Lead Score Case Study Summary

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.

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%.

Solution Summary

Step 1: Reading and Understanding the Data

- We imported important libraries and loaded our dataset.
- Checking the dimension `(9240, 37)` and gathering information

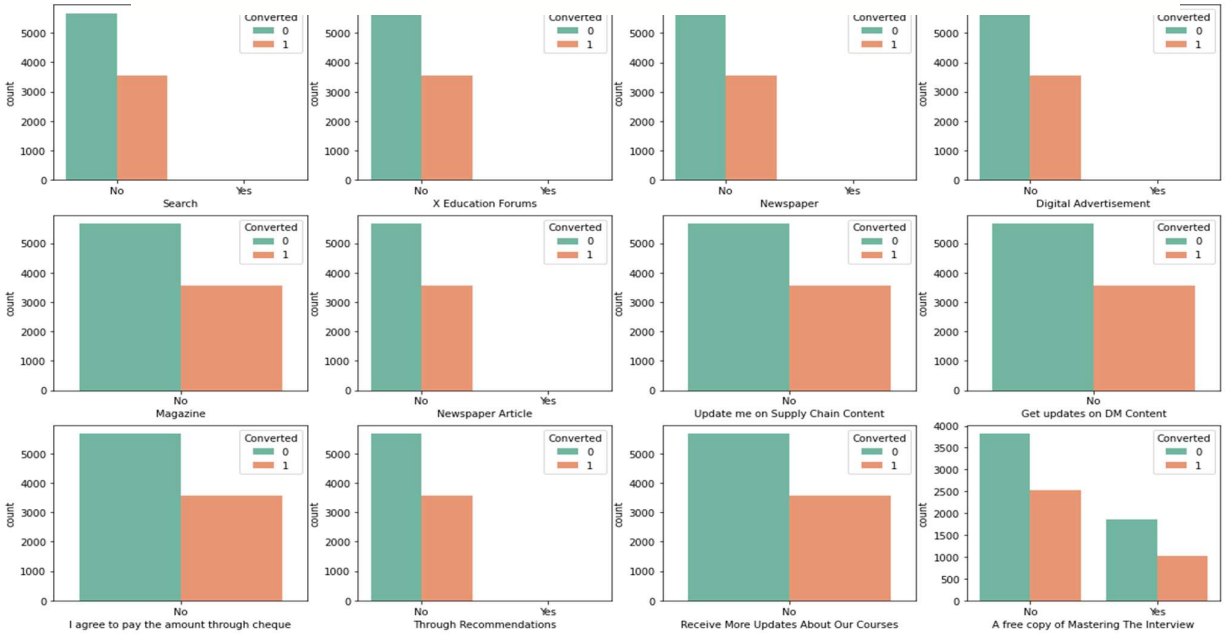
we can see for some categorical variables we have to create dummy variables. Also, we have to handle the null value in coming steps.

Step 2: Data Cleaning

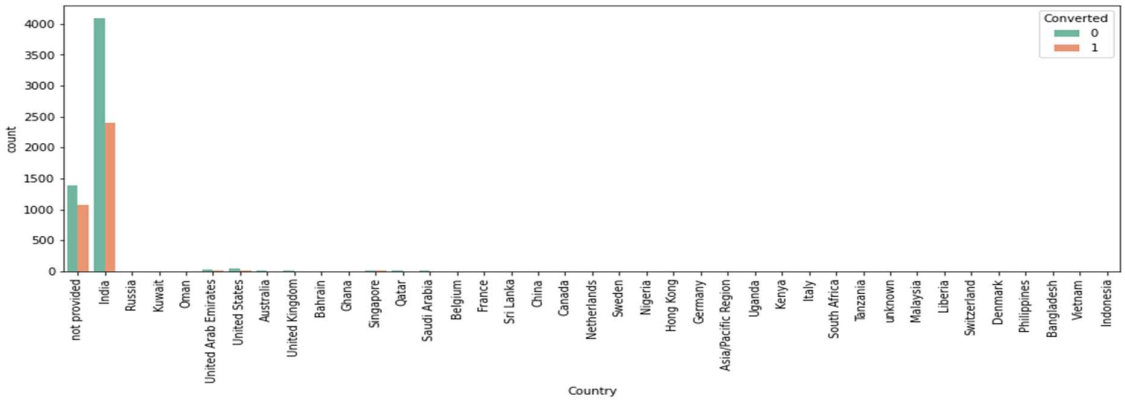
- we can see that no duplicate value in Lead Number and Prospect ID, we are dropping the Lead Number and Prospect ID
- We converted the value to **NaN** since the customer has not selected any option for

