DSC 48 Lead Scoring Case Study

Anushka Saxena

Anjali S Gumne

Satya Ranjan Padhiary

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%.

Business Goal

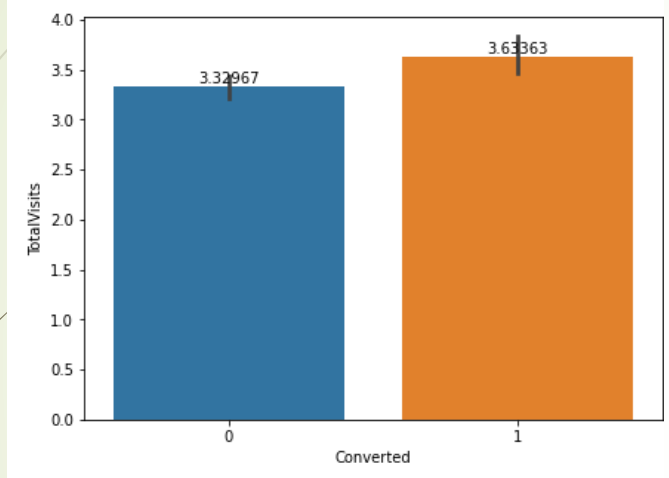
- To select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- Build a model wherein a lead score needs to be assigned to each lead such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The target lead conversion rate has been set to be around 80%.

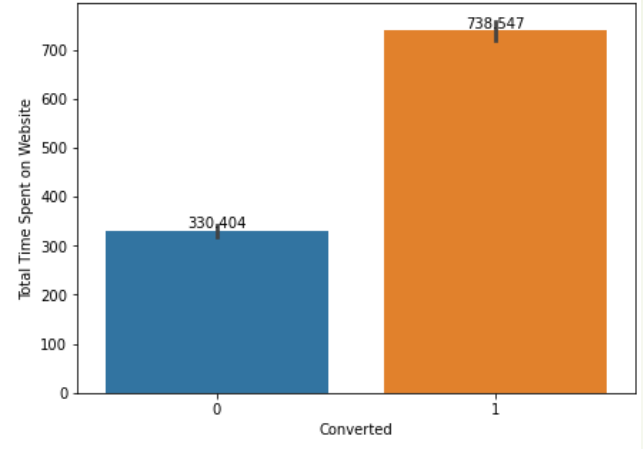
Methodology

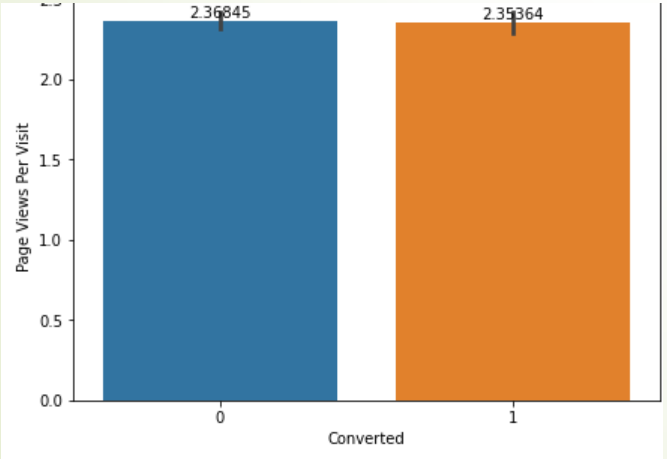
- Data Cleaning
- Exploratory Data Analysis
  - Identifying Data Imbalance
  - Univariate Analysis
  - Bivariate/Multivariate Analysis
- Dividing the data into train and test set
- Feature Scaling of the numerical variables
- Building a model using Logistic Regression and choosing the best one based on VIF, and p values
- Evaluation of the model on train set using metrics such as Accuracy,Sensitivity,Specificity, ROC Curve or Precision and Recall
- Determining the optimal threshold level of probability based on Sensitivity-Specificity trade off or Precision-Recall trade off
- Evaluating the model on the train and test set on the optimal threshold using metrics described above and determining the lead score to determine whether the goal of 80% lead conversion rate has been achieved.

