Lead Score Case Study

Authors

Seethalakshmi & Arun Vikram singh

Problem Statement

AN EDUCATION COMPANY NAMED X EDUCATION SELLS ONLINE COURSES TO INDUSTRY PROFESSIONALS. ON ANY GIVEN DAY, MANY PROFESSIONALS WHO ARE INTERESTED IN THE COURSES LAND ON THEIR WEBSITE AND BROWSE FOR COURSES.

THE COMPANY MARKETS ITS COURSES ON SEVERAL WEBSITES AND SEARCH ENGINES LIKE GOOGLE. ONCE THESE PEOPLE LAND ON THE WEBSITE, THEY MIGHT BROWSE THE COURSES OR FILL UP A FORM FOR THE COURSE OR WATCH SOME VIDEOS. WHEN THESE PEOPLE FILL UP A FORM PROVIDING THEIR EMAIL ADDRESS OR PHONE NUMBER, THEY ARE CLASSIFIED TO BE A LEAD. MOREOVER, THE COMPANY ALSO GETS LEADS THROUGH PAST REFERRALS. ONCE THESE LEADS ARE ACQUIRED, EMPLOYEES FROM THE SALES TEAM START MAKING CALLS, WRITING EMAILS, ETC. THROUGH THIS PROCESS, SOME OF THE LEADS GET CONVERTED WHILE MOST DO NOT. THE TYPICAL LEAD CONVERSION RATE AT X EDUCATION IS AROUND 30%.

Business Goals

There are quite a few goals for this case study.

- ✓ Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- ✓ A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- ✓ There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

APPROACH

- ➢ *Reading the Data*
- ➢ *Data Cleaning & Imputing Missing Values*
- ➢ *Exploratory Data Analysis –Univariate, Bivariate, Multivariate*
- ➢ *Feature Scaling and Dummy Variable Creation*
- ➢ *Model Building (Logistic Regression)*
- ➢ *Model Evaluation*
- ➢ *Conclusions and Recommendations*

