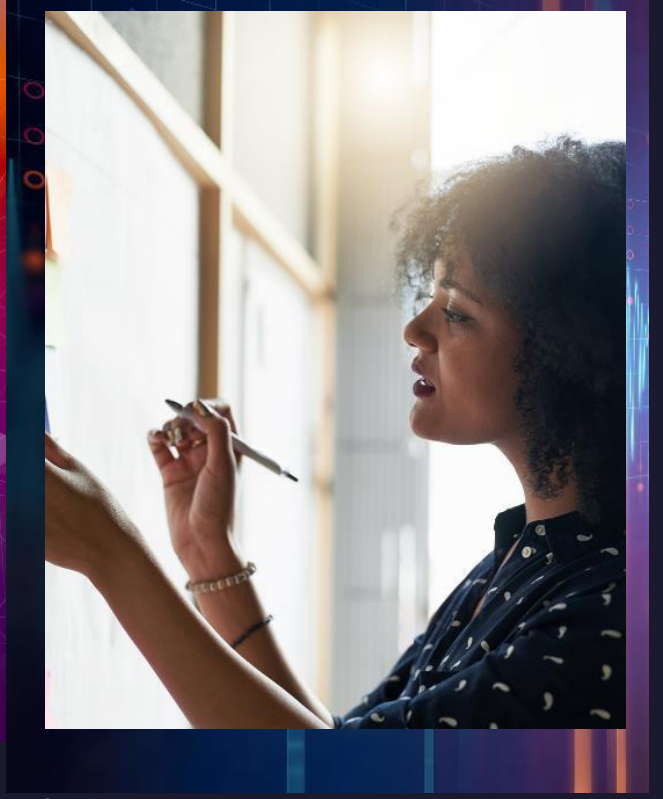
LEAD SCORING CASE STUDY

Business Objective

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.

Social Media Marketing
Selection of Hot Leads
Initial Pool of leads Nurturing
Converted Leads

Goals of the Case Study



1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

- .

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. A typical lead conversion process can be represented using the following funnel:

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads 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 CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

METHODOLOGY

- To build a Logistic Regression model that assigns lead scores to all leads such that the customers with higher lead score have a higher conversion chance and vice versa. Target Lead Conversion Rate ≈ 80%