Insights from Exploratory Data Analysis

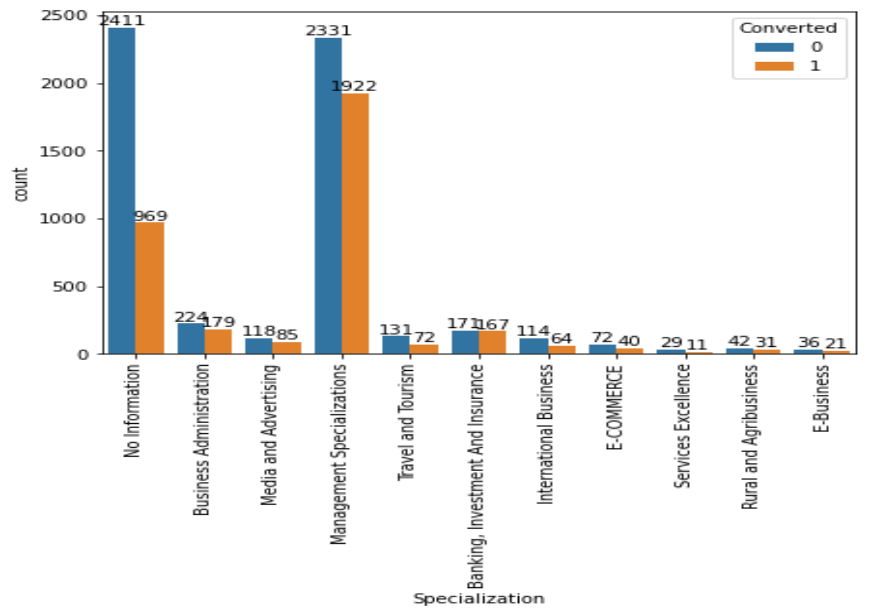
Total Visits, Total Time Spent on Visits, and Page Views Per Visit vs Leads Converted

