Lead Scoring Case Study

Help X Education to target potential leads

DATA EXPLORATION

This dataset has 2 files as explained below:

1. *'Leads.csv'* leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. The target variable, in this case, 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. 'Leads Data Dictionary.xlsx' is data dictionary which describes each attribute in *Leads.csv* dataset.

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.

- Although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

Analysis Overall Approach

- Importing the Data (Leads.csv) file

- Understanding the Data set
  - Handle Missing Values
  - Drop Non significant columns

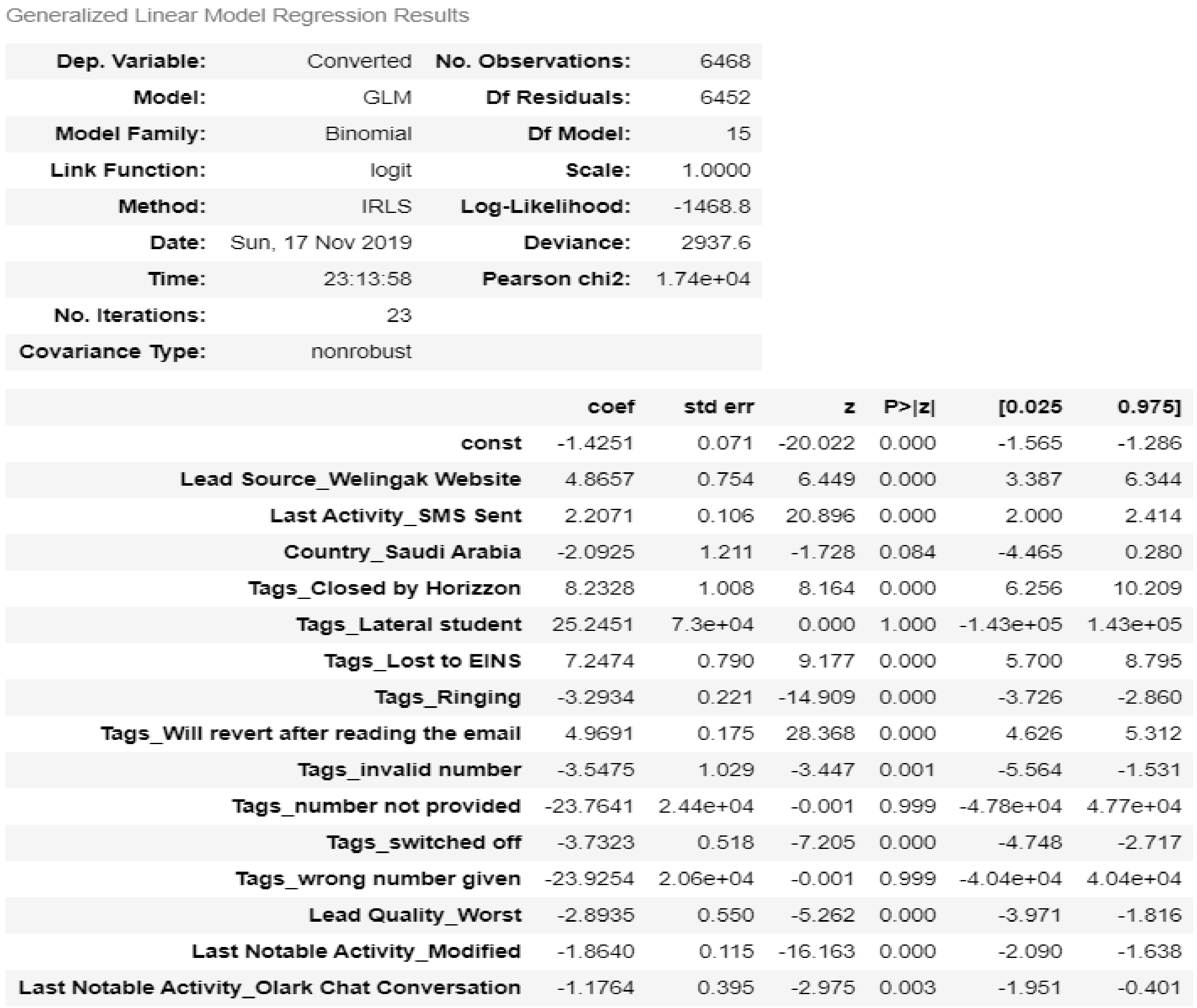
- Scaling the data set
  - Generated Dummy Variables for all categorical columns
  - Scaled data using StandardScaler

