Lead Scoring Case Study

Rishi, Avinash & Taushif

Agenda

- Problem Statement
- Solution Approach
- Data Cleaning and outlier treatment
- EDA
- Preprocessing for Model Building
- Model Building
- Model Evaluation

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%.
- Now, 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’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Solution Approach

Solution will be provided based on a Logistic Regression model which the below steps will be followed

- Data cleaning i.e., treat missing values by the following ways
  - Replace select value as null value.
  - Removing columns where there is more than 30% missing values
  - Filling the missing values with mode values where there less than 30% missing values
  - Removing the rows where the missing percentage is minimal.
- Outlier treatment
  - For continuous values, check if there are outliers and if ant remove the ones greater than 0.99 percentile and less than 0.01 percentile.
- EDA
  - Performance exploratory data analysis to get insights from the data
  - Also, understand which columns which useless for model building i.e. data imbalance etc
  - Analyse where there are multiple levels out of which some of them have minimal data belonging to it .
- Preprocessing for model building
  - Drop the columns which are unnecessary and will not contribute towards Model building for examples like Country where all the prospects are mostly based in India and hence there is no relation to lead conversion.
  - Club the unique values/levels of categorical variables where there is minimal data into Others category. This will reduce the number of dummy variables needed.
  - Create dummy variables for the categorical variables.
  - Split the data into train and test data
  - Scale the continuous variables
  - Proceed with model building

Solution Approach

- Model Building – Logistic Regresssion
  - Use RFE to select the 5 features
  - Fit the model
  - Drop the variables with high p-value and fit the model again.
  - Check the VIF values, drop the variables with high VIF (in this case it was not unnecessary as all VIFs were low enough)
- Model Evaluation
  - Check the accuracy, sensitivity and specificity with different probability cut off to find an optimal value of the probability cutoff using the ROC curve
  - Using an optimal cut off evaluate the model on the train data and check the accuracy, sensitivity and specificity
  - Repeat the above step with test data and check if the metrics are under acceptable limits.
  - Check the model with other metrics like Positive predictive value , negative predictive value, precision and recall.

Data cleaning

- Several columns in the data had more than 30% missing values. These columns were dropped for further analysis

```
Prospect ID                                      0.00
Lead Number                                      0.00
Lead Origin                                      0.00
Lead Source                                      0.00
Do Not Email                                     0.00
Do Not Call                                      0.00
Converted                                        0.00
TotalVisits                                      0.01
Total Time Spent on Website                      0.00
Page Views Per Visit                             0.01
Last Activity                                    0.01
Country                                          0.27
Specialization                                   0.37
How did you hear about X Education               0.78
What is your current occupation                  0.29
What matters most to you in choosing a course    0.29
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
Receive More Updates About Our Courses           0.00
Tags                                             0.36
Lead Quality                                     0.52
Update me on Supply Chain Content                0.00
Get updates on DM Content                        0.00
Lead Profile                                     0.74
City                                             0.40
Asymmetrique Activity Index                      0.46
Asymmetrique Profile Index                       0.46
Asymmetrique Activity Score                      0.46
Asymmetrique Profile Score                       0.46
I agree to pay the amount through cheque         0.00
A free copy of Mastering The Interview           0.00
Last Notable Activity                            0.00
dtype: float64
```

