Lead Scoring Case Study

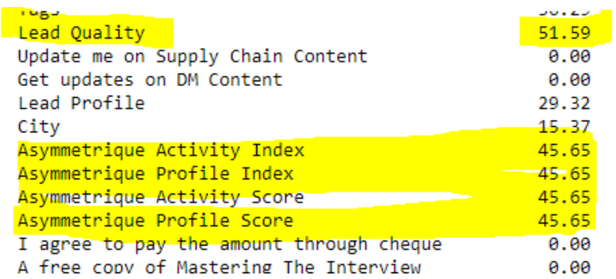
Problem Statement :

The lead conversion case study is about creating a logistic regression model with which

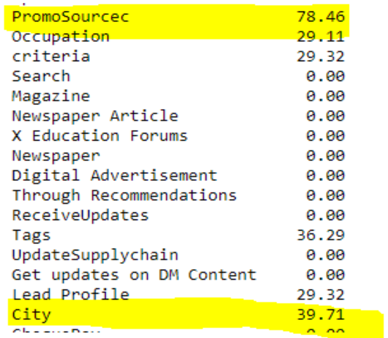
- Looking at the different target variable of a lead company can identify the probability of a lead to get converted.
- The current Lead conversion rate of X-Education is comparatively low because of so many leads flowing to salespersons resulting them to not focus on the potential one.
- The expectation is to identify top target variables that can predict more chances of a lead to get converted so that Team can focus their energy to target only those customers which has more prediction values (hot leads)

Strategy

- Importing Data
  - The file has 9240 data rows and 37 attributes.
- Inspecting Data
  - 4 columns has more than 46% of nulls
  - 4 columns has “Select” as Data Value

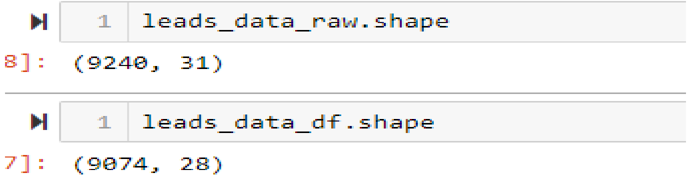


```
Lead Quality                                     51.59
Update me on Supply Chain Content                 0.00
Get updates on DM Content                         0.00
Lead Profile                                     29.32
City                                             15.37
Asymmetrique Activity Index                      45.65
Asymmetrique Profile Index                       45.65
Asymmetrique Activity Score                      45.65
Asymmetrique Profile Score                       45.65
I agree to pay the amount through cheque          0.00
A free copy of Mastering The Interview            0.00
```



```
PromoSourcec                78.46
Occupation                  29.11
criteria                    29.32
Search                       0.00
Magazine                     0.00
Newspaper Article            0.00
X Education Forums           0.00
Newspaper                    0.00
Digital Advertisement        0.00
Through Recommendations      0.00
ReceiveUpdates               0.00
Tags                        36.29
UpdateSupplychain            0.00
Get updates on DM Content    0.00
Lead Profile                29.32
City                        39.71
```

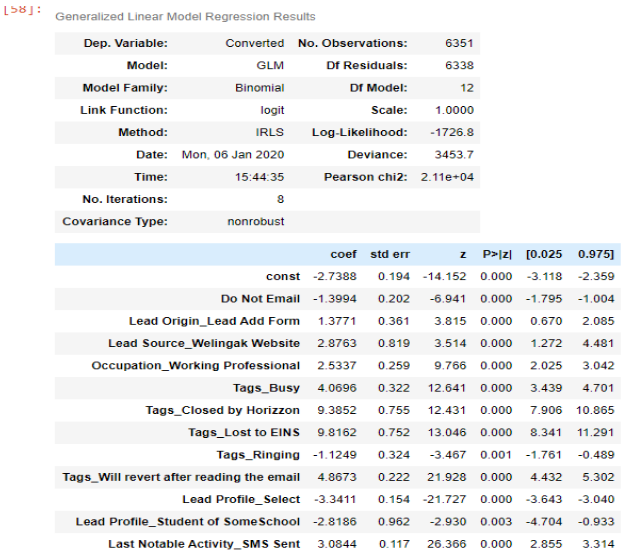
- Cleaning and preparing the data for building model:
  - Replacing the “Select” value with Null
  - Dropping the columns having more than 46% of nulls
  - Dropping the Data which has nulls.
  - Converting the Yes/No columns to (1/0)Identifying and imputing outliers
  - % of Data left After Cleaning



```
leads_data_raw.shape
8]: (9240, 31)

leads_data_df.shape
7]: (9074, 28)
```