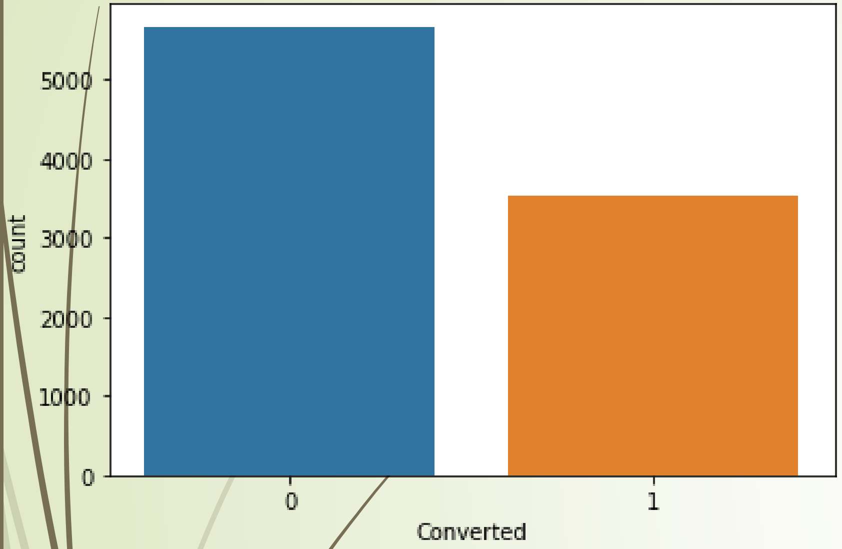
Reading The Data

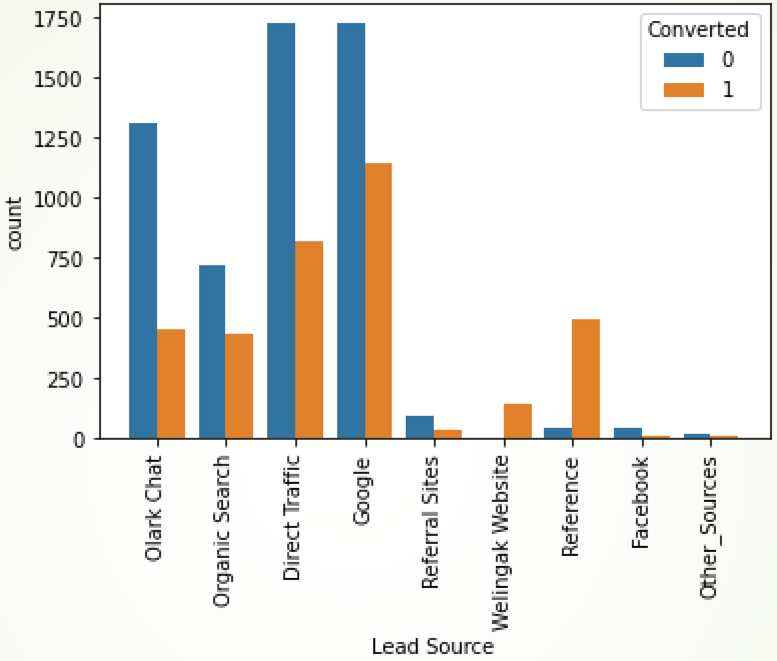
- *Import the Dataset*
- *Import Libraries*
- *Using describe() and info() function to check the data types and null value counts*

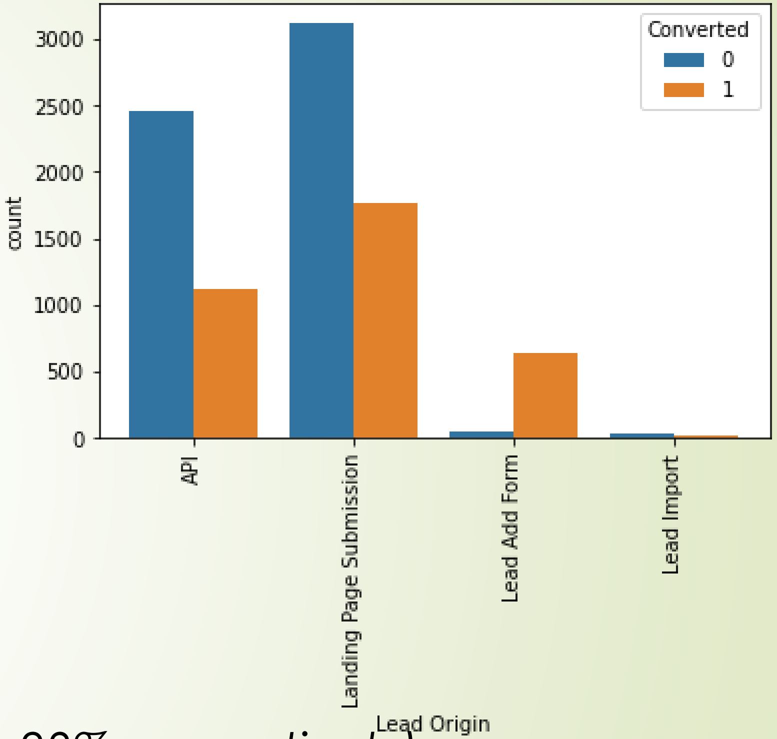
Data Cleaning & Imputing Missing Values

- *CONVERTING THE VALUES YES/NO TO 1/0s*
- *CONVERTING ‘SELECT’ TO NANs*
- *DROPPING THE COLUMNS HAVING MORE THAN 70% NULL VALUES*
- *DROPPING UNNECESSARY COLUMNS*
- *IMPUTING NULL VALUES WITH MEANS/MODE VALUES*