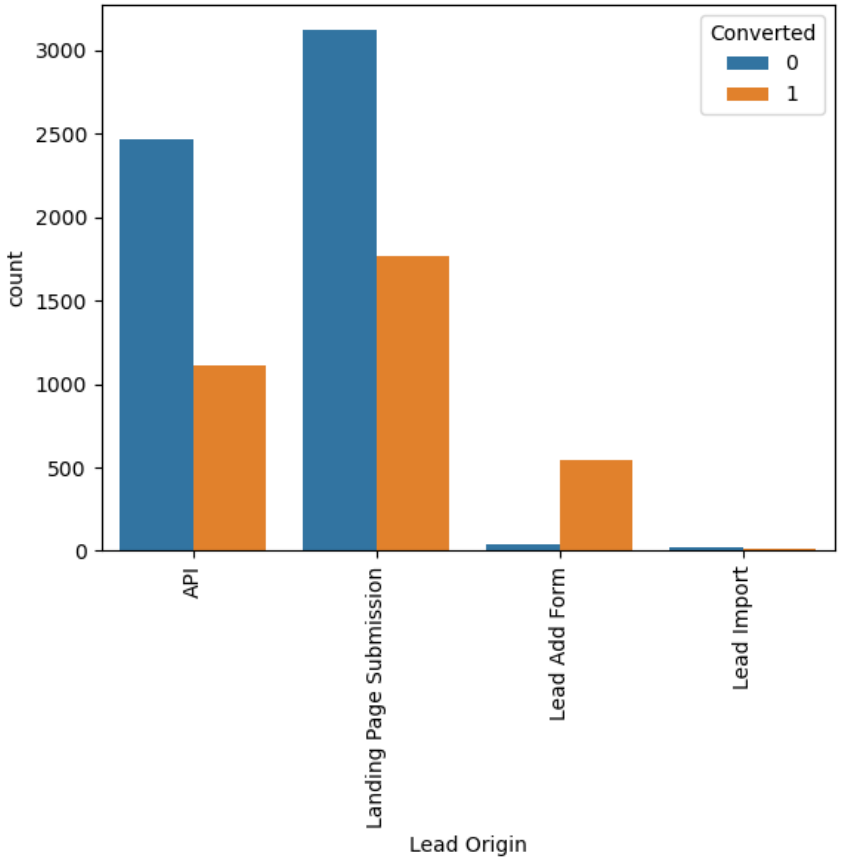
Columns dropped

```
['Specialization',
 'How did you hear about X Education',
 'Tags',
 'Lead Quality',
 'Lead Profile',
```

expand output; double click to hide output

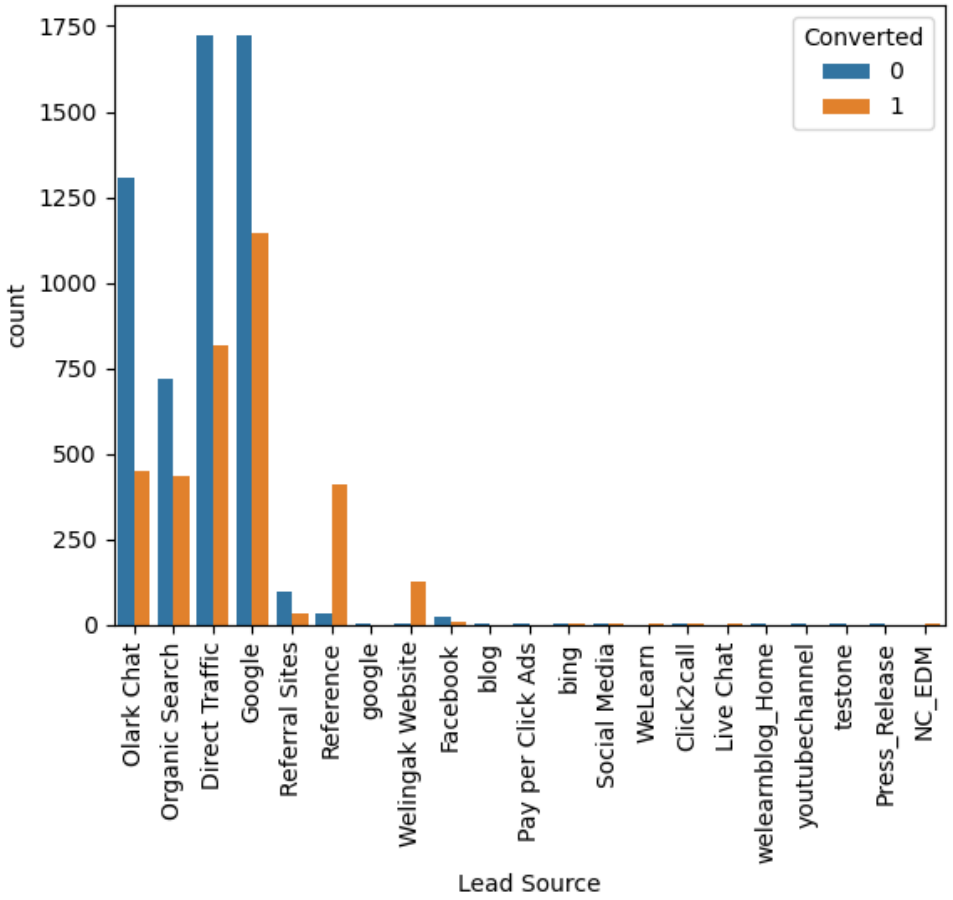
```
 'Asymmetrique Profile Index',
 'Asymmetrique Activity Score',
 'Asymmetrique Profile Score']
```

