Lead Scoring Case Study

- Prepared by Vijaya Salikar and Shruti Pawar

Business Objective

- To help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.
- To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads.

The objective is thus classified into the following sub-goals:

- Create a Logistic Regression model to predict the Lead Conversion probabilities for each lead.
- Decide on a probality threshold value above which a lead will be predicted as converted, whereas not converted if it is below it.
- Multiply the Lead Conversion probability to arrive at the Lead Score value for each lead.

Problem

Solving Methodology

- The approach for this project has been to divide the entire case study into various checkpoints to meet each of the sub-goals. The checkpoints are represented in a sequential flow as below:

Understanding the Data Set & Data Preparation

→

Applying Recursive feature elimination to identify the best performing subset of features for building the model.

→

Building the model with features selected by RFE. Eliminate all features with high p-values and VIF values and finalize the model

↓

Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.

←

Decide on the probability threshold value based on Optimal cutoff point and predict the dependent variable for the training data.

←

Use the model for prediction on the test dataset and perform model evaluation for the test set.

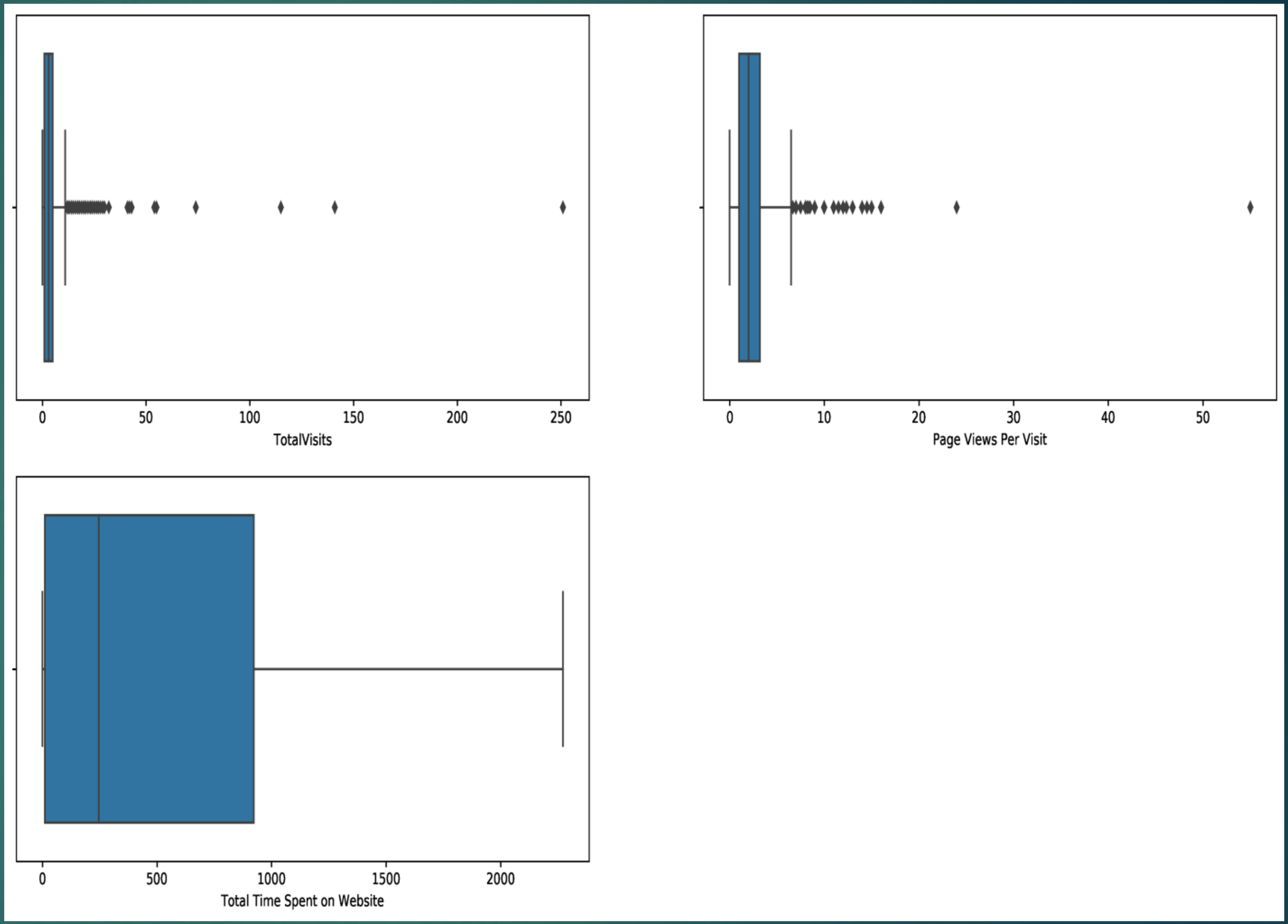
Data Preparation and feature engineering

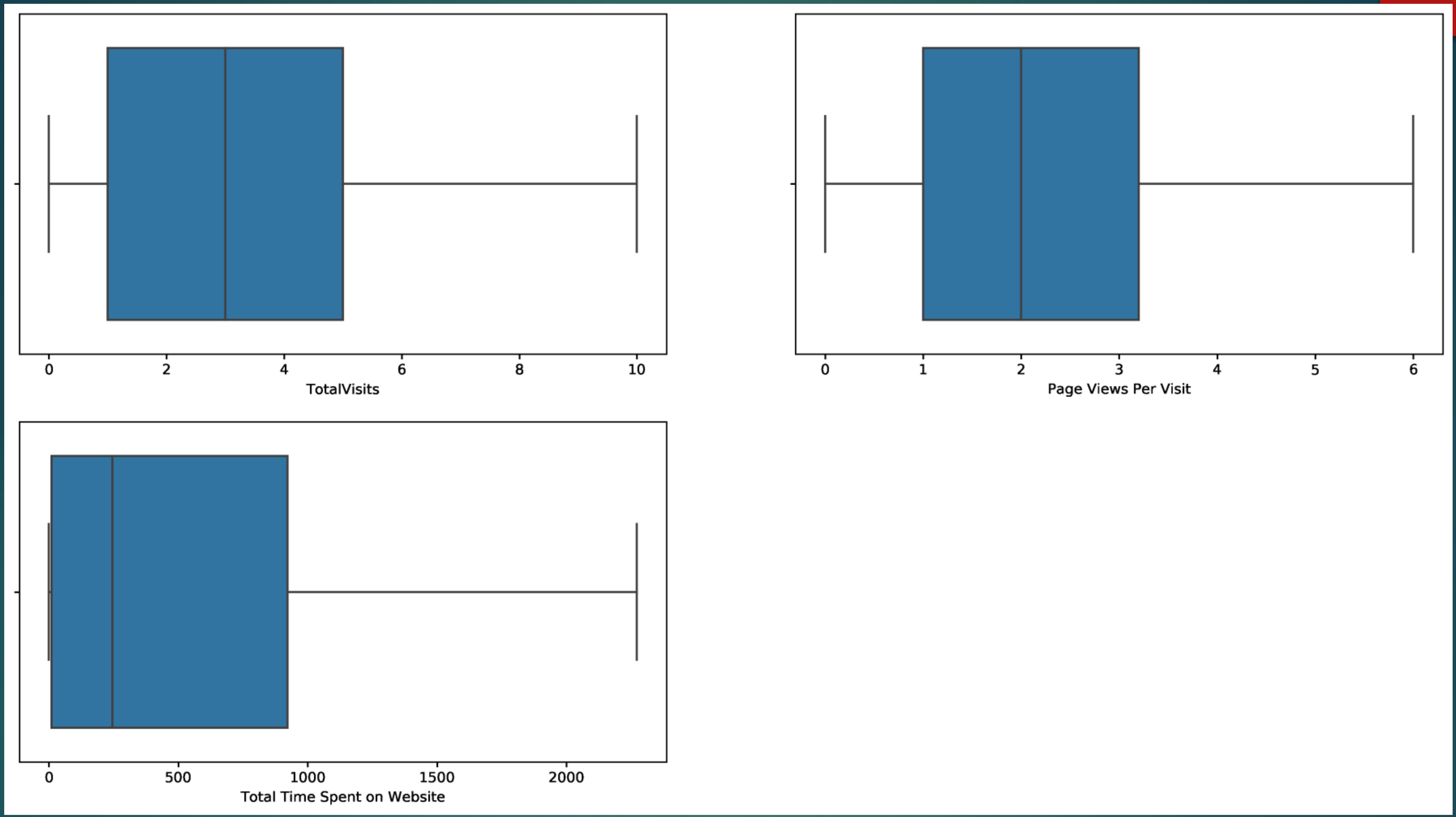
The following data preparation processes were applied to make the data dependable so that it can provide significant business value by improving Decision Making Process:

