LEAD SCORING CASE STUDY

GAUTAM KODUKULA
NAVDEEP R JAGATHKARI

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.

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.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. If they successfully identify this set of leads, the lead conversion rate should go up.

The company requires you to build a model to identify the leads of higher conversion rate and lower conversion rate.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Problem Approach

As a Data Scientist, our job is to analyze the data available and help out the CEO of the company to achieve profits via targeting which leads are to be concentrated to increase conversion percentage.

Since this is a Machine Learning Model.

Data is readily available.

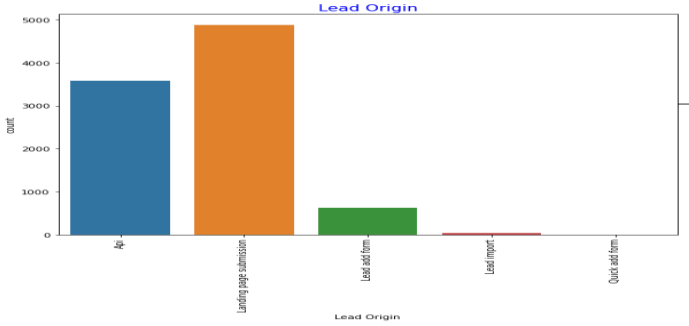
Following are the steps to approach:

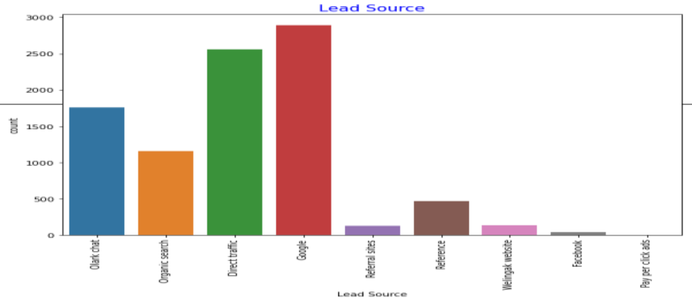
1. Read and Clean the data
2. Visualization and Outlier treatment
3. Data Modelling
4. Data Training
5. Evaluate the model
6. Inference

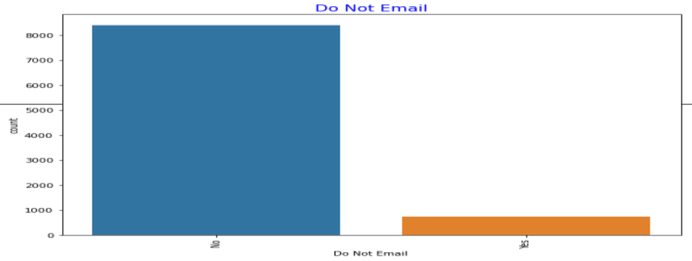
EDA

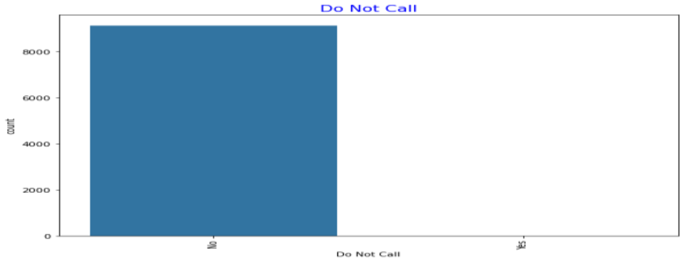
- ✓The Data set is huge and has more variables to be looked at.
- ✓After understanding the variables, Deleted the variables which have high null values(>37%).
- ✓Visualized the Variables based on type of data(Categorical and Numerical).
- ✓Since, we have only 3 numerical variables (after dropping) which contains the outliers and outlier treatment in this data set could lead to loss of data which can harm the further analysis. So skipped outlier treatment.