Strategy

- Splitting the data in train and test
- Feature Scaling to bring all the variables in common scale
- Looking at data correlation to eliminate multicollinearity
- Feature Selection using RFE
    - Fitting the model in RFE for 13 output variables
    - Model assessment with Stats Model
    - Dropped the column having high P values and recreating model
    - Validation of VIF values to be approx. 3 and less.
    - Merging the predicted probabilities with Predicted Flag
    - Decide for a cutoff (0.5) to derive the Predicted Converted Flag
    - Derive the confusion matrix to assess the model for Train data
    - Check for model accuracy, sensitivity , specificity , False +ve Rate , False -ve Rate, Precision and Recall.
    - Plot the ROC Curve to Derive the Tradeoff between Sensitivity and Specificity



[58]:

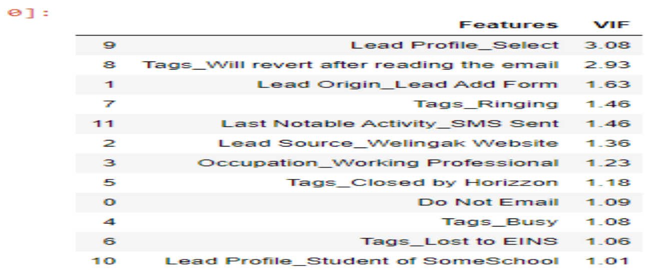
Generalized Linear Model Regression Results

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:	-1726.8
Date:	Mon, 06 Jan 2020	Deviance:	3453.7
Time:	15:44:35	Pearson chi2:	2.11e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7388	0.194	-14.152	0.000	-3.118	-2.359
Do Not Email	-1.3994	0.202	-6.941	0.000	-1.795	-1.004
Lead Origin_Lead Add Form	1.3771	0.361	3.815	0.000	0.670	2.085
Lead Source_Welingak Website	2.8763	0.819	3.514	0.000	1.272	4.481
Occupation_Working Professional	2.5337	0.259	9.766	0.000	2.025	3.042
Tags_Busy	4.0696	0.322	12.641	0.000	3.439	4.701
Tags_Closed by Horizzon	9.3852	0.755	12.431	0.000	7.906	10.865
Tags_Lost to EINS	9.8162	0.752	13.046	0.000	8.341	11.291
Tags_Ringing	-1.1249	0.324	-3.467	0.001	-1.761	-0.489
Tags_Will revert after reading the email	4.8673	0.222	21.928	0.000	4.432	5.302
Lead Profile_Select	-3.3411	0.154	-21.727	0.000	-3.643	-3.040
Lead Profile_Student of SomeSchool	-2.8186	0.962	-2.930	0.003	-4.704	-0.933
Last Notable Activity_SMS Sent	3.0844	0.117	26.366	0.000	2.855	3.314

Strategy

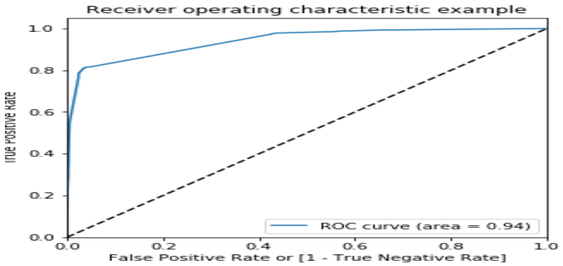
VIF Values:



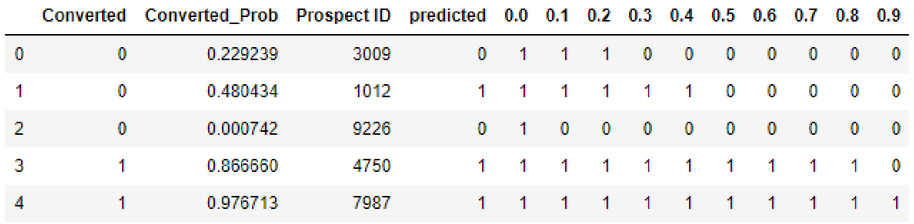
e]:

	Features	VIF
9	Lead Profile_Select	3.08
8	Tags_Will revert after reading the email	2.93
1	Lead Origin_Lead Add Form	1.63
7	Tags_Ringing	1.46
11	Last Notable Activity_SMS Sent	1.46
2	Lead Source_Welingak Website	1.36
3	Occupation_Working Professional	1.23
5	Tags_Closed by Horizzon	1.18
0	Do Not Email	1.09
4	Tags_Busy	1.08
6	Tags_Lost to EINS	1.06
10	Lead Profile_Student of SomeSchool	1.01

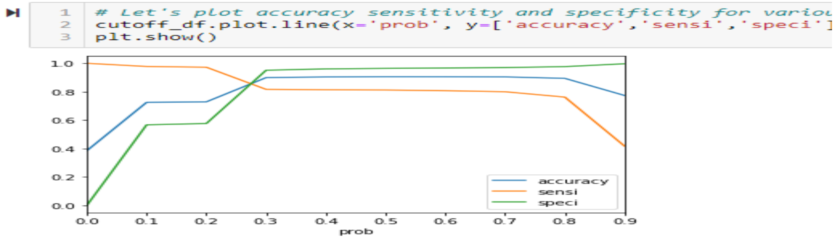
ROC Curve :



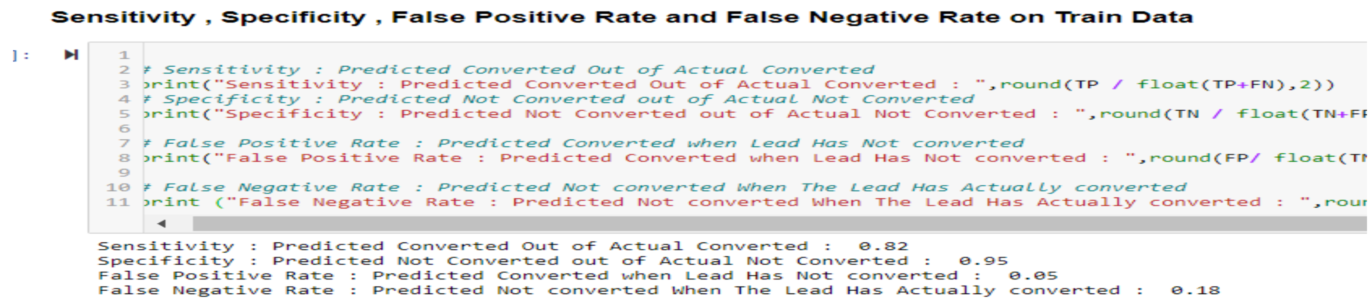
- Find Actual Cutoff point to predict Converted Flag as per the ROC



	Converted	Converted_Prob	Prospect ID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0	0.229239	3009	0	1	1	1	0	0	0	0	0	0	0
1	0	0.480434	1012	1	1	1	1	1	1	0	0	0	0	0
2	0	0.000742	9226	0	1	0	0	0	0	0	0	0	0	0
3	1	0.866660	4750	1	1	1	1	1	1	1	1	1	1	0
4	1	0.976713	7987	1	1	1	1	1	1	1	1	1	1	1



```
# Let's plot accuracy sensitivity and specificity for variou
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci']
plt.show()
```