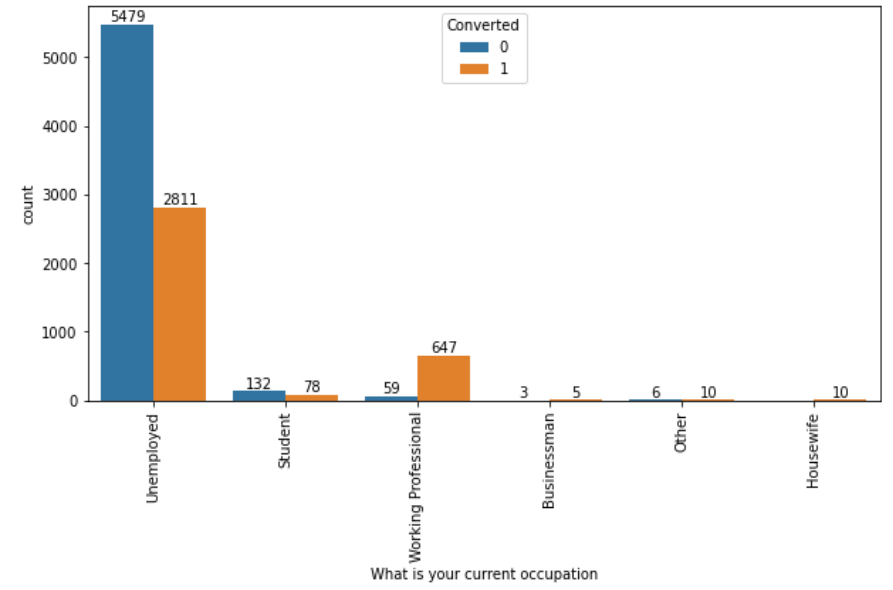






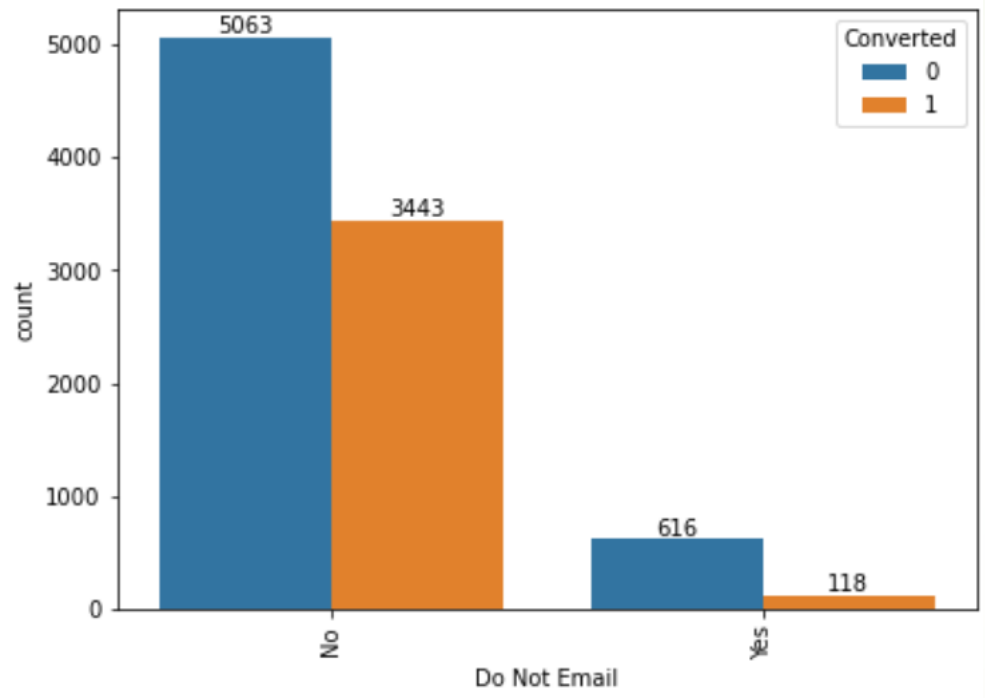
Specialization, Occupation vs Leads Converted





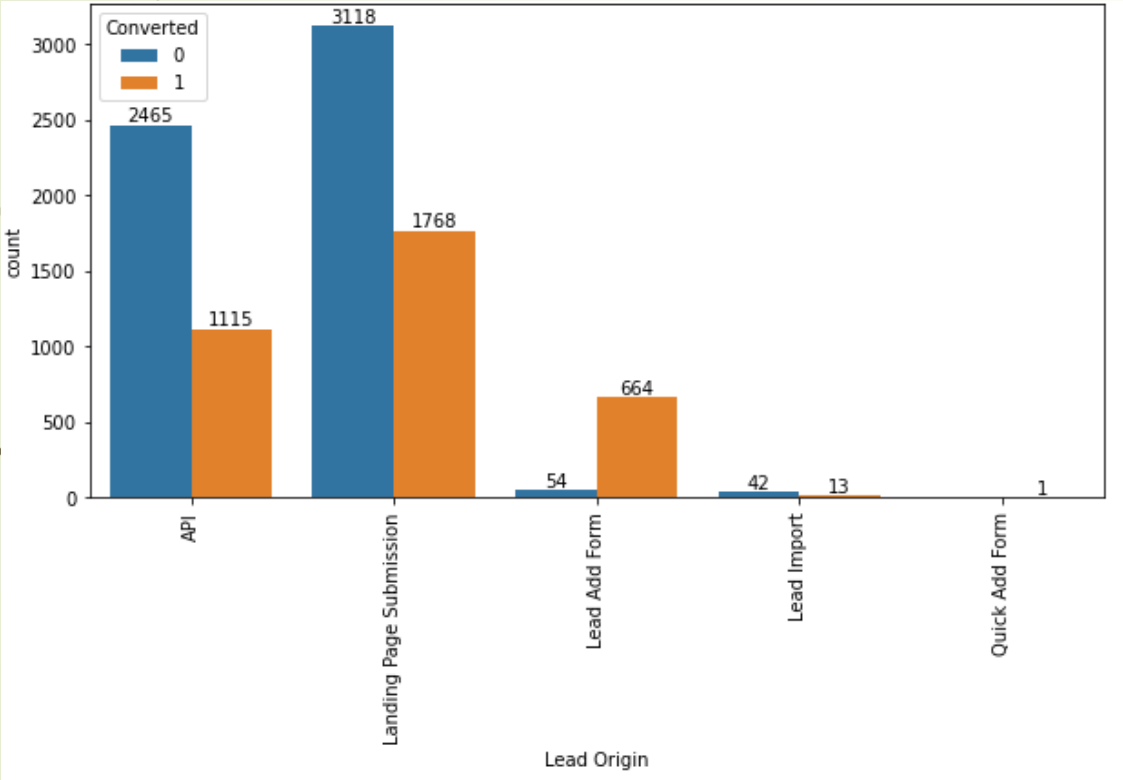
- In the case of Specialization maximum conversion has been observed in the case of Management Specialization
- In the case of Occupation maximum conversion has been observed in the case of Unemployed