- **Remove columns which has only one unique value –highly skewed data**
  - Deleting the following columns as they have only one unique value and hence cannot be responsible in predicting a successful lead case – 'Magazine', 'Receive More Updates About Our Courses' , 'Update me on through cheque'. Supply Chain Content' , 'Update me on Supply Chain Content' and 'I agree to pay the amount
- **Dropping columns having 45% or higher missing values**
- **Imputing NULL values with Mode/Creating a separate category**
  - The columns '**Country**' is a categorical variable with some null values. Also majority of the records belong to the Country 'India'. Thus imputed the null values for this with mode(most occuring value). Then combined rest of category into 'Outside India'. Likewise was done for other columns too.
- **Dropping various sales related columns like Tag, Last_activity etc that are not available while model building**

Outlier treatment

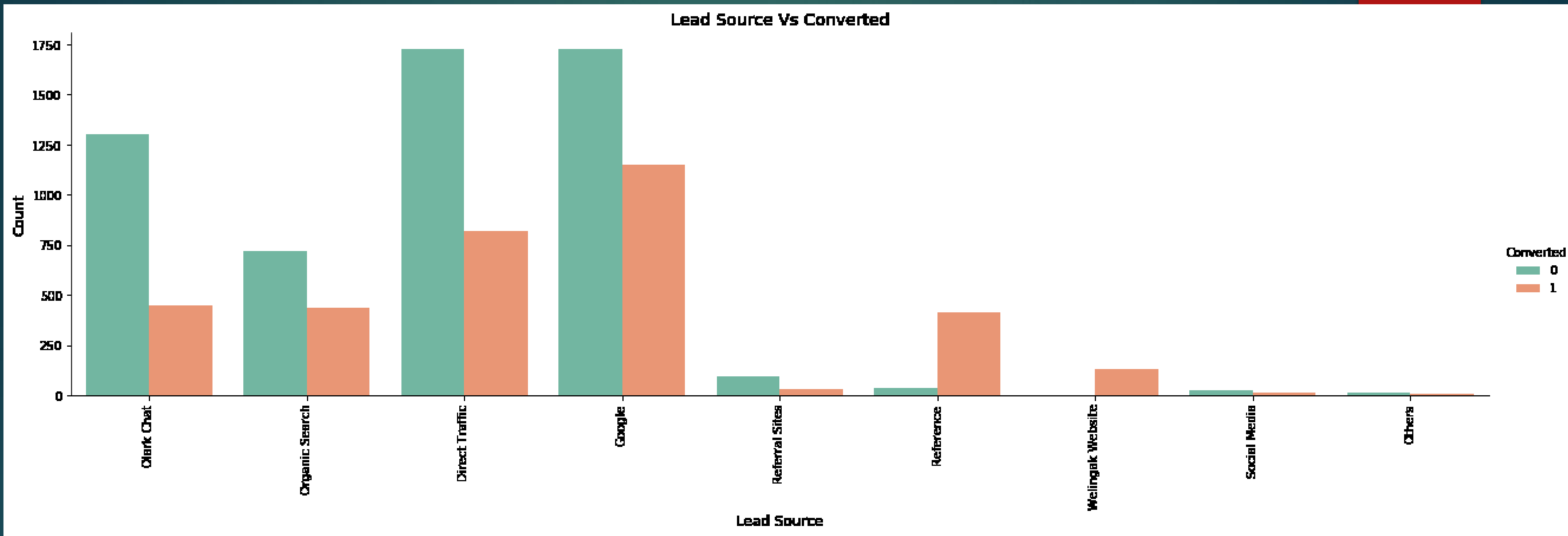
- Observations:
 1) For columns Total Visits and Page Views Per Visit we see there are outliers. For column 'Total Time Spent on Website' there are no outliers. So we will treat columns Total Visits and Page Views Per Visit by capping the data at 95th Percentile





- Observations:
- Around 50% of customers visit the website 3 times , with the highest number of visits being 10
- Around 75 % of customers visits 3-4 pages on the website

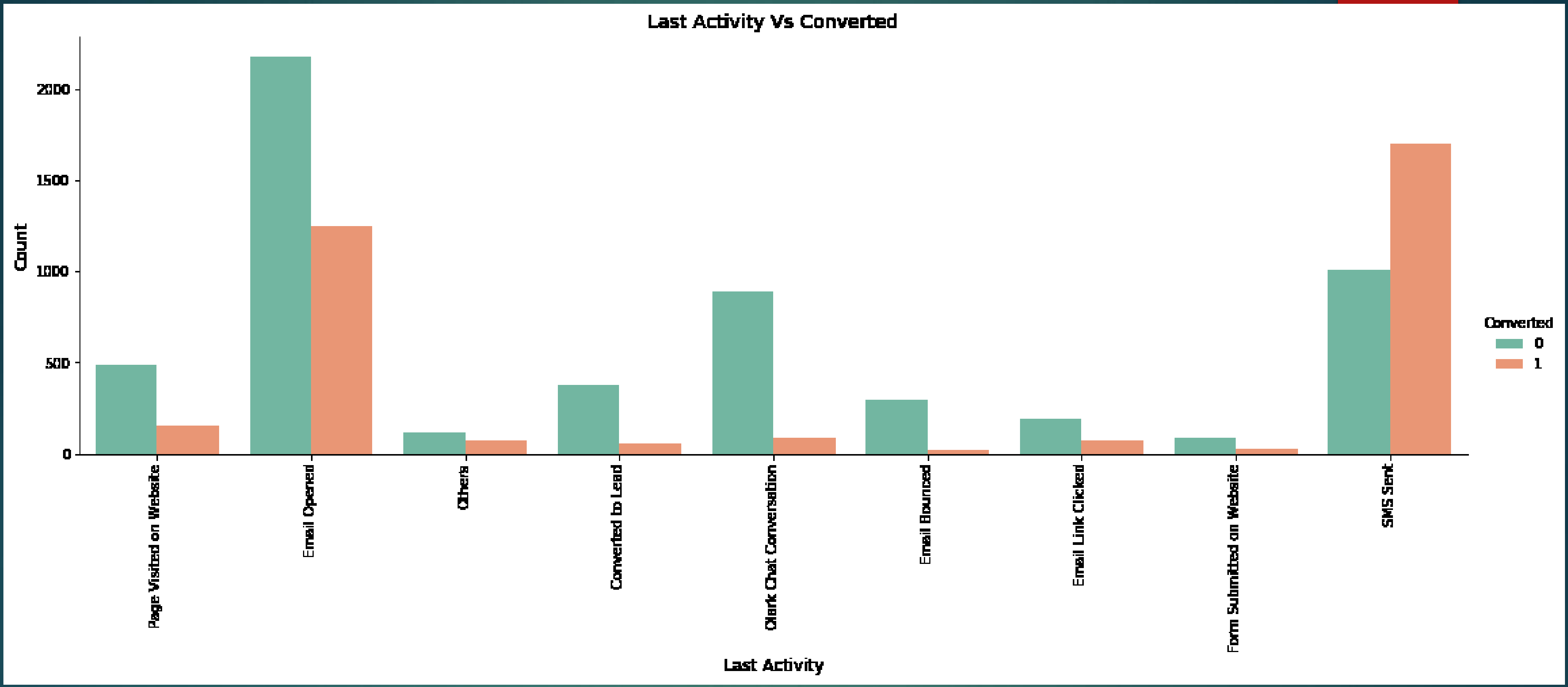
Plotting the variable **“LEAD SOURCES”**