- Modelling
  - Split the data into Test and Train
  - Logistic Regression model
  - Feature Selection using RFE
  - Choose Right Variable for model

Analysis Approach – Handling Data

- Converting the values “Select” to nan values
- Dropped columns with high missing values . Dropped columns “'How did you hear about X Education' ,‘Lead Profile’, 'Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Activity Score','Asymmetrique Profile Score‘ since these have >30% missing values and we couldn’t find anything suitable to handle the missing values
- Handled Missing Values
  - Replaced nan with “Not Sure” in ‘Lead Quality’ column
  - Replaced nan with “Mumbai” in ‘City’ column
  - Replaced nan with “Others” in ‘Specification’ column
  - Replaced nan with “Other issues” in ‘Tags’ column
  - Replaced nan with “Better Career Prospects“ in ‘What matters most to you in choosing a course’ column
  - Replaced nan with “Unemployed” in ‘What is your current occupation’ column
  - Replaced nan with “India” in ‘Country’ column
  - Replaced nan with median in ‘Page Views Per Visit’ column
  - Replaced nan with “Email Opened” in ‘Last Activity’
  - Replaced nan with “Google” in ‘Lead Source’
  - Dropped columns 'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content‘ and 'I agree to pay the amount through cheque‘ since these columns has all values as No and it does not help in our analysis

