LEAD SCORING CASE STUDY

LOGISTIC REGRESSION

UNDERSTANDING THE DATASET

- We read the CSV file and have look at the data
- pandas's method : info, describe, shape, head, etc.,

```
## Describing data

Rawdata.describe()
```

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

From Above description looks like we have missing values oin the data

PREPARING THE DATA

- Treating missing value
  a) Considering `Select` as null
  b) Deleted the columns which have more than 40% missing value
  c) Putting the majority value for the columns

There are many columns where null values are more than 40%. We think they do not add any values hence we can drop them

```
Rawdata=Rawdata.drop(columns=['How did you hear about X Education','Lead Quality','Lead Profile',
                              'Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Activity Score',
                             'Asymmetrique Profile Score'])
```

```
#### As most of the leads are from India we can change missing values to India only
Rawdata['Country']=Rawdata['Country'].replace(np.nan,'India')
```

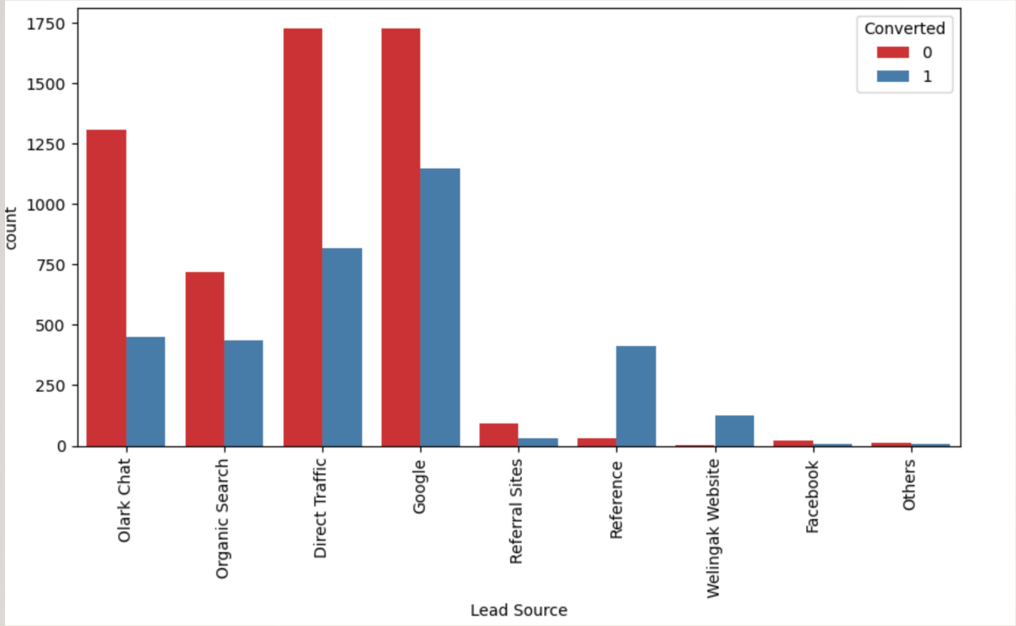
Toggle output scrolling

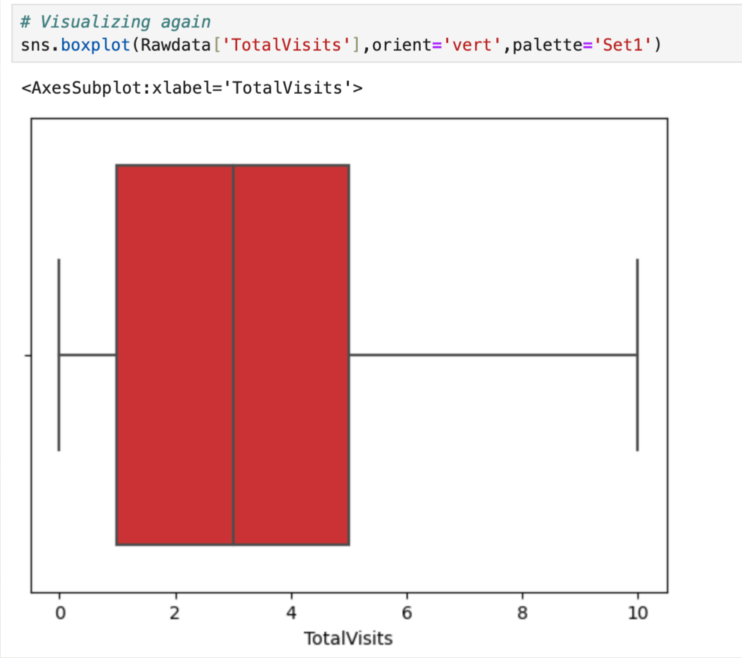
```
# converting the missing data in the 'What is your current occupation' column with 'Unemployed'

Rawdata['What is your current occupation']=Rawdata['What is your current occupation'].replace(np.nan,'Unemployed')
```

DATA VISUALIZATION

- We have done the uni and bi variant analyziz for LeadOrigin, LeadSource, TotalVisits, etc





BUILDING THE MODEL

- We built the model and analyzed the P-Value and VFI of each feature
- We removed the feature one by one based on the P-Value > 0.005 and VFI >= 5

