Palak Gupta
Ravi Kumar
Cohort 12

Analysis – Case Study
For Logistics Regression for Company X

Objective

An education company named X Education sells online courses to industry professionals.

Company X markets their courses through different channels like searching through the Google (or any other search Engine), By referral, Writing Emails etc.

People who landed on their website , they might see videos, search for the courses or fill up the forms. People who fill the forms all called as a lead by providing some required details like Email and Phone number.

Some of these leads gets converted to after sales persons communicate them.
Currently this company is suffering with low conversion rate equals to 30%, and would like to know the factor which could help to increase these leads into HOT leads

By predicting the leads as HOT Leads, Sales team would not spend time to make unnecessary calls to such people who would be not interested at all.

Aim: to find out the driving factors from the dataset (which are received by dataset) to predict leads into HOT Leads. So that Conversion ratio moved from 30% to 80%

Approach

We used below steps during prediction of Leads into HOT Leads

1> Importing Data to the Notebook (used Jupyter)
We found 9240 rows and 37 columns

2> Data Cleaning
Removed fields (columns) which are having high Missing values (Here missing values are not only treated as Null values but also with the data “Select” filled in it

Also removed rows with missing values

After Data cleaning we are left with only 13 fields with 6420 rows.

3> We performed EDA, and found that after Data cleaning lead conversion rate is 48%

4> Dummy variable Creation (for the Categorical fields)
These fields ('Prospect ID', 'Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity', 'What is your current occupation', 'A free copy of Mastering The Interview', 'Last Notable Activity’) are converted into Dummy variables for the further analysis

5> few fields(Prospect ID & Lead Number); which is ID numbers are removed from the Dataset.

6> Perform Test Train split onto the final sets with 70% , 30% division ratio

Approach

We used below steps during prediction of Leads into HOT Leads

7> Perform Min max Scaling to the fields to inline the values into same as range for categorical fields

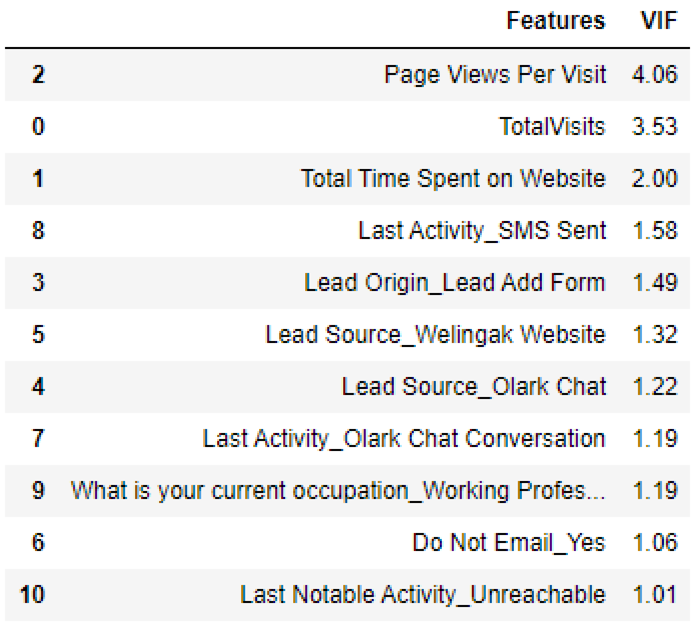
8> Performed Feature selection method (RFE) to create the best model from it.
selected 10 fields with which VIF factor remains under 5 and P values is less than < 0.05