Observation:

- We see that the highest lead source was from Google search and also almost equal number of customers landing directly on the website. Over 1200 cutomers have got converted who came from google followed by Direct traffic source.
- the lowest count belong to 'Others' category.
- Interestingly , the customers who have come through refernce , have high conversion rate

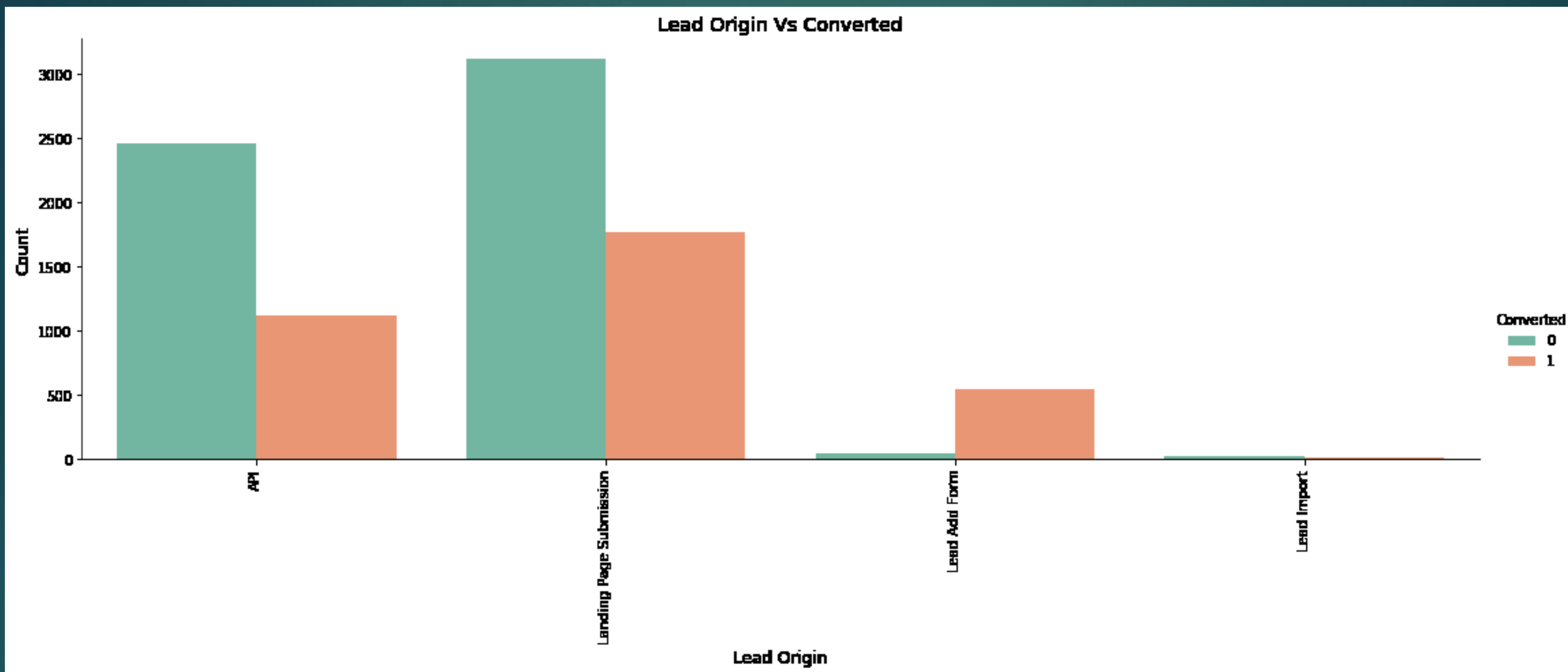
Plotting the variable **"LAST ACTIVITY"**



Observations:

- Over 2000 people had opened the email as their last activity , out of which around 1200 seem to have converted.
- The highest conversion is seen from people who were reached out via sms.

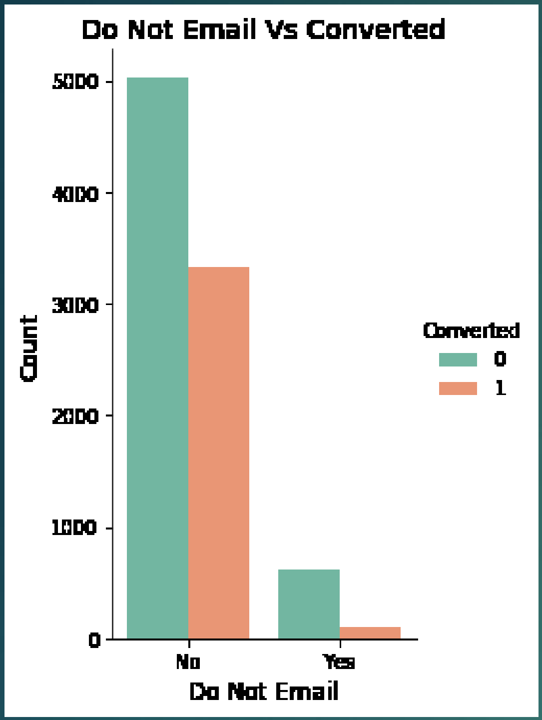
Plotting the variable **“LAST ORIGIN vs CONVERTED”**



Observation:

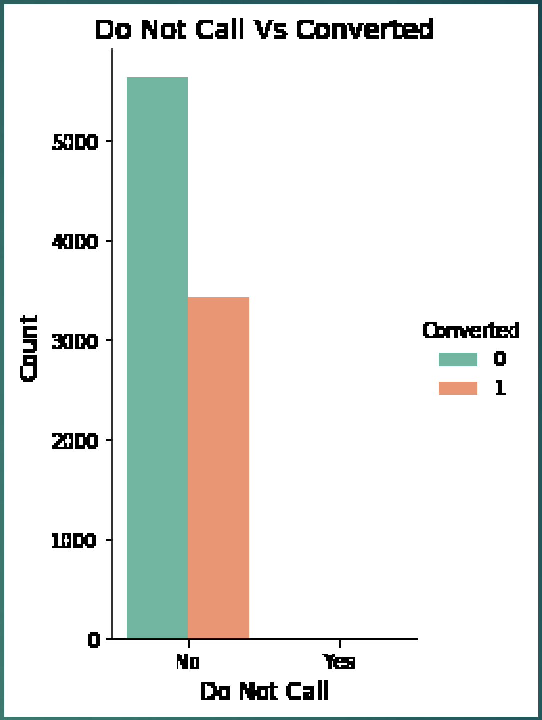
- Landing Page Submission brings higher number of leads as well as conversion followed by API .
- Lead Import gets lowest leads.
- Lead import and Lead Add Form can be foucssed on to generate more leads

Plotting the variable "DO NOT CALL , DO NOT E MAIL AND CURRENT OCCUPATION vs CONVERTED"