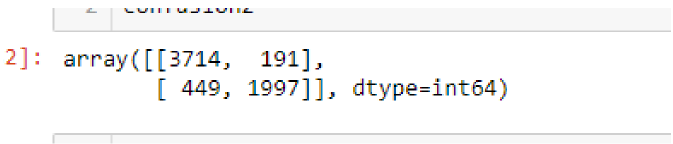


Sensitivity , Specificity , False Positive Rate and False Negative Rate on Train Data

]:

```

# Sensitivity : Predicted Converted Out of Actual Converted
print("Sensitivity : Predicted Converted Out of Actual Converted : ",round(TP / float(TP+FN),2))
# Specificity : Predicted Not Converted out of Actual Not Converted
print("Specificity : Predicted Not Converted out of Actual Not Converted : ",round(TN / float(TN+FF

# False Positive Rate : Predicted Converted when Lead Has Not converted
print("False Positive Rate : Predicted Converted when Lead Has Not converted : ",round(FP/ float(TN

# False Negative Rate : Predicted Not converted When The Lead Has Actually converted
print ("False Negative Rate : Predicted Not converted When The Lead Has Actually converted : ",rour
```

```
Sensitivity : Predicted Converted Out of Actual Converted :  0.82
Specificity : Predicted Not Converted out of Actual Not Converted :  0.95
False Positive Rate : Predicted Converted when Lead Has Not converted :  0.05
False Negative Rate : Predicted Not converted When The Lead Has Actually converted :  0.18
```



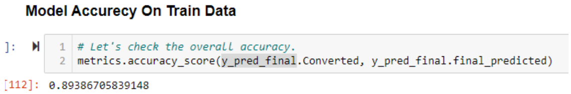
```
confusion2
```

2]:

```
array([[3714,  191],
       [ 449, 1997]], dtype=int64)
```

Strategy

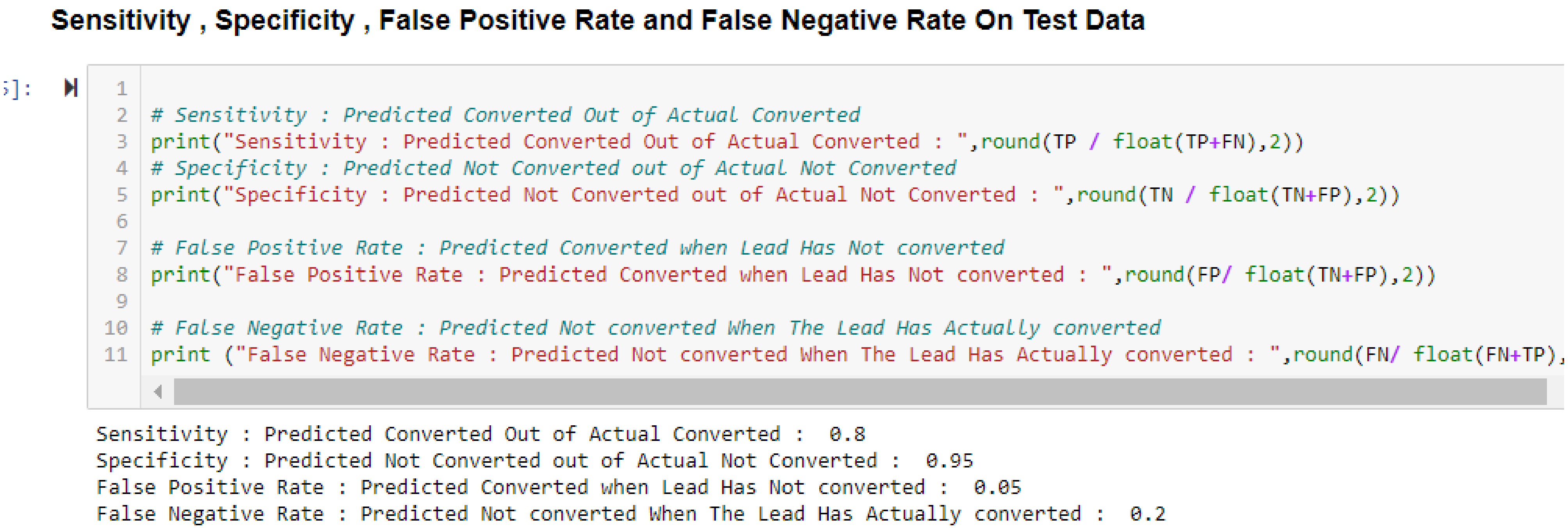
- Apply the model on Test Data to validate the model