1. `Specialization`
2. `How did you hear about X Education`
3. `Lead Profile`
4. `City`

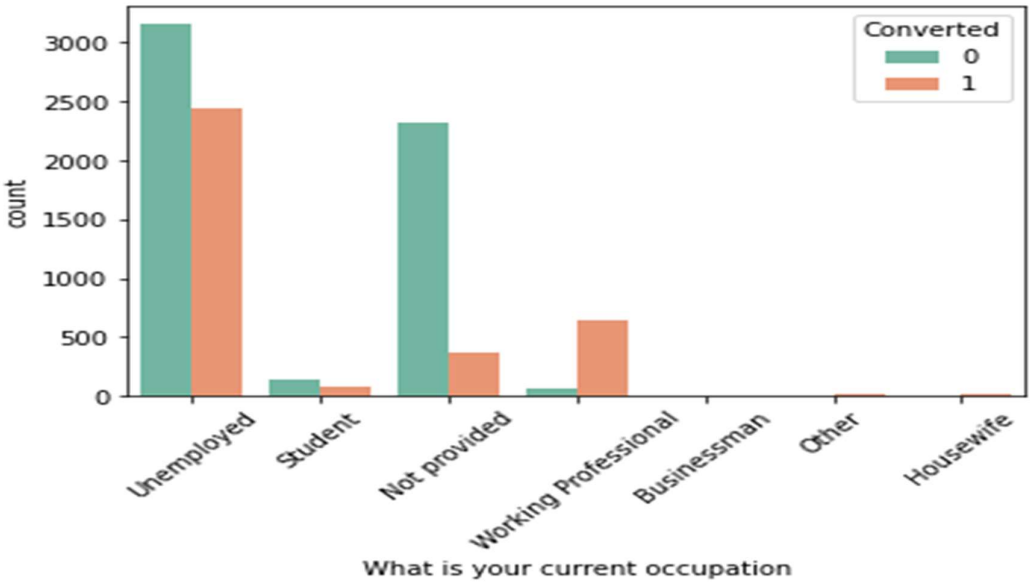
- Data consists of more than 35% missing value, so we are removing columns.
- Categorical Variable analysis.



- For all Columns which are making data imbalance like `'Search'`, `'Newspaper'`, `'Article'`, `'X Education Forums'`,`'Digital Advertisement'`,`'Through Recommendations'`,`'Receive More Updates About Our Courses'`,`'Update me on Supply Chain Content'`,`'Magazine'`,`'Get updates on DM Content'`,`'I agree to pay the amount through cheque'`,`'A free copy of Mastering The Interview'`. So, we going to remove these columns.
- Checking for the country variable where we saw in consists of many countries data along with **India** and **Not Provided**. As we can see that most of the data consists of value **'India'**, no observation can be drawn from this parameter. Hence, we will drop this column.

