Lead Scoring Case Study

16th April, 2023

Objective

- Predict the most potential leads, also known as 'Hot Leads', that led to higher conversion rate

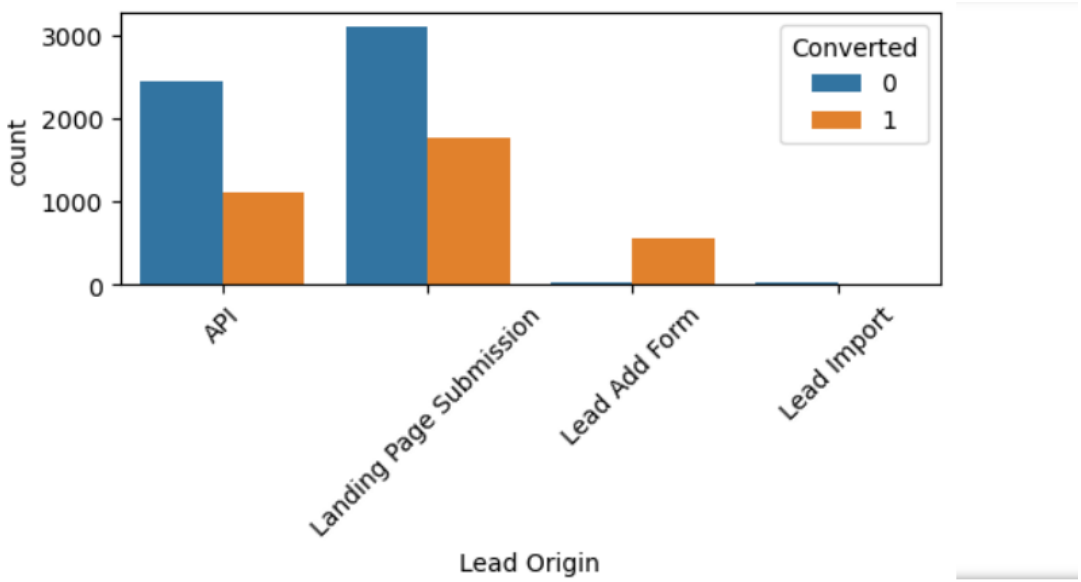
Steps followed

- Build logistic regression model that predicts conversion probability, higher is the probability higher are the chances of potential lead getting converted to conversion
- Followed below steps:
  - **Step 1: Importing and Merging Data**
  - **Step 2: Inspecting the Dataframe**
  - **Step 3: Data Preparation**
  - **Step 4: Data Visualisation**
  - **Step 5: Test-Train Split**
  - **Step 6: Feature Scaling**
  - **Step 7: Looking at Correlations**
  - **Step 8: Model Building**
  - **Step 9:Model evaluation**
  - **Step 10: Plotting the ROC Curve**
  - **Step 11: Finding Optimal Cutoff Point**
  - **Step 12: Making predictions on the test set**

Data Visualisation

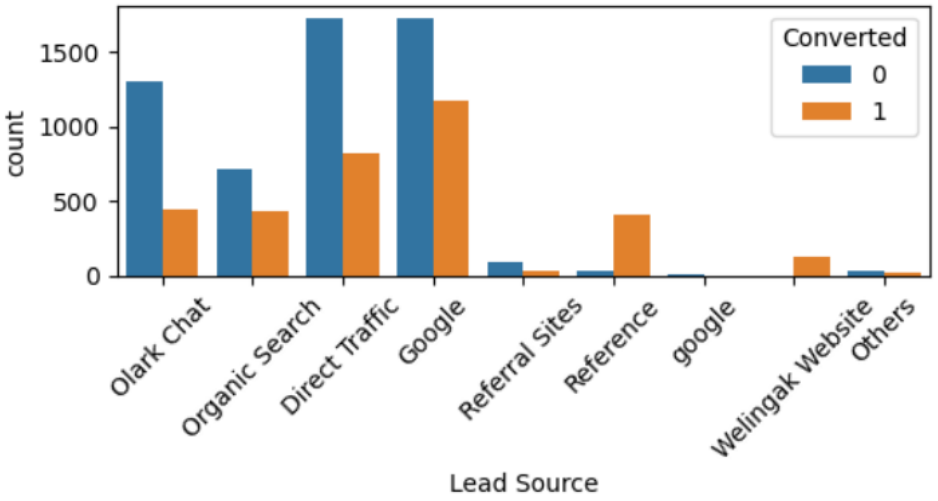
Data Visualisation

- Lead origin Variable:
  - most of the leads originate from API and landing page submission, but lead add form has highest conversion factor
  - we should work on improving the quality of leads coming from API and landing page submission