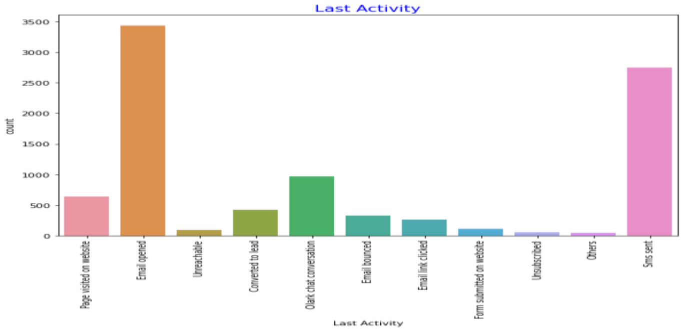
Count vs Categorical Colums

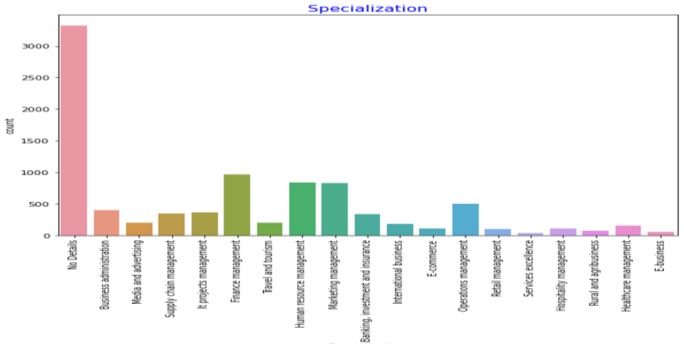


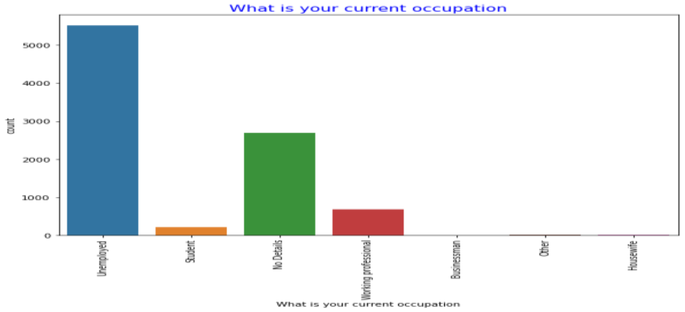


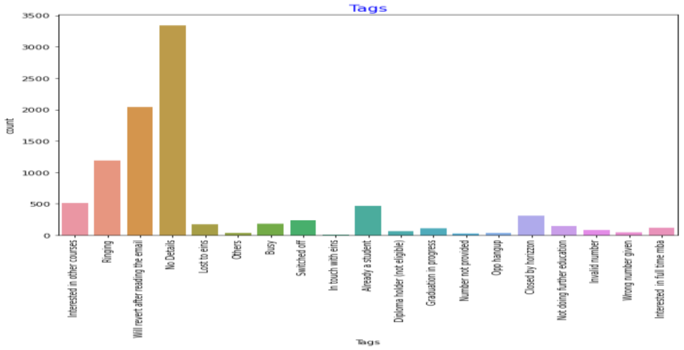


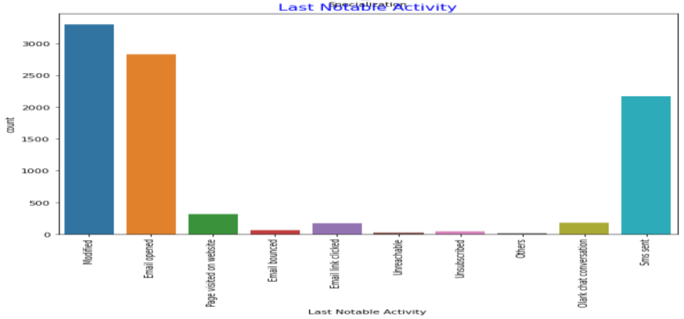




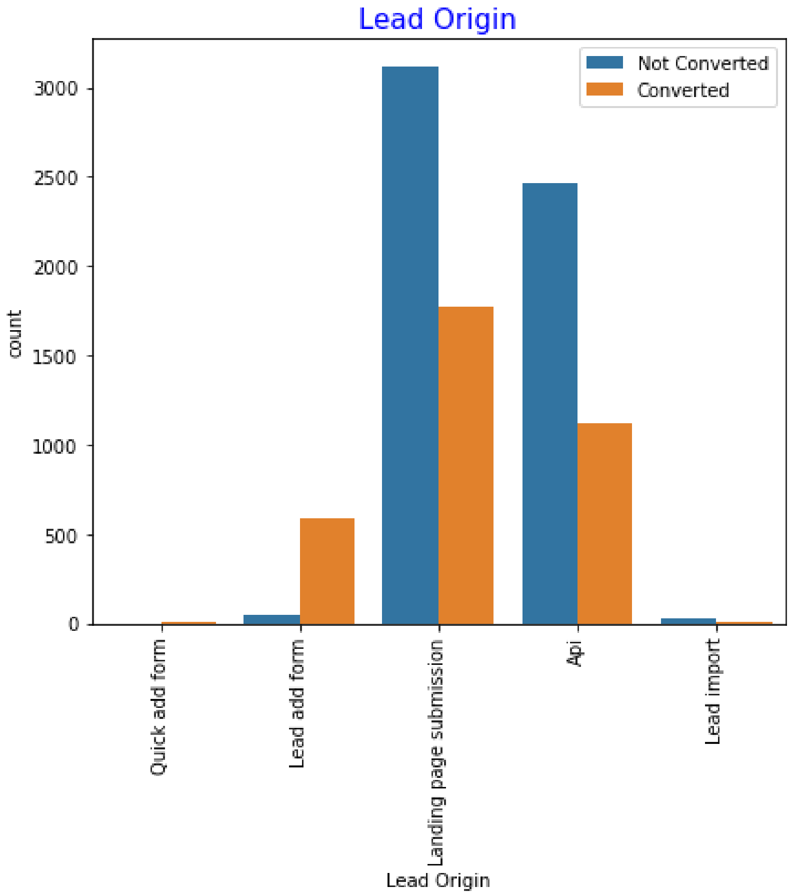




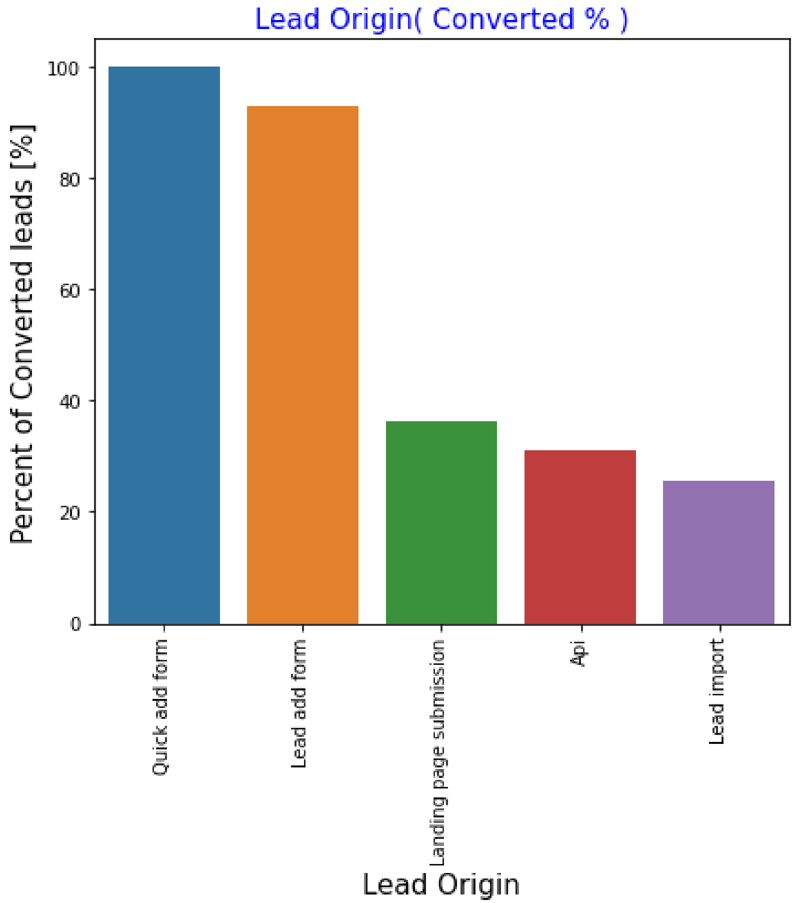




Lead Origin (converted vs not-converted)



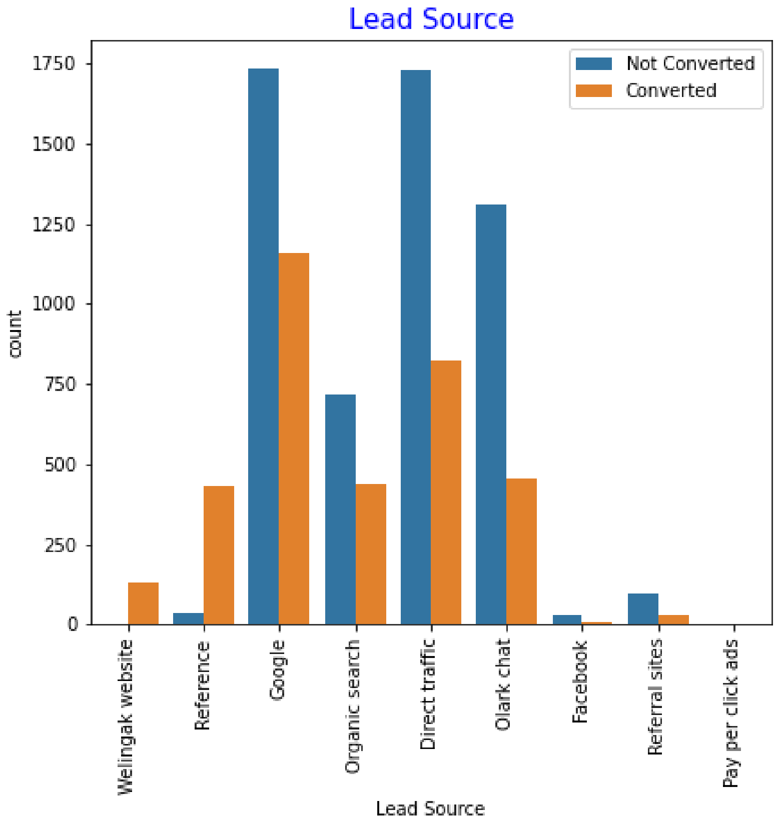
Lead Origin

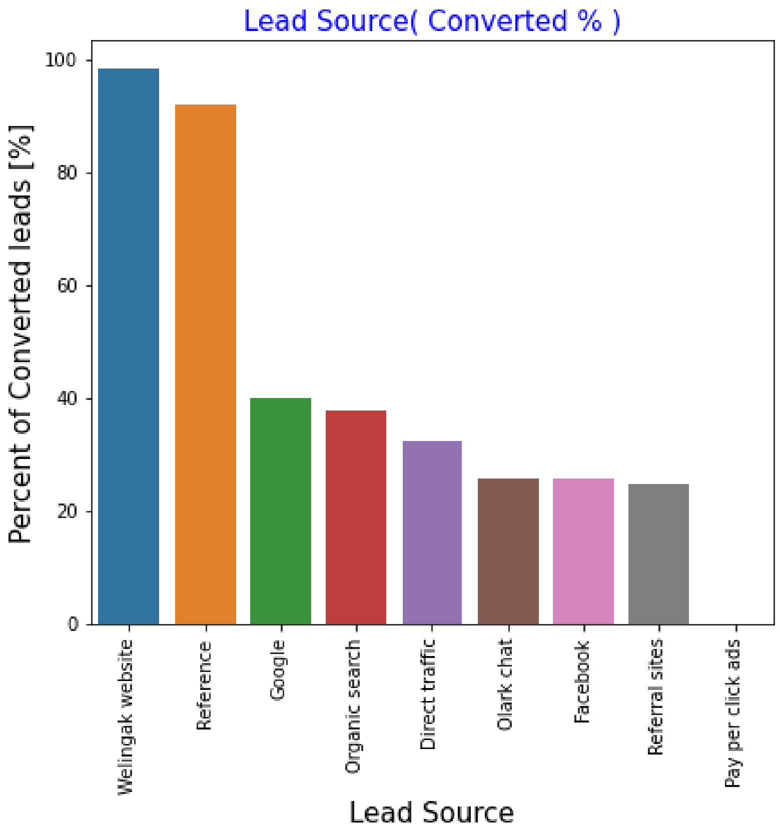


Lead Origin(Converted %)

Highest leads	Highest lead conversion rate
1. Landing page submissions 2. API 3. Lead add	1. Quick lead form 2. Lead add form 3. Landing on page

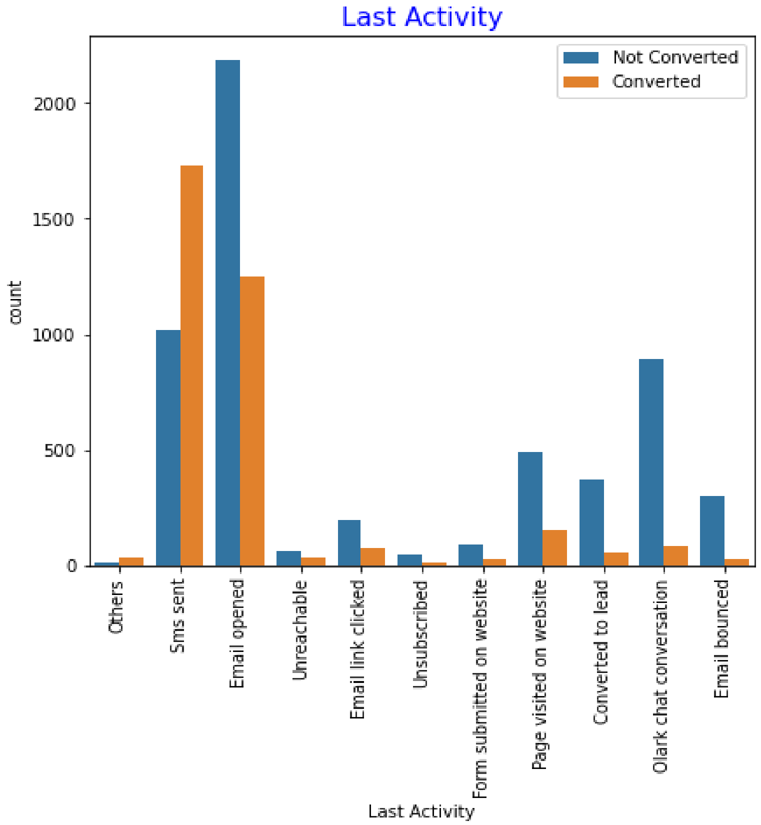
Lead Source (converted vs not-converted)