importing libraries
importing libraries

Reading and Understanding the Data

Importing Libraries

Exploratory Data Analysis

Data Preparation

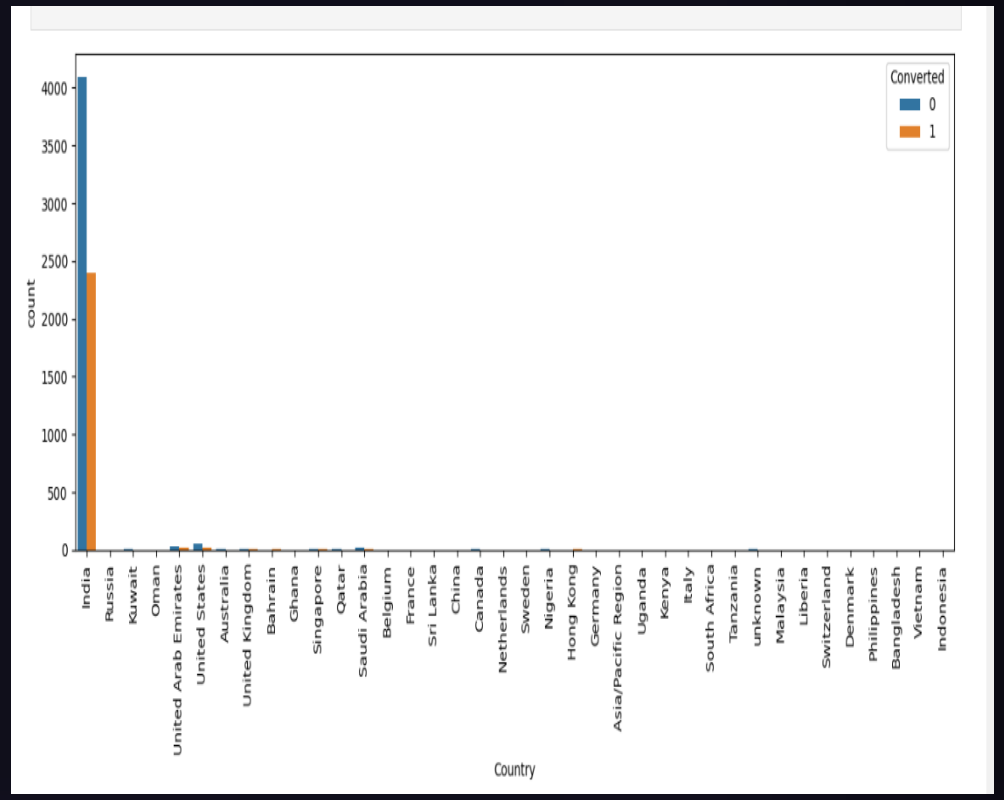
Scaling of Data

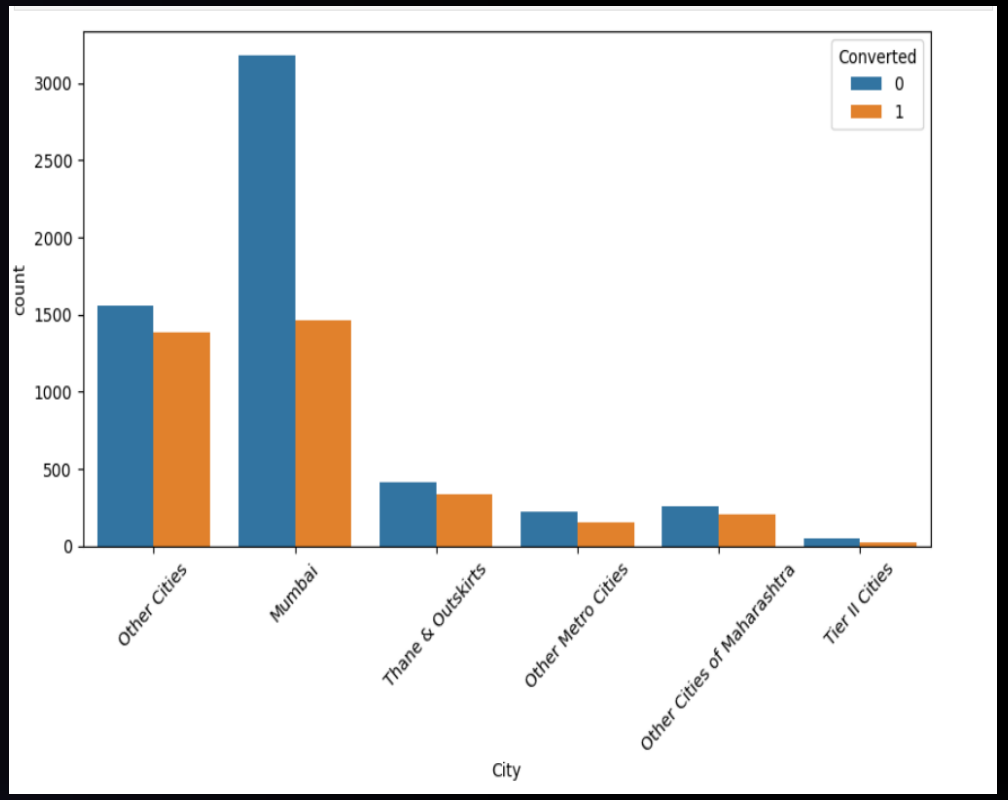