Do Not Email vs Leads Converted



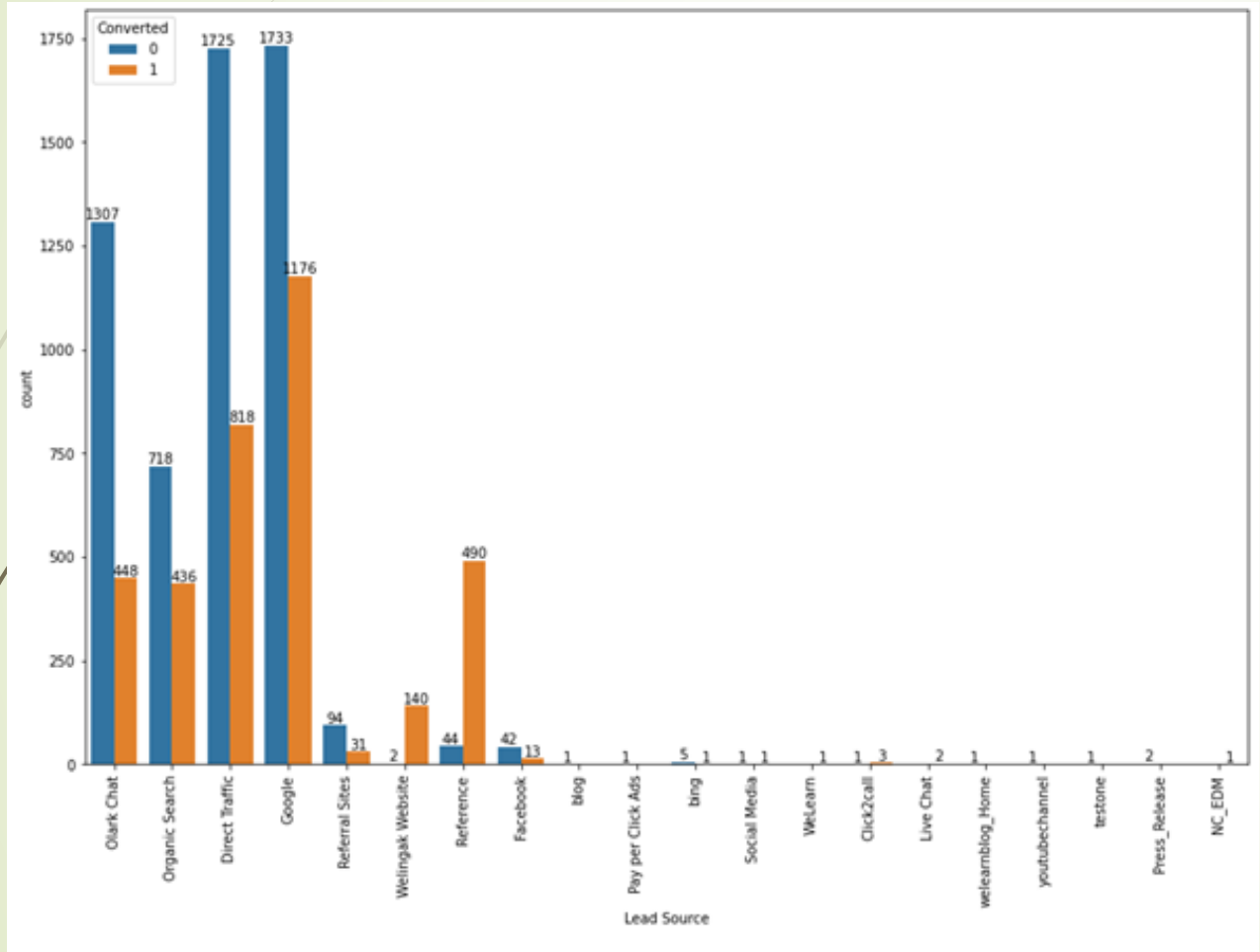
- In the case of Do Not Email maximum conversion has been observed in the case of No