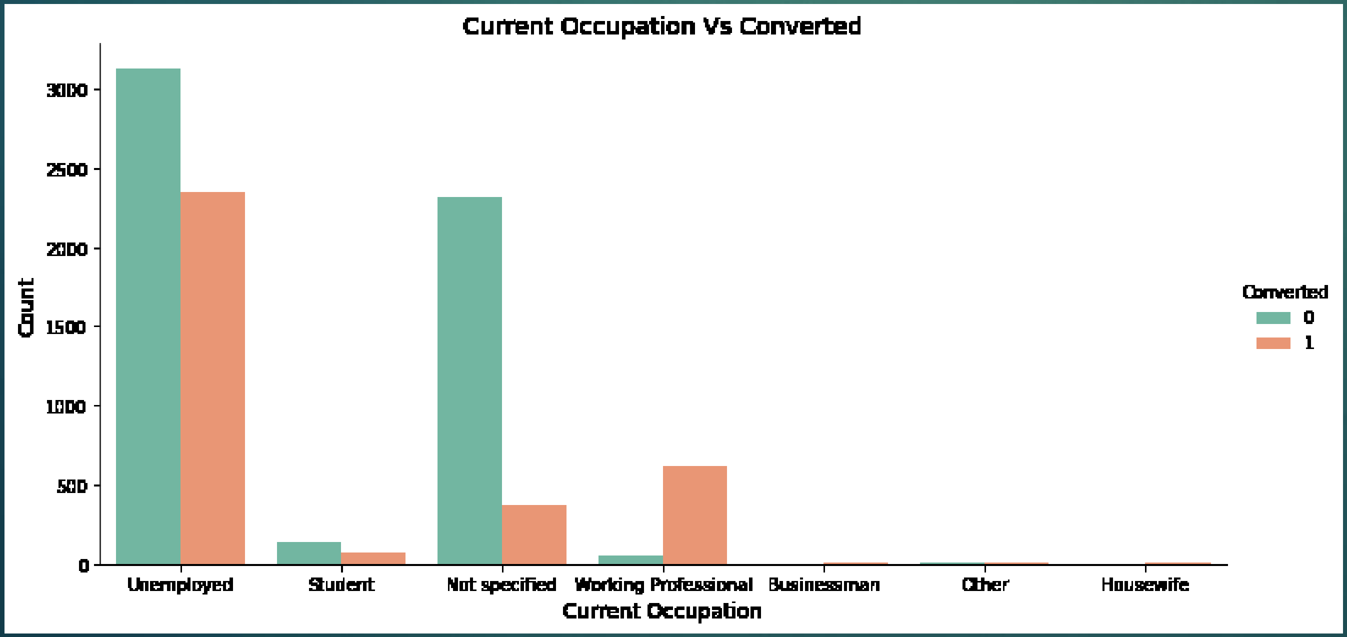
Observation :

The ratio of people opting to be reached out via email is high , we also observe that those who were reached out via email have actually converted. the number is between 3000 -3500



Observation

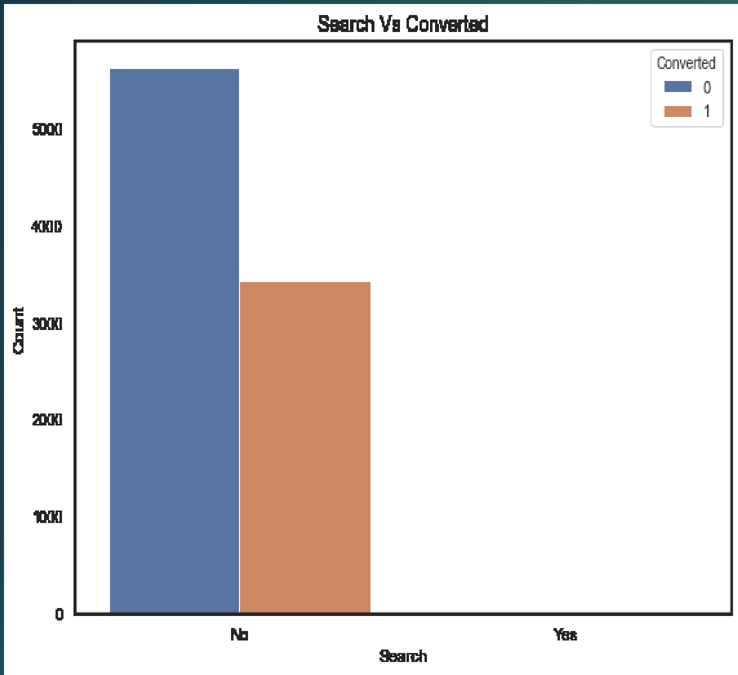
Out of 9072 customers reached out via call is high , more than 3000 people have converted.
- We do not see customers having an issue being reached out via phone calls

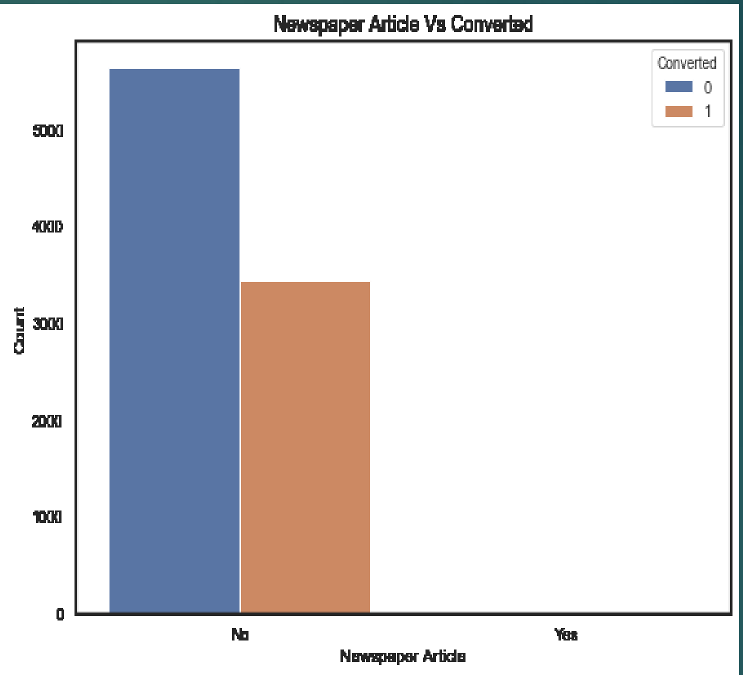


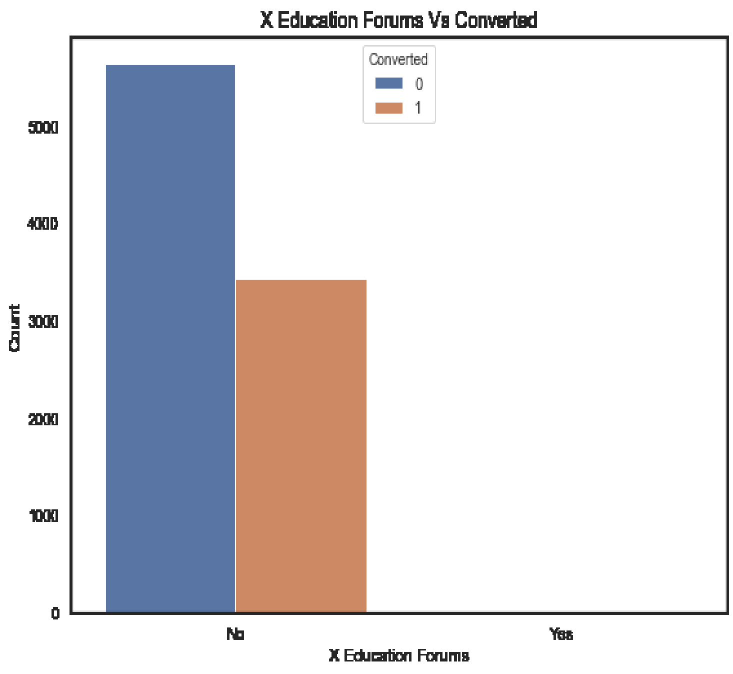
Observation

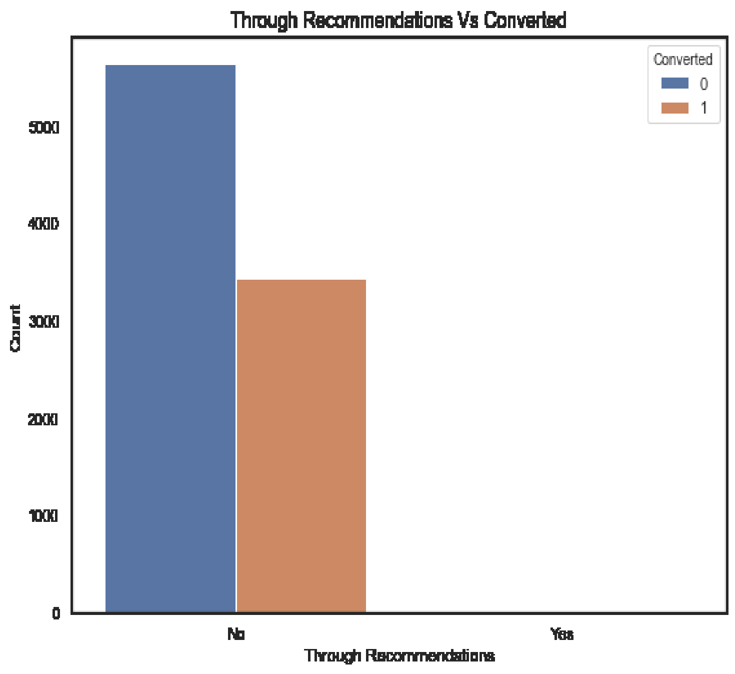
- More than 3000 people belong to the Unemployed category and thisn category also has the highest count of conversion as compared to all other categories
- The ratio of conversion to not conversion is high for 'Working Professionals'
- Customers who have not specified their current occupation have do not tend to convert