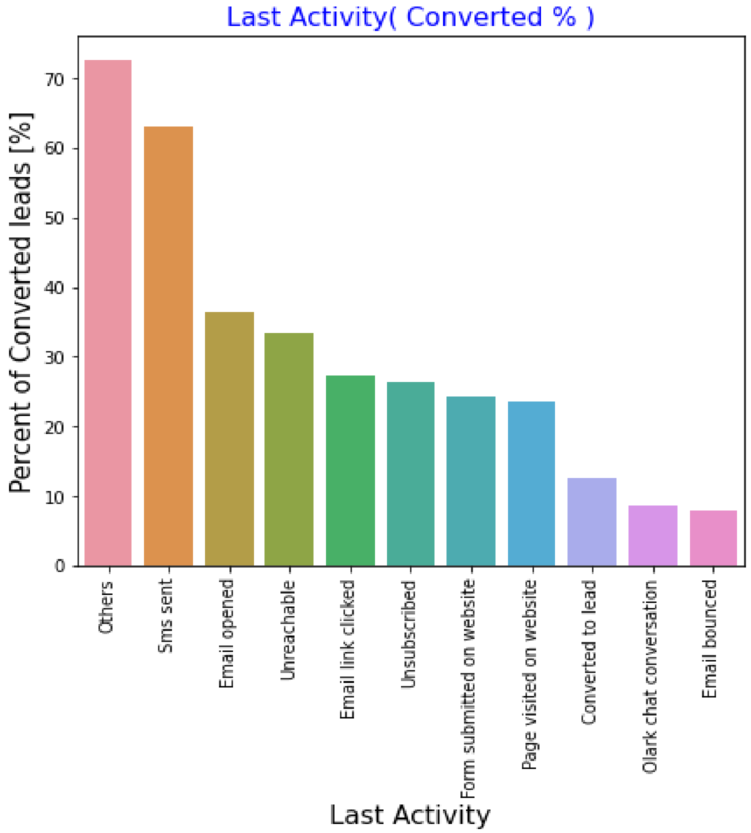


Highest leads	Highest lead conversion rate
1. Google 2. Direct traffic 3. Olark chat	1. Wellington website 2. Reference 3. Google

Last Activity (converted vs not-converted)



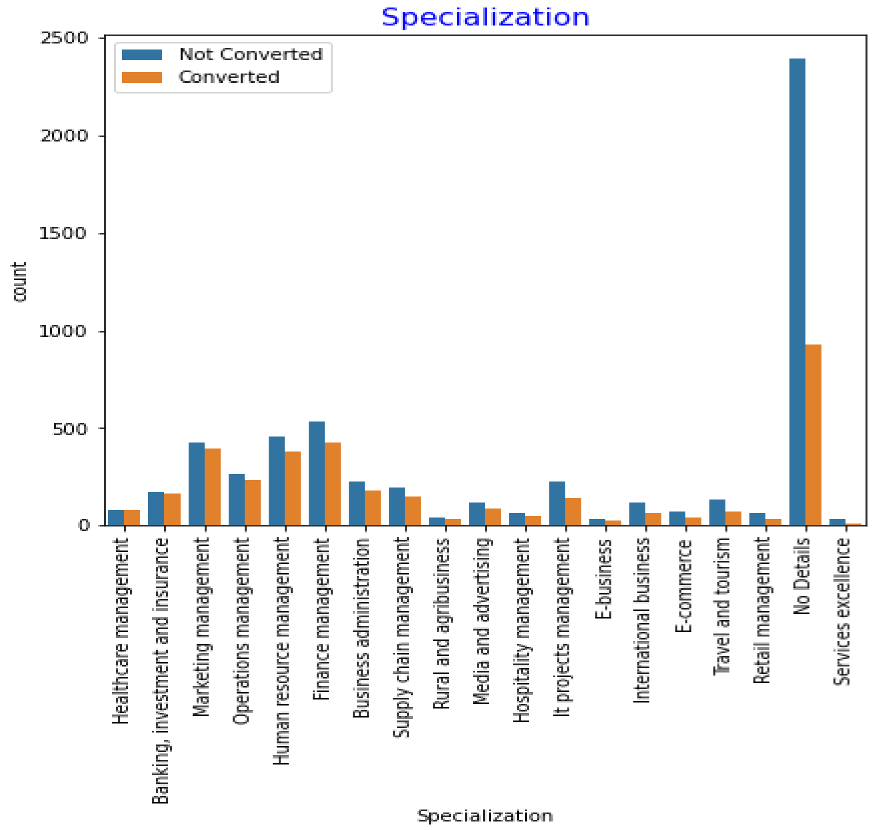
Last Activity

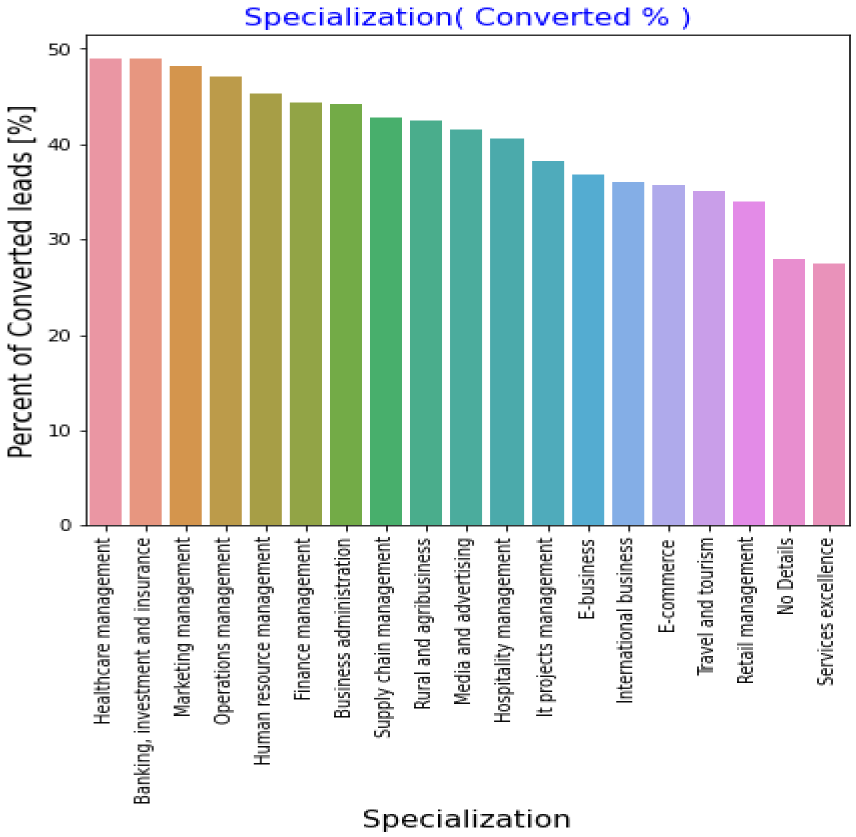


Last Activity(Converted %)

Highest leads	Highest lead conversion rate
1. Email Opened 2. Sms sent 3. Olark chat	1. Sms sent 2. Email sent

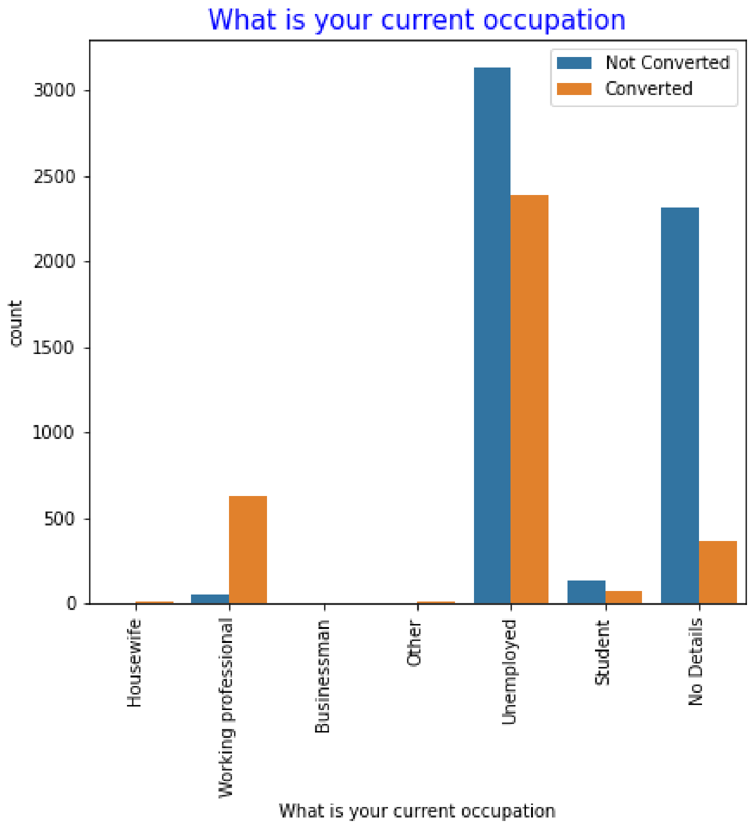
Specialization (converted vs not-converted)