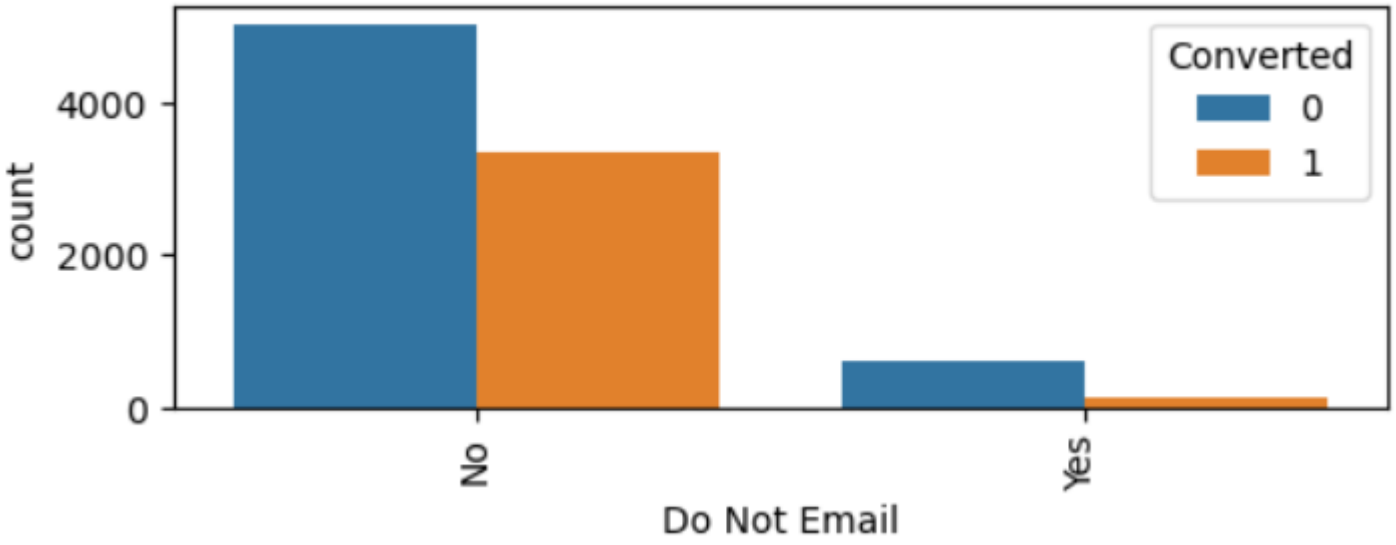
Lead Source

- most of leads are generated by google and direct traffic
- most of the volume is concentrated in top4 lead sources
- Reference has maximum conversion rate followed by wellingak website