EDA



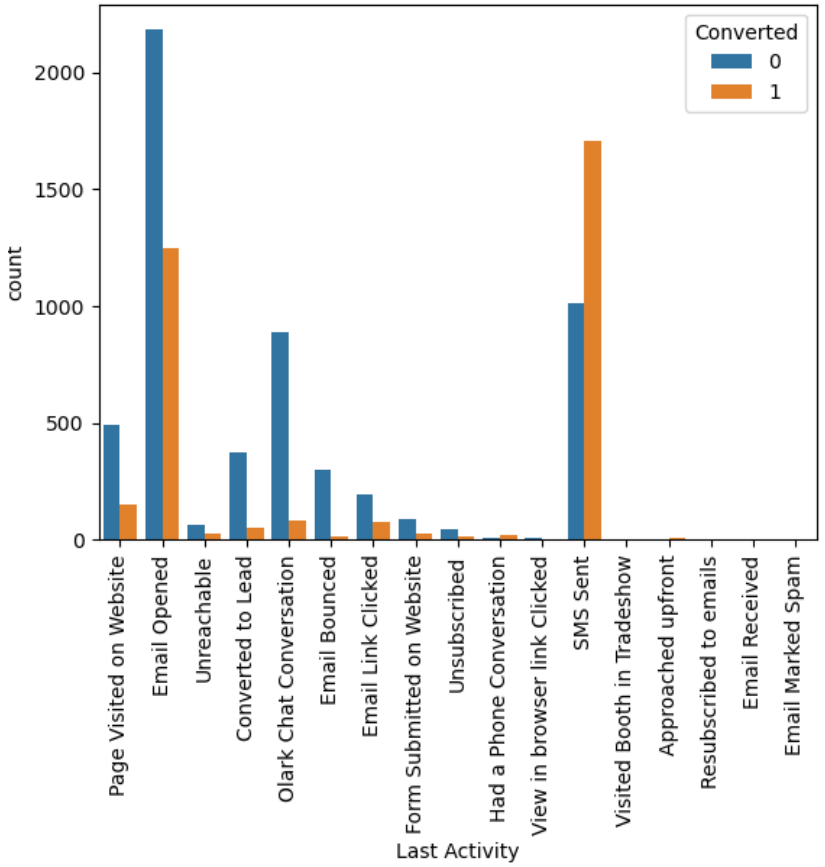
- Lead Origin with Lead Add form has highest conversion rate.