EXPLORATORY DATA ANALYSIS



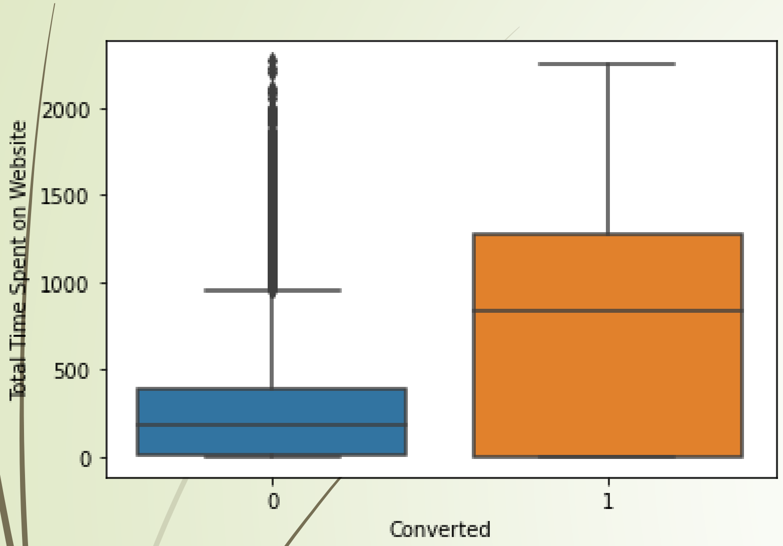




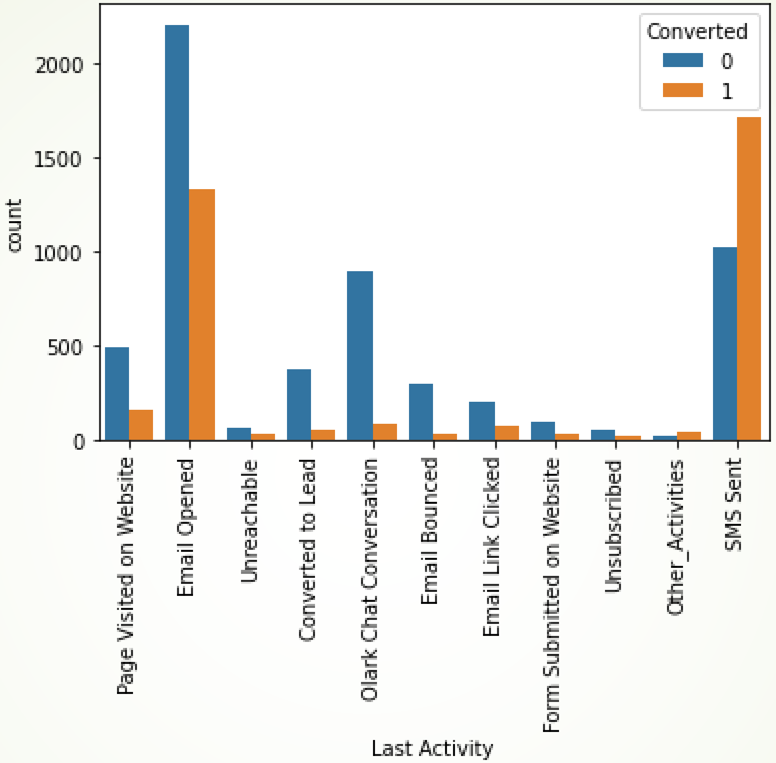
There is overall 38% conversion rate

- ➢ *Google & Direct Traffic have highest conversion rates (32% & 23% respectively)*
- ➢ *Reference , Olark Chat, Organic Search and Welingak Website have less counts but conversion rates are high.*
- ➢ *So we have to focus on Google, Direct Traffic,Reference , Olark Chat, Organic Search and Welingak Website to increase the conversion rates*
- ➢ *Landing Page Submission & API have conversion rate more than 30 % (50% & 31% respectively).*
- ➢ *Lead Add Form has low counts but conversion rate in high as compare to counts.*