Model Building

Model Evaluation

Categorical Attributes Analysis And Null value treatments

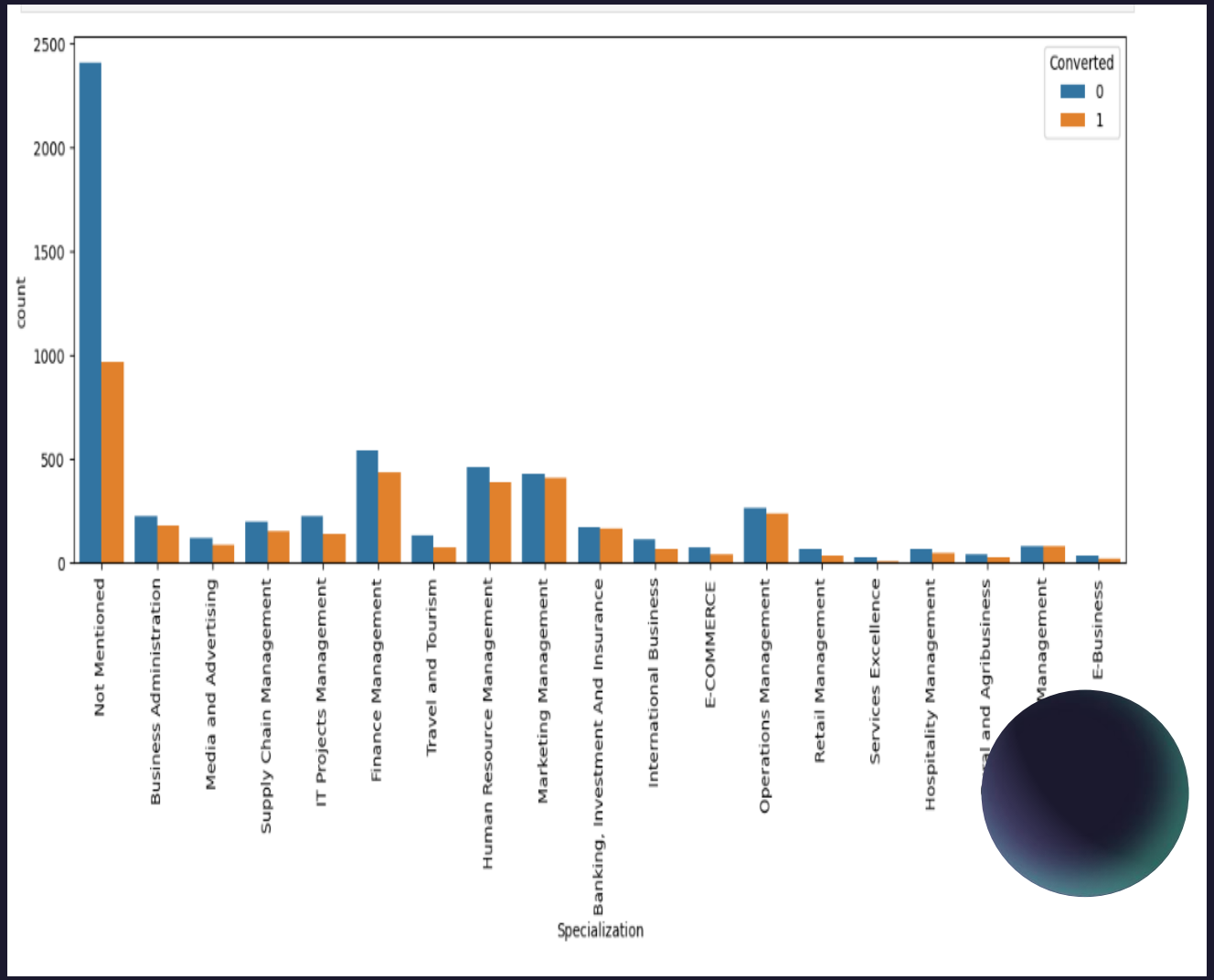
As we can see the Number of Values for India are quite high (approx. 95%+ of the Data), this column can be dropped