univariate categorical analysis

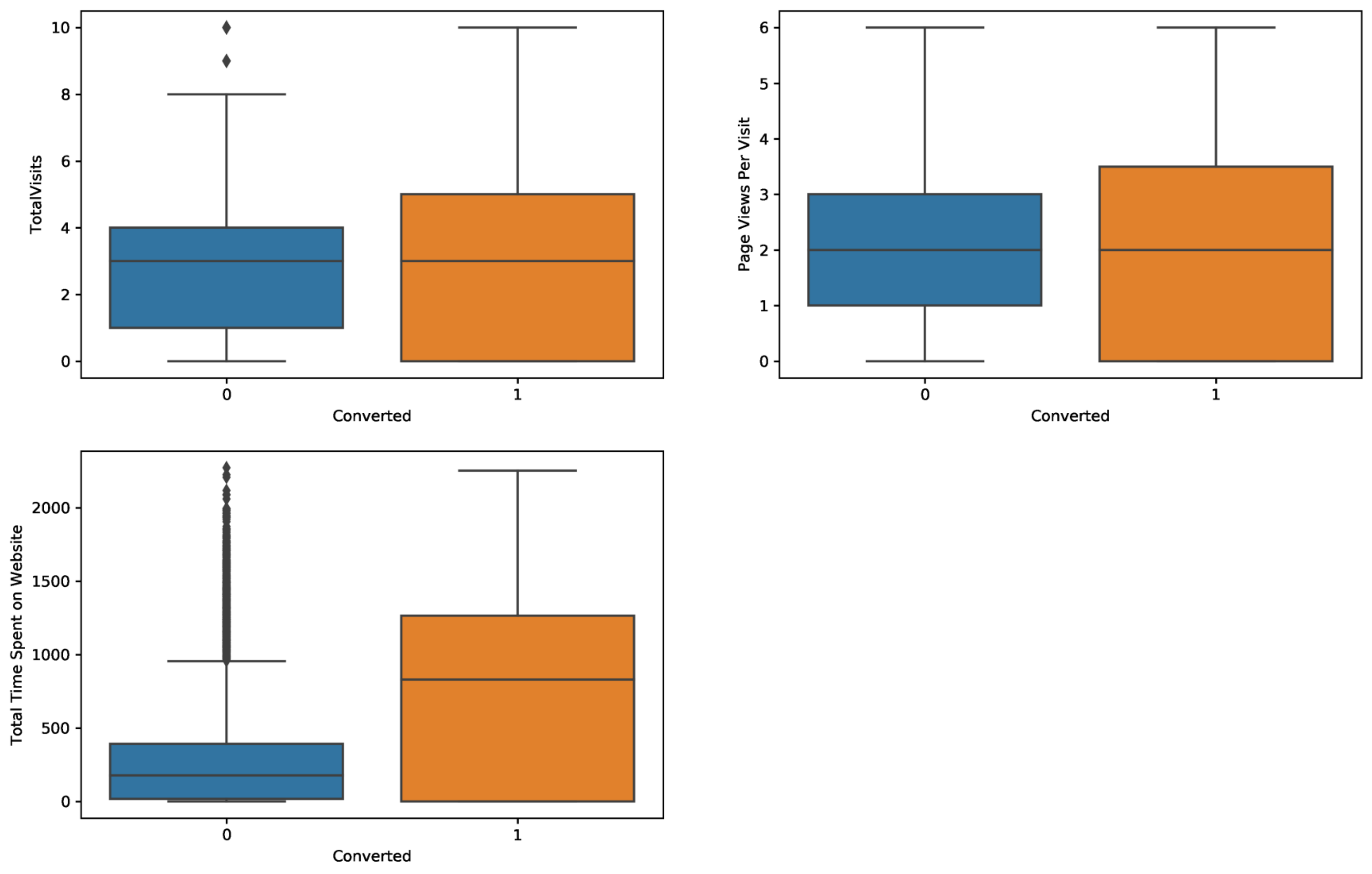








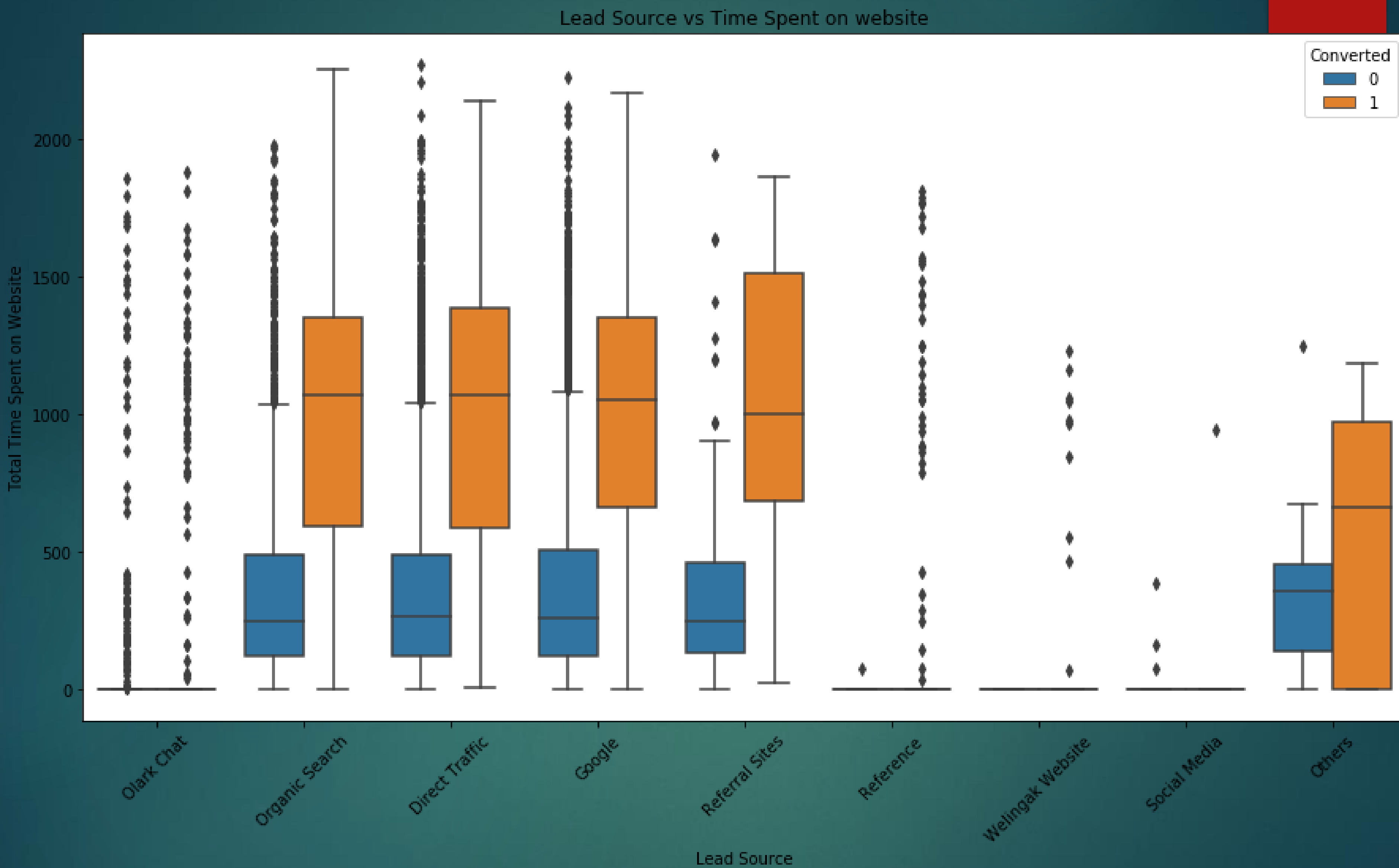
continuous univariate analysis



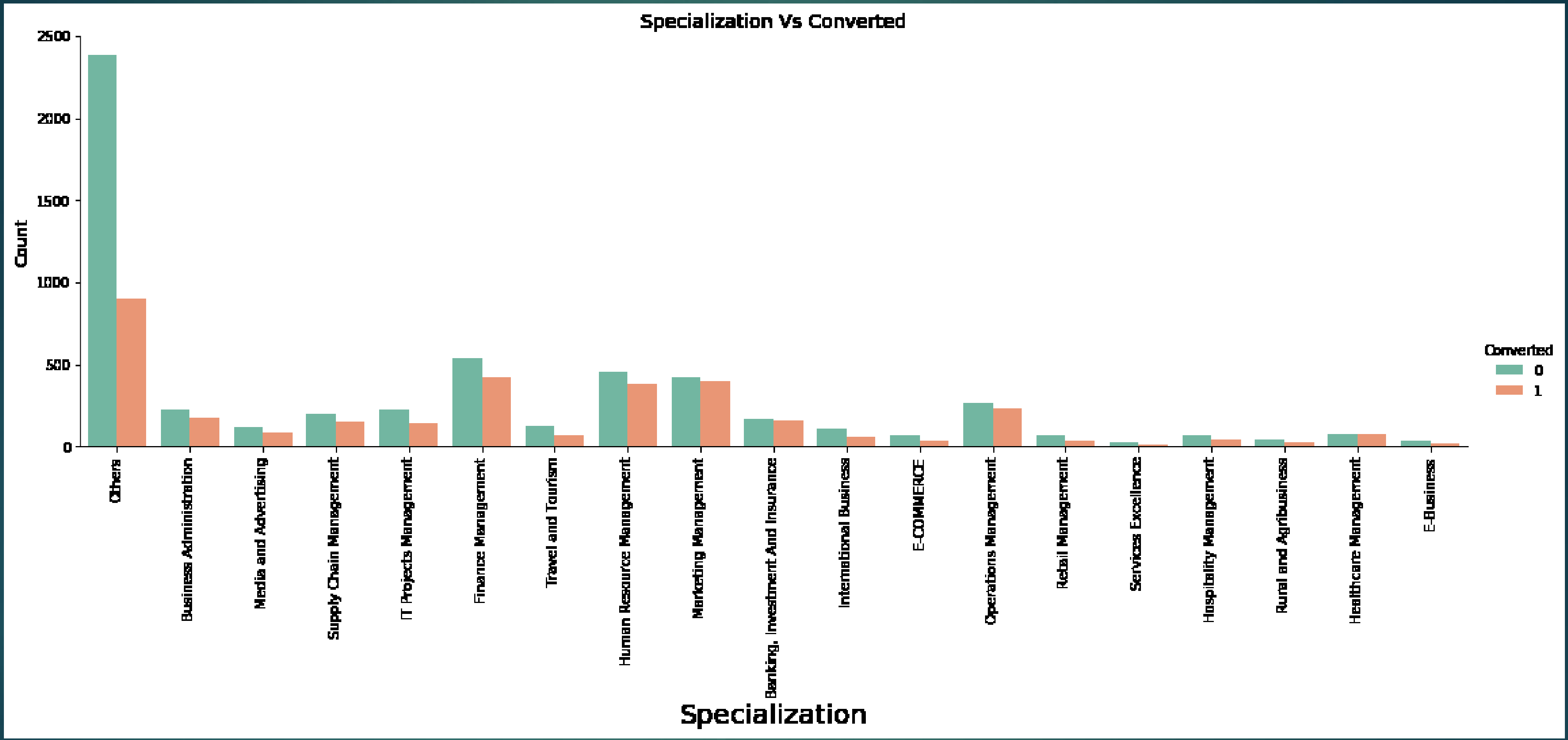
Observation:

- Customers who spend more time on the website have high conversion rate
- We cannot say much about the conversion from the Pages viewed as the means is almost same for Converted , yes and No

