# Lead Score Assignment Summary Report

By
Balaji Raman,
Sharhedha Raghavan

## EDA Analysis

1. first studied the columns
2. converted all the selects to nan
3. calculated percentage of missing values
4. drop columns with missing values more than 40 percent
5. impute some missing category variables with 'others' since they have significant conversions. This is to prevent loss of data and to capture these later better if they are major contributing parameters
6. clubbing some smaller number of contributing categories in each column to make readability better since they are only 1-2 records for each of those categories
7. removing some single value variables since the variables have a constant or near constant value throughout all the 9200 odd records. Such variables do not contribute to our analysis they are rather constants
8. converting binary category variables to 0/1
9. converting categorical variables after trimming some categories into dummy columns
10. merging the data frame with these additional columns and removing the original categorical columns
11. final check of the dataframe before building logistical regression model.

## Logistic Regression

### Training and Model building

1. Scale the numeric variables using standard scaler fit transform
2. Split into test and train
3. Remove the y variable
4. Check the correlation between the variables and remove if there are significant correlations
5. Build regression model using 'sm.GLM' fit for and initial view. Remove variables with high probabilities and proceed for rfe
6. Use RFE logistic regression GET 15 columns to perform regression again
7. Repeat step 5
8. Get the result column
   Apply VIF and remove variables with high VIF' if necessary

   Make a new dataframe with calculated probability of conversion , actual conversion (actual y) for each customer

Now with a random probability threshold, calculate predicted conversion (predicted y) and merge with the above dataframe.

9. Calculate accuracy ie predicted vs real conversion

10. Now plot ROC  curve and the area under ROC curve gives the model performance. i.e. the higher the area the better is the model.
11. Finding optimal cutoff : plot the accuracy , sensitivity and specificity at various probabilities. Select the probability at the which all curves intersect as the optimal cut off.
12. calculate predicted conversion (predicted y) with optimal cut off and make final data frame.
13. Calculated with final dataframe **metrics.**

## Testing the model

1. Repeat same on test set by just transforming scaling and then applying the earlier result from the final train fit
2. We apply the learnt cutoff to calculate predicted test column
3. Compare the metrics for test and train sensitivity specificity and accuracy. If above 0.9 it is an excellent fit.