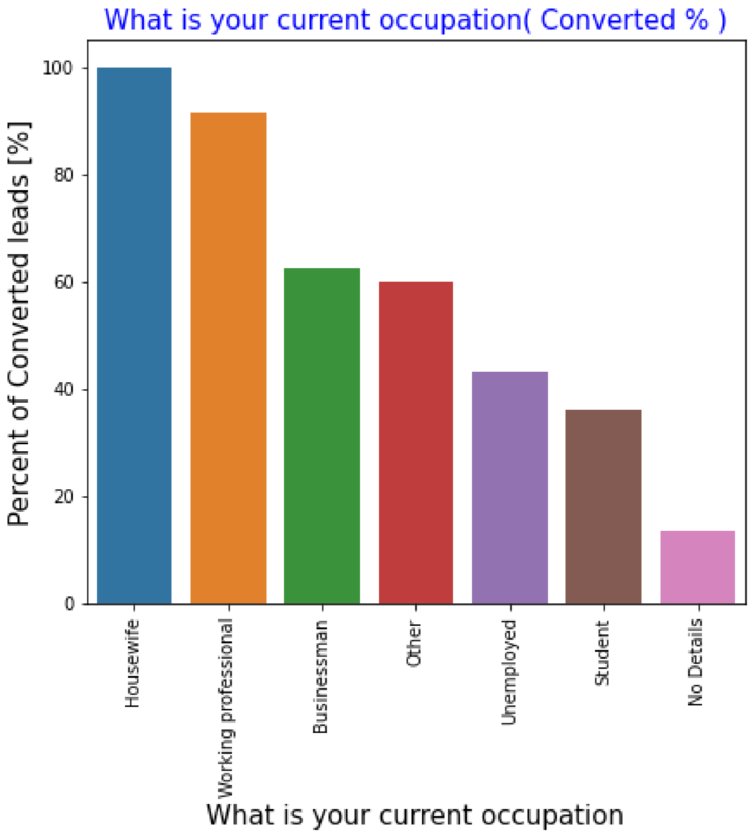




Highest leads	Highest lead conversion rate
1. Finanace 2. Marketing 3. Operations	1. Health-care 2. Banking,Investment and Insurance 3. Marketing

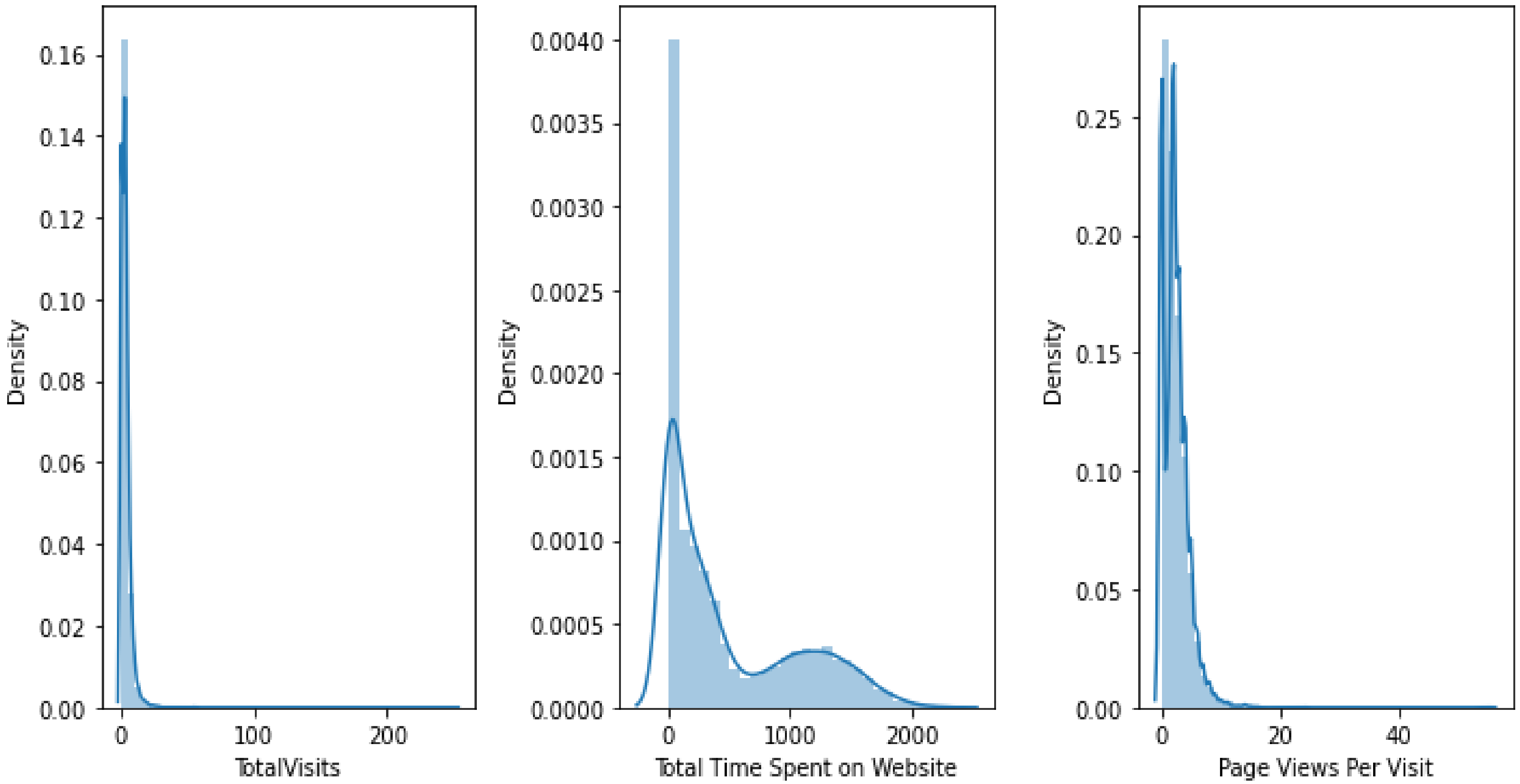
Occupation (converted vs not-converted)





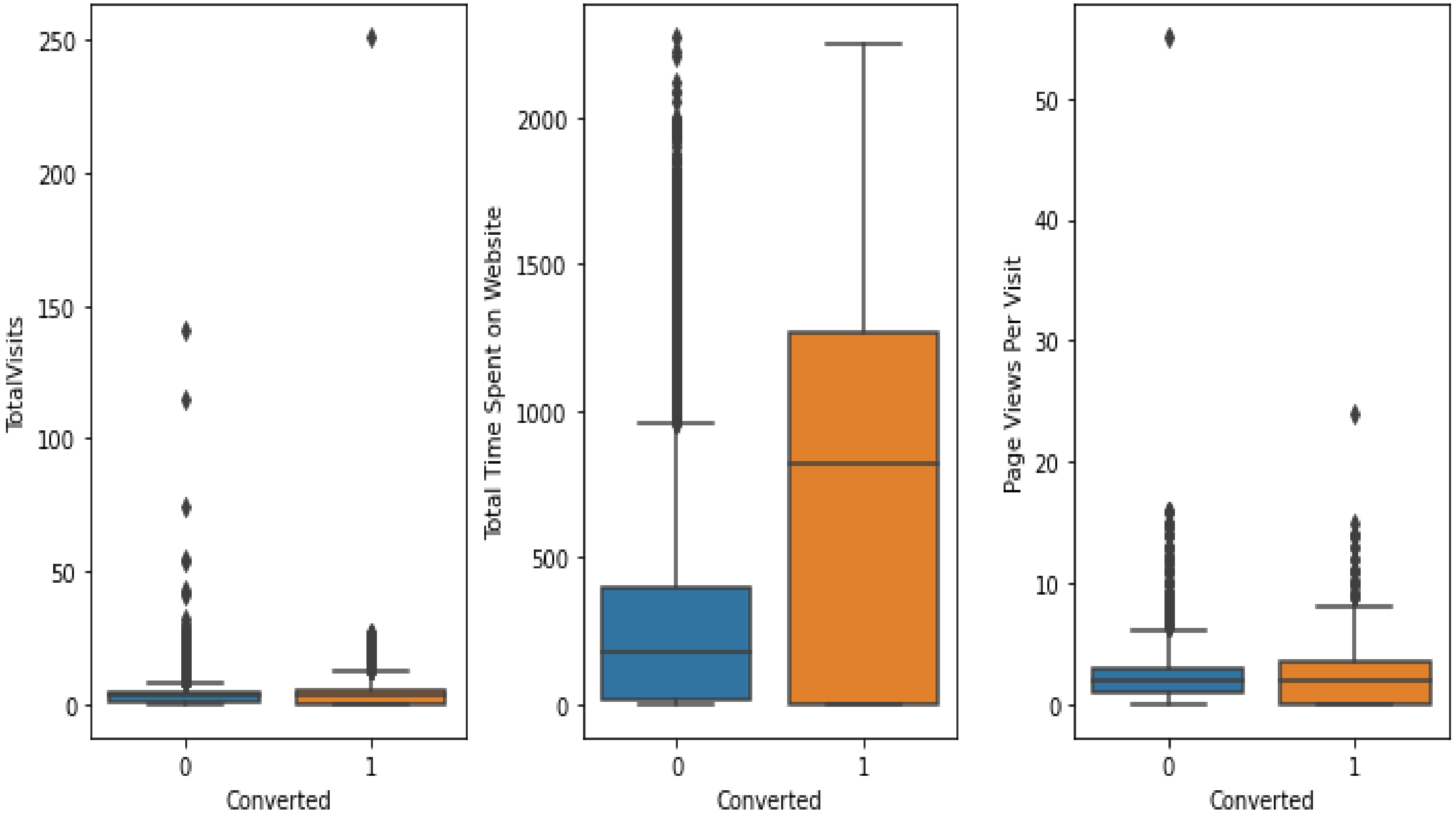
Highest leads	Highest lead conversion rate
1. Unemployed 2. Working professionals 3. Students	1. Housewife 2. Working Professional 3. Businessman

