Lead scoring case study report – Ravi Edla

Objective:

To build a model such that a lead score is assigned to each of the leads in such a way that the customers with higher lead score have a higher conversion chance, and the customers with lower lead score have a lower conversion chance. And the target lead conversion rate is around 80%.

Analysis methodology:

1. **Loading and Understanding Data:**
   From the given case study, we have the leads dataset from the past with around 9240 data points. The shape of the dataset is (9240, 37). The target variable here is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.
2. **Data Cleaning:**

a) On inspecting, we can see that several columns are having select as a value, which is the result of the prospect that is not selecting any value in that field, this means it is as good as having Null. Therefore, we are replacing all select values with Null.
b) Null value checks: we have dropped 10 columns whose percentage of null values is more than 30%.
c) On inspecting further, we have seen that most of the columns are having 1 or 2 unique values and the majority of them are having almost 100% of rows containing the same value. Hence dropped all those columns, which are around 16 columns.
d) We have then filtered few of the nulls from the rows.
e) We removed outliers from the columns 'Total visit' and 'Total time spent on website'.
f) In the end, we are now left with 9 columns and 8901 data points.

3. **Univariate analysis:**
   We tried visualizing the original data variables to look for any pattern or correlation. Few of the observations are that – the 'Olark chat conversation' has the highest successful conversion count compared to other last notable activities.
   Major lead origins are from 'Landing Page Submission' and 'API'. Major source of leads is 'Google', 'Direct traffic', 'Olark Chat' with most traffic coming from 'India'.
   We can see that there are few columns where 'Select' values are high in numbers. These data points are not adding any valuable information so we can treat them as NULL.

4. **Dummy Creation:**
   We have created the dummy variables for all the categorical variables.

5. We have converted 'A free copy of Mastering The Interview' – binary variables (Yes/No) to 0/1.

6. **Test- train split:** we have Split the data with 70% of the data as the train data set.

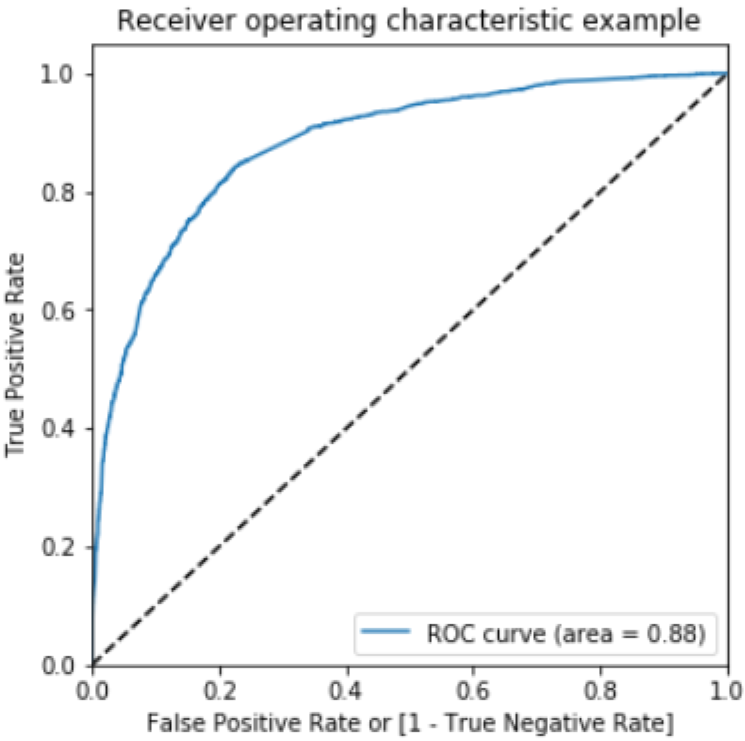
7. **Scaling the data:** Standardised the columns- ‘TotalVisits’, 'Total Time Spent on Website' and 'Page Views Per Visit'.

8. **Model building:**

a. As an initial step, we ran the stats of GLM model on train dataset. We identified that many features are not significant as it had large p- values.

b. We then chose top 16 features using RFE and then ended up with 14 features after manual elimination.

c. ROC curve was plotted and the cutoff point was chosen which is 0.3 in this case.



d. The model parameters for test and train sets were consistent and hence the same model was fitted with the entire data to obtain final result.

e. It can be seen that “Lead Source_Welingak Website”, “Lead Origin_Lead Add Form” and “What is your current occupation_Working Professional” are the variables which have the highest positive effect on conversion.