LEAD SCORE CASE STUDY

BY

SAI LALITH SISTLA &

PRIYANKA AKAVARAM

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%.

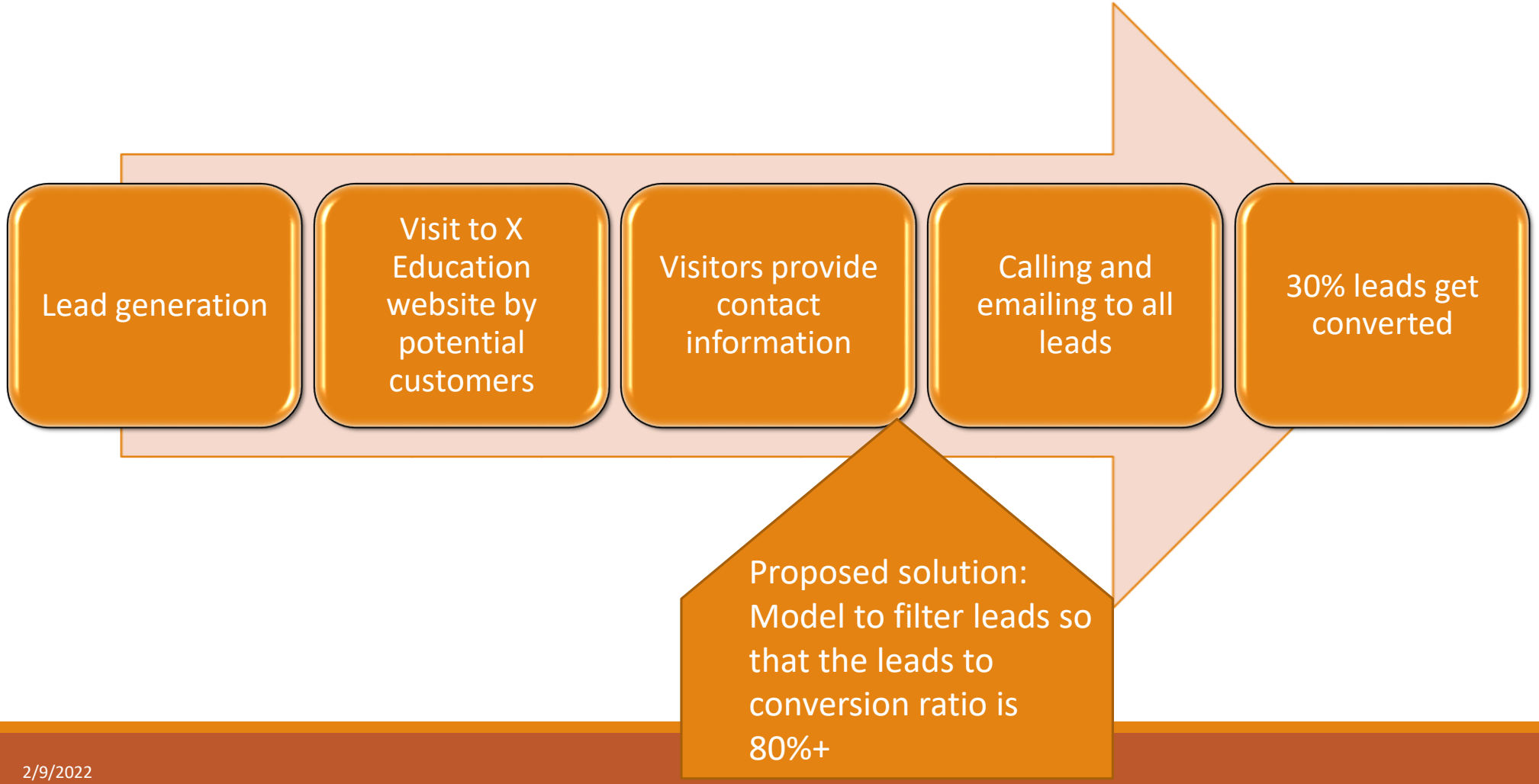
Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If 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.

Goals of the Case Study

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.

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

Lead conversion process



Strategy

Data cleaning & Manipulation

- Loading and observing past data
- Check and handle duplicate values, missing data, Nan values. Drop columns if it contains high amount of missing values. Imputations of values. Check and handle outliers if any.

EDA

- Univariate data analysis: value count, distribution of variables.
- Bivariate data analysis: correlation coefficients and pattern between the variables.

Data preparation

- Creating dummy variables.
- Feature scaling on features that have a huge range.
- Splitting the data into test and train data sets.

Strategy

Model building and improvement

- Selection of top features using RFE.
- Reduction of columns and model rebuilding.
- Building logistic regression model and calculate lead score.

Final model Evaluation

- Check for the ROC curve.
- Evaluation of model and finding optimal probability cut-off based on Specificity, Sensitivity, Precision and Recall.
- Checking the same on test data.

Conclusion

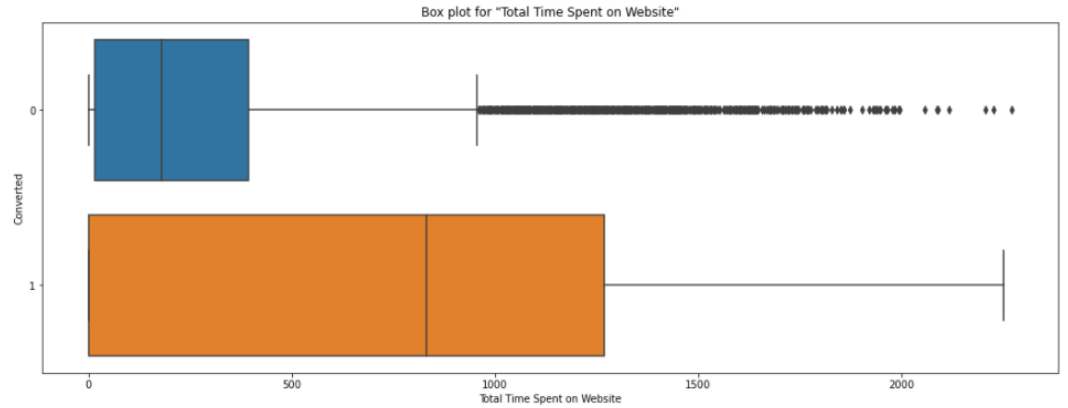
- Making recommendations based on our model for clients requirement.