Lead Origin vs Leads Converted



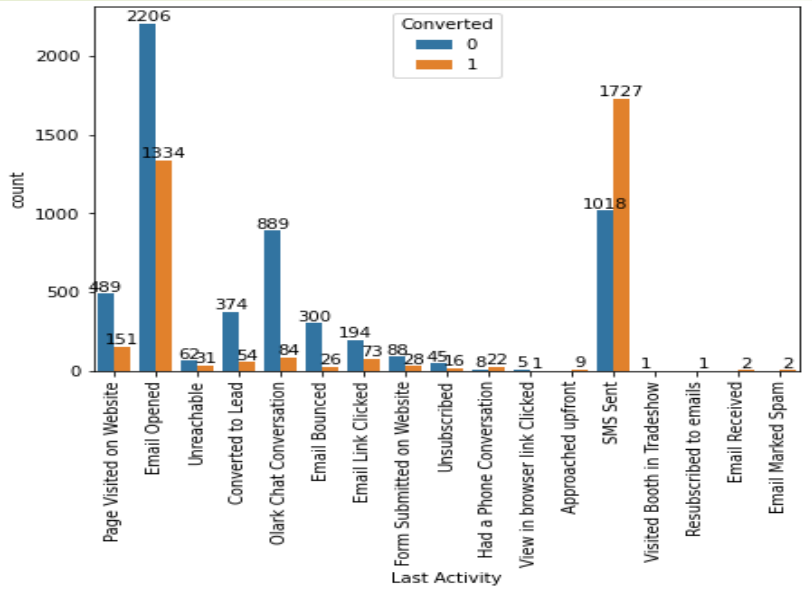
In the case of Lead Origin maximum conversion has been observed in the case of Landing Page Submission.

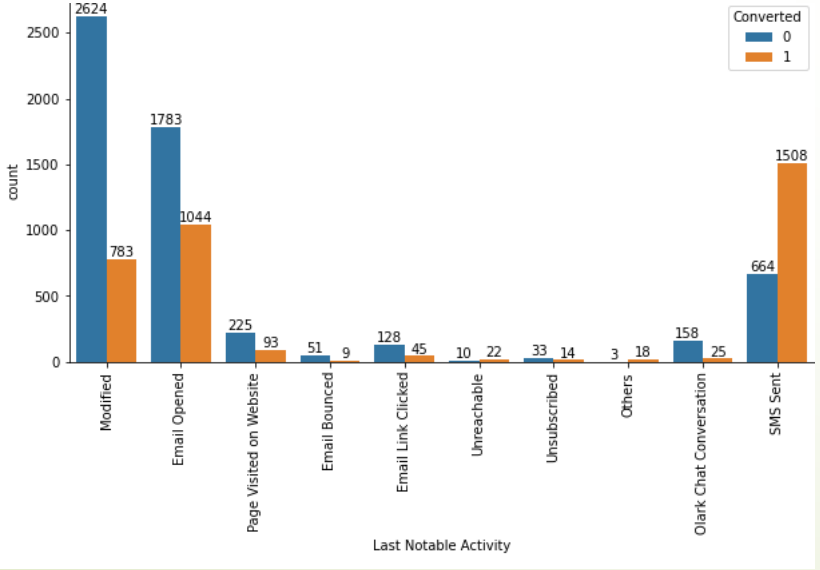
Lead Source vs Leads Converted



In the case of Lead Source maximum conversion has been observed in the case of Google