EDA



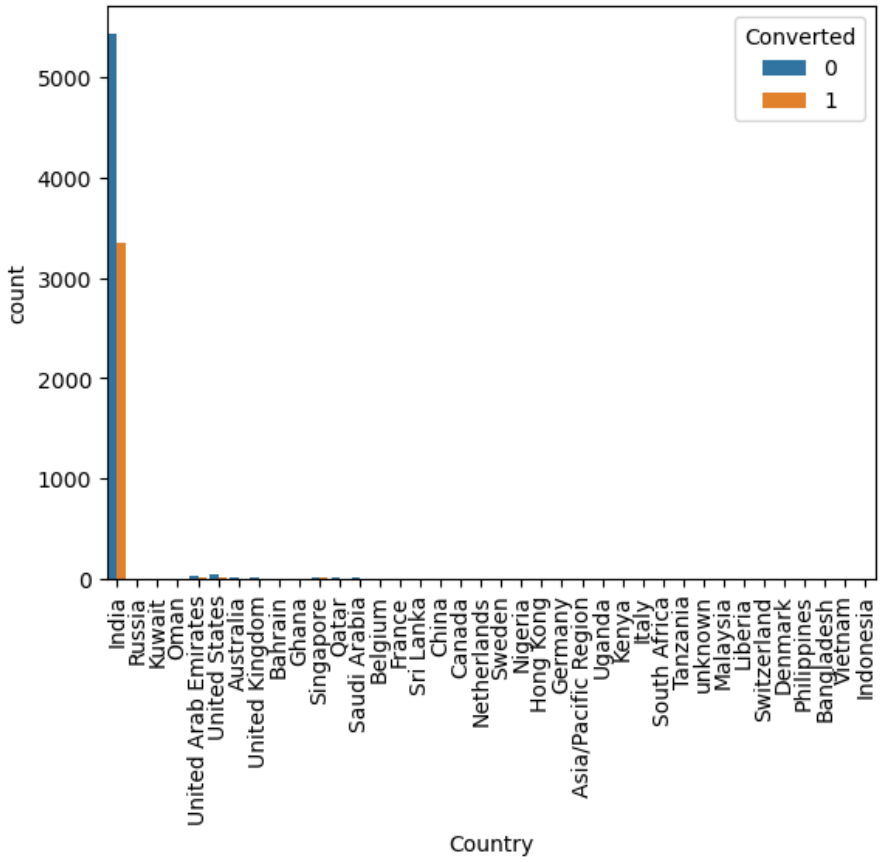
- Some of the Lead Sources are form the more than 80% of the sources and most of the conversions are also from those sources.
- The rest of the sources can be merged together in category Others (this will also save with a lot of dummy variables in model building)

EDA



- Last Activity SMS sent has one of the most conversions
- Many of the last activity levels are insignificant and hence can be clubbed in Others.