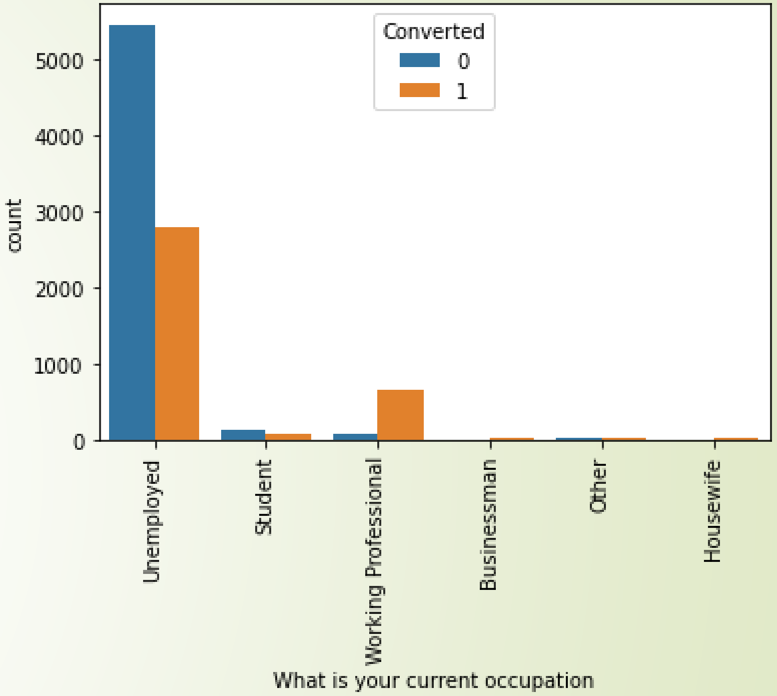
EXPLORATORY DATA ANALYSIS



There is significant increase in conversion rate with Total Time Spent on Website.
So website should be designed to be more attractive and informative , so that people spend more time and get more information and this will also help in increasing the conversion rate

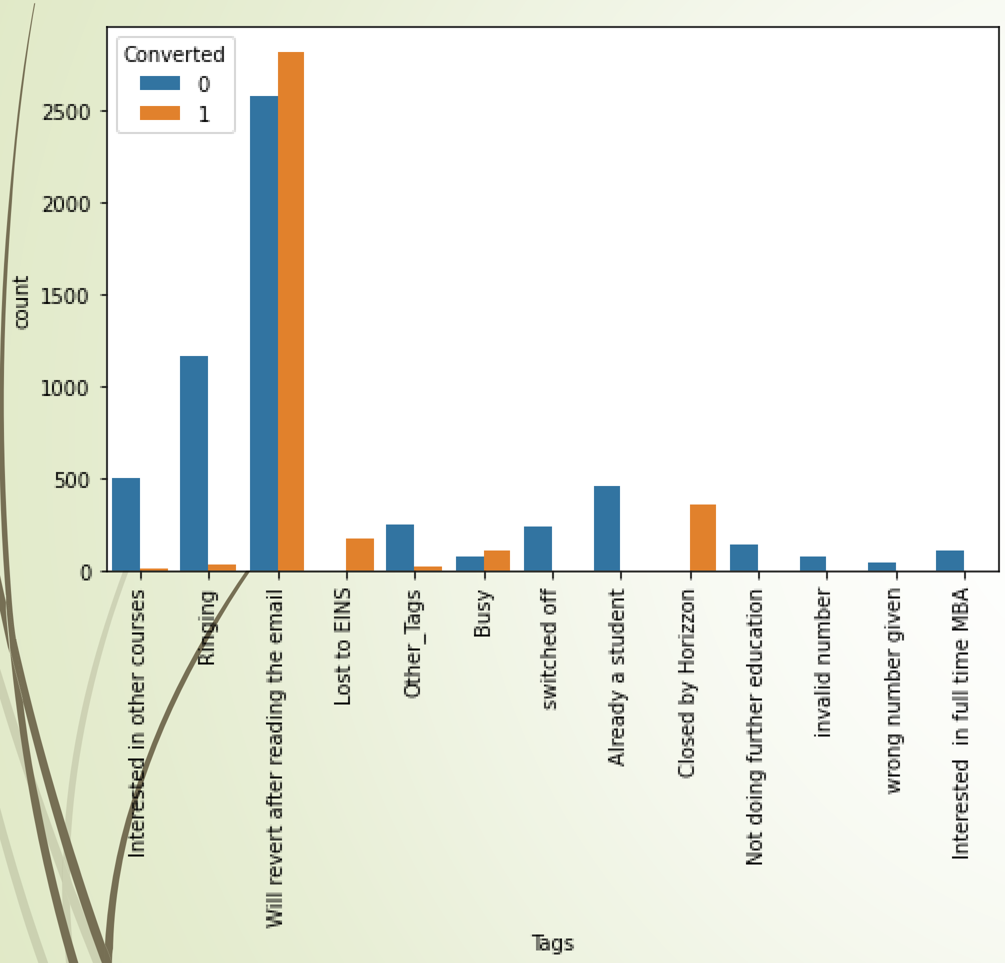


SMS Sent & Email Opened have the maximum conversion rates (48% and 37% respectively)



Working Professionals have high conversion rate.
2) 89% of Total leads are coming from Unemployed and more than 78% of Total conversions are from Unemployed.