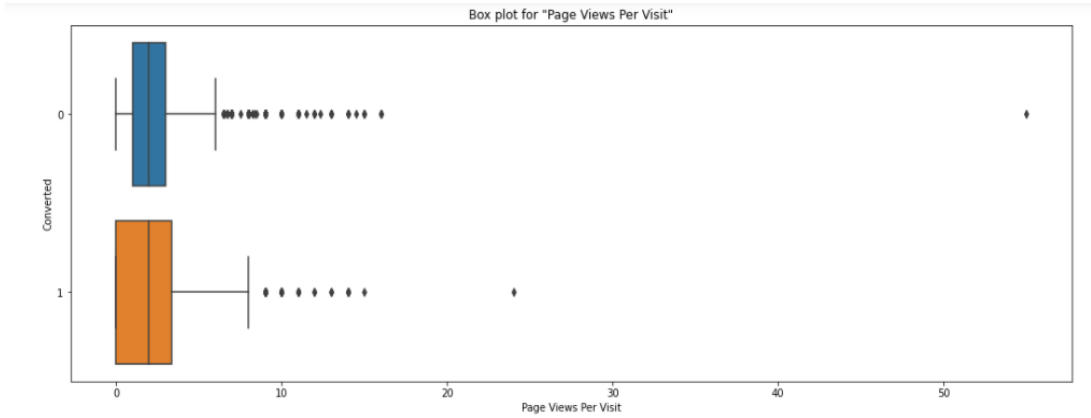
EDA Univariate data analysis

Time spent on website and Page views per visit

Observation:

1. Clearly there are outliers in Data
2. Mean of page visits are nearly same for people who converted and who didn't
3. People who spent more time on pages per visit have more chances of conversion
4. Mean of page views per visit is same for people who convert and who don't
5. Outlier treatment is done using the median.

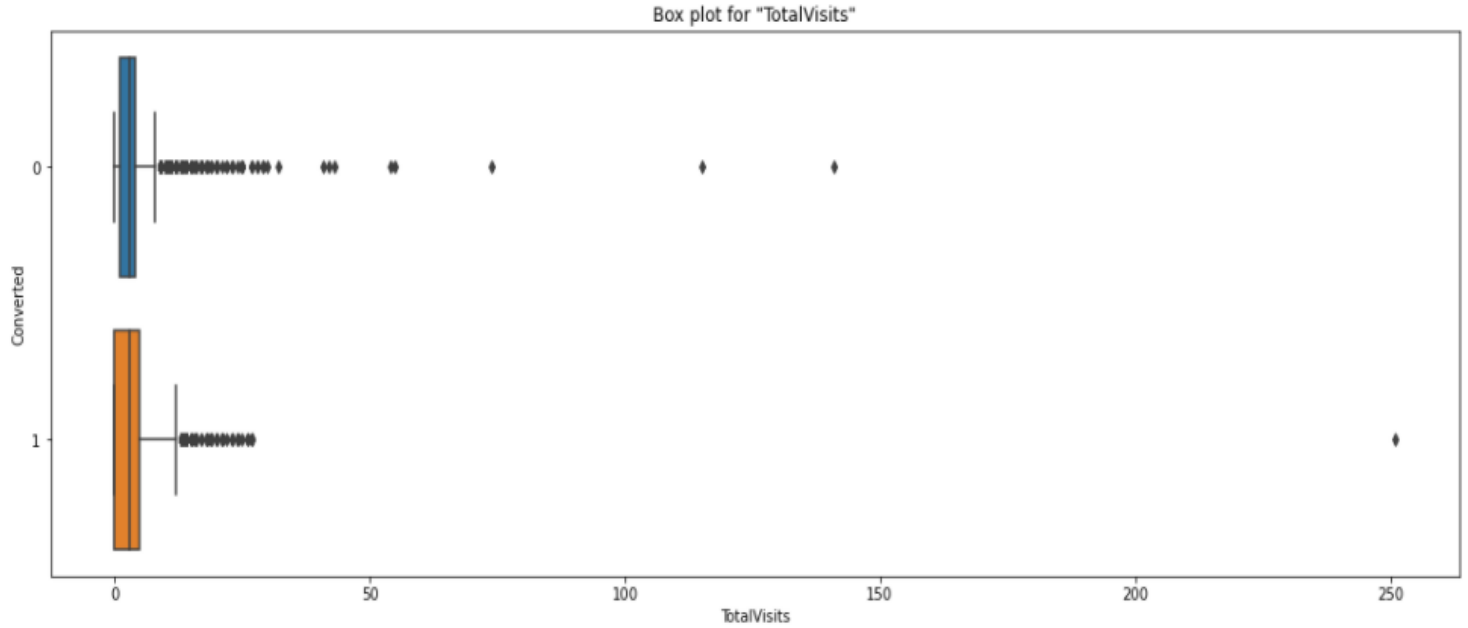




Total visits

Observation:

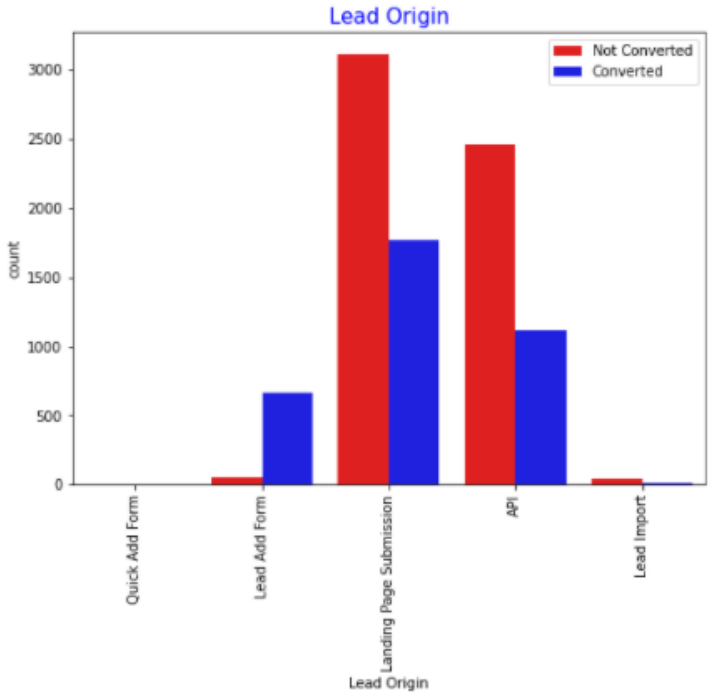
1. There are outliers present, and the mean of converted and not converted is almost the same.
2. Outlier treatment is done using median.