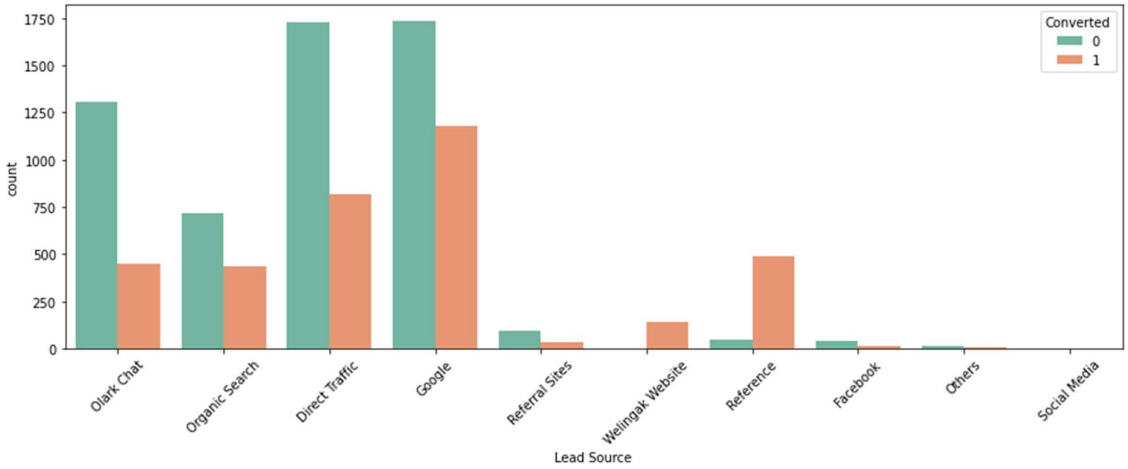


- As there is no information provided regarding occupation, we will replace missing values with new category 'Not provided'. Conversion rate of working professionals is high.

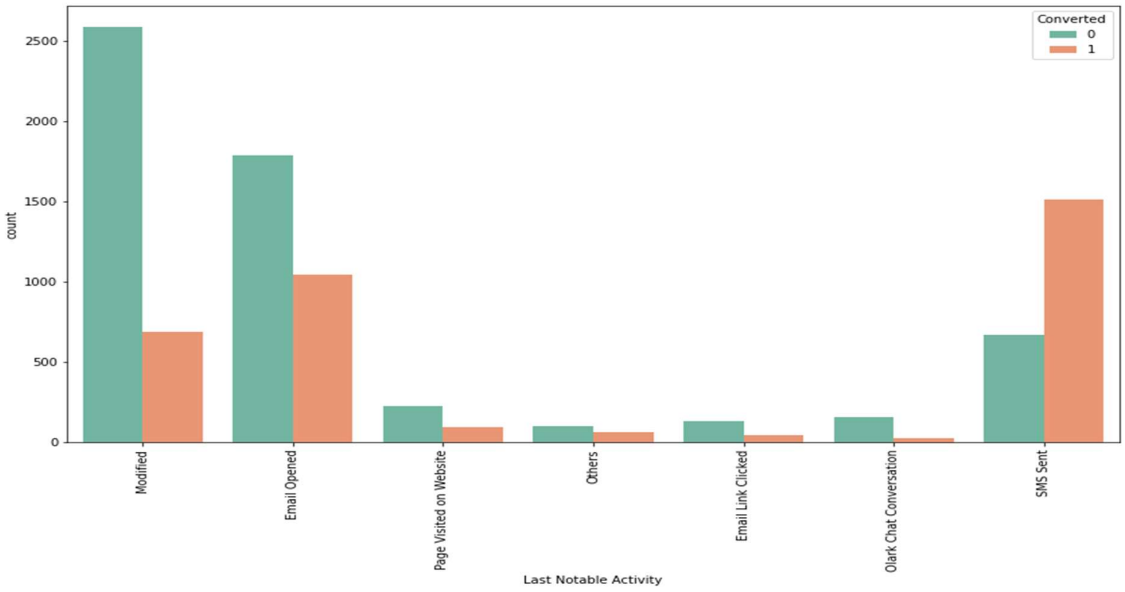


Maximum leads generated are **unemployed** and its conversion rate is >50%.

- As we can see 'Google' having highest number of occurrences, hence we will impute the missing values with label 'Google'. Conversion rate of Reference leads and Welinkgak Website leads is very high. Most Leads are generated by **'Google'** and **'Direct Traffic'**.



- SMS sent as last activity has high conversion rate. Maximum leads are generated having last activity as email opened but conversion rate is not so good.