Model Accurecy On Train Data

```
# Let's check the overall accuracy.
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
[112]: 0.89386705839148
```



Sensitivity , Specificity , False Positive Rate and False Negative Rate On Test Data

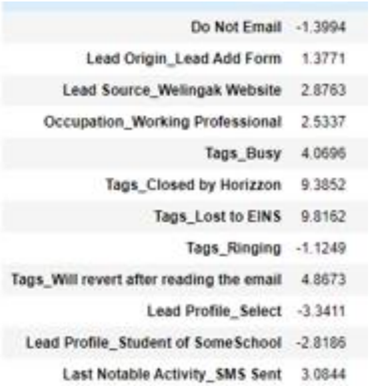
```

# Sensitivity : Predicted Converted Out of Actual Converted
print("Sensitivity : Predicted Converted Out of Actual Converted : ",round(TP / float(TP+FN),2))
# Specificity : Predicted Not Converted out of Actual Not Converted
print("Specificity : Predicted Not Converted out of Actual Not Converted : ",round(TN / float(TN+FP),2))

# False Positive Rate : Predicted Converted when Lead Has Not converted
print("False Positive Rate : Predicted Converted when Lead Has Not converted : ",round(FP/ float(TN+FP),2))

# False Negative Rate : Predicted Not converted When The Lead Has Actually converted
print ("False Negative Rate : Predicted Not converted When The Lead Has Actually converted : ",round(FN/ float(FN+TP),
```

```
Sensitivity : Predicted Converted Out of Actual Converted :  0.8
Specificity : Predicted Not Converted out of Actual Not Converted :  0.95
False Positive Rate : Predicted Converted when Lead Has Not converted :  0.05
False Negative Rate : Predicted Not converted When The Lead Has Actually converted :  0.2
```

Conclusion

- As per the model after dropping null columns, null values, Outlier Treatment , Scaling below are the 12 major features that contribute to lead conversion prediction.

- As per the current cutoff of 0.3 % ,The model has approx. 89 % accuracy with 80% sensitivity and 95% Specificity.
  Hence, model can be applied to predict approx. 80% chances of a lead to get converted. Salesperson can focus on these leads to increase the Lead Conversion Rate of XEducation.



Do Not Email	-1.3994
Lead Origin_Lead Add Form	1.3771
Lead Source_Welingak Website	2.8763
Occupation_Working Professional	2.5337
Tags_Busy	4.0696
Tags_Closed by Horizzon	9.3852
Tags_Lost to EINS	9.8162
Tags_Ringing	-1.1249
Tags_Will revert after reading the email	4.8673
Lead Profile_Select	-3.3411
Lead Profile_Student of SomeSchool	-2.8186
Last Notable Activity_SMS Sent	3.0844

- We can decrease the cutoff to 0.2, to predict more leads that has chances to get converted. In this case with more staff sales staff, we can focus on more leads.

- Similarly, in we can increase the cutoff to be conservative about the cases which to get converted so that sales team can focus on the strategy for the next year.

Thank You