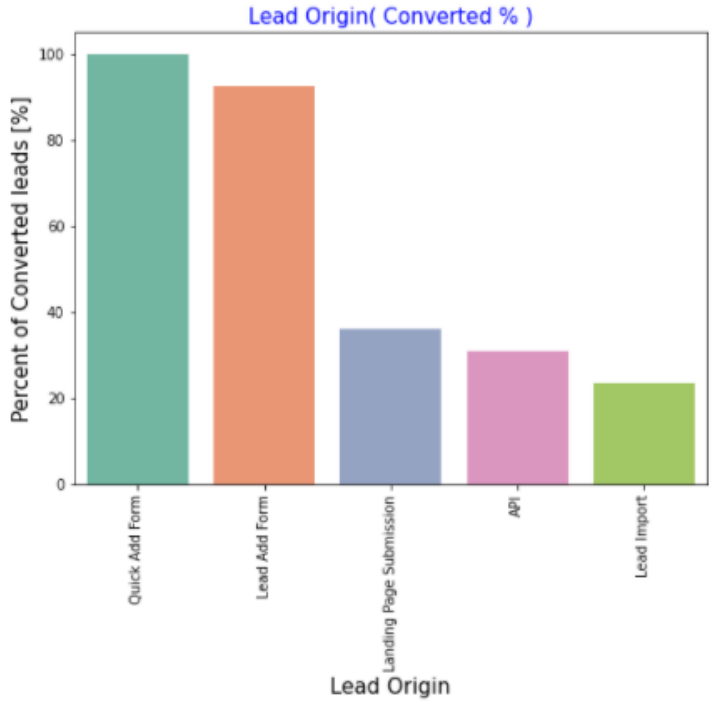


Lead origin

Observation:

1. Quick Add form and Lead Add Form has high conversion rate. But the total leads obtained through these are very less compared to other.
2. Landing page submission and API have given more number of leads, the conversion rate is less than half but still some thing we can work with
3. Lead import has less number of leads and also the less conversion rate

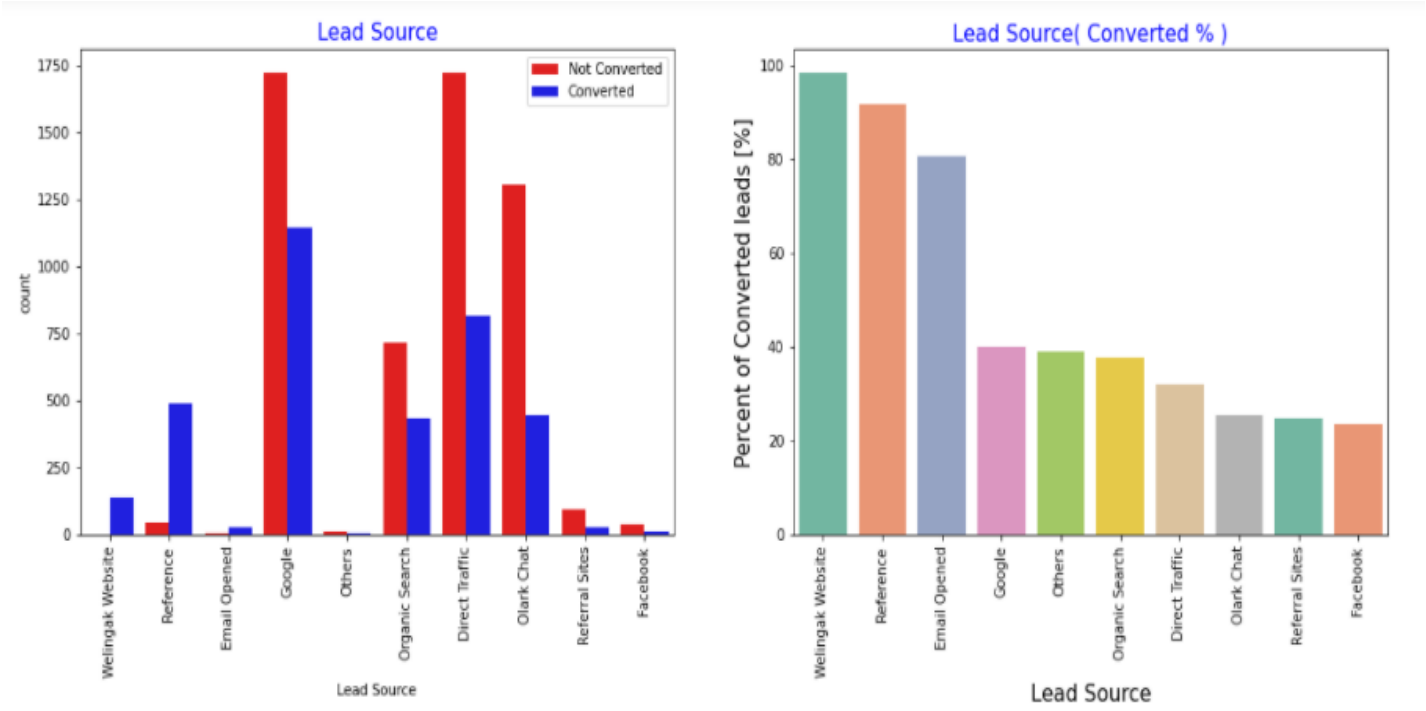




Lead source

Observation:

1. Google and Direct traffic are the two sources generating high leads. Their lead conversion rate is near 40% levels.
2. Welingak Website and Reference have conversion rates but the leads are few, Should focus on the website more and More referrals from the existing users must be taken for more hot leads.