Analysis Approach – Critical Decisions: **Feature Selection Using RFE**

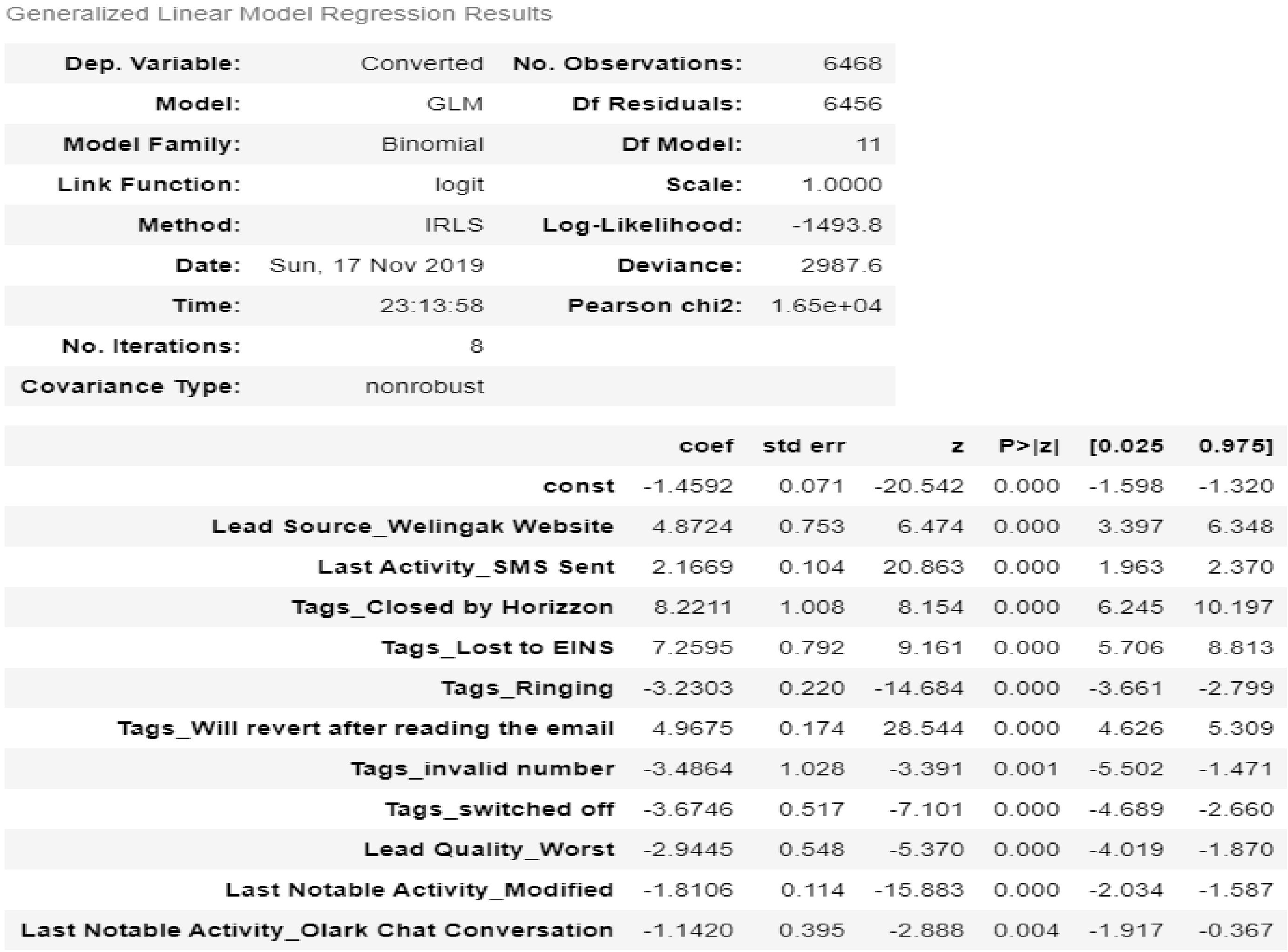


```
Generalized Linear Model Regression Results

       Dep. Variable:          Converted   No. Observations:       6468
               Model:                GLM       Df Residuals:       6452
        Model Family:           Binomial           Df Model:         15
       Link Function:              logit              Scale:     1.0000
              Method:               IRLS     Log-Likelihood:    -1468.8
                Date:   Sun, 17 Nov 2019           Deviance:     2937.6
                Time:           23:13:58      Pearson chi2:    1.74e+04
      No. Iterations:                 23
     Covariance Type:          nonrobust

                                              coef    std err         z   P>|z|      [0.025     0.975]
                                    const  -1.4251      0.071   -20.022   0.000      -1.565     -1.286
              Lead Source_Welingak Website   4.8657      0.754     6.449   0.000       3.387      6.344
                    Last Activity_SMS Sent   2.2071      0.106    20.896   0.000       2.000      2.414
                     Country_Saudi Arabia  -2.0925      1.211    -1.728   0.084      -4.465      0.280
                    Tags_Closed by Horizzon  8.2328      1.008     8.164   0.000       6.256     10.209
                     Tags_Lateral student  25.2451    7.3e+04     0.000   1.000   -1.43e+05   1.43e+05
                        Tags_Lost to EINS   7.2474      0.790     9.177   0.000       5.700      8.795
                             Tags_Ringing  -3.2934      0.221   -14.909   0.000      -3.726     -2.860
         Tags_Will revert after reading the email   4.9691      0.175    28.368   0.000       4.626      5.312
                      Tags_invalid number  -3.5475      1.029    -3.447   0.001      -5.564     -1.531
                 Tags_number not provided -23.7641   2.44e+04    -0.001   0.999   -4.78e+04   4.77e+04
                        Tags_switched off  -3.7323      0.518    -7.205   0.000      -4.748     -2.717
                  Tags_wrong number given -23.9254   2.06e+04    -0.001   0.999   -4.04e+04   4.04e+04
                      Lead Quality_Worst  -2.8935      0.550    -5.262   0.000      -3.971     -1.816
            Last Notable Activity_Modified  -1.8640      0.115   -16.163   0.000      -2.090     -1.638
  Last Notable Activity_Olark Chat Conversation  -1.1764      0.395    -2.975   0.003      -1.951     -0.401
```