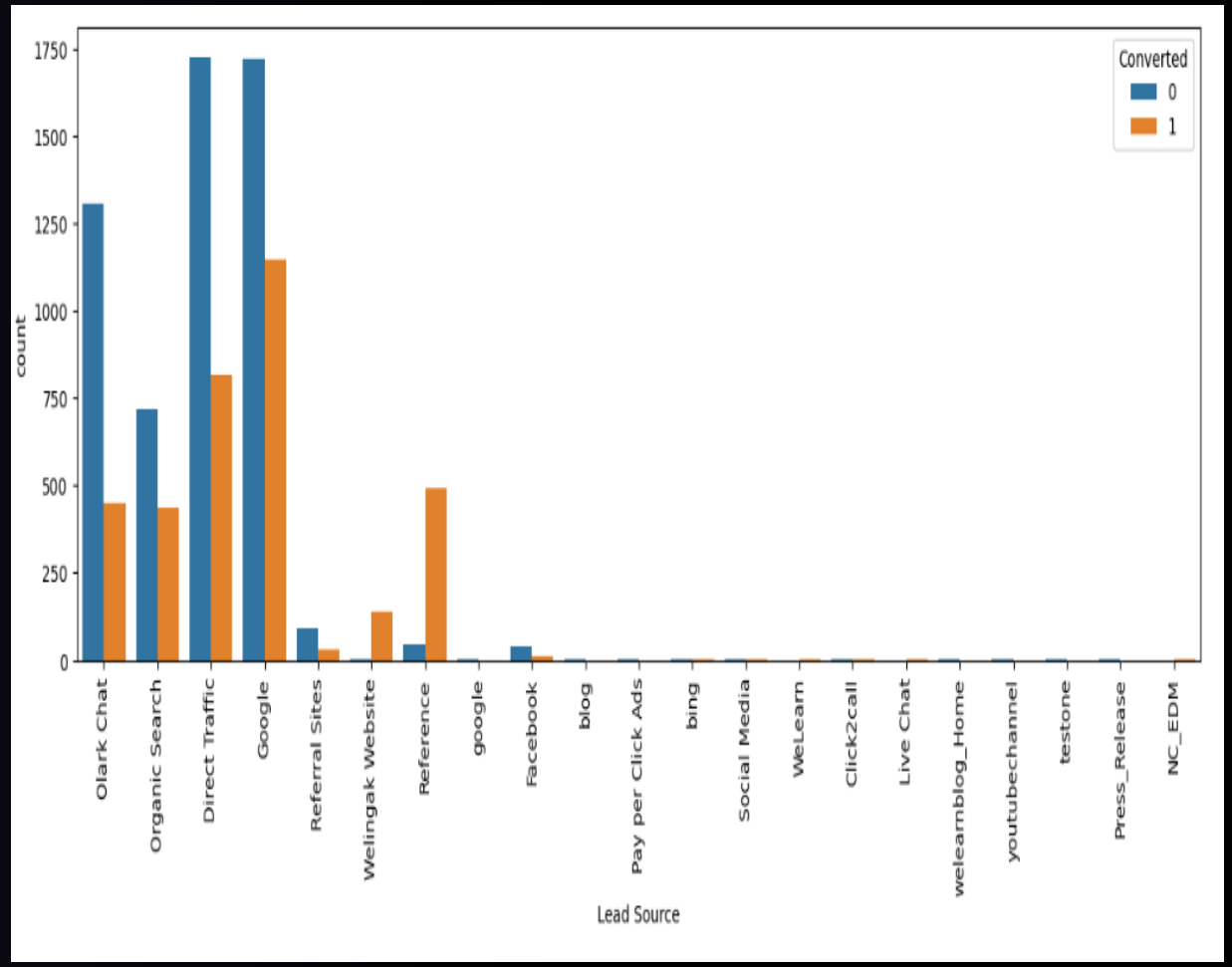
Categorical Attributes Analysis And Null value treatments

We observe that specializations with a focus on management exhibit a notably higher number of generated leads, and a substantial proportion of these leads are successfully converted. Consequently, it becomes evident that this 'Specialization' variable holds significant importance and should not be considered for elimination.