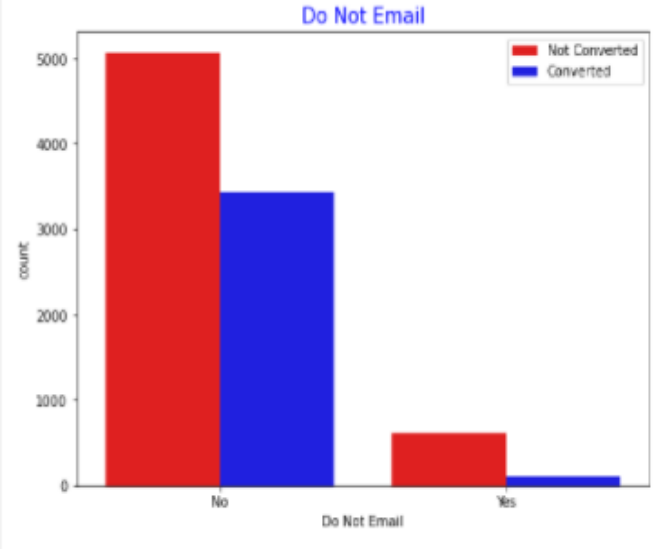


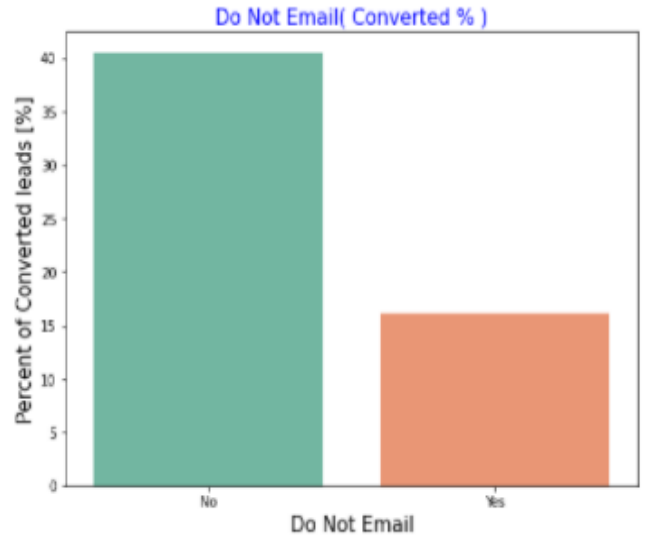
Do not email

Observation:

1. Most of the leads are not willing to receive emails
2. Nearly 40% of people who choose not to receive mail still converted.
3. While there is only 15% conversion rate with those who opted to receive mail

	Do Not Email	Total Leads	% Values
0	No	8506	92.056277
1	Yes	734	7.943723



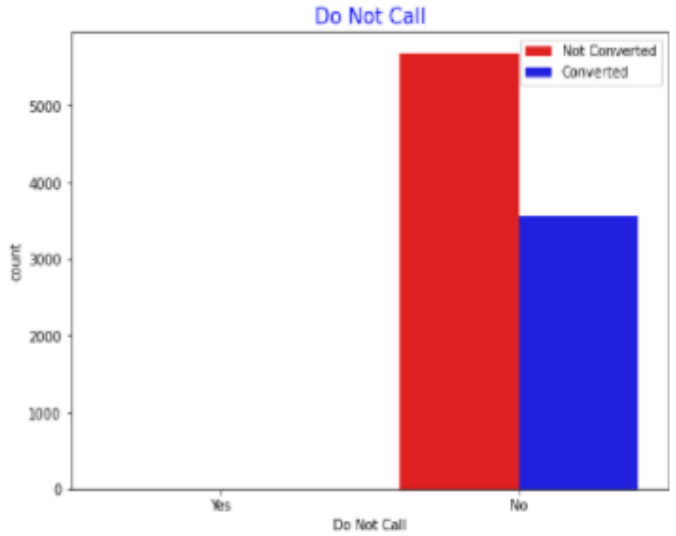


Do not call

Observation:

1. The data looks highly skewed towards No for 'Do Not Call' and will not be helpful in analysis.
2. Hence dropping this column

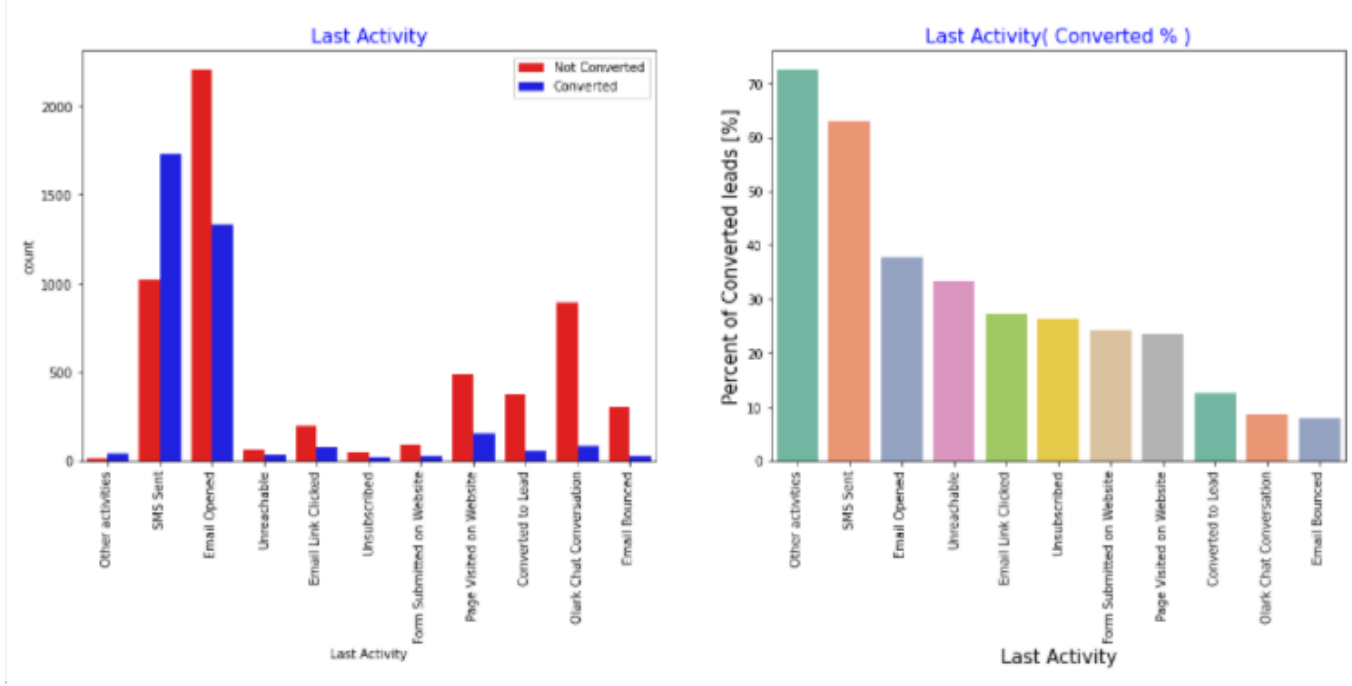
	Do Not Call	Total Leads	% Values
0	No	9238	99.978355
1	Yes	2	0.021645



