data into test and train – in the ratio of 30-70 %

As per the problem statement we need lead score against each lead number so we need to keep this lead number, but we don't want to process this on model so made this column an index.

Performed Feature scaling was required for some continuous variable. Looked for correlation between variables. Then proceed with model building:

Used RFE to get top 15 feature and passed these variables to the stats model, checked the p- value and VIF for each variable after creation of 1st model. Dropped the variable with high p-value and then again create model then drop the variable with high VIF and get all the metrics value after each model creation.

Here, sensitivity is the important metrics so we need high sensitivity.

After tuning the model we get sensitivity greater than 80%.

And the as per the requirement:

Prepared a list of lead score against each lead number.