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:**	-2610.5
Date:	Mon, 17 Jul 2023	**Deviance:**	5221.0
Time:	18:43:07	**Pearson chi2:**	6.53e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4001
Covariance Type:	nonrobust		

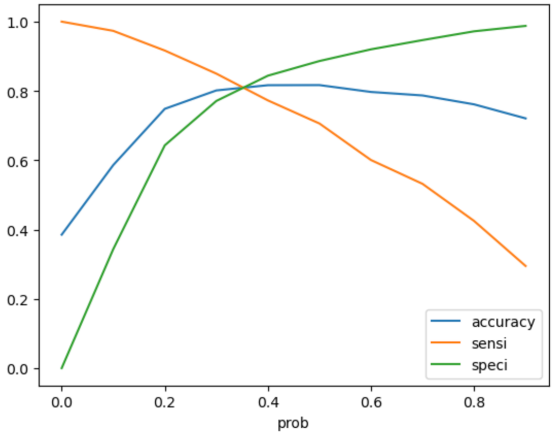
Covariance Type: nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0376	0.125	-0.300	0.764	-0.283	0.208
Do Not Email	-1.5218	0.177	-8.611	0.000	-1.868	-1.175
Total Time Spent on Website	1.0954	0.040	27.225	0.000	1.017	1.174
Lead Origin_Landing Page Submission	-1.1940	0.128	-9.360	0.000	-1.444	-0.944
Lead Source_Olark Chat	1.0819	0.122	8.847	0.000	0.842	1.322
Lead Source_Reference	3.3166	0.241	13.747	0.000	2.844	3.789
Lead Source_Welingak Website	5.8115	0.728	7.981	0.000	4.384	7.239
Last Activity_Olark Chat Conversation	-0.9613	0.171	-5.610	0.000	-1.297	-0.625
Last Activity_Other_Activity	2.1751	0.463	4.699	0.000	1.268	3.082
Last Activity_SMS Sent	1.2942	0.075	17.308	0.000	1.148	1.441
Specialization_Others	-1.2025	0.125	-9.582	0.000	-1.448	-0.957
What is your current occupation_Working Professional	2.6083	0.194	13.454	0.000	2.228	2.988
Last Notable Activity_Modified	-0.9004	0.081	-11.097	0.000	-1.059	-0.741

	Features	VIF
9	Specialization_Others	2.16
3	Lead Source_Olark Chat	2.03
11	Last Notable Activity_Modified	1.78
2	Lead Origin_Landing Page Submission	1.69
6	Last Activity_Olark Chat Conversation	1.59
8	Last Activity_SMS Sent	1.56
1	Total Time Spent on Website	1.29
4	Lead Source_Reference	1.24
10	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.13
5	Lead Source_Welingak Website	1.09
7	Last Activity_Other_Activity	1.01

CHOOSING THE OPTIMIZED PROBABILITY

- To get the y-pred we have to choose the optimum Probaility value
- We draw graph for variouse value of p b/w (0 to 1) for Sensitivity, Specificity and accruacy
- Optimized probability is 0.34



As you can see that around 0.34, you get the optimal values of the three metrics. So let's choose 0.34 as our cutoff now.

MODEL METRICS

- We calculated the accracy, specificity and senstivity with prob value 0.34

```
# Positive and Negative predictive value
print("Positive Predictive Value :",TP / float(TP+FP))
print("Negative Predictive Value : ",TN / float(TN+ FN))
```

```
Positive Predictive Value : 0.7261169633127498
Negative Predictive Value :  0.8757643135075042
```

Sensitivity : 0.8172526573998364 and Specificity : 0.8069142125480153 look good to go

MODEL INTERPRETATION

We have higher postive correlation with
- LeadSource_WelingakWebsite
- LeadSource_Reference

We have higher negative correlation with
- DoNotEmail
- SpecializationOthers

In other words:
- Sources WelingakWebiste and Reference
have higher conversion
- Peoople who don’t want to contact or others profession as have less conversion

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0376	0.125	-0.300	0.764	-0.283	0.208
Do Not Email	-1.5218	0.177	-8.611	0.000	-1.868	-1.175
Total Time Spent on Website	1.0954	0.040	27.225	0.000	1.017	1.174
Lead Origin_Landing Page Submission	-1.1940	0.128	-9.360	0.000	-1.444	-0.944
Lead Source_Olark Chat	1.0819	0.122	8.847	0.000	0.842	1.322
Lead Source_Reference	3.3166	0.241	13.747	0.000	2.844	3.789
Lead Source_Welingak Website	5.8115	0.728	7.981	0.000	4.384	7.239
Last Activity_Olark Chat Conversation	-0.9613	0.171	-5.610	0.000	-1.297	-0.625
Last Activity_Other_Activity	2.1751	0.463	4.699	0.000	1.268	3.082
Last Activity_SMS Sent	1.2942	0.075	17.308	0.000	1.148	1.441
Specialization_Others	-1.2025	0.125	-9.582	0.000	-1.448	-0.957
What is your current occupation_Working Professional	2.6083	0.194	13.454	0.000	2.228	2.988
Last Notable Activity_Modified	-0.9004	0.081	-11.097	0.000	-1.059	-0.741

THANK YOU