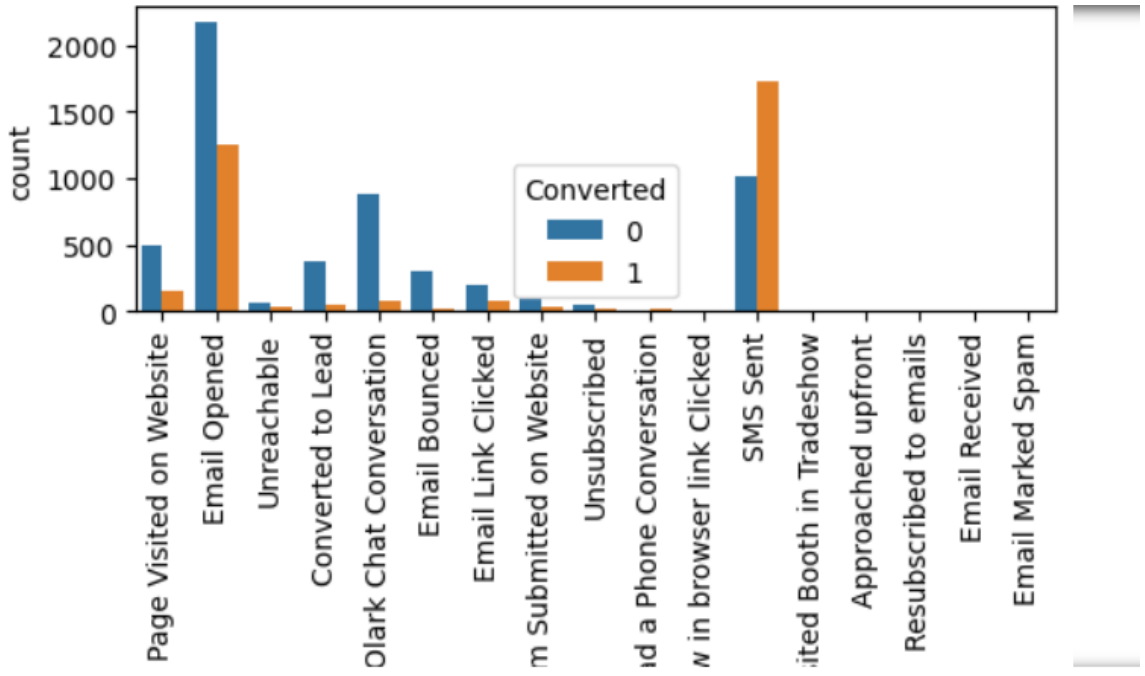
Do not email

- NO response has generated most of the leads and conversion rate is similar across these categories



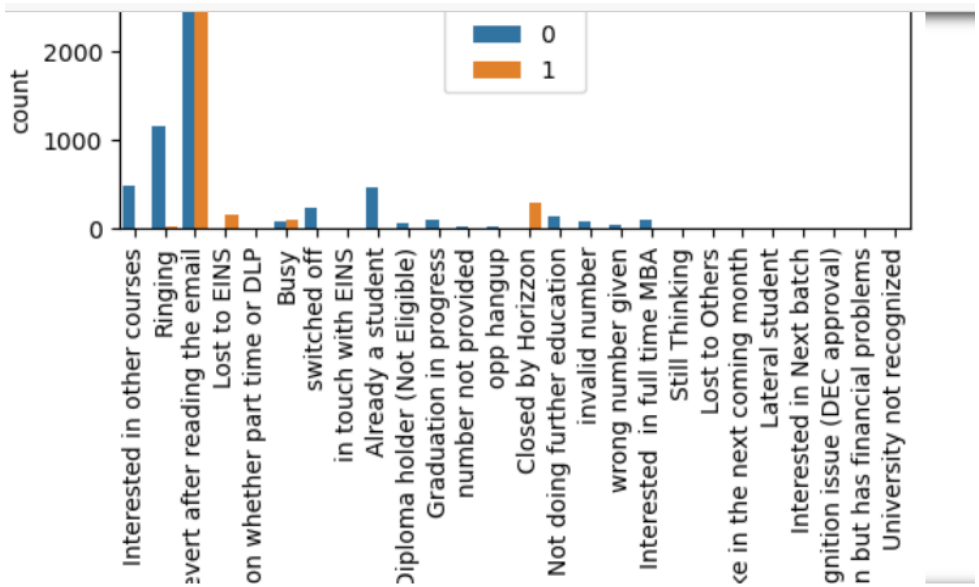
Last activity

- Customers whose lst activity is email_opened and sms sent has highest number of leads
- but the conversion rate in sms sent is highest



Tags

- most of the volume is in will revert after reading email category and it has very good conversion rate as well



Data Building and Evaluation

Model building

- First logistic model with all the variables: not so efficient model, multicollinearity

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.6869	1.562	-1.720	0.085	-5.748	0.374
Do Not Email	-1.5019	0.258	-5.815	0.000	-2.008	-0.996
TotalVisits	0.3410	0.062	5.486	0.000	0.219	0.463
Total Time Spent on Website	1.1149	0.051	21.835	0.000	1.015	1.215
Page Views Per Visit	-0.2966	0.069	-4.277	0.000	-0.433	-0.161
A free copy of Mastering The Interview	-0.2058	0.133	-1.550	0.121	-0.466	0.055
Lead Origin_Landing Page Submission	-0.8096	0.166	-4.890	0.000	-1.134	-0.485
Lead Origin_Lead Add Form	3.1835	0.818	3.891	0.000	1.580	4.787
Lead Origin_Lead Import	1.9445	0.852	2.282	0.022	0.275	3.614
Lead Source_Google	0.1616	0.142	1.140	0.254	-0.116	0.440
Lead Source_Olark Chat	0.8757	0.199	4.410	0.000	0.487	1.265
Lead Source_Organic Search	0.3029	0.160	1.896	0.058	-0.010	0.616

Optimized model with VIF<5 and P value<0.05

And based on the model output following are the top 3 variables:Tags_Closed by Horizzon, Tags_Lost to EINS,Tags_Will revert after reading the emai