Last Activity, Last Notable Activity vs Leads Converted

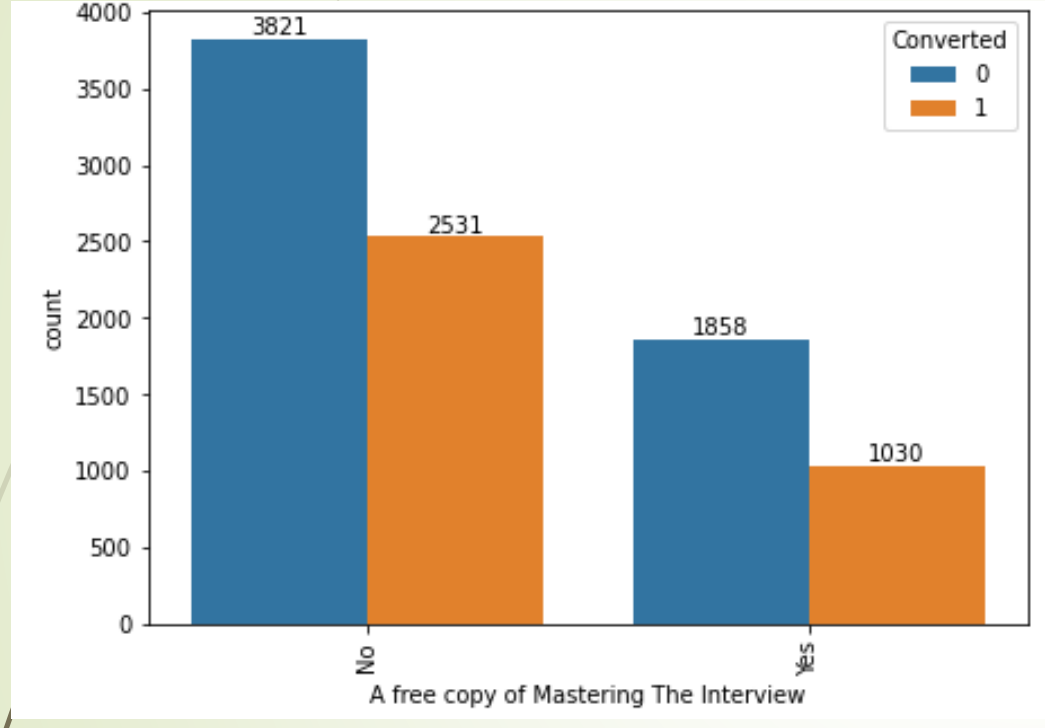




In the case of Last Activity maximum conversion has been observed in the case of SMS Sent

In the case of Last Notable Activity maximum conversion has been observed in the case of SMS Sent

A free copy of mastering the interview vs Leads Converted



In the case of A free copy of mastering the interview maximum conversion has been observed in the case of No

Regression Analysis-Results

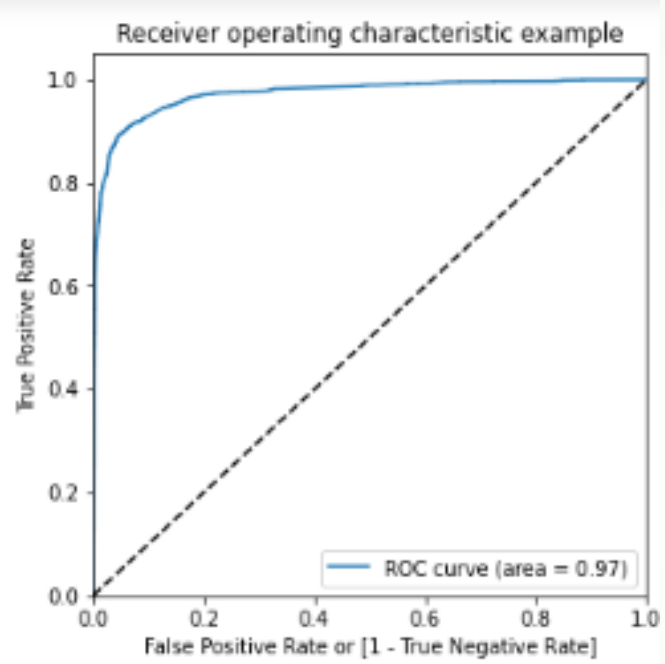
Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6447
Model Family:	Binomial	Df Model:	20
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1259.1
Date:	Tue, 21 Mar 2023	Deviance:	2518.1
Time:	10:04:44	Pearson chi2:	1.08e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.6094
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4387	0.267	-5.398	0.000	-1.961	-0.916
TotalVisits	6.2833	1.776	3.537	0.000	2.802	9.765
Total Time Spent on Website	4.3093	0.246	17.513	0.000	3.827	4.792
LeadOrigin_Landing Page Submission	-0.7983	0.134	-5.955	0.000	-1.061	-0.536
LeadOrigin_Lead Add Form	1.1354	0.379	2.994	0.003	0.392	1.879
LeadSource_Olark Chat	0.7618	0.164	4.633	0.000	0.439	1.084
LeadSource_Welingak Website	4.0108	0.820	4.893	0.000	2.404	5.618
DoNotEmail_Yes	-0.6762	0.238	-2.845	0.004	-1.142	-0.210
LastActivity_Page Visited on Website	-0.6028	0.243	-2.480	0.013	-1.079	-0.126
LastActivity_SMS Sent	1.9544	0.118	16.523	0.000	1.723	2.186
Whatisyourcurrentoccupation_Working Professional	0.8386	0.369	2.274	0.023	0.116	1.561
Tags1_Closed by Horizzon	6.7275	1.047	6.428	0.000	4.676	8.779
Tags1_Interested in other courses	-2.4854	0.425	-5.843	0.000	-3.319	-1.652
Tags1_Lost to EINS	5.6726	0.768	7.389	0.000	4.168	7.177
Tags1_Not Specified	-0.6070	0.236	-2.569	0.010	-1.070	-0.144
Tags1_Other_Tags	-2.8846	0.298	-9.674	0.000	-3.469	-2.300
Tags1_Ringing	-3.9524	0.314	-12.606	0.000	-4.567	-3.338
Tags1_Will revert after reading the email	3.8703	0.293	13.226	0.000	3.297	4.444
LastNotableActivity_Email Link Clicked	-1.3266	0.455	-2.914	0.004	-2.219	-0.434
LastNotableActivity_Modified	-1.6697	0.125	-13.384	0.000	-1.914	-1.425
LastNotableActivity_Olark Chat Conversation	-1.7145	0.434	-3.947	0.000	-2.566	-0.863