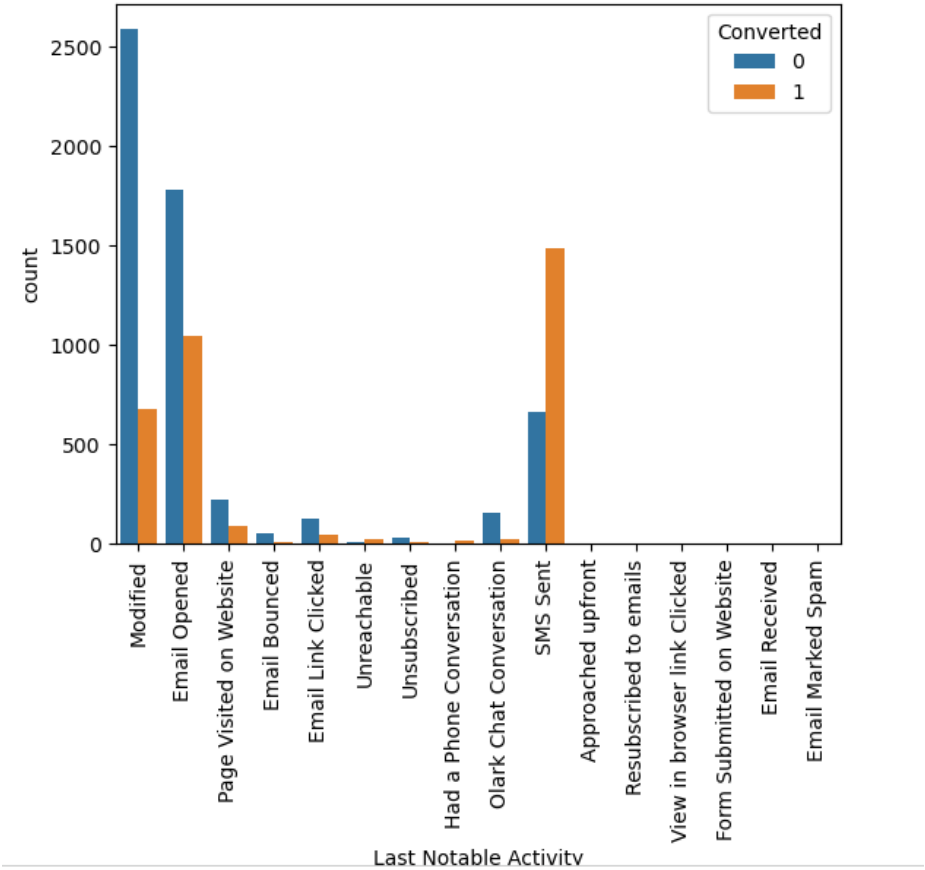
EDA



- Mainly the prospects belong to one country i.e. India and hence would not bear any relationship to the conversion of the lead.
- We would drop this feature/variable
- Several others like country are imbalance and hence not necessary for model building

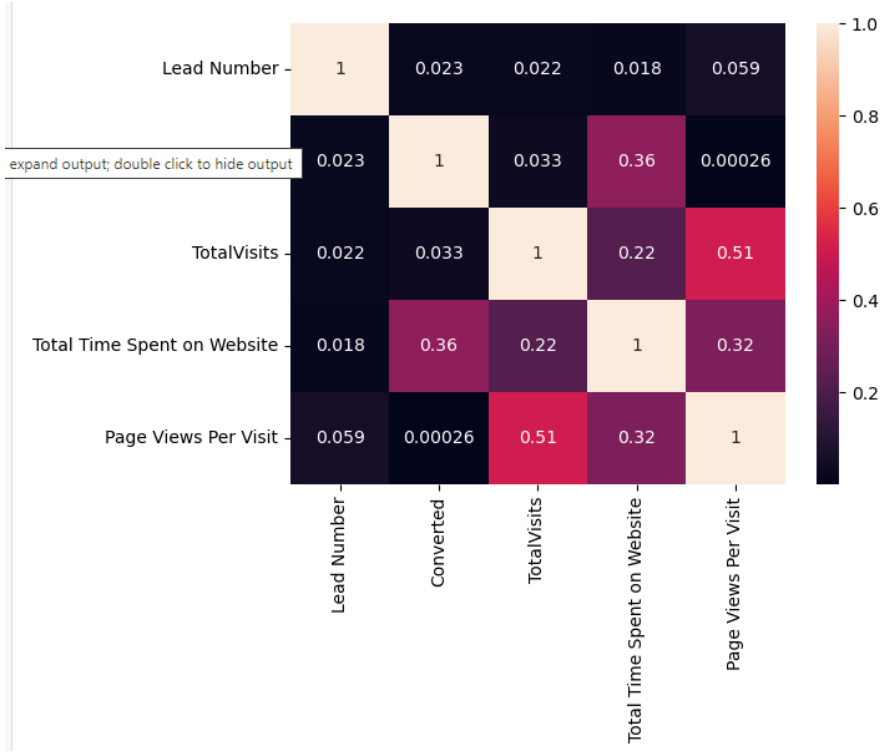
```
['Do Not Call',
 'Country',
 'What matters most to you in choosing a course',
 'Search',
 'Magazine',
 'Newspaper Article',
 'X Education Forums',
 'Newspaper',
 'Digital Advertisement',
 'Through Recommendations',
 'Receive More Updates About Our Courses',
 'Update me on Supply Chain Content',
 'Get updates on DM Content',
 'I agree to pay the amount through cheque']
```