bivariate analysis of Continuous-Categorical columns



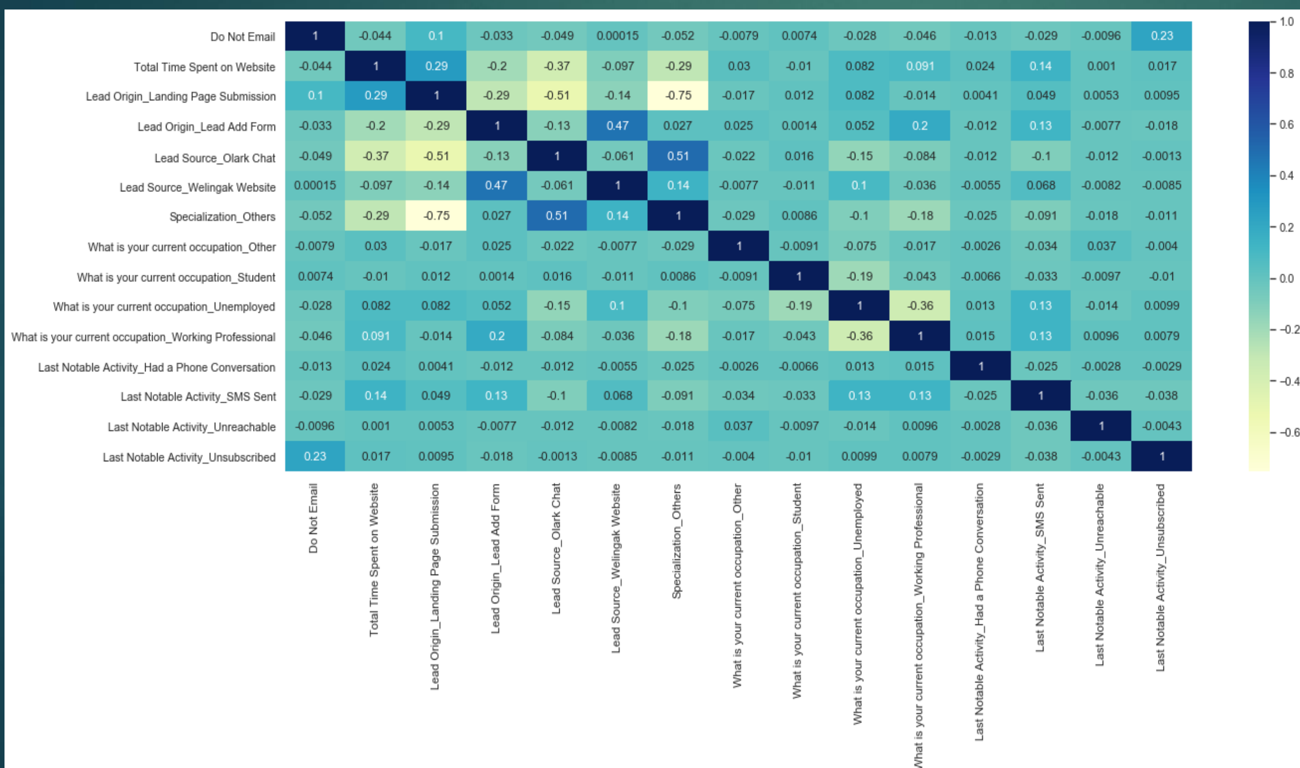
- Leads coming in through referral sites , Google, Direct traffic, Organic Search spend more time on the website and also have a good conversion rate plotting spread of Specialization columnn after replacing NaN values