Variables that have an impact on conversion rate of leads:

- Total Visits
- Total Time Spent on Website
- Lead Origin-Landing Page Submission
- Lead Origin-Lead Add Form
- Lead Source-Welingak Website
- Do Not Email-Yes
- Last Activity-Page Visited on Website
- Occupation-Working Professional
- Tags-Closed by Horizon
- Tags-Lost to ENIS
- Tags-Not Specified
- Tags-Other Tags
- Tags-Ringing
- Tags-Will Revert after reading the mail
- Last Notable Activity-Email Link Clicked
- Last Notable Activity-Modified
- Last Notable Activity-Olark Chat Conversation

Model Evaluation-Train Set

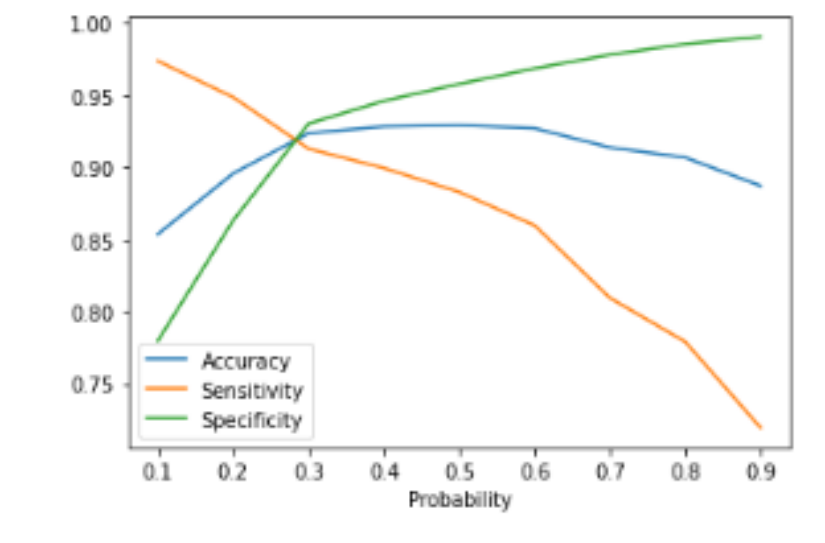


Confusion Matrix(at probability=0.5)

Actual/ Predicted	Not Converted	Converted
Not Converted	3833	169
Converted	289	2177

Accuracy-92.92%
Sensitivity-88.28%
Specificity-95.77%
Precision-92.79%
Recall-88.28%

Model Evaluation(Sensitivity-Specificity Trade Off)-Train Set



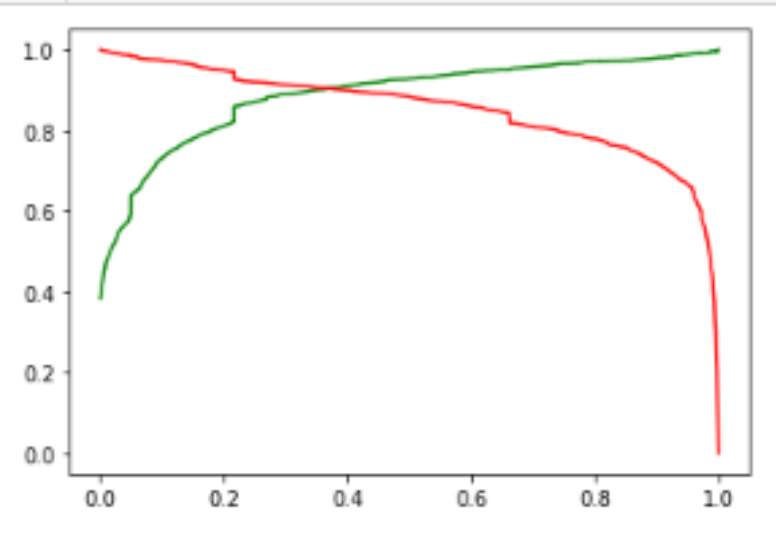
The optimal probability threshold according to the sensitivity-specificity trade off is 0.3.