EDA



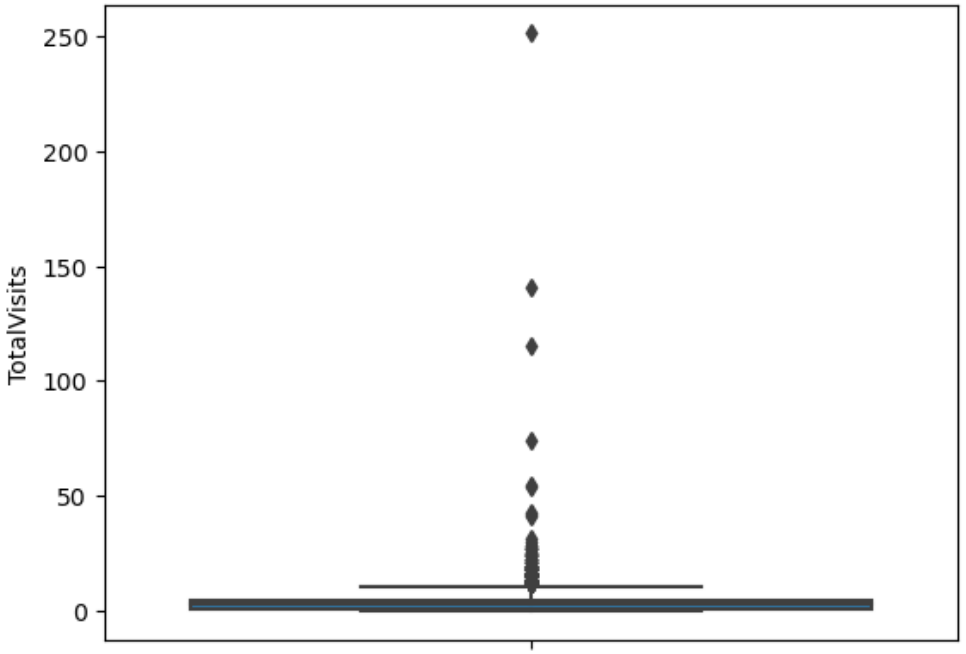
- Last Notable Activity SMS sent has a higher conversion rate
- Some of the last notable activity are insignificant and hence can be clubbed as Others

EDA – Correlation

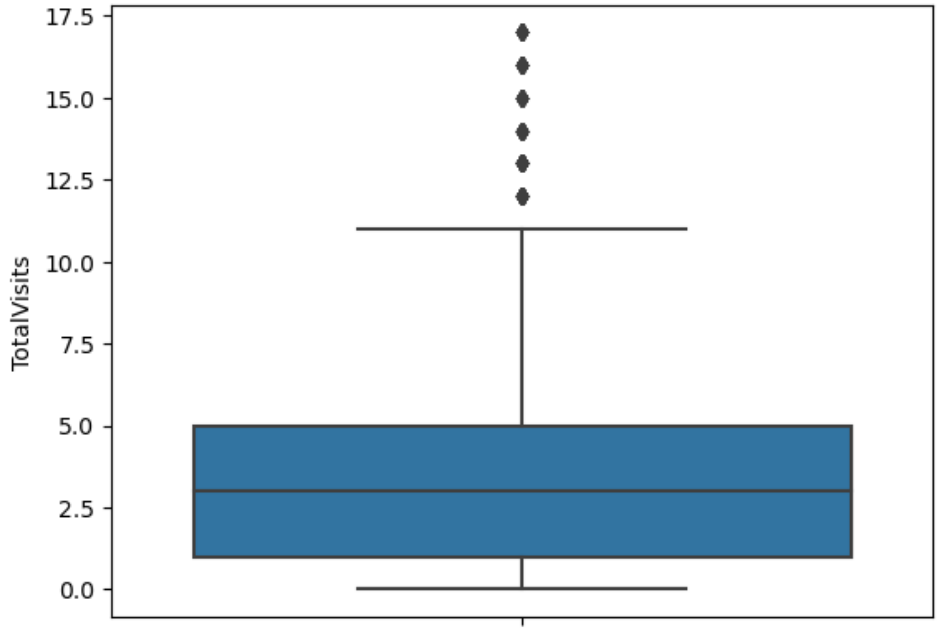


- Totals Visits and Total time spent on website are corelated but we will drop one of the columns as part of the model building exercise

EDA – Outlier Analysis



- Totals Visits clearly has outliers.
- Remove the records above 99 percentile and below 1 percentile.