Last activity

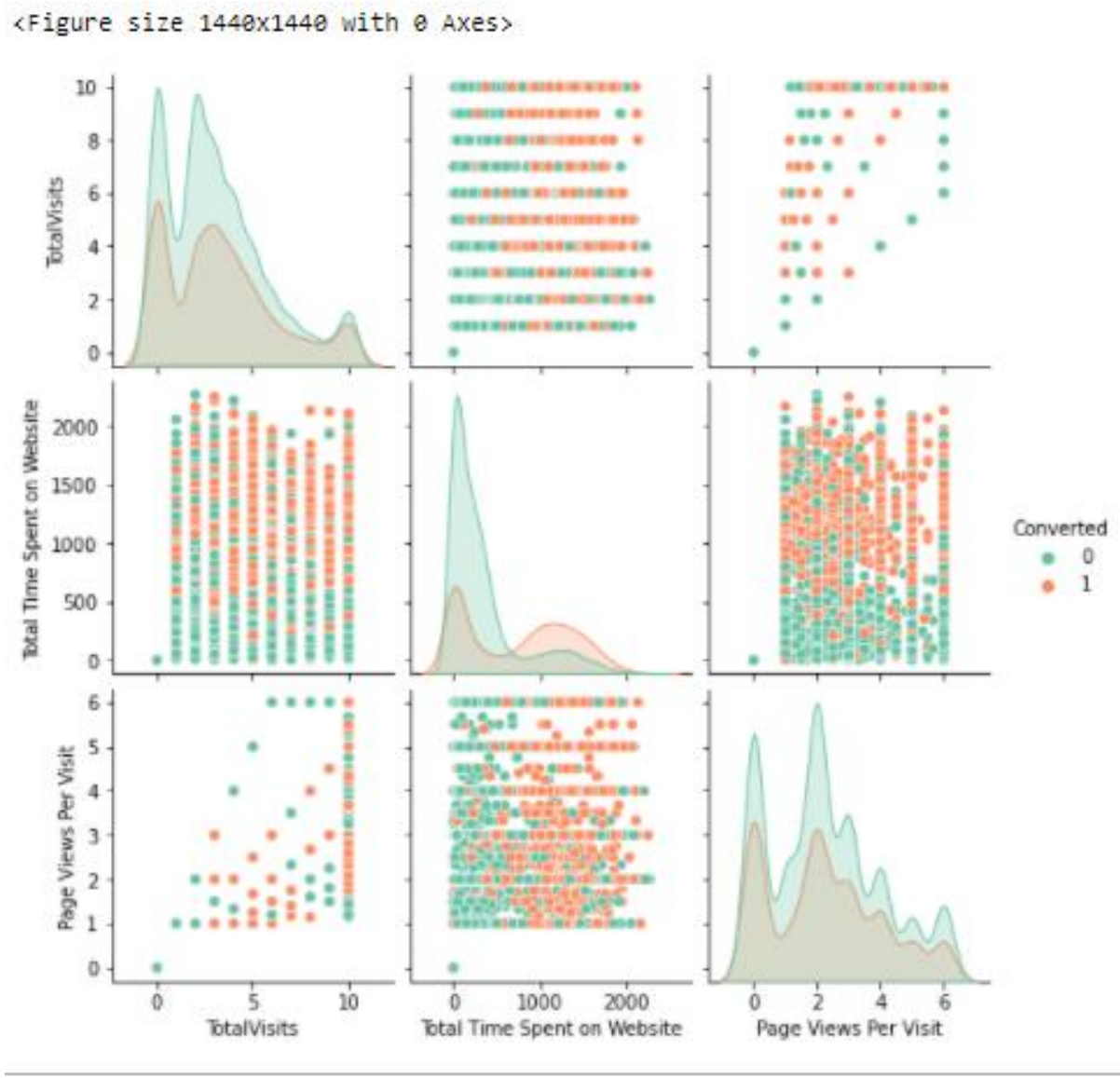
Observation:

1. Other activities have a high conversion rate , but have only few frequency.
2. Emailed Opened has highest frequency of them all but has conversion rate near 40%.
3. SMS Sent is second most frequent , and has near ^0% conversion rate.
4. Leads with email bounced have the least conversion rate.