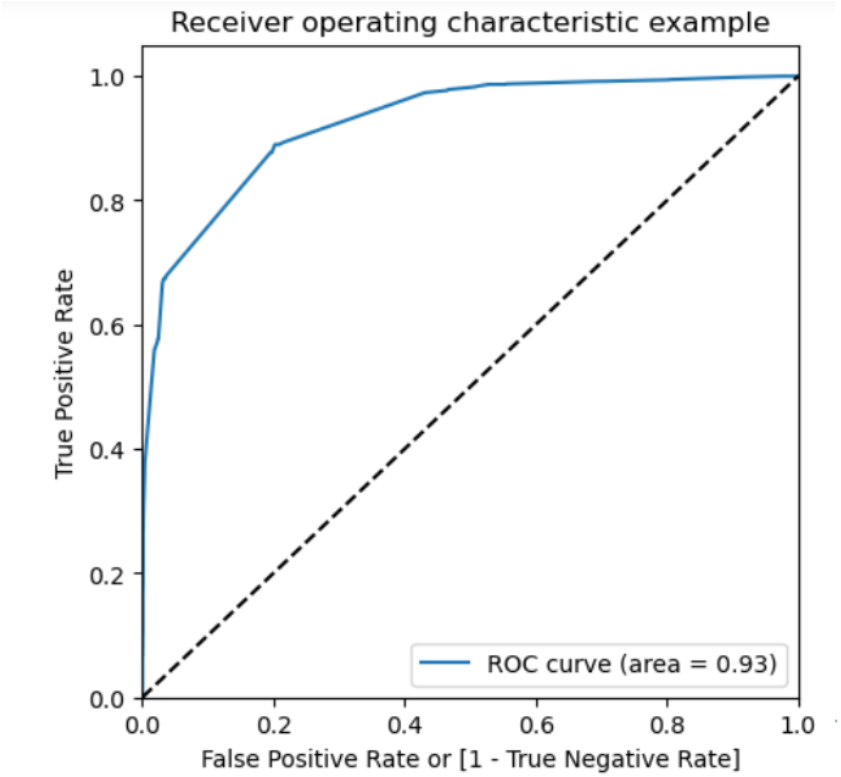
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2068.7
Date:	Sun, 16 Apr 2023	**Deviance:**	4137.5
Time:	17:33:51	**Pearson chi2:**	1.03e+04
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.4926
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.7858	0.190	-19.938	0.000	-4.158	-3.414
Do Not Email	-1.4895	0.179	-8.334	0.000	-1.840	-1.139
Lead Origin_Lead Add Form	2.5902	0.288	8.980	0.000	2.025	3.156
Lead Source_Welingak Website	1.7819	0.789	2.260	0.024	0.236	3.328
Last Activity_Had a Phone Conversation	1.9406	0.728	2.665	0.008	0.513	3.368
Specialization_Missing	-0.9713	0.082	-11.787	0.000	-1.133	-0.810
What is your current occupation_Working Professional	2.6031	0.233	11.177	0.000	2.147	3.060
Tags_Busy	2.9014	0.287	10.126	0.000	2.340	3.463
Tags_Closed by Horizzon	8.0042	0.740	10.823	0.000	6.555	9.454
Tags_Lost to EINS	8.1556	0.743	10.983	0.000	6.700	9.611
Tags_Ringing	-1.2711	0.298	-4.266	0.000	-1.855	-0.687
Tags_Will revert after reading the email	3.4754	0.191	18.169	0.000	3.101	3.850
Tags_switched off	-1.5894	0.622	-2.553	0.011	-2.809	-0.369
Last Notable Activity_SMS Sent	2.5752	0.106	24.355	0.000	2.368	2.782

	Features	VIF
10	Tags_Will revert after reading the email	1.89
1	Lead Origin_Lead Add Form	1.59
4	Specialization_Missing	1.50
12	Last Notable Activity_SMS Sent	1.49
2	Lead Source_Welingak Website	1.35
5	What is your current occupation_Working Profes...	1.21
7	Tags_Closed by Horizzon	1.16
9	Tags_Ringing	1.14
0	Do Not Email	1.05
6	Tags_Busy	1.04
11	Tags_switched off	1.03
3	Last Activity_Had a Phone Conversation	1.01
8	Tags_Lost to EINS	1.01