Numerical columns vs Density



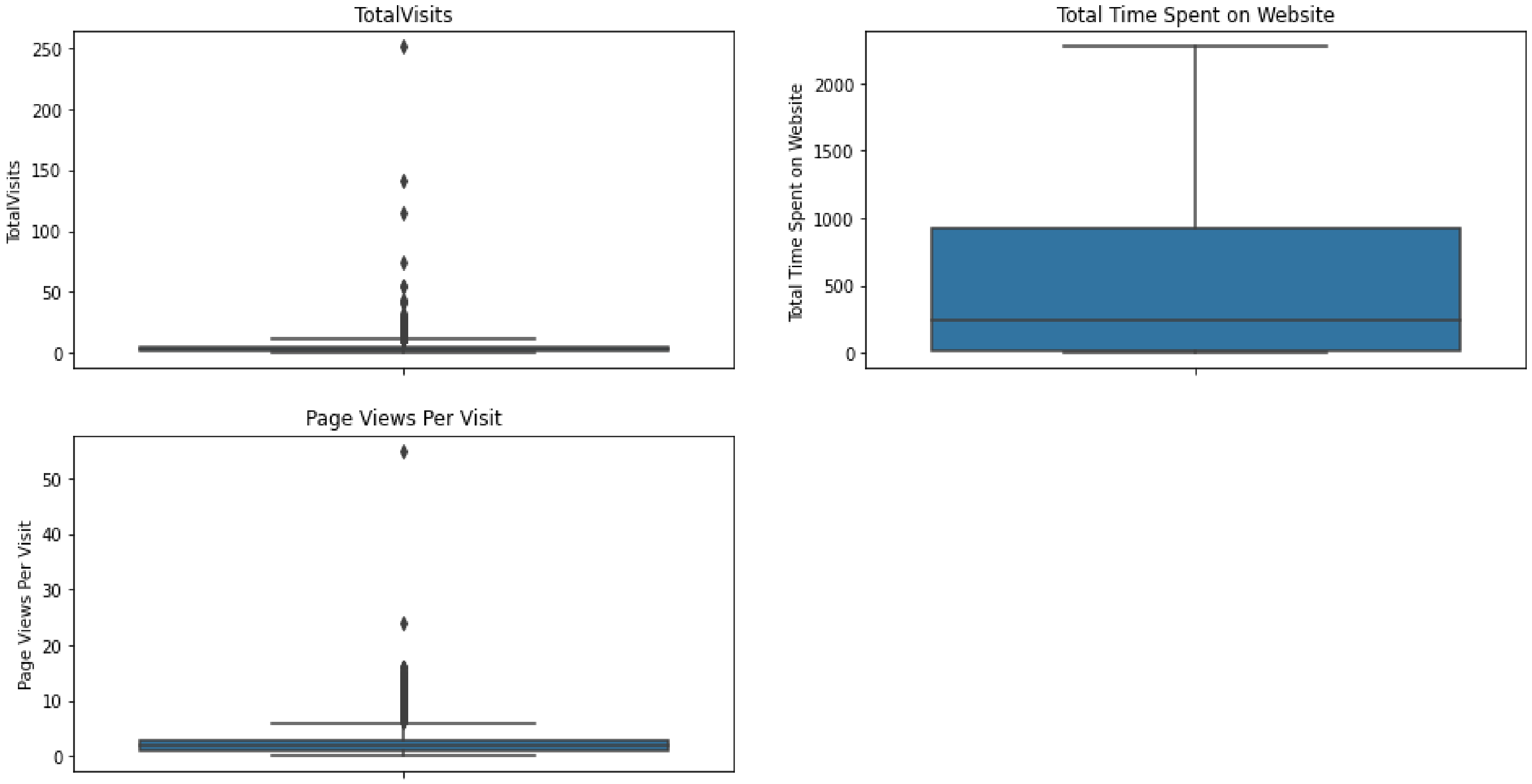
- **Since data is normalized and is clearly skewed to left side**

Numerical Columns vs Conversion



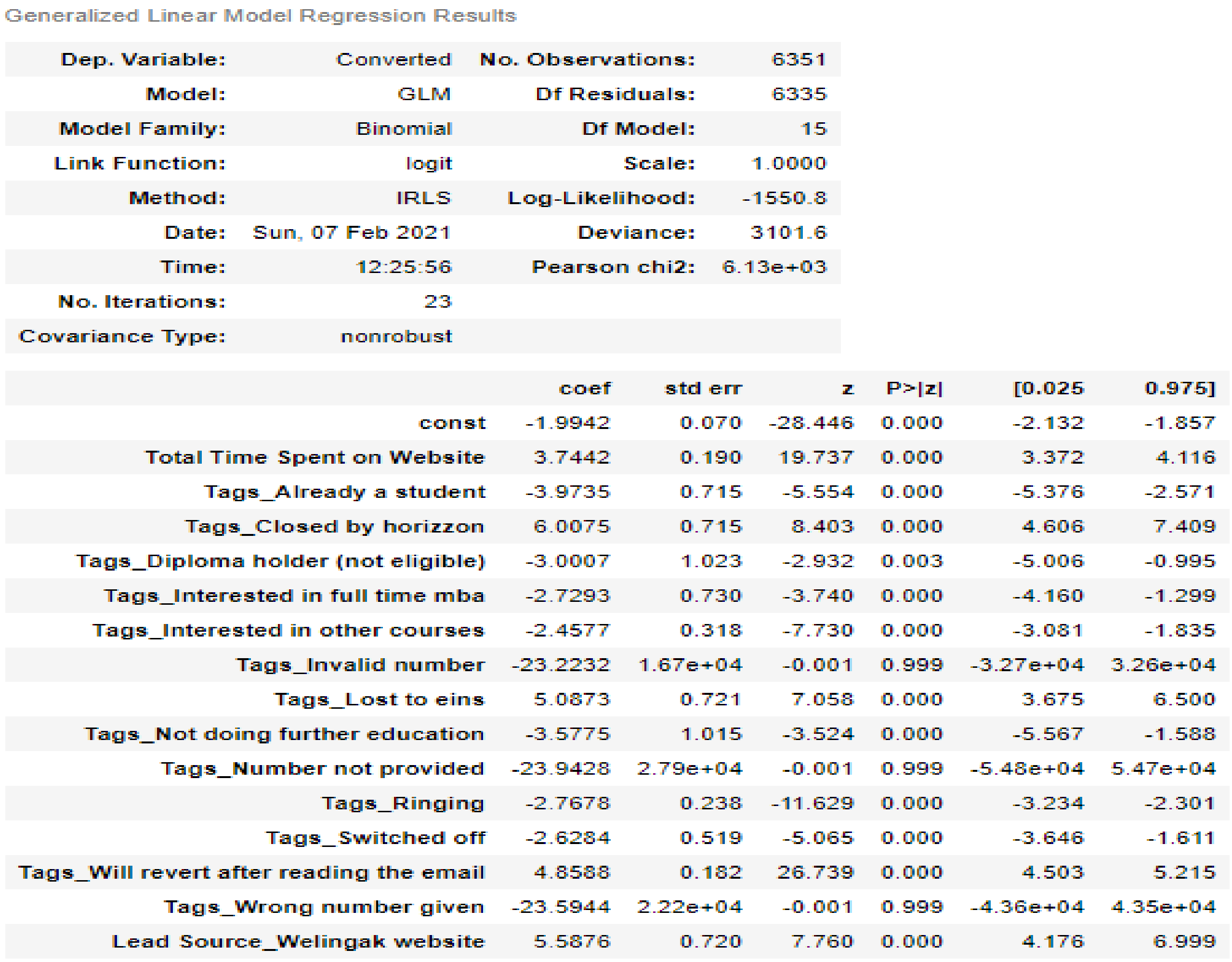
- **Conversion rate is more in Total time spent on website.**