- We deal with outlier of for Total Visits, Total Time Spent on Website and Page Views Per Visit variables. We saw that the conversion rate is high for **Total Visits, Total Time Spent on Website and Page Views Per Visit variables**.

Step 3: Data Preparation

In this step we are converting data as per requirement like changing data into binary variable (Yes:1, No:0) and also changed categorical value into dummies.

Step 4: Train- Test Split

In this step we are converting data into 70%(train) and 30% (test). Train dataset will transform and trained on our model. Test dataset will help us to check whether our dataset is preforming well.

Step 5: Scaling of features

Feature scaling is the process of normalising the range of features in a dataset. Real-world datasets often contain features that are varying in degrees of magnitude, range and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling. So, we are converting the datasets for Scaling of features.

We can drop highly correlated dummy variables.

Step 6: Building a model using Stats Model and RFE

1. We can see that p-value of variable What is your current occupation_Housewife is high (0.999), so we can drop it.
2. As we can see that p-value of variable "Lead Source_Welingak Website" is high((0.490), hence we can drop it.
3. We can see that variable 'What is your current occupation_Businessman' has high p-value (1.475), so we need to drop it.
4. We can see that variable 'What is your current occupation_Other' has high p-value (0.039), so we need to drop it.

Generalized Linear Model Regression Results

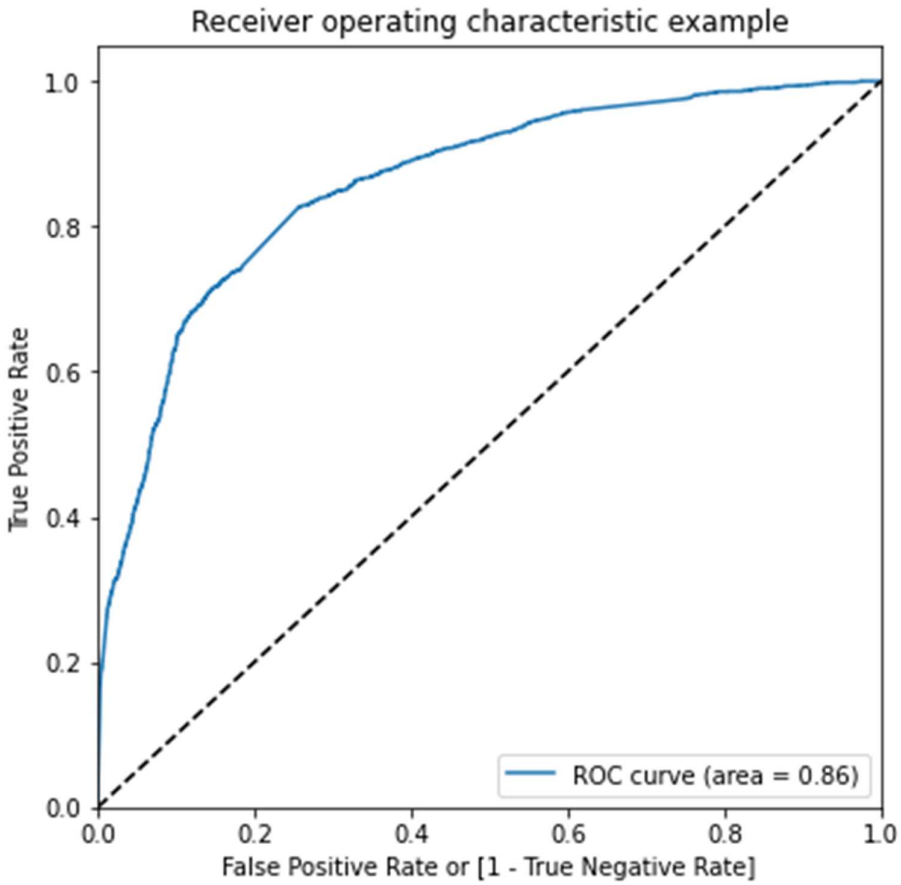
Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6360
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2875.6
Date:	Mon, 20 Mar 2023	**Deviance:**	5751.2
Time:	02:48:25	**Pearson chi2:**	6.43e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3464
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2020	0.094	-12.723	0.000	-1.387	-1.017
Do Not Email	-0.3600	0.043	-8.348	0.000	-0.445	-0.276
Total Time Spent on Website	1.1023	0.038	28.710	0.000	1.027	1.178
Lead Origin_Lead Add Form	4.6119	0.523	8.816	0.000	3.587	5.637
Lead Source_Direct Traffic	-1.0496	0.107	-9.783	0.000	-1.260	-0.839
Lead Source_Google	-0.7804	0.102	-7.615	0.000	-0.981	-0.580
Lead Source_Organic Search	-0.8639	0.124	-6.987	0.000	-1.106	-0.622
Lead Source_Reference	-1.7425	0.564	-3.089	0.002	-2.848	-0.637
Lead Source_Referral Sites	-1.3749	0.336	-4.094	0.000	-2.033	-0.717
What is your current occupation_Student	1.1342	0.224	5.057	0.000	0.695	1.574
What is your current occupation_Unemployed	1.2613	0.082	15.384	0.000	1.101	1.422
What is your current occupation_Working Professional	3.7575	0.189	19.919	0.000	3.388	4.127