These set of fields are

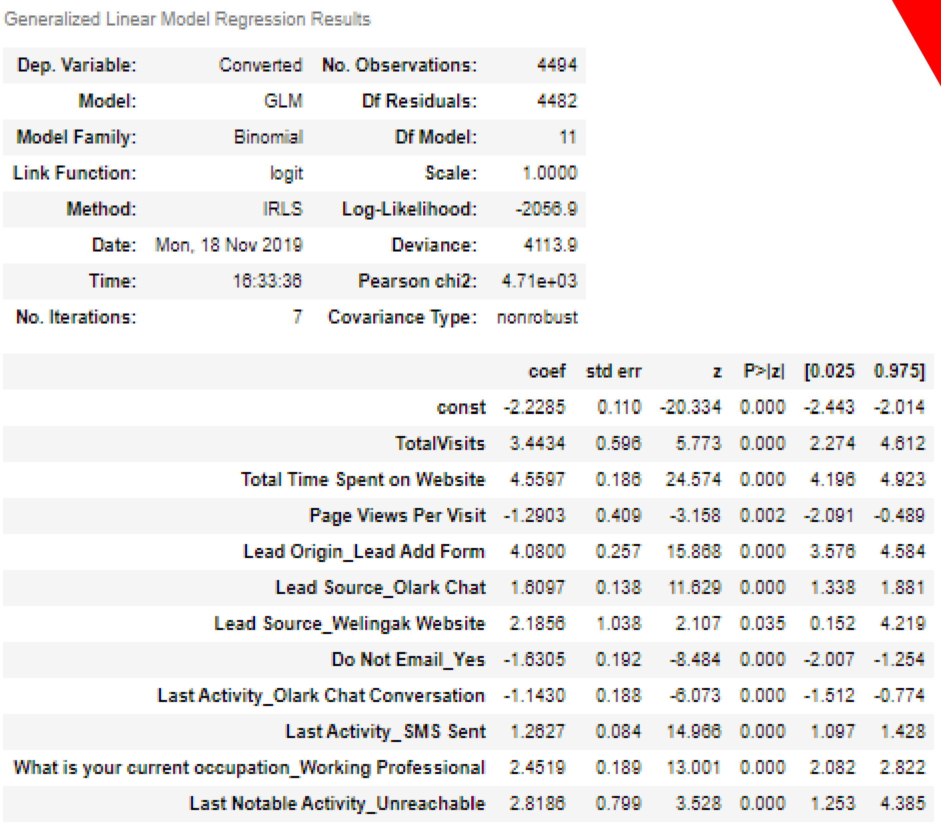
" Page Views Per Visit
TotalVisits
Total Time Spent on Website
Last Activity_SMS Sent
Lead Origin_Lead Add Form
Lead Source_Welingak Website
Lead Source_Olark Chat
Last Activity_Olark Chat Conversation
What is your current occupation_Working Professionals
Do Not Email_Yes
Last Notable Activity_Unreachable"

Approach

VIF and P values are indicated as below for these fields



	Features	VIF
2	Page Views Per Visit	4.06
0	TotalVisits	3.53
1	Total Time Spent on Website	2.00
8	Last Activity_SMS Sent	1.58
3	Lead Origin_Lead Add Form	1.49
5	Lead Source_Welingak Website	1.32
4	Lead Source_Olark Chat	1.22
7	Last Activity_Olark Chat Conversation	1.19
9	What is your current occupation_Working Profes...	1.19
6	Do Not Email_Yes	1.06
10	Last Notable Activity_Unreachable	1.01



```
Generalized Linear Model Regression Results
   Dep. Variable:          Converted    No. Observations:       4494
           Model:                GLM        Df Residuals:       4482
    Model Family:           Binomial            Df Model:         11
   Link Function:              logit               Scale:     1.0000
          Method:               IRLS      Log-Likelihood:    -2056.9
            Date:   Mon, 18 Nov 2019            Deviance:     4113.9
            Time:           16:33:36        Pearson chi2:   4.71e+03
  No. Iterations:                  7     Covariance Type:  nonrobust

                                                     coef  std err        z  P>|z|  [0.025  0.975]
                                          const   -2.2285    0.110  -20.334  0.000  -2.443  -2.014
                                    TotalVisits    3.4434    0.596    5.773  0.000   2.274   4.612
                    Total Time Spent on Website    4.5597    0.186   24.574  0.000   4.196   4.923
                           Page Views Per Visit   -1.2903    0.409   -3.158  0.002  -2.091  -0.489
                      Lead Origin_Lead Add Form    4.0800    0.257   15.868  0.000   3.576   4.584
                         Lead Source_Olark Chat    1.6097    0.138   11.629  0.000   1.338   1.881
                   Lead Source_Welingak Website    2.1856    1.038    2.107  0.035   0.152   4.219
                               Do Not Email_Yes   -1.6305    0.192   -8.484  0.000  -2.007  -1.254
          Last Activity_Olark Chat Conversation   -1.1430    0.188   -6.073  0.000  -1.512  -0.774
                         Last Activity_SMS Sent    1.2627    0.084   14.966  0.000   1.097   1.428
What is your current occupation_Working Professional    2.4519    0.189   13.001  0.000   2.082   2.822
              Last Notable Activity_Unreachable    2.8186    0.799    3.528  0.000   1.253   4.385
```