EXPLORATORY DATA ANALYSIS



'Will revert after reading the email' & 'Ringing' have the highest conversion rates (79% and 10% respectively)

<u>Overall Summary from Exploratory Data Analysis</u>

1) Google & Direct Traffic have highest conversion rates (32% & 23% respectively).Reference , Olark Chat, Organic Search and Welingak Website have less counts but conversion rates are high. So we need to focus on these Lead Sources to improve the conversion rate.

2) Landing Page Submission & API have conversion rate more than 30 % (50% & 31% respectively). Lead Add Form has low counts but conversion rate in high as compare to counts. So we have to focus on Landing Page Submission & API & Lead Add Form for increase in conversion rates.

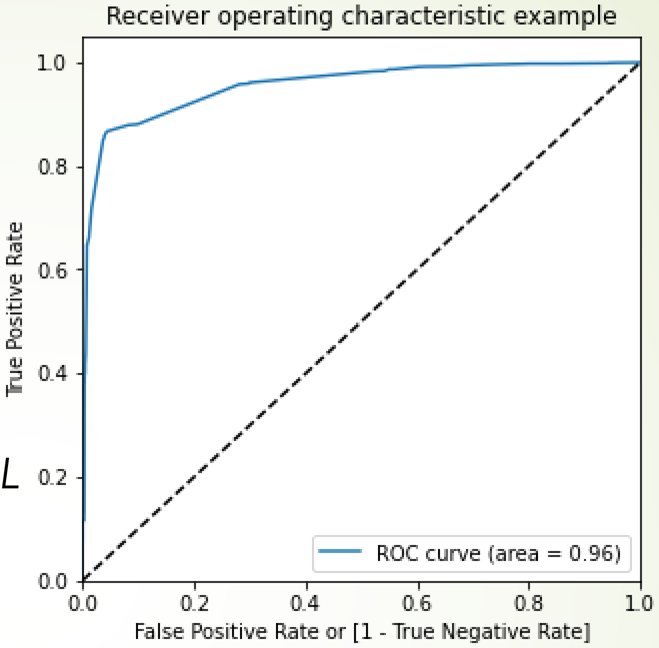
3) SMS Sent & Email Opened have the maximum conversion rates (48% and 37% respectively). So we have to focus to increase conversion rate from SMS Sent & Email Opened.

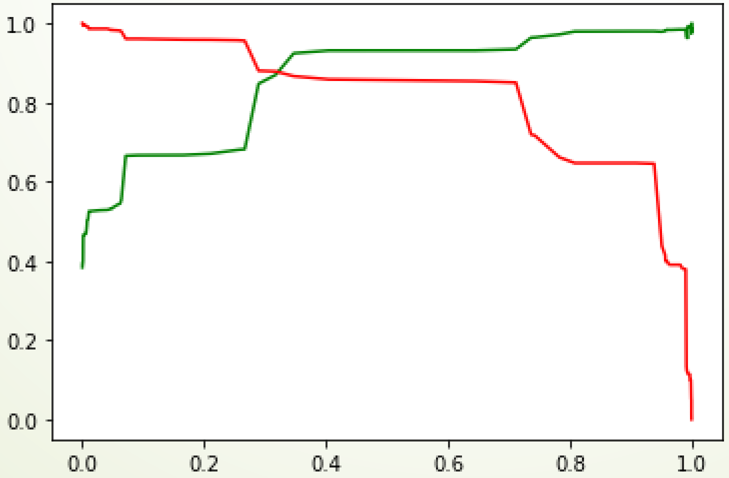
4) Working Professionals have high conversion rate.89% of Total leads are coming from Unemployed and more than 78% of Total conversions are from Unemployed. So we have to focus on Working Professionals and Unemployed to increase the conversion rate.