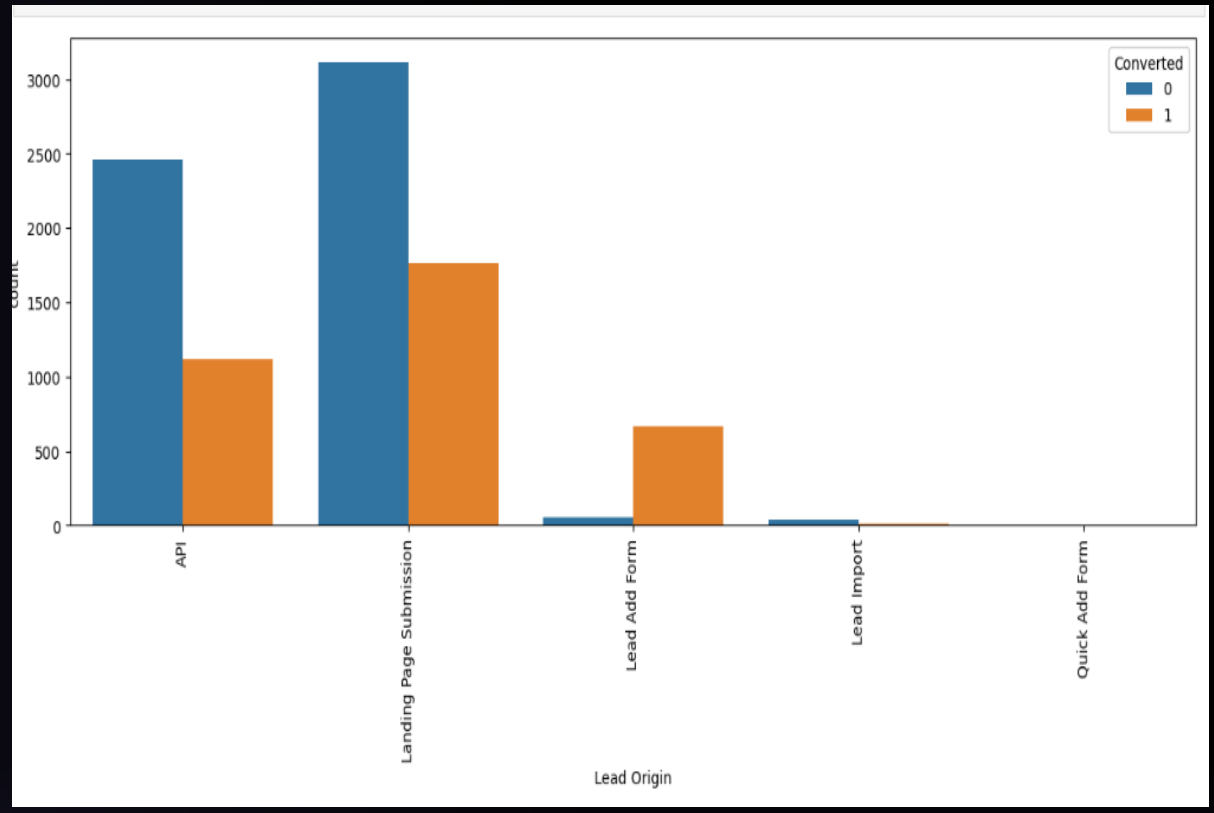
Categorical Attributes Analysis And Null value treatments

- Google and Direct traffic sources generate the highest number of leads.
- The conversion rate for reference leads and leads through the Welingak website is notably high.
- To enhance the overall lead conversion rate, prioritize improving the conversion rates for Olark chat, organic search, direct traffic, and Google leads.
- Additionally, aim to increase lead generation from reference sources and the Welingak website.