Confusion Matrix(at optimal probability threshold)

Actual/ Predicted	Not Converted	Converted
Not Converted	3723	279
Converted	215	2251

Accuracy-92.36%
Sensitivity-91.28%
Specificity-93.03%
Precision-88.97%
Recall-91.28%
Lead Conversion Rate-91.28%

Model Evaluation(Precision-Recall)-Train Set



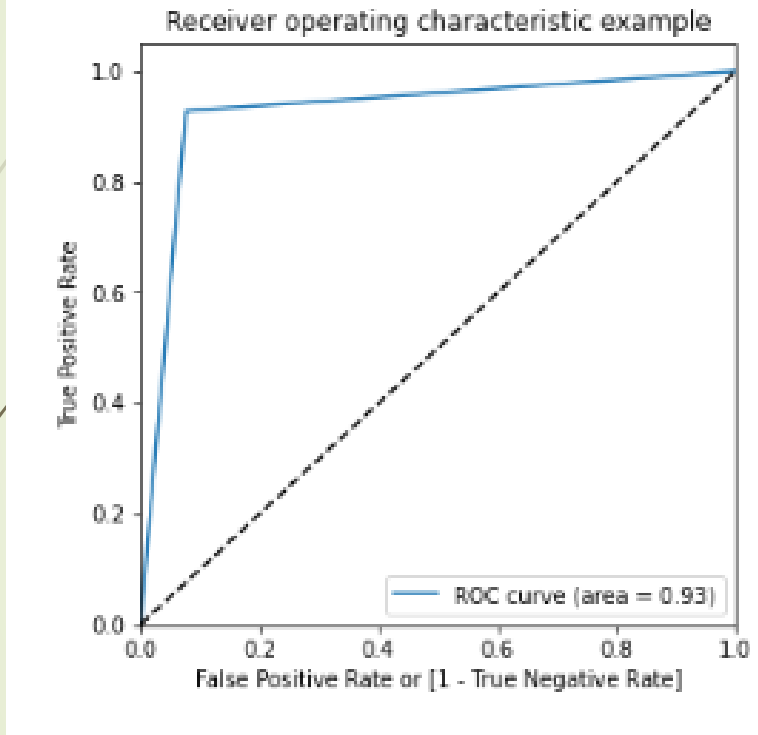
The optimal probability threshold according to the precision-recall is 0.38.

Confusion Matrix(at optimal probability threshold)

Actual/ Predicted	Not Converted	Converted
Not Converted	3769	233
Converted	243	2223

Accuracy-92.64%
Sensitivity-90.15%
Specificity-94.17%
Precision-90.51%
Recall-90.15%

Model Evaluation(Sensitivity-Specificity Trade Off)-Test Set



Confusion Matrix(at optimal probability threshold)

Actual/ Predicted	Not Converted	Converted
Not Converted	1549	128
Converted	79	1016

Accuracy-92.53%
Sensitivity-92.79%
Specificity-92.37%
Precision-88.81%
Recall-92.79%
Lead Conversion Rate-92.79%

Conclusions and Implications

- Considered the sensitivity-specificity trade off for the optimal probability threshold(0.3) for the test data.

Metrics	Train test	Test Set
AUC of ROC	0.97	0.93
Accuracy	92.36%	92.53%
Sensitivity	91.28%	92.79%
Specificity	93.03%	92.37%
Precision	88.97%	88.81%
Recall	91.28%	92.79%
Lead Conversion Rate	91.28%	92.79%

The goal of a lead conversion rate has been met with a lead conversion rate of 91.28% for the train set and 92.79% for the test set.

Conclusions and Implications

- The top three variables which contribute most towards the probability of a lead getting converted include:
    - Closed by Horizzon (from Tags) (coefficient 6.7275)
    - Total Visits (coefficient 6.2833)
    - Lost to ENIS (from Tags) (coefficient 5.6726)
- The top three categorical/dummy variables which contribute most towards the probability of a lead getting converted include:
    - Closed by Horizzon (from Tags) (coefficient 6.7275)
    - Lost to ENIS (from Tags) (coefficient 5.6726)
    - Welingak Website (from Lead Source) (coefficient 4.010)
- Other important variables consist of Total Time Spent on the Website, Tags(Ringing),Tags(Will revert after reading the mail),Tags(Interested in other courses),Last Activity(SMS Sent),Lead Origin(Lead Add Form),Lead Source(Olark Chat),Occupation(Working Professional),Last Activity(SMS Sent),Last Notable Activity(Olark Chat Conversation)