MODEL BUILDING

Steps for Model Buliding

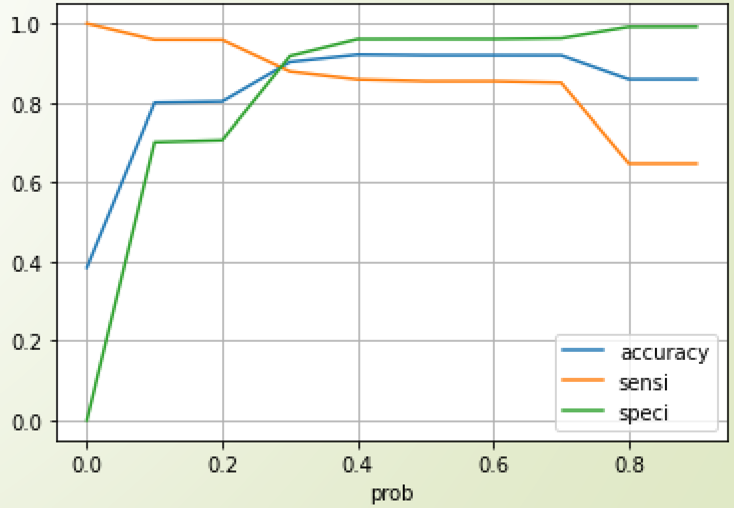
1) *SPLITTING THE DATA INTO TRAIN AND TEST SETS*
2) *TRAIN-TEST SPLIT RATIO=70:30*
3) *USING RFE (WITH 15 VARIABLES)*
4) *USING VIF TO CHECK HIGH VIF VALUE AND STATSMODEL.API TO CHECK HIGH P-VALUE IN MODEL*
5) *PREDICTION ON TRAIN DATA SET*
6) *PREDICTION ON TEST DATA SET*





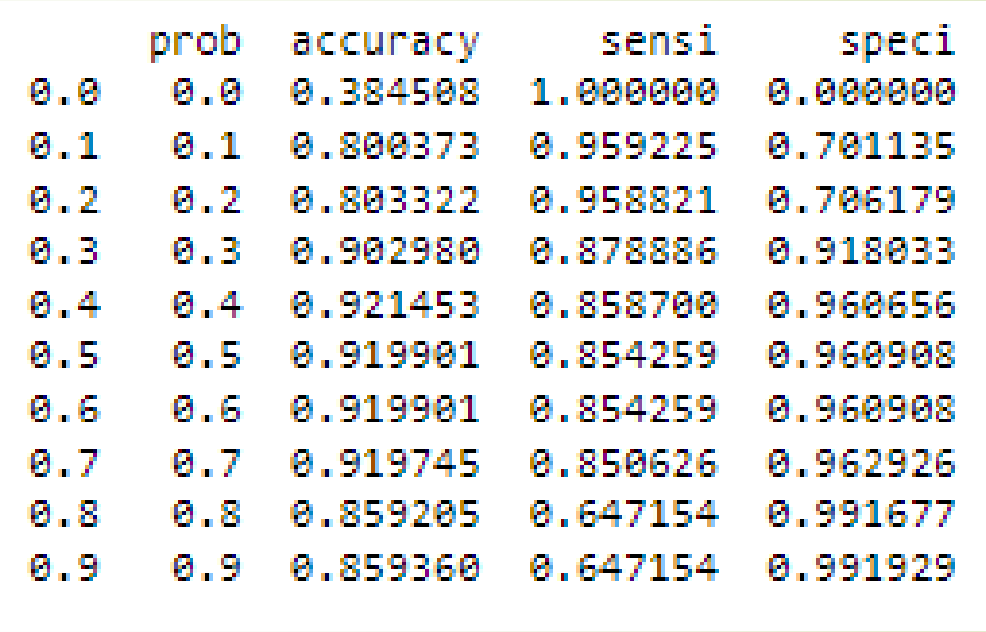
PRECISION RECALL CURVE

CUT OFF THRESHOLD=0.28



MODEL EVALUATION

1) *ACCURACY, SENSITIVITY AND SPECITIVITY ARE CALCULATED FOR VARIOUS PROBABILITY CUT OFFS*
2) *WE CAN SEE THAT OPTIMAL POINT IS 0.28*



```
     prob  accuracy     sensi     speci
0.0   0.0  0.384508  1.000000  0.000000
0.1   0.1  0.800373  0.959225  0.701135
0.2   0.2  0.803322  0.958821  0.706179
0.3   0.3  0.902980  0.878886  0.918033
0.4   0.4  0.921453  0.858700  0.960656
0.5   0.5  0.919901  0.854259  0.960908
0.6   0.6  0.919901  0.854259  0.960908
0.7   0.7  0.919745  0.850626  0.962926
0.8   0.8  0.859205  0.647154  0.991677
0.9   0.9  0.859360  0.647154  0.991929
```