- Identified top 15 variables which are of significant importance

- We will drop variable one by one so that P value is close to zero

Analysis Approach – Critical Decisions – Finalizing Variables



Generalized Linear Model Regression Results

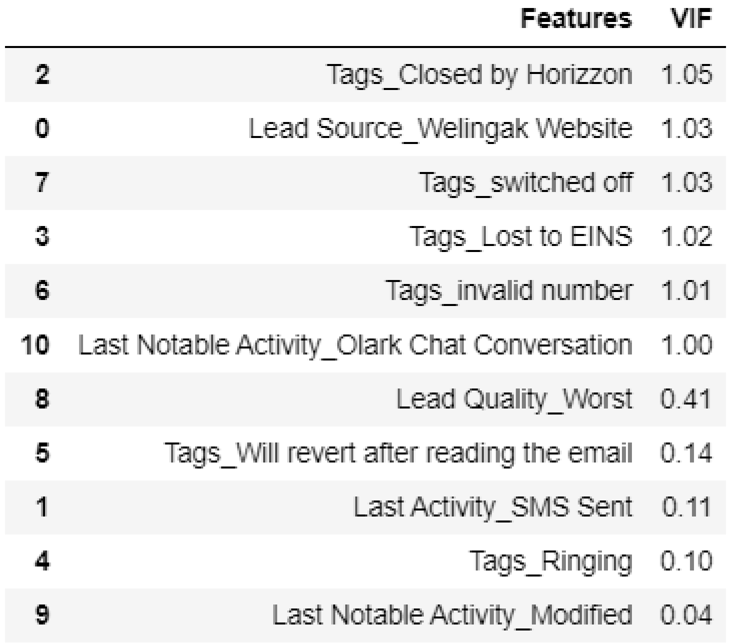
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6456
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1493.8
Date:	Sun, 17 Nov 2019	**Deviance:**	2987.6
Time:	23:13:58	**Pearson chi2:**	1.65e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-1.4592	0.071	-20.542	0.000	-1.598	-1.320
Lead Source_Welingak Website	4.8724	0.753	6.474	0.000	3.397	6.348
Last Activity_SMS Sent	2.1669	0.104	20.863	0.000	1.963	2.370
Tags_Closed by Horizzon	8.2211	1.008	8.154	0.000	6.245	10.197
Tags_Lost to EINS	7.2595	0.792	9.161	0.000	5.706	8.813
Tags_Ringing	-3.2303	0.220	-14.684	0.000	-3.661	-2.799
Tags_Will revert after reading the email	4.9675	0.174	28.544	0.000	4.626	5.309
Tags_invalid number	-3.4864	1.028	-3.391	0.001	-5.502	-1.471
Tags_switched off	-3.6746	0.517	-7.101	0.000	-4.689	-2.660
Lead Quality_Worst	-2.9445	0.548	-5.370	0.000	-4.019	-1.870
Last Notable Activity_Modified	-1.8106	0.114	-15.883	0.000	-2.034	-1.587
Last Notable Activity_Olark Chat Conversation	-1.1420	0.395	-2.888	0.004	-1.917	-0.367

- Started out with 15 variables and by referring to P values, dropped variables with high P value 1 by 1