Preprocessing for Model Building

- Categorical variables are processed to create dummy variables.
- Continuous variables are scaled using Scaler package

	Do Not Email	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	What is your current occupation_Housewife	What is your current occupation_Other	...	LastAct_Form Submitted on Website	I
0	0	0.0	0	0.0	0	0	0	0	0	0	...	0	
1	0	5.0	674	2.5	0	0	0	0	0	0	...	0	
2	0	2.0	1532	2.0	1	1	0	0	0	0	...	0	
3	0	1.0	305	1.0	0	1	0	0	0	0	...	0	
4	0	2.0	1428	1.0	0	1	0	0	0	0	...	0	

LastAct_Olark Chat Conversation	LastAct_Page Visited on Website	LastAct_SMS Sent	LastNotAct_Email Link Clicked	LastNotAct_Email Opened	LastNotAct_Modified	LastNotAct_Olark Chat Conversation	LastNotAct_Page Visited on Website	LastNotAct_SMS Sent
0	1	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	0

Model Building

Final Model

Dep. Variable:	Converted	**No. Observations:**	6246
Model:	GLM	**Df Residuals:**	6231
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2597.2
Date:	Mon, 22 May 2023	**Deviance:**	5194.4
Time:	01:02:57	**Pearson chi2:**	6.59e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3915
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2436	0.070	3.483	0.000	0.107	0.381
Do Not Email	-1.5592	0.206	-7.566	0.000	-1.963	-1.155
Total Time Spent on Website	1.1113	0.040	27.452	0.000	1.032	1.191
Lead Origin_Lead Add Form	3.9045	0.225	17.324	0.000	3.463	4.346
What is your current occupation_Working Professional	2.7140	0.188	14.455	0.000	2.346	3.082
LSource_Olark Chat	1.1400	0.103	11.115	0.000	0.939	1.341
LSource_Welingak Website	1.8681	0.757	2.466	0.014	0.384	3.353
LastAct_Converted to Lead	-1.1720	0.228	-5.132	0.000	-1.620	-0.724
LastAct_Email Bounced	-1.3536	0.420	-3.220	0.001	-2.178	-0.530
LastAct_Olark Chat Conversation	-1.4491	0.198	-7.323	0.000	-1.837	-1.061
LastNotAct_Email Link Clicked	-1.9336	0.256	-7.556	0.000	-2.435	-1.432
LastNotAct_Email Opened	-1.4906	0.088	-16.882	0.000	-1.664	-1.318
LastNotAct_Modified	-1.7149	0.101	-16.926	0.000	-1.913	-1.516
LastNotAct_Olark Chat Conversation	-1.4927	0.367	-4.072	0.000	-2.211	-0.774
LastNotAct_Page Visited on Website	-1.8375	0.204	-8.994	0.000	-2.238	-1.437