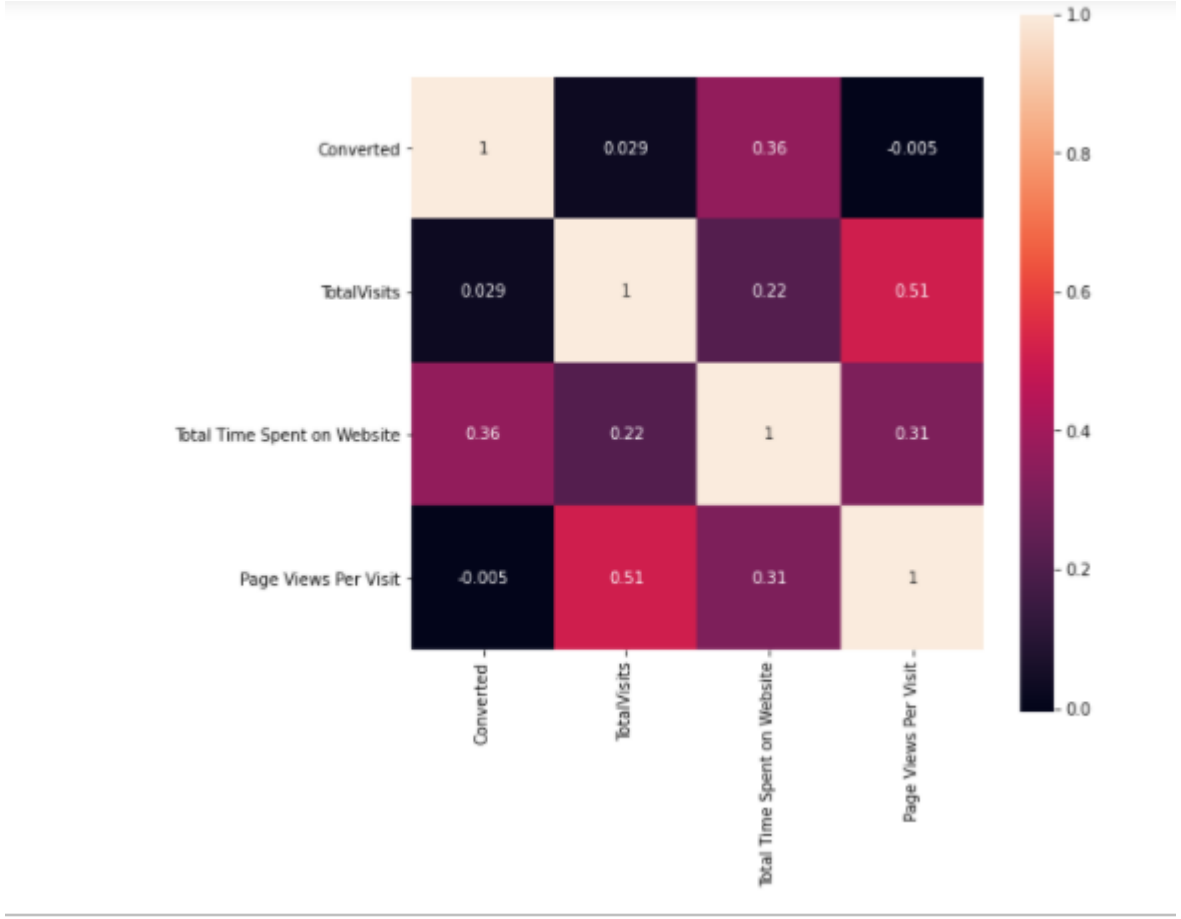
Bivariate analysis

1. Pair plot to understand the converted and not converted across various variables.
2. There are no visible correlations. The data seems to be normally distributed.



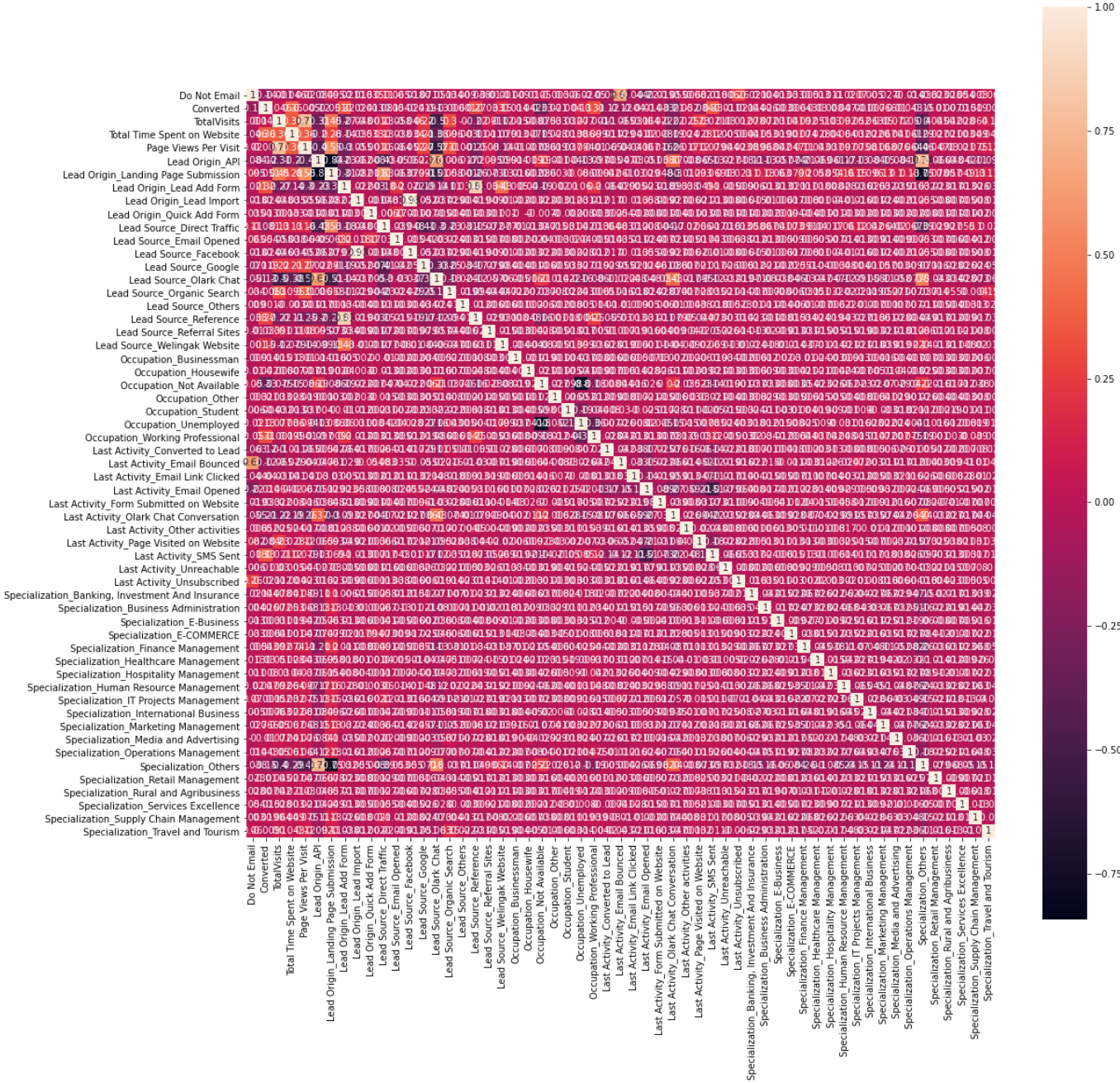
Heat map of data

Observation: Numerical columns have outliers and need outlier treatment , There is also a presence of normal distribution relation ship between features



Correlation

Correlation between features after dummy variable creation.