- As seen above , the specialization column has 19 categories , so to reduce these categories for performing one hot encoding , we will combine all management related specializations in one broad category called as 'Management'

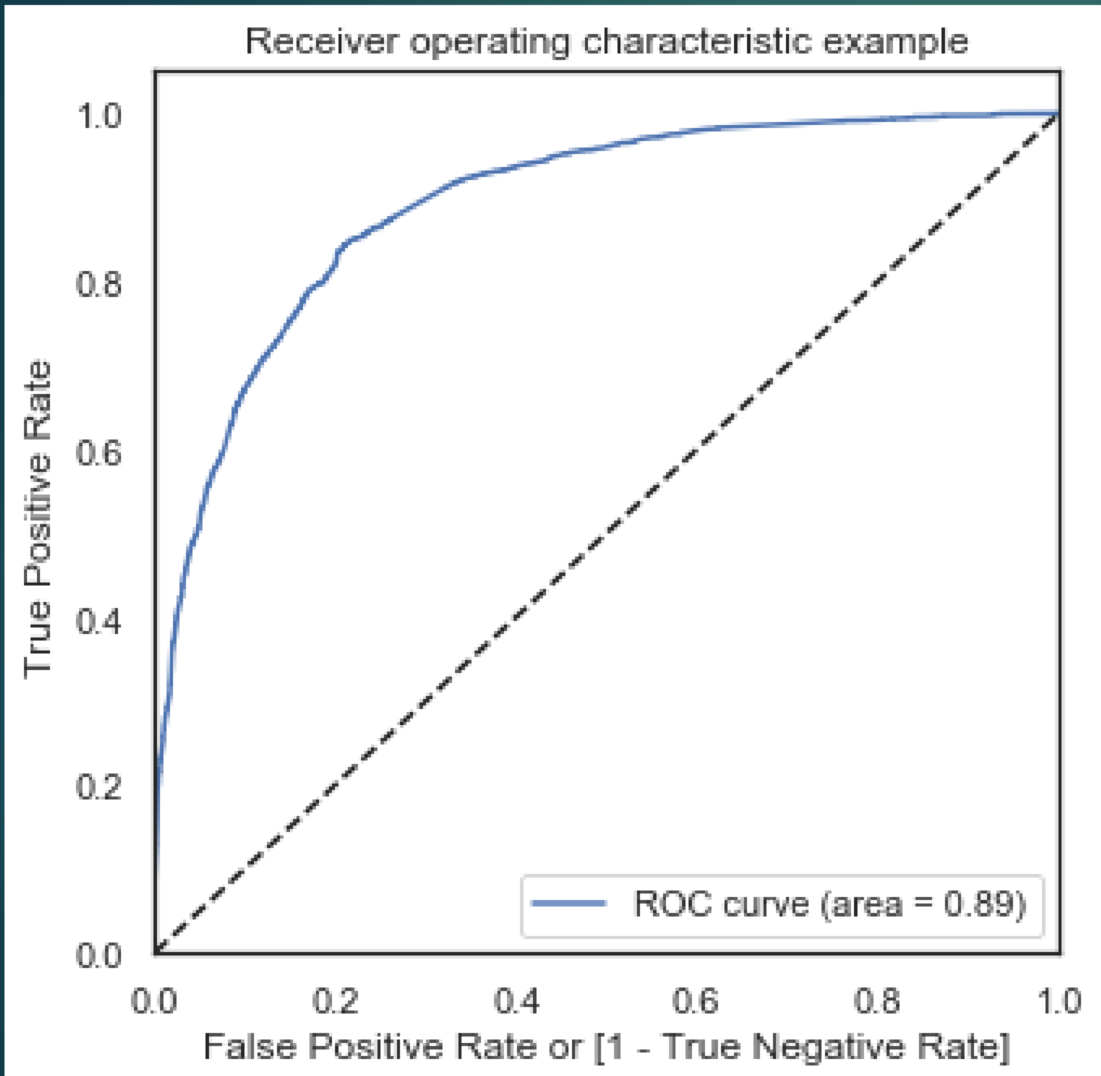
•Generalized Linear Models from StatsModels is used to build the Logistic Regression model.
•The model is built initially with the 20 variables selected by RFE.
•Unwanted features are dropped serially after checking p values (< 0.5) and VIF (< 5) and model is built multiple times.
•The final model with 16 features, passes both the significance test and the multi-collinearity test.

Building the model



•A heat map consisting of the final 16 features proves that there is no significant correlation between the independent variables.

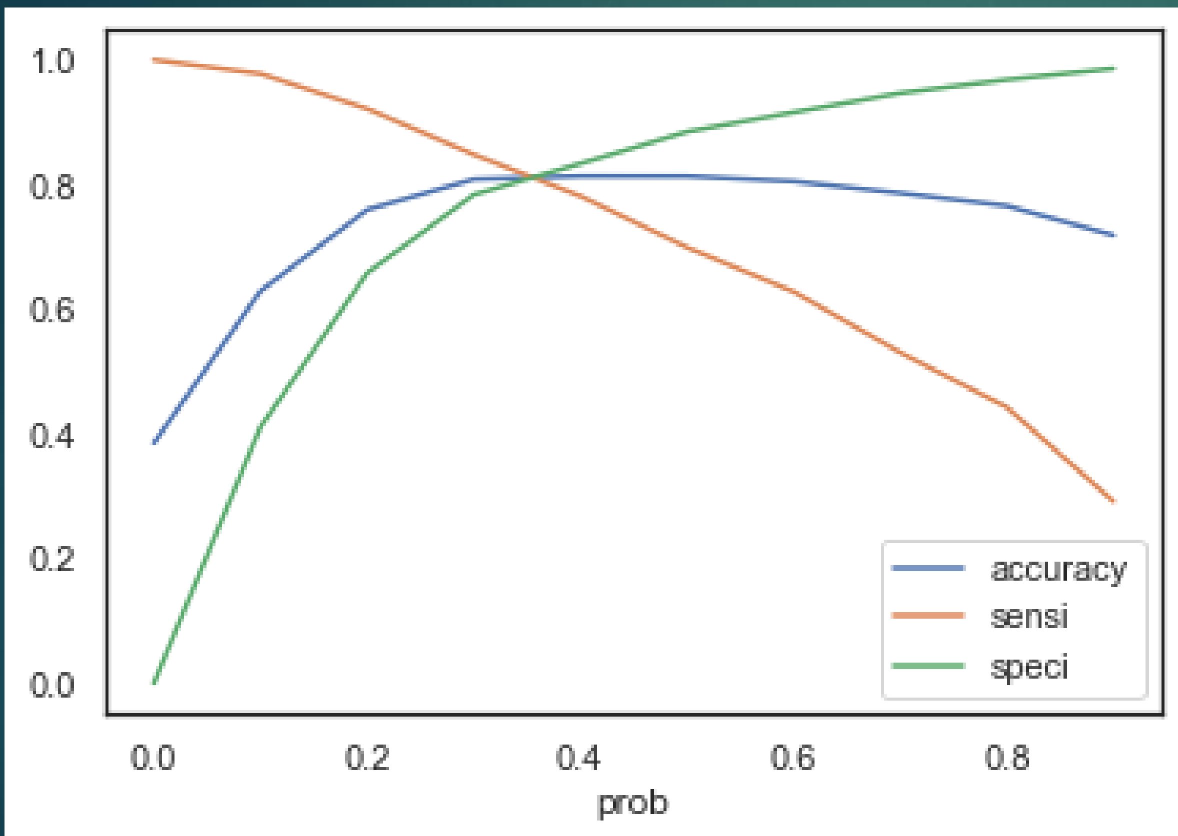
Plotting the ROC Curve



An ROC curve demonstrates several things:

It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

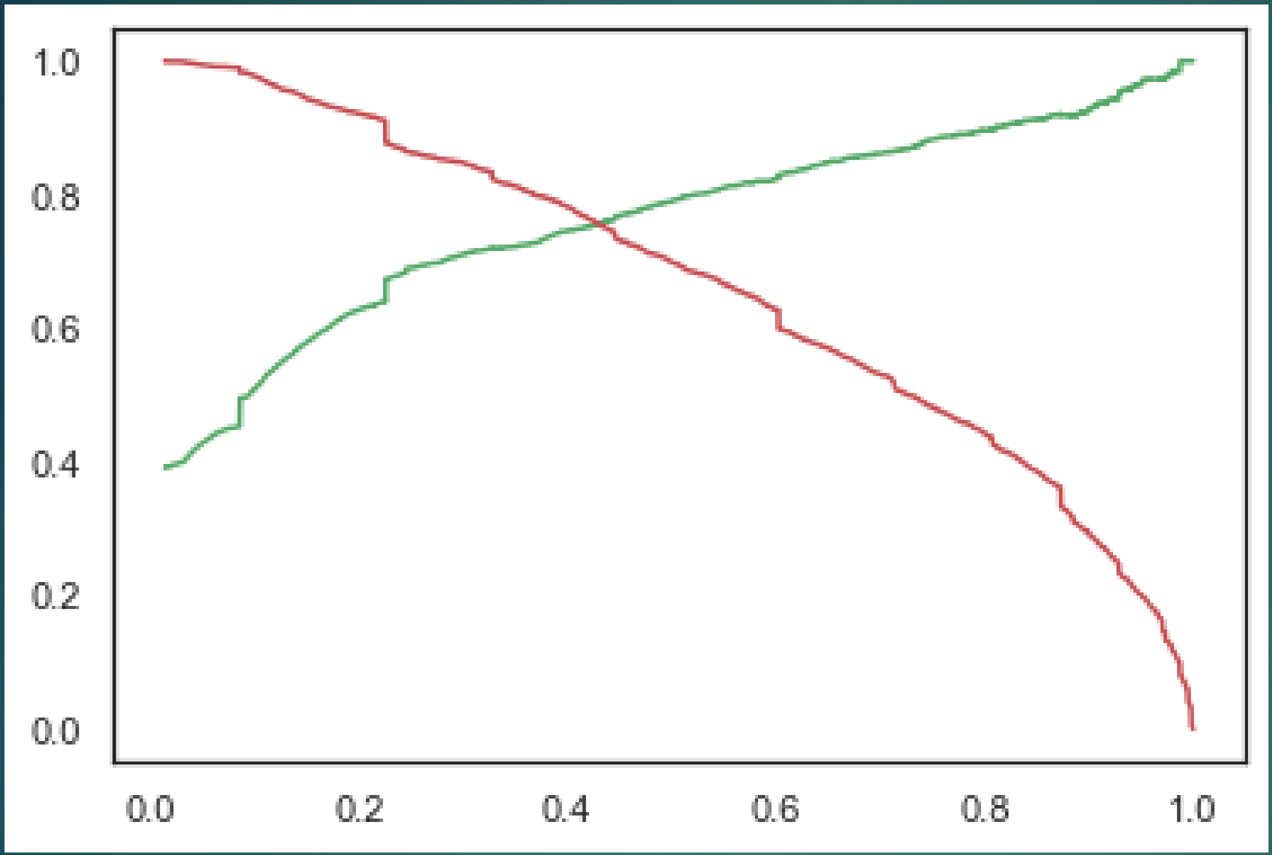
plotting accuracy sensitivity and specificity
for various probabilities.



	prob	accuracy	sensi	speci
0.0000	0.0000	0.3851	1.0000	0.0000
0.1000	0.1000	0.6292	0.9775	0.4110
0.2000	0.2000	0.7594	0.9215	0.6579
0.3000	0.3000	0.8084	0.8479	0.7836
0.4000	0.4000	0.8133	0.7813	0.8333
0.5000	0.5000	0.8131	0.6991	0.8845
0.6000	0.6000	0.8049	0.6280	0.9157
0.7000	0.7000	0.7862	0.5303	0.9465
0.8000	0.8000	0.7655	0.4424	0.9680
0.9000	0.9000	0.7185	0.2915	0.9859

- From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

Precision and recall tradeoff



Evaluating model on train set

Confusion Matrix

# **Predicted** / # **Actual**	Not Converted	Converted
Not Converted	3454	451
Converted	736	1710

converted rate = 0.37

Accuracy TP +TN/ (TP+TN+FN+FP)	Sensitivity TP / (TP+FN)	Specificity TN / (TN+FP)	False Positive Rate FP/ (TN+FP)	Positive PredictiveValue TP / (TP+FP)
0.80	0.84	0.78	0.21	0.71

Negative PredictiveValue TN / (TN+ FN)	Precision TP / TP + FP	Recall TP / TP + FN	*F1 score = 2×(Precision*Recall)/(Precision+Recall)*	Area under the cuve
0.89	0.71	0.84	0.77	0.89

Making predictions of test set

•The final model on the train dataset is used to make predictions for the test dataset
•The train data set was scaled using the scaler.transform function that was used to scale the train dataset.
•The Predicted probabilities were added to the leads in the test dataframe.
•Using the probability threshold value of 0.33, the leads from the test dataset were predicted if they will convert or not.

	Converted	Lead Number	Converted_Prob	final_predicted
0	0	3271	0.0428	0
1	1	1490	0.9612	1
2	0	7936	0.0364	0
3	1	4216	0.8842	1
4	0	3830	0.0444	0

•The Conversion Matrix was calculated based on the Actual and Predicted 'Converted' columns.

Evaluating model on test set

The following evaluation metrices were recorded for the test dataset.

Metric	Formula	Value
Accuracy	TP +TN/ (TP+TN+FN+FP)	0.80
Sensitivity	TP / (TP+FN)	0.83
Specificity	TN / (TN+FP)	0.78
Negative PredictiveValue	TN / (TN+ FN)	0.89
Precision	TP / TP + FP	0.71
Recall	TP / TP + FN	0.84
F1 score	*= 2×(Precision*Recall)/(Precision+Recall)*	0.77

Formula for Lead Score calculation

Lead Score is calculated for all the leads in the original dataframe.

Formula for Lead Score calculation is:

Lead Score = 100 * Conversion Probability

LeadID	Lead Number	Converted	Conversion_Prob	final_predicted	Lead_Score
0	660737	0	0.03	0	3
1	660728	0	0.01	0	1
2	660727	1	0.80	1	80
3	660719	0	0.01	0	1
4	660681	1	0.96	1	96
5	660680	0	0.08	0	8
6	660673	1	0.96	1	96
7	660664	0	0.08	0	8
8	660624	0	0.08	0	8
9	660616	0	0.08	0	8

The figure showing Lead Score for top 10 records from the data set.

- The train and test dataset is concatenated to get the entire list of leads available.
- The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.
- Higher the lead score, higher is the probability of a lead getting converted and vice versa,
- Since, we had used 0.3 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of 33
or above will have a value of '1' in the final_predicted column.

Conclusion:

We have successfully built a Logistic Regression Model with below Evaluation scores:

Train Set:
- sensitivity 84%
- specificity 78%
- Accuracy 80%

Test Set:
- sensitivity 83%
- specificity 78%
- Accuracy 80%

1) In Order to increase the lead conversion rate ,the sales team can follow up with customers who were reached out by phone call as the last activity.
2) Customers who belong to the working Professional as their current occupation
3) Concentrate on customers for whom the lead source was Welingak Website
4) Customers for whom the lead source was reference
5) If Customers opt for 'Do not email' as Yes , then their conversion rate is low , as they do not want to be reached out via email, so the possibility of conversion is low

In combination with the lead score and above-mentioned columns , the conversation rate for X education company can be increased.

Scalability:

The threshold value can be tweaked to increase or decrease the sensitivity of the model. Depending on the business scenario we can choose the metrics of accuracy, sensitivity, specificity etc