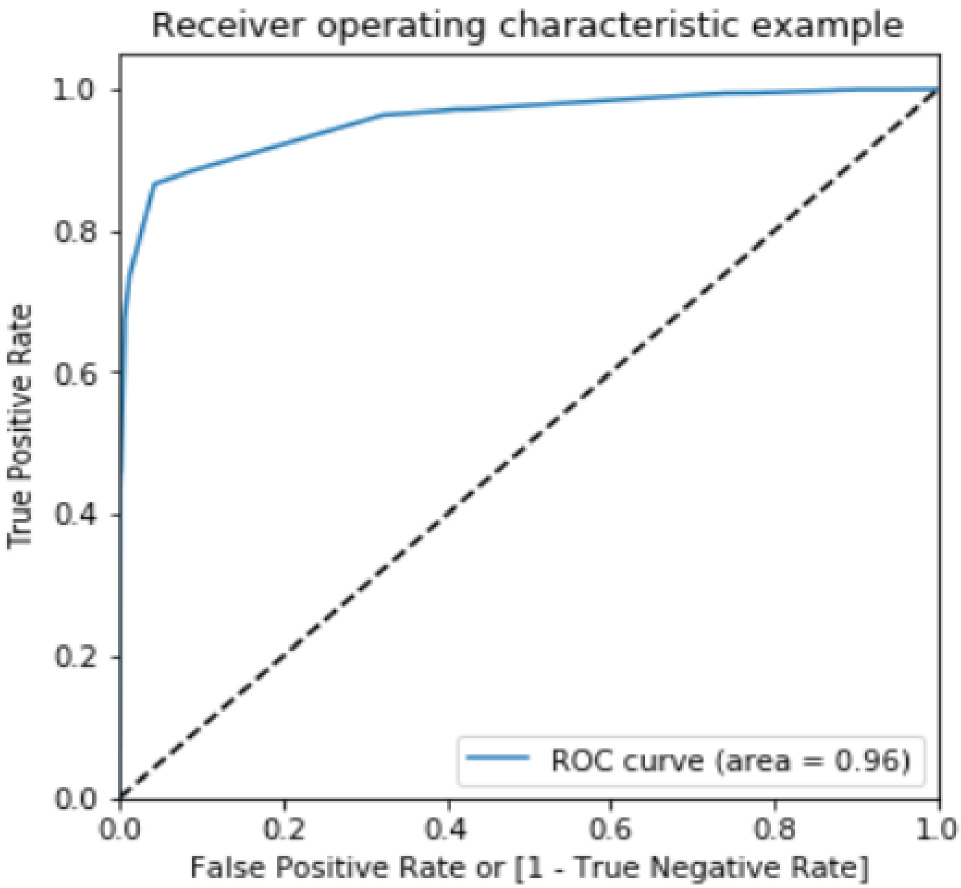
- Finalized 11 variables

- VIF is also under 5 for these variables



	Features	**VIF**
2	Tags_Closed by Horizzon	1.05
0	Lead Source_Welingak Website	1.03
7	Tags_switched off	1.03
3	Tags_Lost to EINS	1.02
6	Tags_invalid number	1.01
10	Last Notable Activity_Olark Chat Conversation	1.00
8	Lead Quality_Worst	0.41
5	Tags_Will revert after reading the email	0.14
1	Last Activity_SMS Sent	0.11
4	Tags_Ringing	0.10
9	Last Notable Activity_Modified	0.04