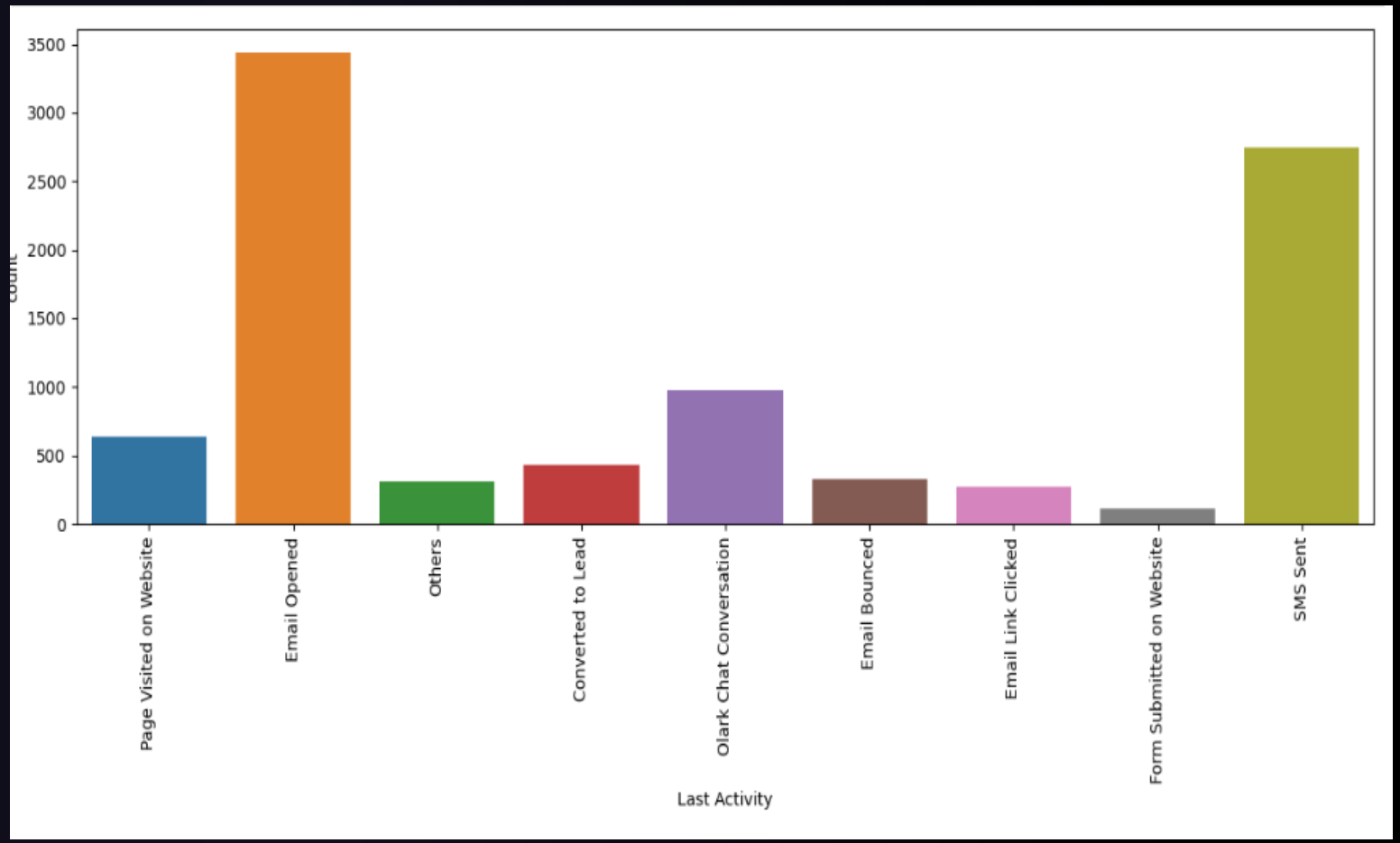


Categorical Attributes Analysis And Null value treatments

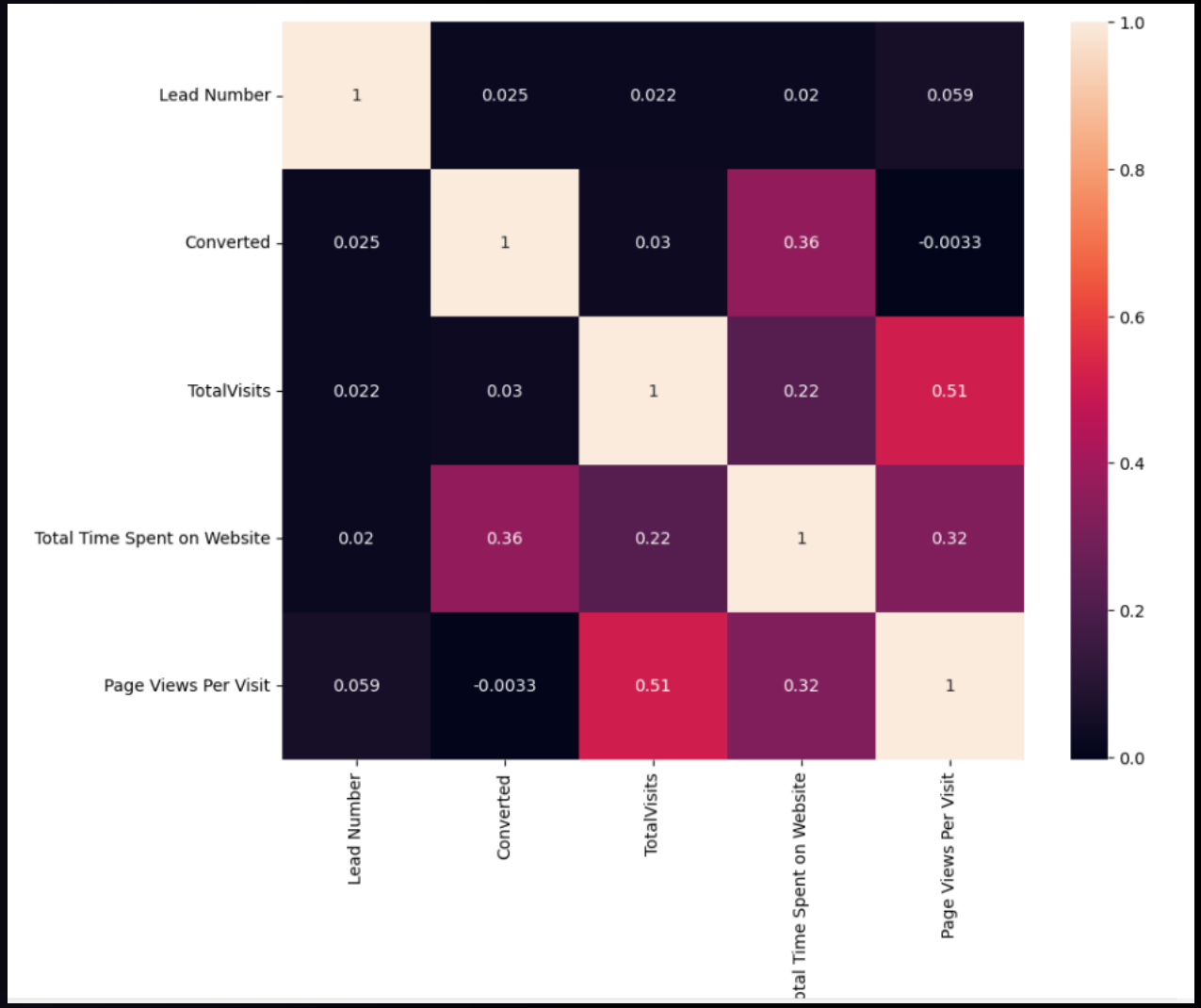
- API and Landing Page Submission yield a high number of leads and conversions.
- Lead Add Form has a remarkable conversion rate, even though the total lead count is not very high.
- Lead Import and Quick Add Form generate comparatively fewer leads.
- To enhance the overall lead conversion rate, focus on improving the conversion of leads from API and Landing Page Submission sources.
- Prioritize efforts to generate more leads through the Lead Add Form.