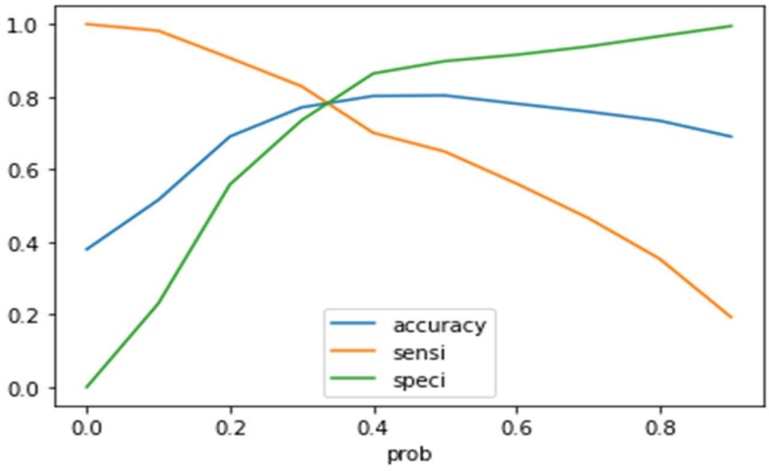
5. All variables have a good VIF value. So, we don't need to drop any more variables and we will proceed with this model only.

	Features	VIF
2	Lead Origin_Lead Add Form	3.81
6	Lead Source_Reference	3.63
9	What is your current occupation_Unemployed	2.58
4	Lead Source_Google	1.70
3	Lead Source_Direct Traffic	1.67
5	Lead Source_Organic Search	1.31
10	What is your current occupation_Working Profes...	1.29
1	Total Time Spent on Website	1.12
8	What is your current occupation_Student	1.05
0	Do Not Email	1.03
7	Lead Source_Referral Sites	1.02

6. The ROC Curve should be a value close to 1. We got a good value of 0.86 indicating a good predictive model.



7. From the below curve, 0.3 is the optimum point to take it as a cut-off probability.



8. As we can see above the model is performing well. The ROC curve has a value of 0.86, which is good value. We got the following values for the Train Data:

- Accuracy: 77.05%
- Sensitivity :82.89%
- Specificity: 73.49%

9. After running the model on the Test Data, we got following figures:
    - Accuracy: 77.52%
    - Sensitivity :83.01%
    - Specificity: 74.13%

During checking both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which is closer to values calculated using trained set.

Also, the lead score calculated in the trained dataset shows the conversion rate on the final predicted model is around 80%.

Hence overall this model seems to be good.

CONCLUSION:

Important features for good conversion rate or the ones' which contributes towards the probability of a lead getting converted are

- Lead Origin Lead Add Form
- What is your current occupation_Working Professional
- Total Time Spent on Website