Top 10 correlations

Lead Source_Facebook	Lead Origin_Lead Import	0.981709
Lead Source_Reference	Lead Origin_Lead Add Form	0.853237
Lead Origin_Landing Page Submission	Lead Origin_API	0.842492
Occupation_Unemployed	Occupation_Not Available	0.794875
Page Views Per Visit	TotalVisits	0.767585
Specialization_Others	Lead Origin_Landing Page Submission	0.748263
	Lead Origin_API	0.740470
Last Activity_Email Bounced	Do Not Email	0.618470
Lead Source_Olark Chat	Lead Origin_API	0.607716
	Page Views Per Visit	0.573334

dtype: float64

Model Building and Evaluation on Train data set

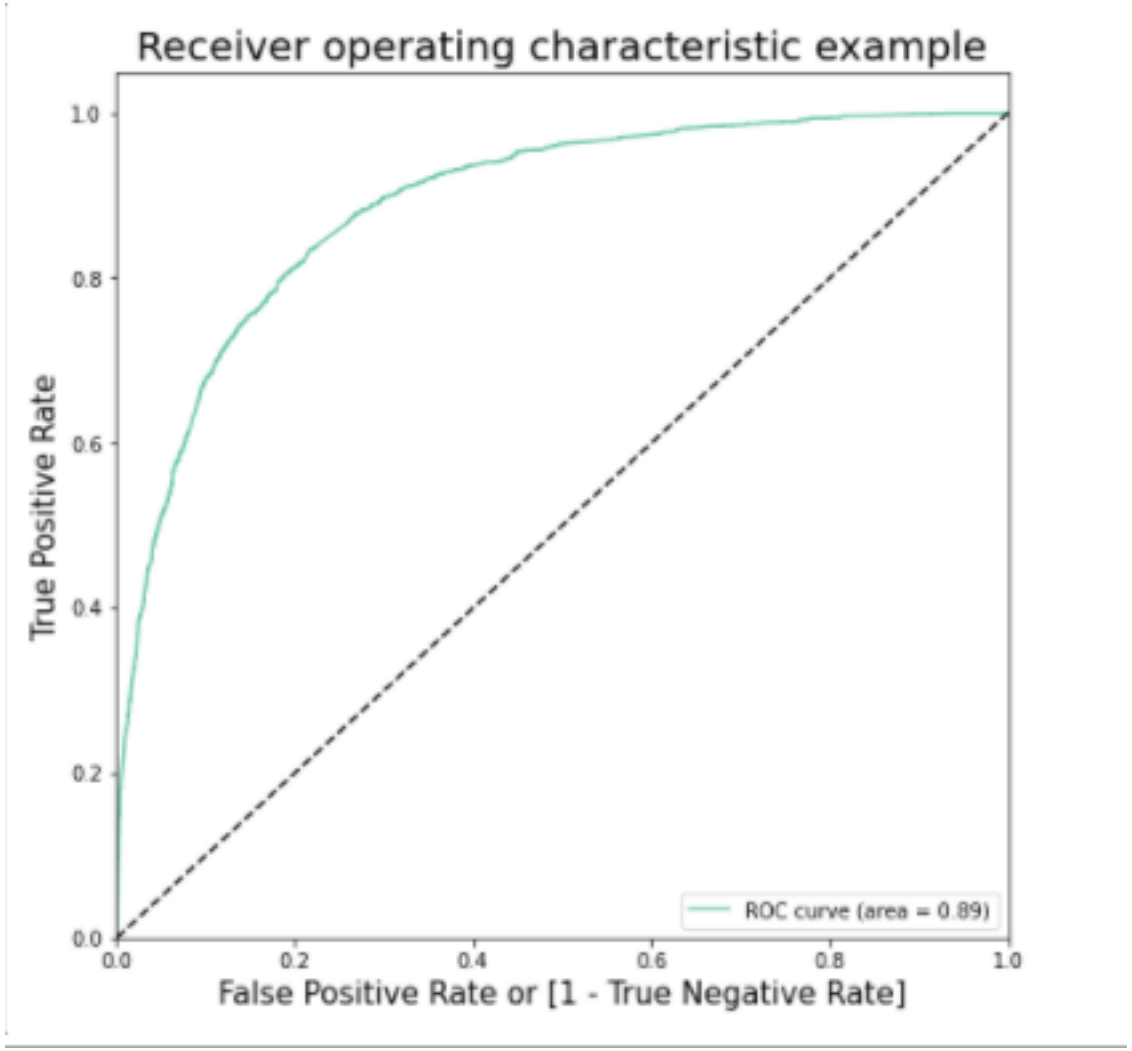
- First step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Using RFE for feature selection, with 20 variables as output.
- Building model by removing the variable whose p-value is greater than 1 and the VIF is greater than 2.
- Overall accuracy is 82%.
- Precision is 80%.
- Recall is 68%.
- Confusion matrix:

Model Evaluation on Test data set

- Accuracy is 80% and Specificity is 90%.
- Precision is 82%.
- Recall is 64%.
- Confusion matrix:

ROC Curve

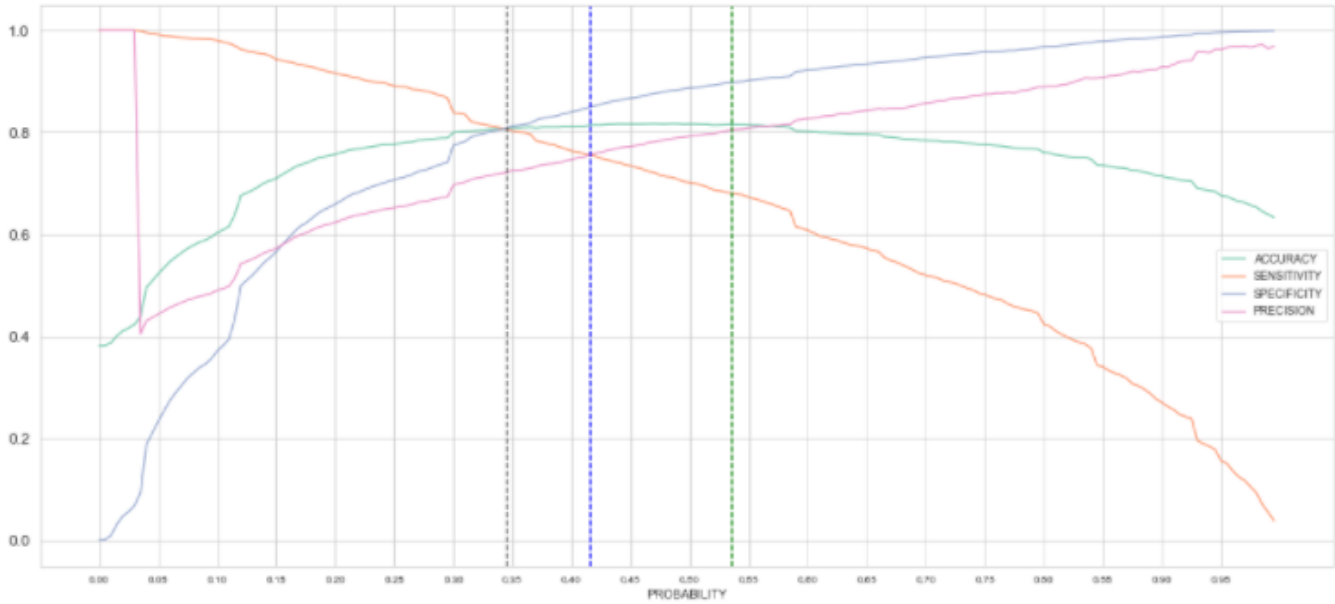
Area under ROC Curve is 0.89, which is good for a model.