Email Opened and SMS sent are 2 most activity seen by the leads.



Numerical Attributes Analysis and Null value treatment



Scaling of Data:

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
4798	0.058824	0.141535	0.111111
6733	0.117647	0.732172	0.111111
5301	0.176471	0.043005	0.166667
6251	0.000000	0.000000	0.000000
445	0.294118	0.360915	0.555556

Model Building using Stats Model & RFE:

	const	Total Time Spent on Website	Lead Origin_Lead Add Form	Tags_Already a student	Tags_Closed by Horizzon	Tags_Diploma holder (Not Eligible)	Tags_Interested in full time MBA	Tags_Interested in other courses	Tags_Lost to EINS	Tags_Not doing further education
4798	1.0	0.141535	0	0	0	0	0	0	0	0
6733	1.0	0.732172	0	0	0	0	0	0	0	0
5301	1.0	0.043005	0	0	0	0	0	0	0	0
6251	1.0	0.000000	0	0	0	0	0	0	0	0
445	1.0	0.360915	0	0	0	0	0	0	0	0

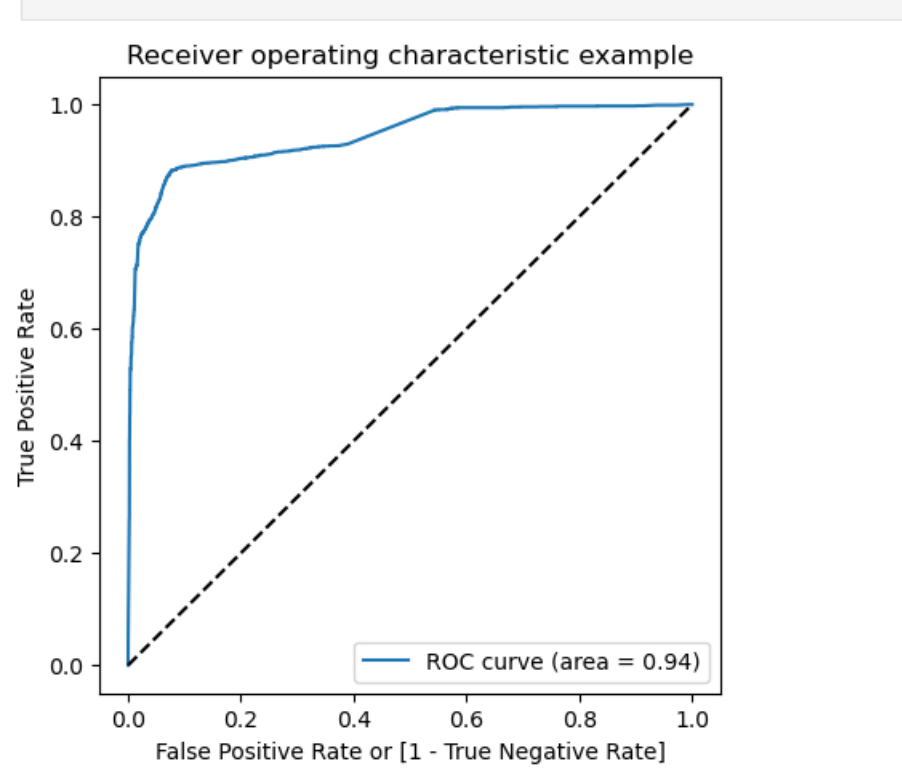
OLS Regression Results

Dep. Variable:	Converted	**R-squared:**	0.649
Model:	OLS	**Adj. R-squared:**	0.648
Method:	Least Squares	**F-statistic:**	871.0
Date:	Wed, 11 Oct 2023	**Prob (F-statistic):**	0.00
Time:	01:09:43	**Log-Likelihood:**	-1198.1
No. Observations:	7090	**AIC:**	2428.
Df Residuals:	7074	**BIC:**	2538.
Df Model:	15		
Covariance Type:	nonrobust		

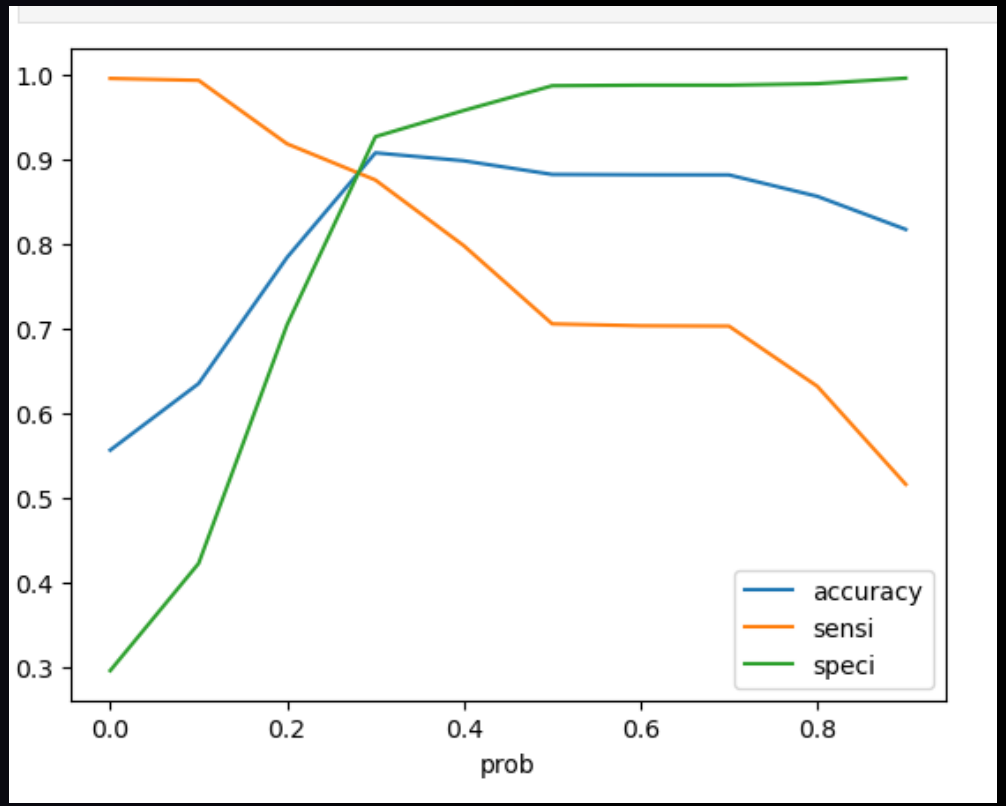
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1792	0.006	29.765	0.000	0.167	0.191
Total Time Spent on Website	0.3280	0.013	24.983	0.000	0.302	0.354
Lead Origin_Lead Add Form	0.2906	0.015	19.149	0.000	0.261	0.320
Tags_Already a student	-0.2299	0.016	-14.385	0.000	-0.261	-0.199
Tags_Closed by Horizzon	0.5811	0.021	27.877	0.000	0.540	0.622
Tags_Diploma holder (Not Eligible)	-0.2226	0.041	-5.441	0.000	-0.303	-0.142
Tags_Interested in full time MBA	-0.2160	0.030	-7.111	0.000	-0.276	-0.156
Tags_Interested in other courses	-0.2251	0.015	-14.530	0.000	-0.255	-0.195
Tags_Lost to EINS	0.6473	0.026	24.976	0.000	0.596	0.698
Tags_Not doing further education	-0.2437	0.028	-8.745	0.000	-0.298	-0.189
Tags_Ringing	-0.2301	0.011	-21.377	0.000	-0.251	-0.209
Tags_Will revert after reading the email	0.6117	0.010	63.016	0.000	0.593	0.631
Tags_invalid number	-0.2240	0.035	-6.456	0.000	-0.292	-0.156
Tags_number not provided	-0.2705	0.059	-4.601	0.000	-0.386	-0.155
Tags_switched off	-0.2301	0.022	-10.687	0.000	-0.272	-0.188
Tags_wrong number given	-0.2384	0.050	-4.747	0.000	-0.337	-0.140

Omnibus:	1457.820	**Durbin-Watson:**	1.985
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3195.610
Skew:	1.184	**Prob(JB):**	0.00
Kurtosis:	5.283	**Cond. No.**	18.6

The ROC Curve should be a value close to 1. We are getting a good value of 0.94 indicating a good predictive model.