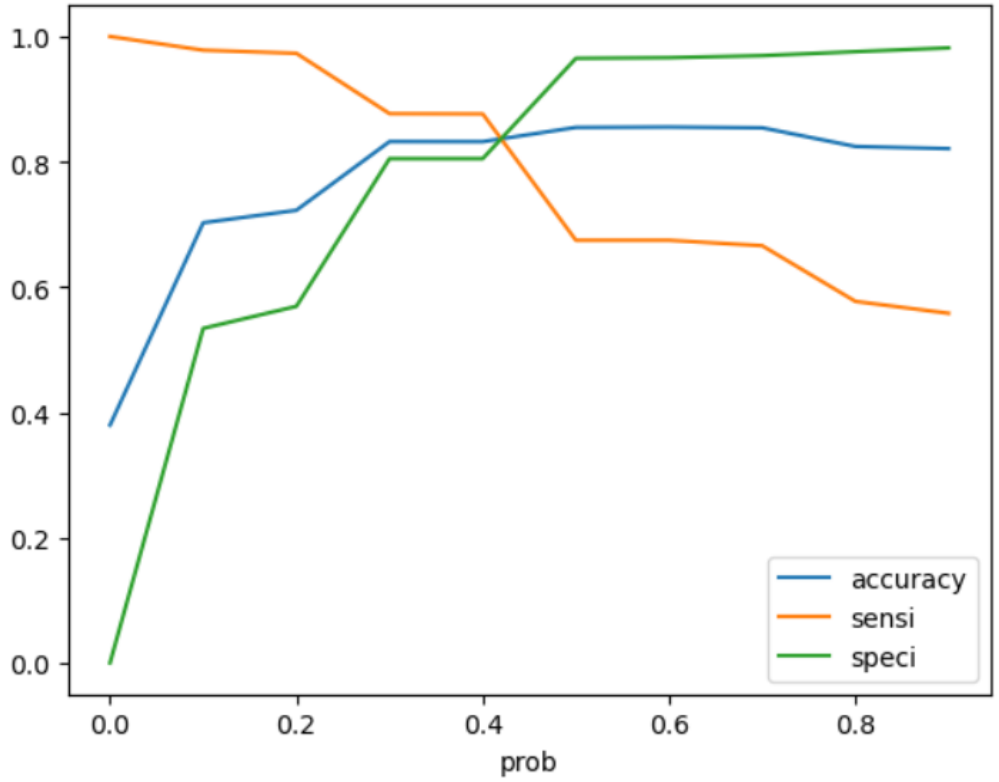
ROC curve

- area under the curve is 0.93, which is quite good
- earlier we have choosed the 0.5 threshold like this only, now we will find optimal cutoff



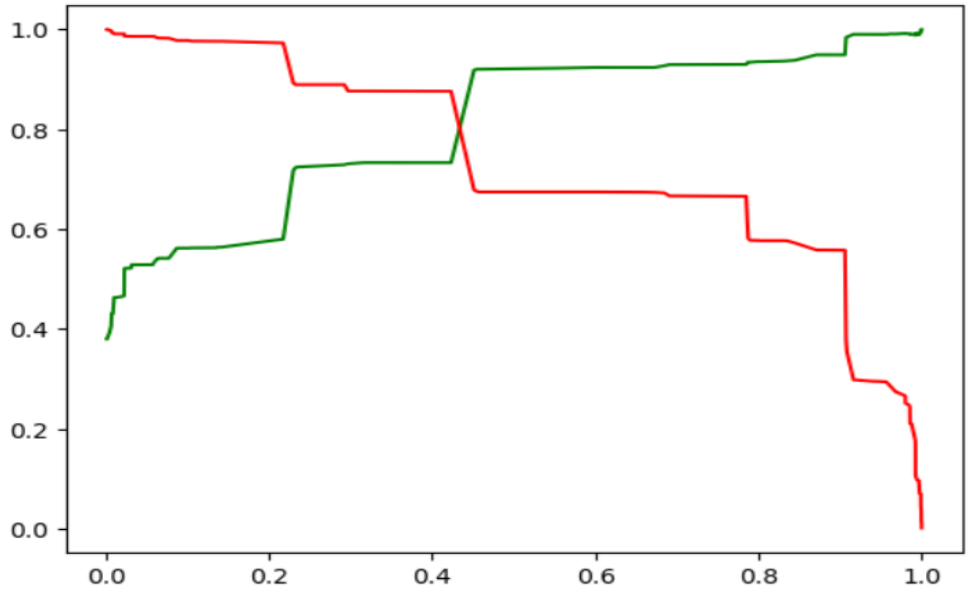
Accuracy, Sensitivity and Specificity

- From the curve below, 0.45 is the optimum point to take it as a cutoff probability



Precision and recall trade-off

- from above curve 0.45 is the optimum point to take as a cutoff probability using Precision-Recall



Classification report on test data

- out of all the predicted lead, 94% got converted F1 score 0.78 close to 1, shows model does a good job

	precision	recall	f1-score	support
0	0.82	0.98	0.89	1689
1	0.94	0.66	0.78	1042
accuracy			0.85	2731
macro avg	0.88	0.82	0.83	2731
weighted avg	0.87	0.85	0.85	2731