Analysis Approach – Critical Decisions – ROC Curve

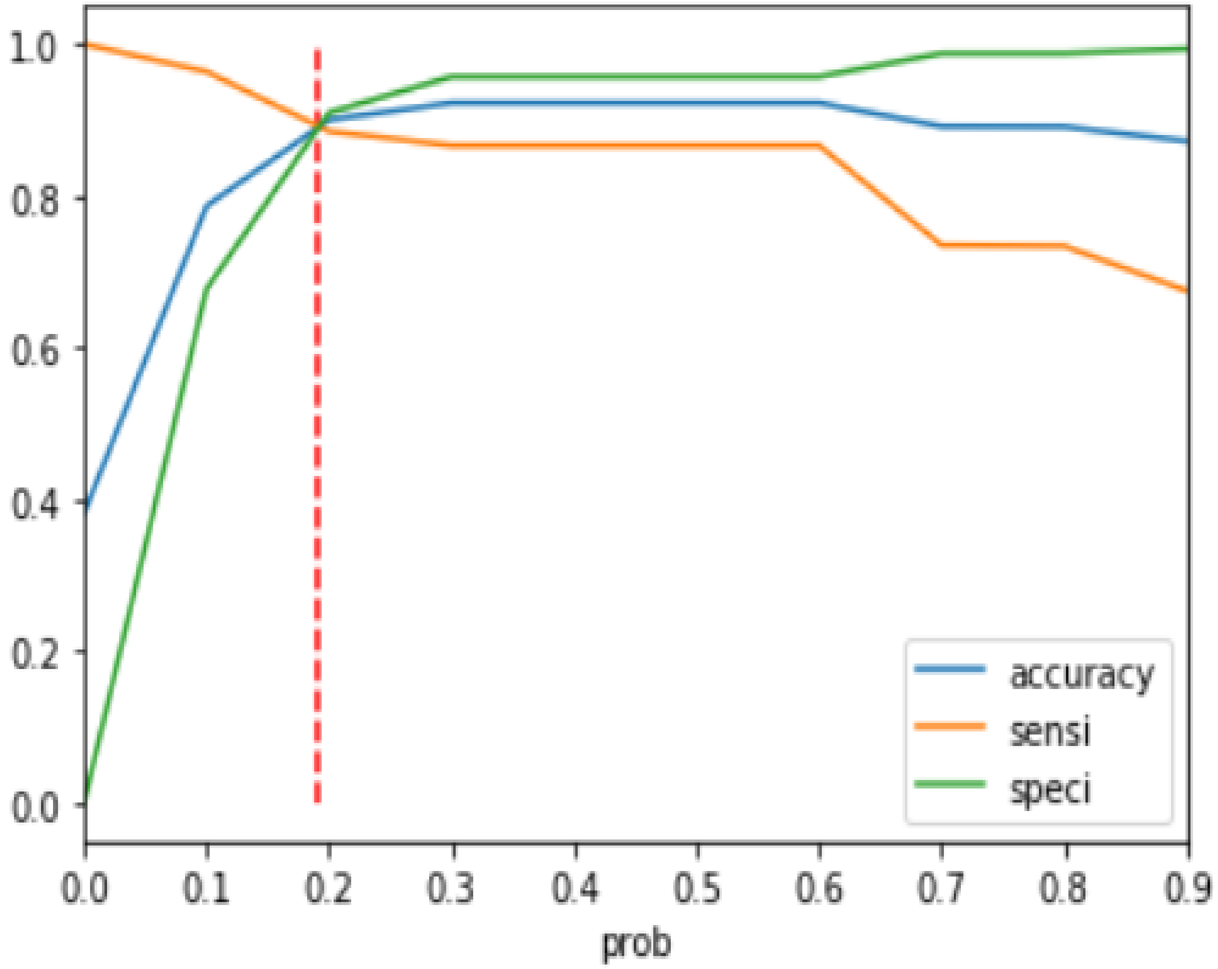


<u>ROC Curve</u>

- ROC curve is more towards the upper-left corner of the graph, it means that the model is very good

- Area under the ROC curve is 0.96 which detonates it’s a good model

Analysis Approach – Critical Decisions – Finding Optimal Cut Off



- Based on Accuracy, Sensitivity and Specificity, we have considered the cut off probability as 0.19. it’s a point where all three metrics meet
- Accuracy @ 90%. Accuracy is expected to measure how well the test predicts both categories
- Sensitivity @ 88%. Sensitivity indicates, how well the test predicts one category(True Positives or True Leads)
- Specificity @ 91%. Specificity measures how well the test predicts the other category(True Negatives)
- Precision @ 93% and Recall @ 87%

Based on the business needs we need to change the cut off. If business goal is to predict the true leads then we should change the prob cut off close to 0. As you could see in the graph orange line i.e sensitivity gradually drops as prob increases that’s why the cut off should be close to 0

Analysis Approach – Critical Decisions – Model Results & Conclusion

Variable	Coef.
Tags_Closed by Horizzon	8.22
Tags_Lost to EINS	7.26
Tags_Will revert after reading the email	4.97
Lead Source_Welingak Website	4.87
Last Activity_SMS Sent	2.17
Last Notable Activity_Olark Chat Conversation	-1.14
Last Notable Activity_Modified	-1.81
Lead Quality_Worst	-2.94
Tags_Ringing	-3.23
Tags_invalid number	-3.49
Tags_switched off	-3.67

Focus Areas:

- Tags : Ringing, Invalid Number and Switched off are having –ve coefficient. If we could restrict invalid phone number entry in forms through mobile number verification will help and also capturing alternate number will help

- Advertise more in Welingak Website

- 11 variables which are of high significance in determining a lead customer

- Left side table detonates the how significant the variable is (listed in descending order)

- Sensitivity @ 88%. Sensitivity indicates, how well the test predicts one category(True Positives)

- Specificity @ 95%. Specificity measures how well the test predicts the other category(True Negatives)

Thank You

Prepared by

- Dayanand Sagar Kukkala
- A Anto Selvanathan