Lead Scoring Case Study

In order to increase the lead conversion rate by identifying the hot leads from the large number of leads coming from various sources, we have developed a logistic regression machine learning model. This model predicts the probability of a lead to be converted which is called as Lead Score. With this we can filter the most promising leads which have high score. We can set a threshold value for lead score above which leads are considered as hot leads. And the company now focus on targeting these hot leads for selling their products.

The approach for the building a logistic regression model is that it is simple, linear and easily interpretable. We first cleaned the data by identifying and handling missing values, invalid values and outliers. After that we prepared the data for modeling which included creating dummy variables for categorical variable, scaling continuous variables and then performing train test split.

Since logistic regression is a linear model, we have to hold and check all the assumptions required for the linear model.

As we created multiple dummy variables the number of dimensions increases a lot, we identified variables which have only one category or very low variability like Magazine, Newspaper and dropped them as they will not contribute in the model building.

Next, we checked which predictor variables are highly correlated and dropped those variables as they would lead to multicollinearity.

Then for model building we took a hybrid approach of building a automated model and then fine tuning the model by manually iteratively dropping / adding the variables in order to increase the significance of the model and also the accuracy of the model.

Once we were satisfied with the model, we used the model on the test dataset to predict the lead score.

We evaluated the model performance using confusion matrix and other metrics

With the help of ROC curve, we decided the threshold probability which categorizes the leads into converted or not. We then calculated the lead conversion rate.

Then we made predictions using our model with the test data set. Here, the lead score matched the score obtained with the train data set.

And we could achieve the desired lead score i.e. 80%.