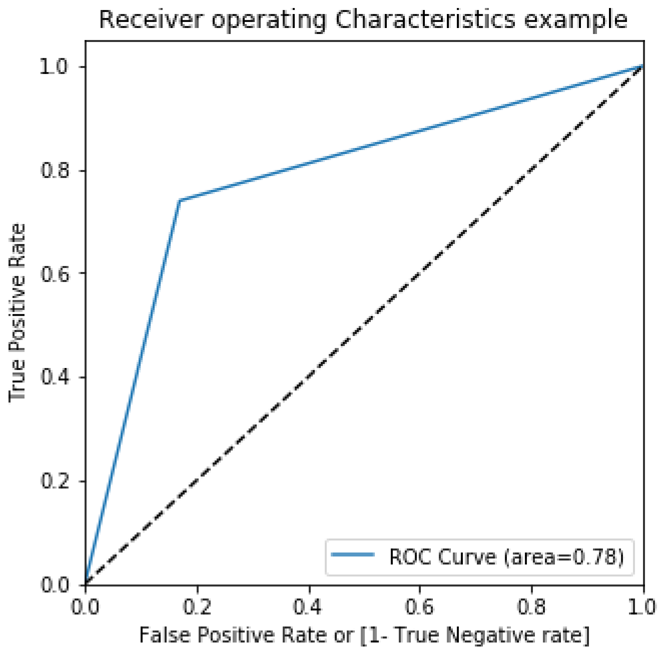
Approach

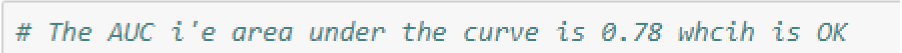
We used below steps during prediction of Leads into HOT Leads

9> We did Model evaluation using onto the train dataset with conversion rate as 0.5

10> Later we used Confusion matix to find the optimal cut-off to decide which leads can be created as Hot Leads

11> Created the ROC curve and value is coming as 0.78 for AUC which is good to start



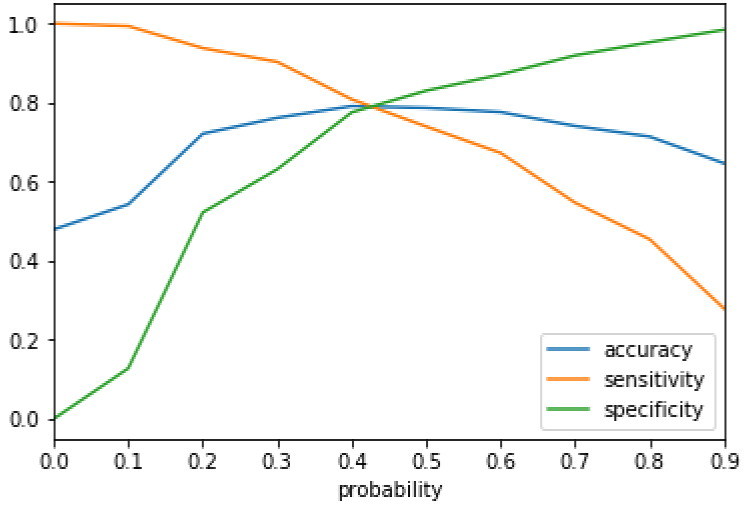


```
# The AUC i'e area under the curve is 0.78 whcih is OK
```

Approach

We used below steps during prediction of Leads into HOT Leads

9> Using Confusion matix we found that the cut-off values could be 0.42 as shown into the figure below



10> Pedicting onto the dataset with cutoff value 0.44 and nearby values to decide optimum values for Precison and Recall, performing few steps more with nearby values, we found cutoff values 0.46 is optimum

with Precision values = 0.78
and Recall = 0.77

Suggestion

As we saw these fields are used to decide prediction parameter for these dataset

" Page Views Per Visit
TotalVisits
Total Time Spent on Website
Last Activity_SMS Sent
Lead Origin_Lead Add Form
Lead Source_Welingak Website
Lead Source_Olark Chat
Last Activity_Olark Chat Conversation
What is your current occupation_Working Professionals
Do Not Email_Yes
Last Notable Activity_Unreachable"

Sales team focus on those candidate who provide to who spent more time on Website; are Professionals