Outliers for numerical columns



- **Since there are only 3 numerical columns left.**
- **Outliers are present in the variables and are important for the analysis**

Model Obtained after RFE



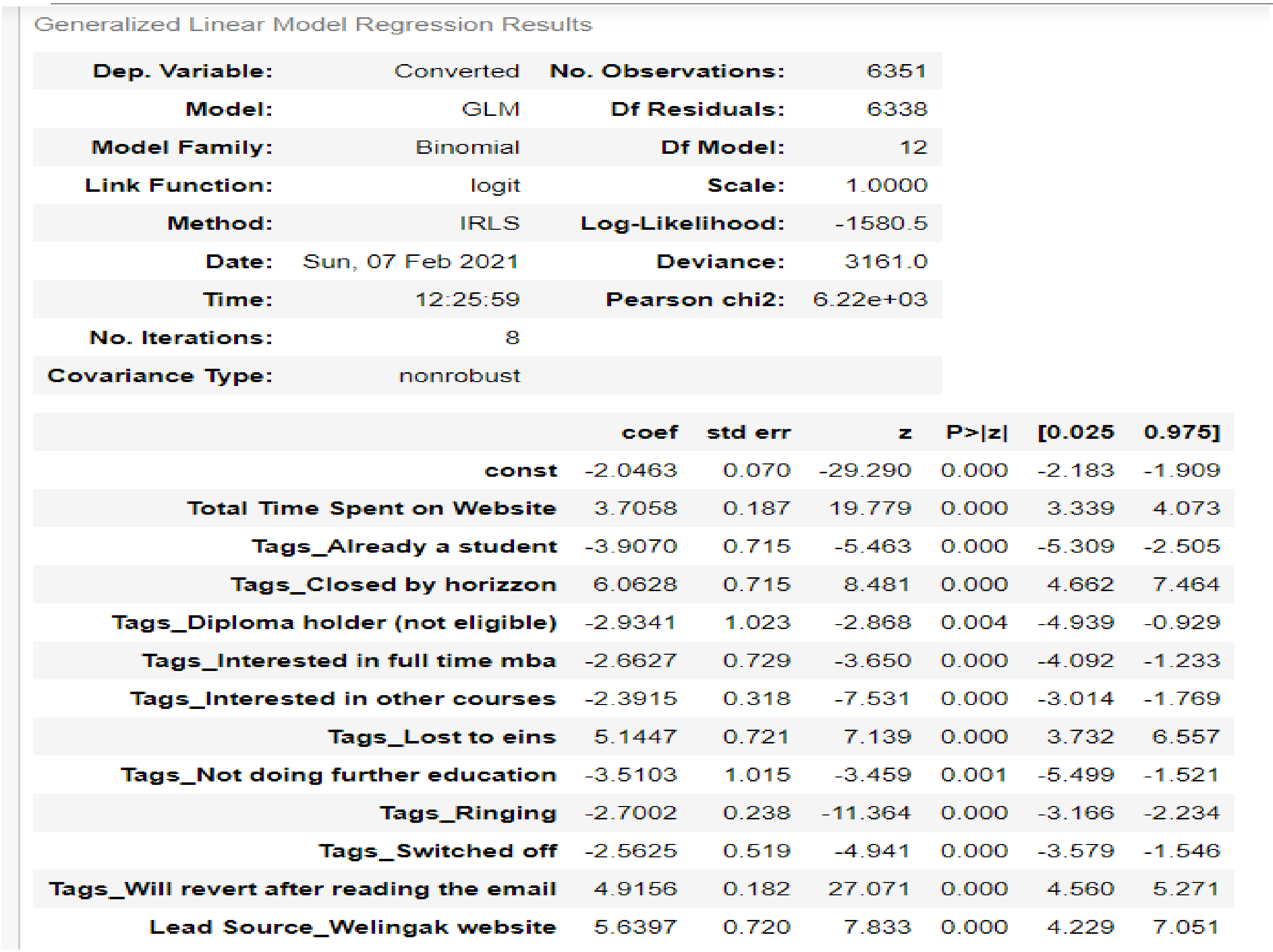
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6335
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1550.8
Date:	Sun, 07 Feb 2021	Deviance:	3101.6
Time:	12:25:56	Pearson chi2:	6.13e+03
No. Iterations:	23		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9942	0.070	-28.446	0.000	-2.132	-1.857
Total Time Spent on Website	3.7442	0.190	19.737	0.000	3.372	4.116
Tags_Already a student	-3.9735	0.715	-5.554	0.000	-5.376	-2.571
Tags_Closed by horizzon	6.0075	0.715	8.403	0.000	4.606	7.409
Tags_Diploma holder (not eligible)	-3.0007	1.023	-2.932	0.003	-5.006	-0.995
Tags_Interested in full time mba	-2.7293	0.730	-3.740	0.000	-4.160	-1.299
Tags_Interested in other courses	-2.4577	0.318	-7.730	0.000	-3.081	-1.835
Tags_Invalid number	-23.2232	1.67e+04	-0.001	0.999	-3.27e+04	3.26e+04
Tags_Lost to eins	5.0873	0.721	7.058	0.000	3.675	6.500
Tags_Not doing further education	-3.5775	1.015	-3.524	0.000	-5.567	-1.588
Tags_Number not provided	-23.9428	2.79e+04	-0.001	0.999	-5.48e+04	5.47e+04
Tags_Ringing	-2.7678	0.238	-11.629	0.000	-3.234	-2.301
Tags_Switched off	-2.6284	0.519	-5.065	0.000	-3.646	-1.611
Tags_Will revert after reading the email	4.8588	0.182	26.739	0.000	4.503	5.215
Tags_Wrong number given	-23.5944	2.22e+04	-0.001	0.999	-4.36e+04	4.35e+04
Lead Source_Welingak website	5.5876	0.720	7.760	0.000	4.176	6.999

This is the Model-1 which have a few unwanted and invalid parameters(high p-value and VIF) which are to be reduced manually.

Final Model after RFE and Manual Reduction

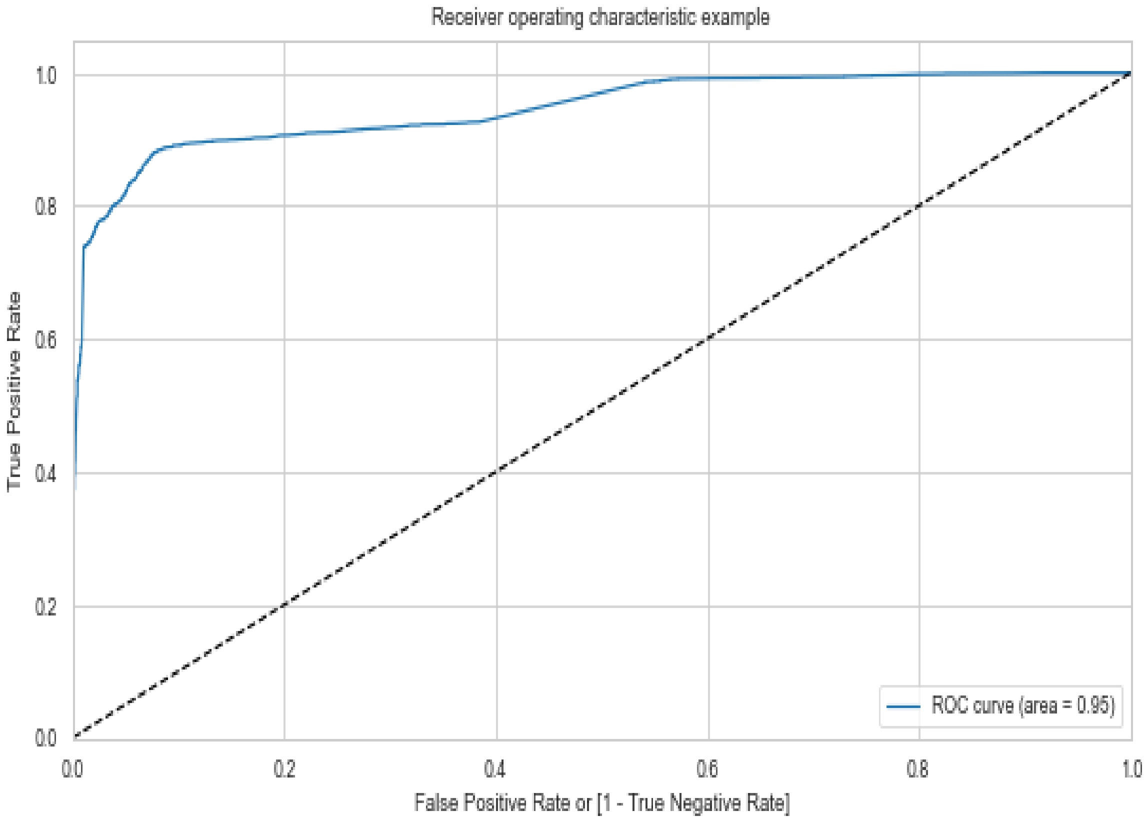


```
Generalized Linear Model Regression Results
     Dep. Variable:        Converted   No. Observations:     6351
            Model:              GLM       Df Residuals:     6338
     Model Family:         Binomial           Df Model:       12
    Link Function:            logit              Scale:   1.0000
           Method:             IRLS     Log-Likelihood:  -1580.5
             Date: Sun, 07 Feb 2021           Deviance:   3161.0
             Time:          12:25:59      Pearson chi2: 6.22e+03
   No. Iterations:                8
  Covariance Type:        nonrobust

                                        coef  std err        z  P>|z|  [0.025  0.975]
                              const  -2.0463    0.070  -29.290  0.000  -2.183  -1.909
          Total Time Spent on Website  3.7058    0.187   19.779  0.000   3.339   4.073
              Tags_Already a student  -3.9070    0.715   -5.463  0.000  -5.309  -2.505
              Tags_Closed by horizzon  6.0628    0.715    8.481  0.000   4.662   7.464
     Tags_Diploma holder (not eligible) -2.9341  1.023   -2.868  0.004  -4.939  -0.929
        Tags_Interested in full time mba -2.6627  0.729   -3.650  0.000  -4.092  -1.233
       Tags_Interested in other courses -2.3915   0.318   -7.531  0.000  -3.014  -1.769
                     Tags_Lost to eins  5.1447    0.721    7.139  0.000   3.732   6.557
        Tags_Not doing further education -3.5103  1.015   -3.459  0.001  -5.499  -1.521
                          Tags_Ringing -2.7002    0.238  -11.364  0.000  -3.166  -2.234
                     Tags_Switched off -2.5625    0.519   -4.941  0.000  -3.579  -1.546
   Tags_Will revert after reading the email  4.9156  0.182  27.071  0.000   4.560   5.271
           Lead Source_Welingak website  5.6397   0.720    7.833  0.000   4.229   7.051
```