FOR TRAIN DATA SET

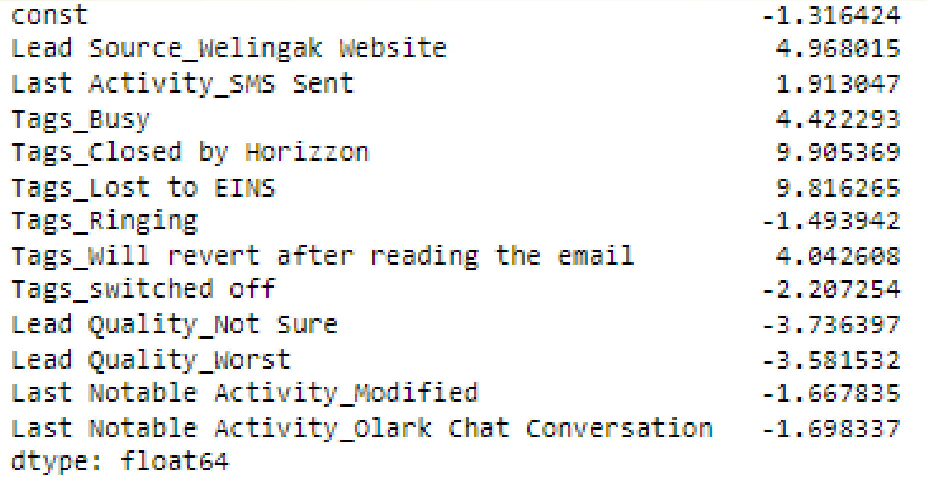
ATTRIBUTES	PERCENTAGES
ACCURACY	*89.32%*
PRECISION	*84.76%*
SENSITIVITY	*88.05%*
SPECIFICITY	*90.11%*

CONFUSUION MATRIX ON TRAIN DATA SET

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	*3573*	*392*
CONVERTED	*296*	*2181*

MODEL PREDICTION

TOP FEATURES IN MODEL



```
const                                             -1.316424
Lead Source_Welingak Website                       4.968015
Last Activity_SMS Sent                             1.913047
Tags_Busy                                          4.422293
Tags_Closed by Horizzon                            9.905369
Tags_Lost to EINS                                  9.816265
Tags_Ringing                                      -1.493942
Tags_Will revert after reading the email           4.042608
Tags_switched off                                 -2.207254
Lead Quality_Not Sure                             -3.736397
Lead Quality_Worst                                -3.581532
Last Notable Activity_Modified                    -1.667835
Last Notable Activity_Olark Chat Conversation     -1.698337
dtype: float64
```

CONFUSUION MATRIX ON TEST DATA SET

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	*1540*	*167*
CONVERTED	*131*	*924*

