Summary for Lead scoring case study

By Sai Priya and Raoof

An education company named sells online, Once people land on the website and fill the form with they are marked as lead, employees from the sales team starts contacting the leads, then if the lead enrolls into the course they are classified as converted, The percentage of lead conversion is about 30, So to make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads' and thereby directing their resources towards the hot leads, create a model which could help to do, And make this model deployable in future.

We began with **understanding the data** where we looked upon various parameters like shape, columns, info etc. to gain a basic understanding of dataset, and then removed the unwanted columns,

Then we moved onto **cleaning the dataset** by removing any column which contained more than 40 percent of missing values, while imputing the null values for all other columns with appropriate values, we did so by replacing those missing values by a measure of central tendency (mean) for numerical data and mode for categorical data where missing values were less, as for columns with higher values we imputed with labels such as 'undefined' and 'other',

Once missing data was dealt with, we moved onto treating outliers by clipping the values to some appropriate limit, we analyzed the same using boxplot,

Then we performed univariate and bivariate analysis **(EDA)** on all features, for categorical data we drew count-plots and for numerical data we constructed heatmap, boxplots and bar-charts

After EDA we started data preprocessing for model development, where we created dummy variables, scaled the numerical variables and then performed train-test split,

Then we started training the model by using GLM, we used RFE to get the top 15 column upon which we performed fine tunning using p-values and VIF, as a result we got our final model with 13 variables,

For our final model we calculated Accuracy, sensitivity and all other relevant metrics, then we drew ROC curve and plotted Trade-off curve to find optimal point, using which we formulated lead scoring based upon which we predicted the lead as hot or cold, the we also conducted Precision and Recall test on our model and reached the following conclusions,

The more the lead is avoiding the contact higher are their chance of failing of becoming cold lead. The more doubt the lead is showing, the higher are their chances of becoming a cold lead. If the lead is already engaged in some other course, they will become cold heat and,

If the lead has originated via Add form, they are most likely to convert.

If the last activity is done via SMS, there is high probability of lead conversion, those leads should be target whom were interested before and people who are visiting website for longer amount of time.