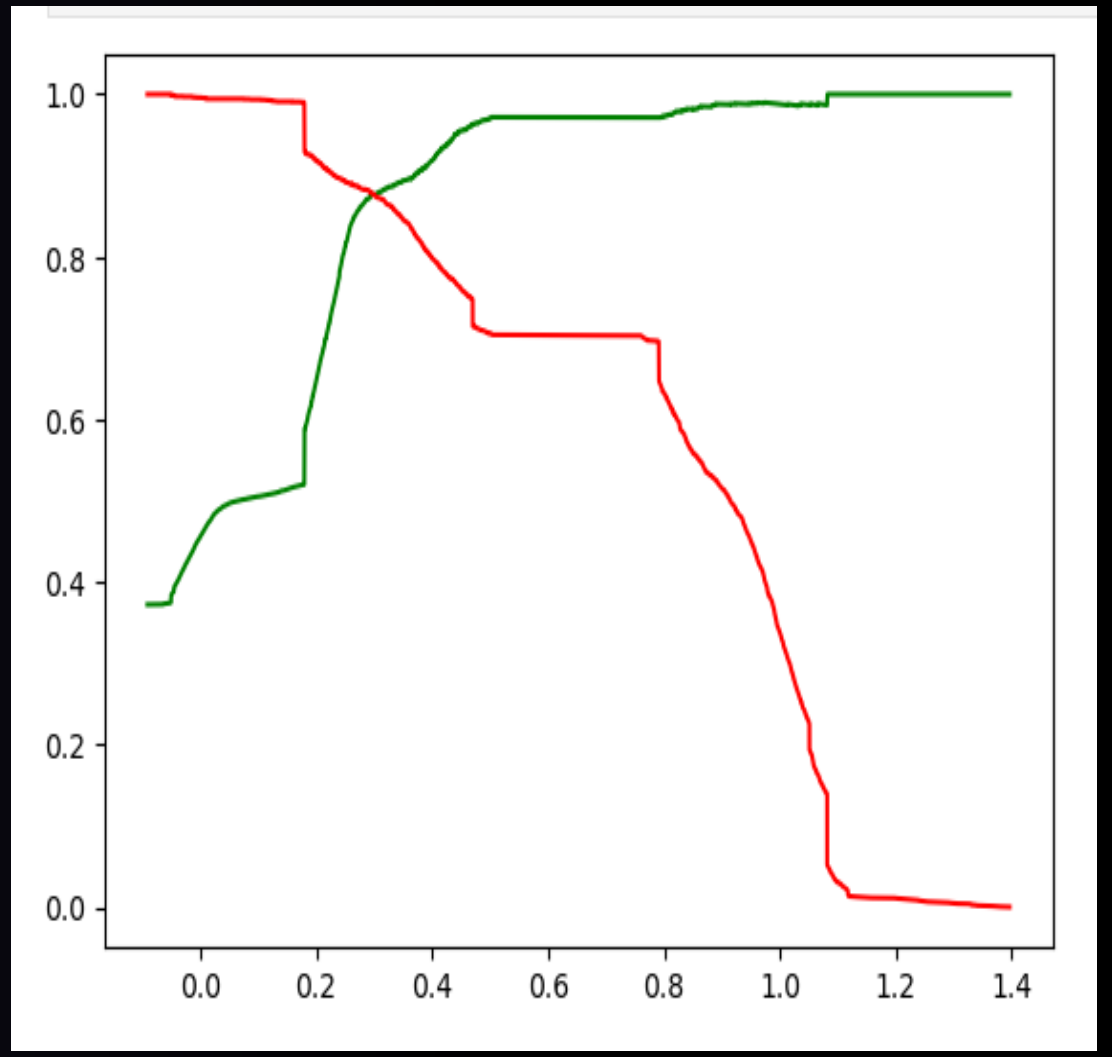


The below graph display an approximate cutoff at 0.3



As we can see the model seems to be performing well. The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data 1.66%

- Accuracy : 90.78%
- Sensitivity :70.56%
- Specificity: 98.71%



Modal evaluation

- After running the model on the Test Data these are the figures we obtain:

Accuracy: 91.03%

Sensitivity:88.20%

Specificity:92.85%

Modal evaluation

Train Data
- Accuracy : 90.78%
- Sensitivity :70.56
- Specificity: 98.7

Test Data
- Accuracy: 91.03%
- Sensitivity:88.20
- Specificity:92.85%1%

Conclusion

While we have checked both sensitivity-specificity as well as Precision & recall metrics, we have considered the optimal cut off based on sensitivity & specificity for calculating the final prediction

Accuracy, Sensitivity & specificity values of test set are around 91%,88%,92% which are approximately closer to Values calculated using Trained Data Set Lead Score Calculated for the conversion rate final model on Train & Test dataset is 90.78 % 70.56 % & 98.7 respectively. Hence, Overall Model seems to be Good

Thank You

TEAM

Debank Kundu

Shwetha

Leema Soaphy