VIF Analysis

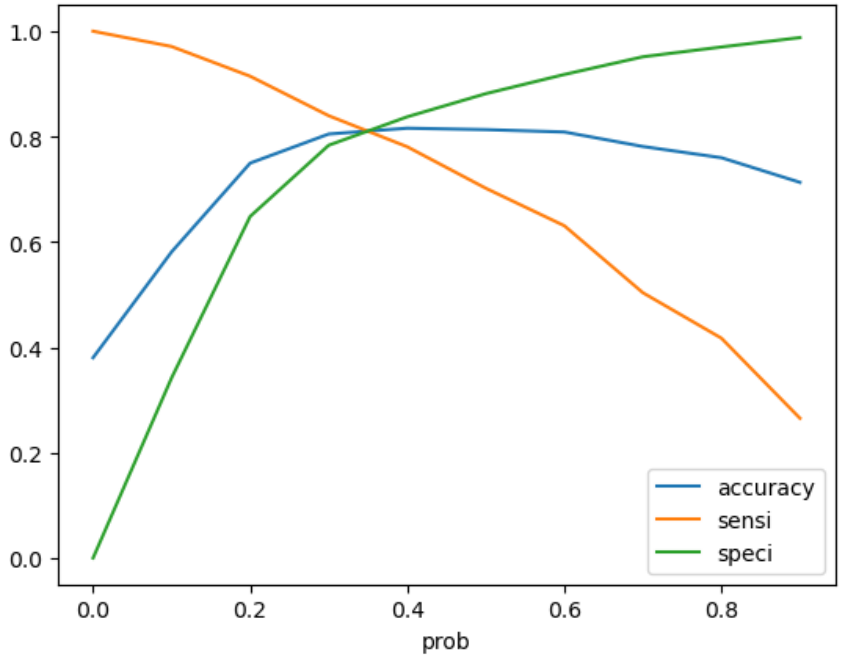
	Features	VIF
8	LastAct_Olark Chat Conversation	2.05
11	LastNotAct_Modified	1.94
0	Do Not Email	1.81
7	LastAct_Email Bounced	1.81
4	LSource_Olark Chat	1.66
2	Lead Origin_Lead Add Form	1.46
12	LastNotAct_Olark Chat Conversation	1.37
5	LSource_Welingak Website	1.29
6	LastAct_Converted to Lead	1.26
1	Total Time Spent on Website	1.23
3	What is your current occupation_Working Profes...	1.12
10	LastNotAct_Email Opened	1.11
9	LastNotAct_Email Link Clicked	1.02
13	LastNotAct_Page Visited on Website	1.02

Model Evaluation

Model metric with different Probability Threholds

	prob	accuracy	sensi	speci
0.0	0.0	0.380243	1.000000	0.000000
0.1	0.1	0.581172	0.970947	0.342030
0.2	0.2	0.749600	0.914526	0.648411
0.3	0.3	0.804835	0.839579	0.783518
0.4	0.4	0.815882	0.780632	0.837510
0.5	0.5	0.813160	0.701895	0.881426
0.6	0.6	0.808678	0.630737	0.917851
0.7	0.7	0.781140	0.503579	0.951434
0.8	0.8	0.759846	0.417263	0.970034
0.9	0.9	0.713096	0.265263	0.987858

ROC curve



To get a balance of sensitivity and specificity, probability threshold of 0.3 can be selected.

Model Evaluation

Train set

```
# Let's check the overall
metrics.accuracy_score(y_t
```

```
0.8048350944604546
```

```
# Let's see the sensitivity of c
TP / float(TP+FN)
```

```
0.8395789473684211
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.783518470679411
```

```
# Calculate false postive rate -
print(FP/ float(TN+FP))
```

```
0.216481529320589
```

```
# Positive predictive value
print (TP / float(TP+FP))
```

```
0.7040960451977402
```

```
# Negative predictive value
print (TN / float(TN+ FN))
```

```
0.8884007029876977
```

```
precision_score(y_train_pred_fin
```

```
0.784101599247413
```

```
recall_score(y_train_pred_final.
```

```
0.7018947368421052
```

Test set

```
# Let's check the overal
metrics.accuracy_score(y
```

```
0.7923823749066468
```

```
# Let's see the sensitivi
TP / float(TP+FN)
```

```
0.8289738430583501
```

```
# Let us calculate specif
TN / float(TN+FP)
```

```
0.7707838479809976
```

```
precision_score(y_pred_fi
```

```
0.6809917355371901
```

```
recall_score(y_pred_final
```

```
0.8289738430583501
```

Thank You