Probability

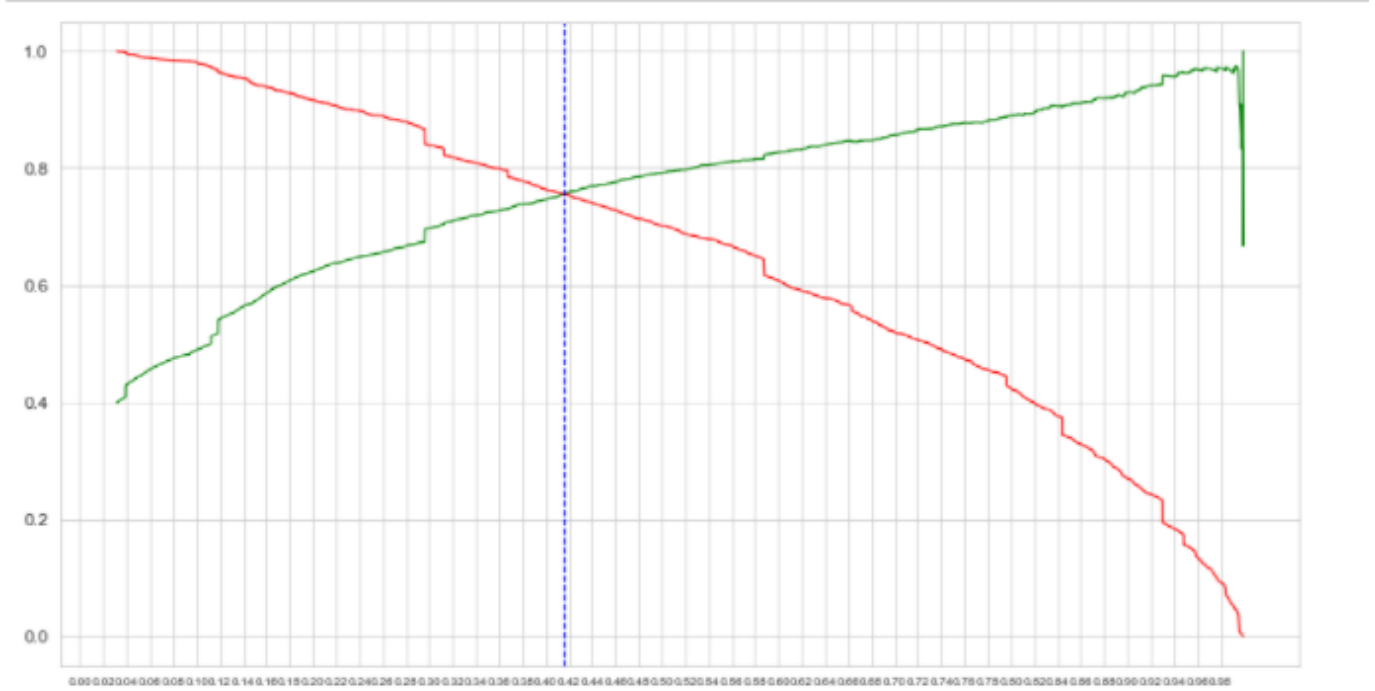
Observation:

From the graph, probability at 0.535 has an ideal trade off between Precision and Recall



Precision-Recall curve

The green curve is precision and the red curve is recall. The probability at the intersection is at 0.415.



Model Evaluation comparing metrics

The model did not memorise the train data set and is neither under fitted. Thus we conclude the model is a good fit. As the precision and accuracy meets the requirements of the X-Education company, we can say that the model is a good fit.

	Train	Test
Accuracy	0.82	0.8
Precision	0.8	0.82
Recall	0.68	0.65
Specificity	0.9	0.9

Log Odds

- log(p/(1-p)) = 0.394648 * const + 2.230094 * Lead Origin_Lead Add Form + 1.060354 * Total Time Spent on Website+ 2.459470 * Occupation_Working Professional + 1.220911 * Last Activity_SMS Sent - 1.394574 * Do Not Email - 1.722602 * Lead Source_Direct Traffic - 1.656503 * Lead Source_Facebook - 1.250309 * Lead Source_Google - 1.406921 * Lead Source_Organic Search- 1.270930 * Lead Source_Referral Sites - 1.141158 * Occupation_Not Available - 1.172836 * Last Activity_Converted to Lead - 1.198687 * Last Activity_Olark Chat Conversation+ 1.603956 * Last Activity_Other activities- 0.322992 * Specialization_Others

Conclusion

X-Education company has a better chance of converting a potential lead when:

1. Occupation_Working Professional: Working professionals have a near 90% conversion and Housewives have near 100% conversion rate however they have low in number of leads. Hence people from these two categories should be more approached.
2. Lead Origin_Lead Add Form: Landing page submission and API have given more number of leads even if the conversion rate is less than half. Hence targeting the people submitting forms come on priority especially if they are working professionals.
3. Last Activity_Other activities: Targeting the Customers with other activities is good as they have a high conversion rate even if they are with less frequency.
4. Total Time Spent on Website: Based on EDA customers spending more time on website have more chances of conversion, hence targeting them is recommended.
5. Last Activity_SMS Sent: SMS Sent is second most frequent , and has near 60% conversion rate.

Recommendations-1

Question: Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Answer: As there is no need for aggressive lead conversion and it is ok to lose out on few hot leads, the best strategy would be to go with high precision. In terms of lead score, try taking up leads with highest score (increase the cut-off lead score around 90).

Recommendations-2

Question: X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

Answer: A good strategy here would be to focus on both precision and recall. Precision to accurately find the potential leads, while recall is for picking out all of the potential leads in the data. Hence, to maintain a good precision and recall, we will choose the lead score from the point of intersection of those two curves. From this strategy, the cut-off lead score is 41.5.