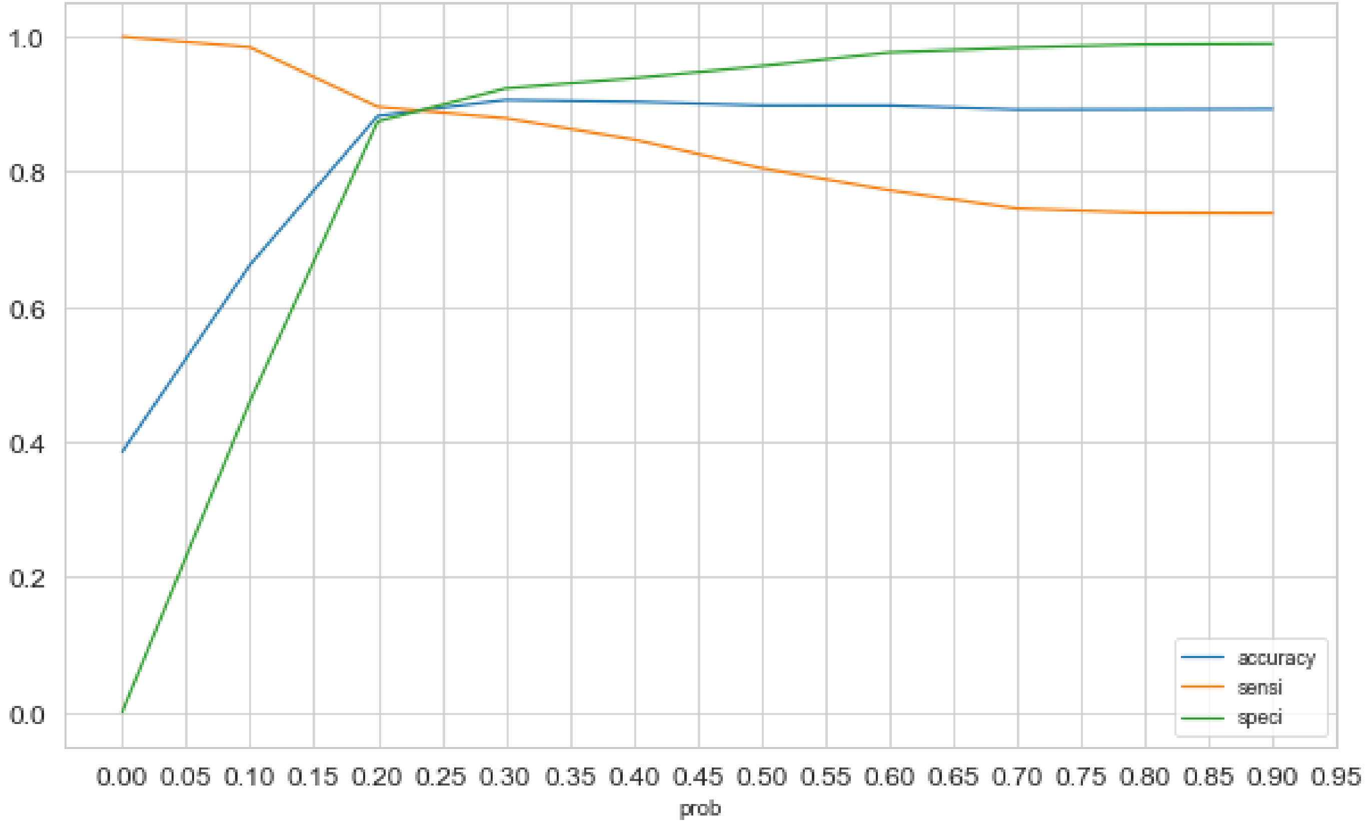
- **The Model-4 after all necessary reductions.**
- **This is our final model with significant parameters for business development**

ROC curve train-set



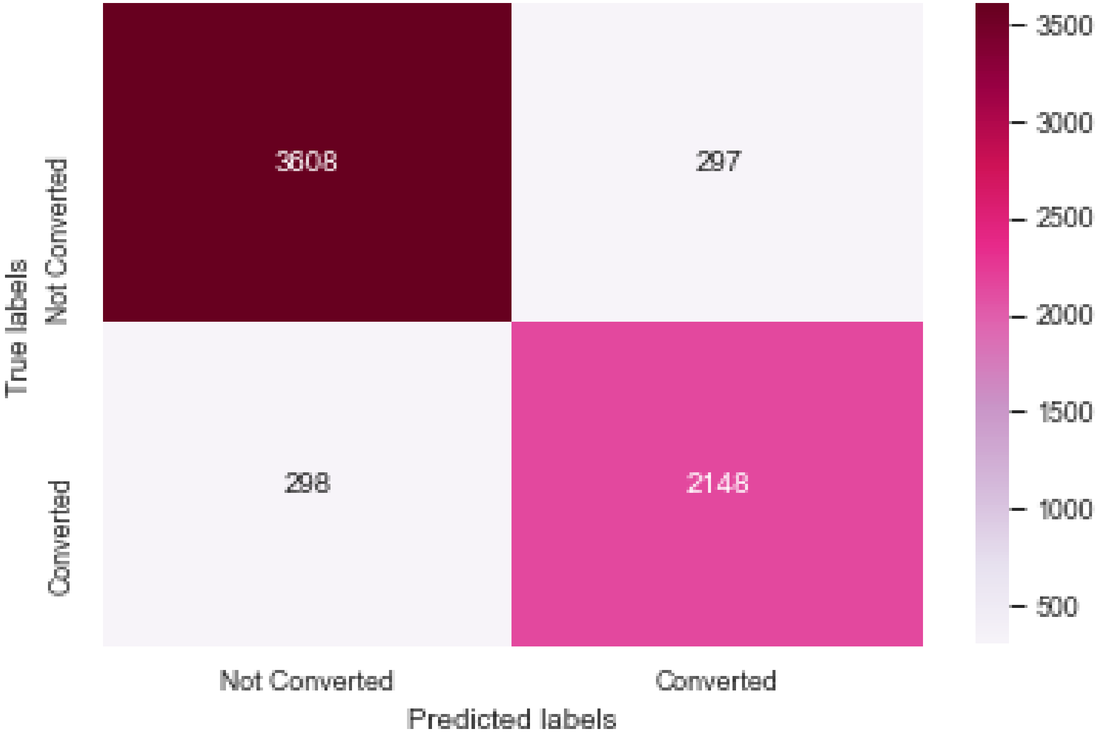
The ROC curve gives area =0.95 which is exceptional value w.r.t model obtained.