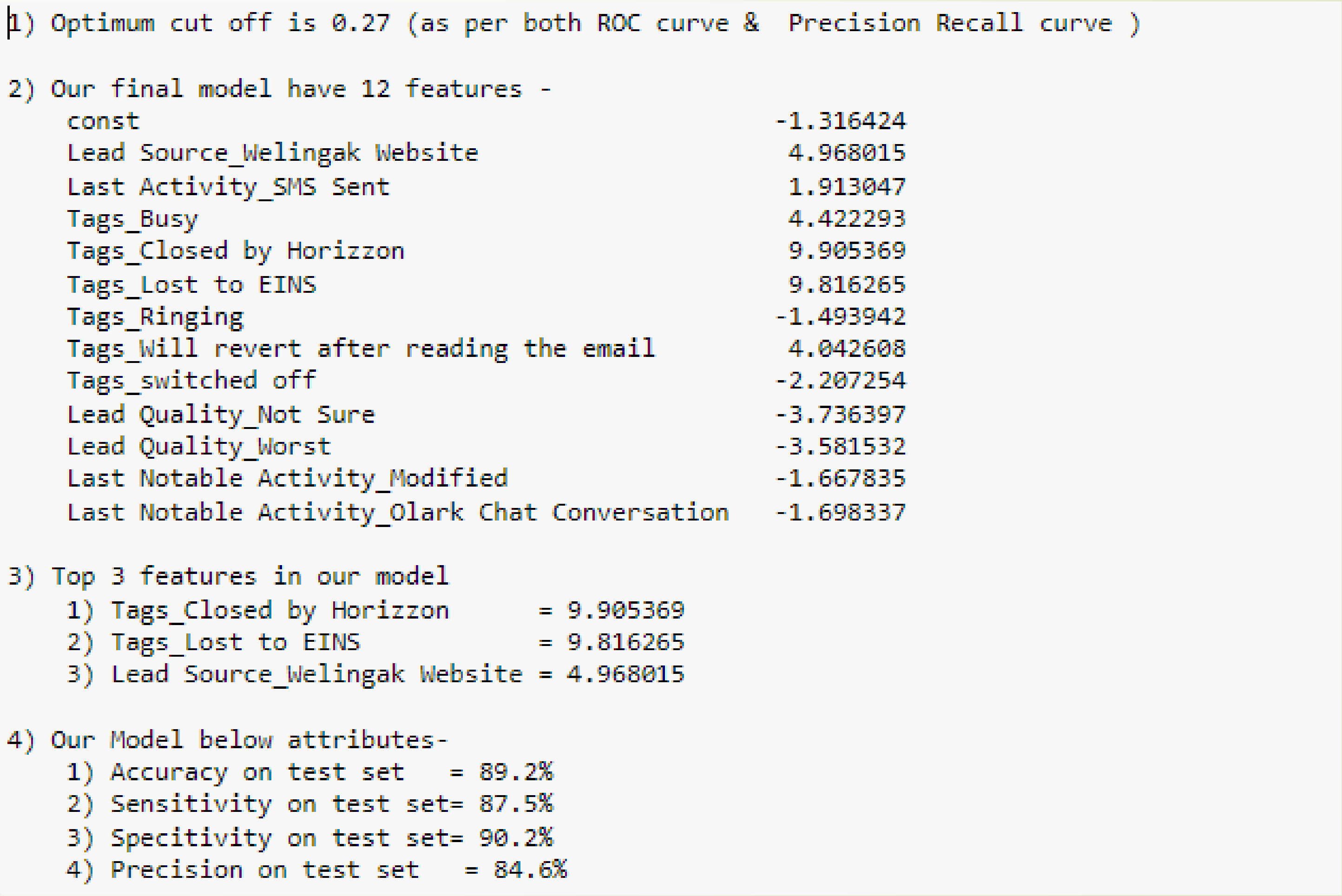
FOR TEST DATA SET

ATTRIBUTES	PERCENTAGES
ACCURACY	*89.21%*
PRECISION	*84.69%*
SENSITIVITY	*87.58%*
SPECIFICITY	*90.21%*

CONCLUSION

OUR MODEL HAS PREDICTED 88.05% FOR TRAIN DATA SET AND FOR TEST DATA SET IT HAS PREDICTED 87.58%

SO OUR MODEL SHOWING GOOD RESULT



```
1) Optimum cut off is 0.27 (as per both ROC curve &  Precision Recall curve )

2) Our final model have 12 features -
    const                                             -1.316424
    Lead Source_Welingak Website                       4.968015
    Last Activity_SMS Sent                             1.913047
    Tags_Busy                                          4.422293
    Tags_Closed by Horizzon                            9.905369
    Tags_Lost to EINS                                  9.816265
    Tags_Ringing                                      -1.493942
    Tags_Will revert after reading the email           4.042608
    Tags_switched off                                 -2.207254
    Lead Quality_Not Sure                             -3.736397
    Lead Quality_Worst                                -3.581532
    Last Notable Activity_Modified                    -1.667835
    Last Notable Activity_Olark Chat Conversation     -1.698337

3) Top 3 features in our model
    1) Tags_Closed by Horizzon      = 9.905369
    2) Tags_Lost to EINS            = 9.816265
    3) Lead Source_Welingak Website = 4.968015

4) Our Model below attributes-
    1) Accuracy on test set    = 89.2%
    2) Sensitivity on test set= 87.5%
    3) Specitivity on test set= 90.2%
    4) Precision on test set    = 84.6%
```