Trade-off curve

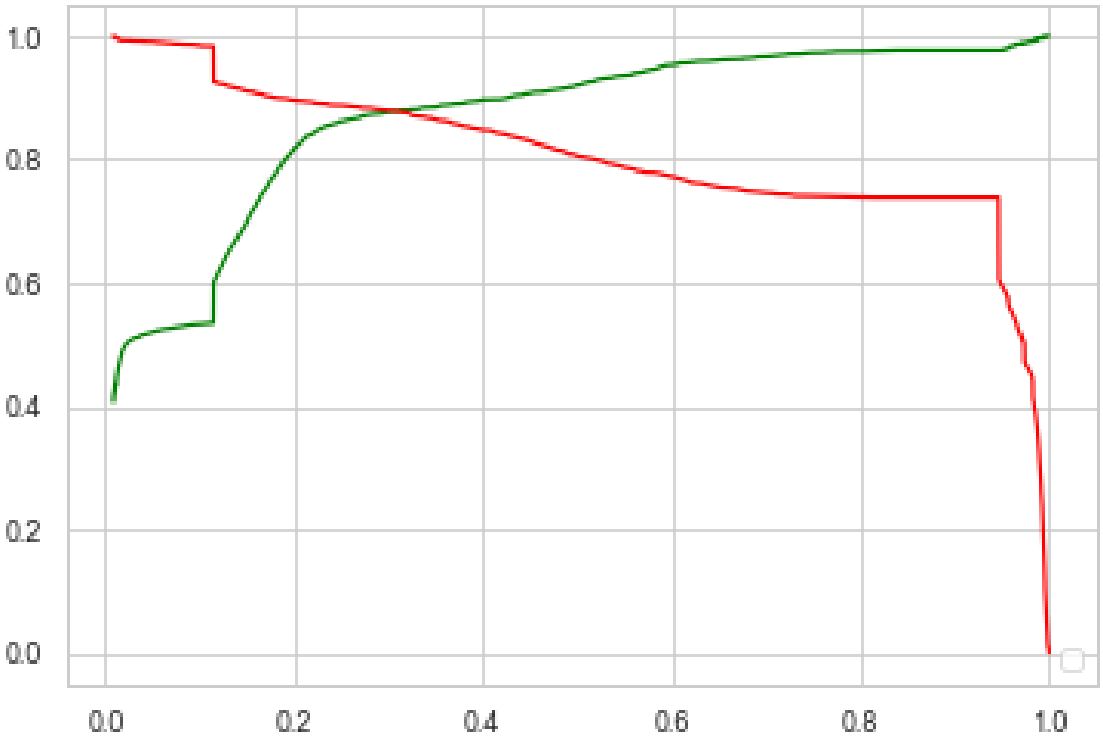


- **Trade-off point is 0.24, since it is very low conversion rate**
- **We tried doing the analysis with 0.33 trade-off which come out to be a good. So considered the later for the same.**

Confusion Matrix (after trade-off)

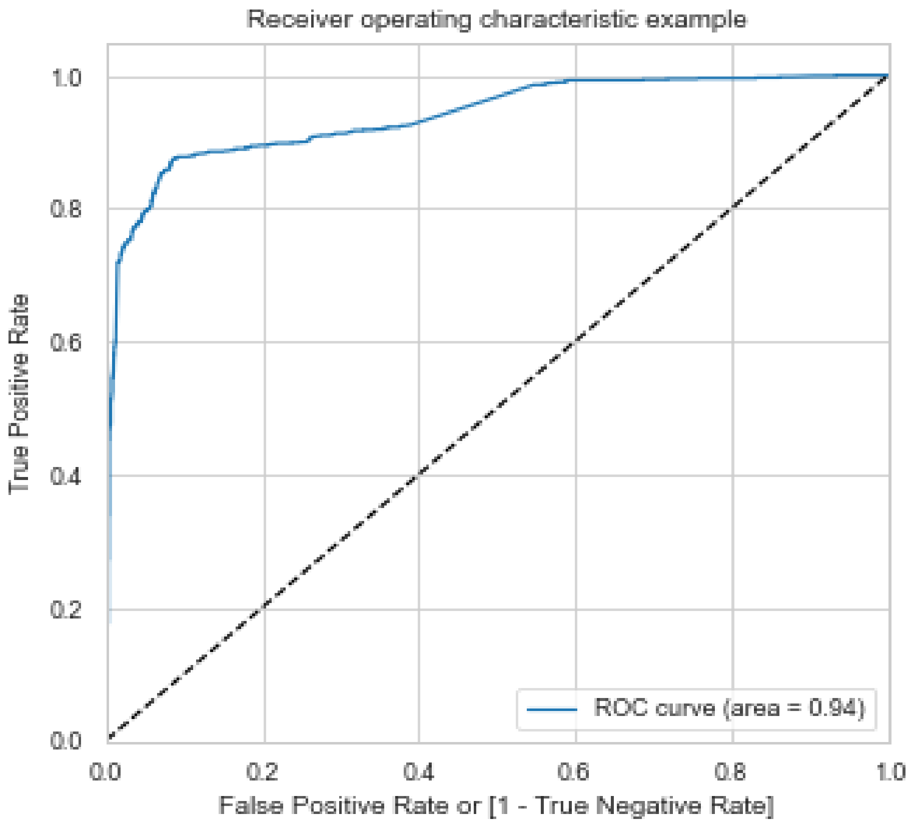


Precision recall curve



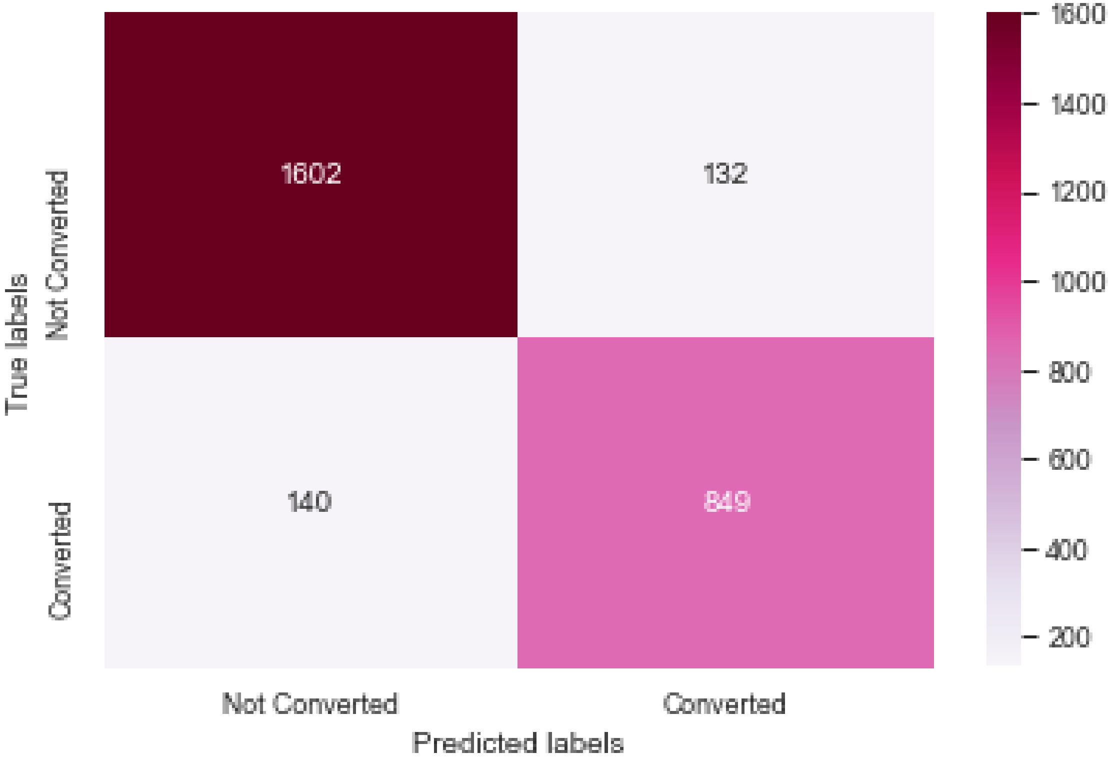
- **The precision-recall curve gives an exceptional accuracy of around 0.88.**
- **Which indicates the model is Good**

ROC curve test-set



The ROC curve gives area =0.94 which is exceptional value w.r.t model obtained.

Confusion matrix on test-set



Metrics train-set vs test-set

Metrics after trade off point on train-set:

Accuracy : 0.905

Sensitivity : 0.805

Specificity : 0.956

Metrics for test-set:

Accuracy : 0.900

Sensitivity : 0.858

Specificity : 0.923

- **The metrics of in two sets are almost similar, so the model is good.**
- **The Metrics are above the expected business perspective**

Conclusion

The CEO of the company expected the model to be 80% accurate. As expected model is more than 80% accurate. So model is good to go for the business development.

Here are the top five parameters which need to be concentrated more by the sales to increase conversion rate.

1. Total Time Spent on Website

2. Tags_Will revert after reading the email

3. Tags_Ringing

4. Tags_Closed by horizzon

5. Tags_Lost to eins