LEAD SCORING CASE STUDY

LOGISTIC REGRESSION

Suyash Sharma
Sri Ram Srikakolapu
Manjeet Sudhanshu

PROBLEM STATEMENT

X Education is an organization which provides online courses for industry professional. The company marks its courses on several popular websites like google.

X Education wants to select most promising leads that can be converted to paying customers.

Although the company generates a lot of leads only a few are converted into paying customers, wherein the company wants a higher lead conversion. Leads come through numerous modes like email, advertisements on websites, google searches etc.

The company has had 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversions.

BUSINESS GOAL

The company requires a model to be built for selecting most promising leads.

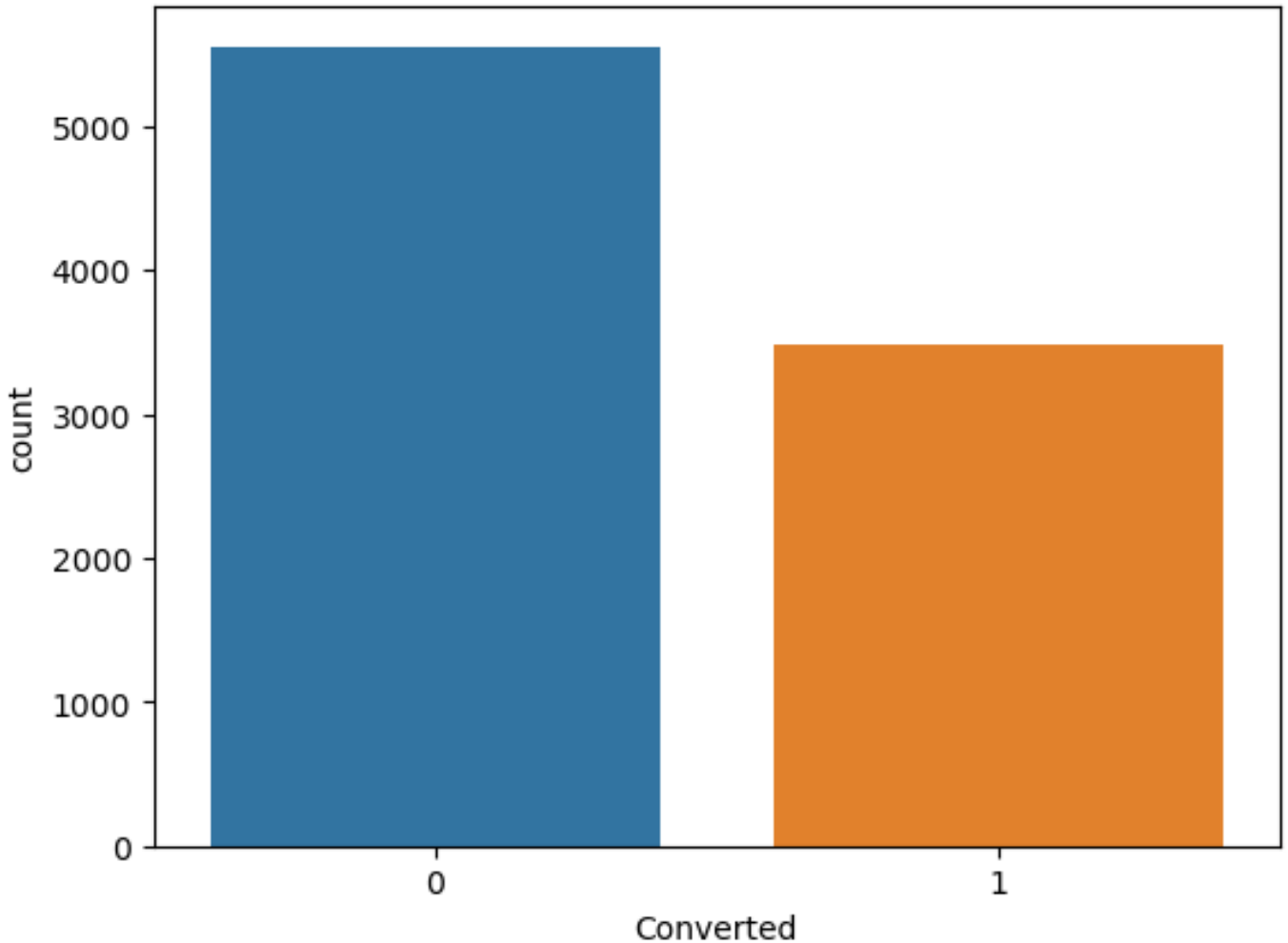
Lead score to be given to each leads such that it indicates how promising the lead could be. The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion.

The model to be built in lead conversion rate around 80%.

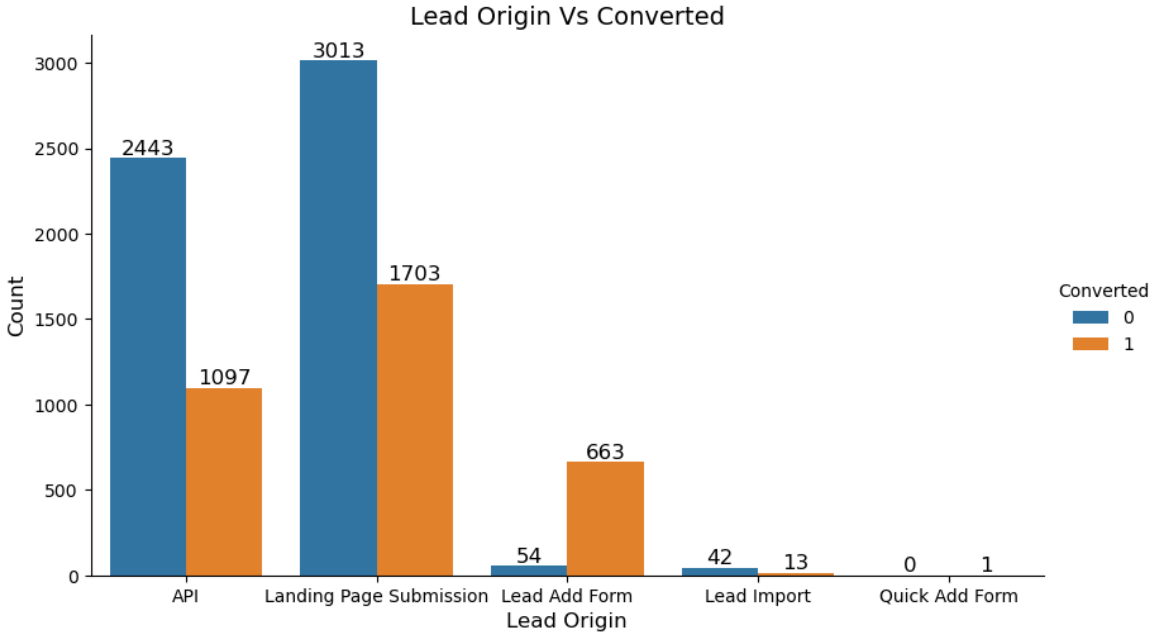
Strategy

1. Reading the data and data understanding
2. Cleaning the data
3. Data analysis
4. Data preparation (Dummy creation)
5. Train-Test split
6. Rescaling the features
7. Model building
8. Feature selection (using RFE)
9. ROC curve
10. Finding optimal cutoff
11. Predictions on test set

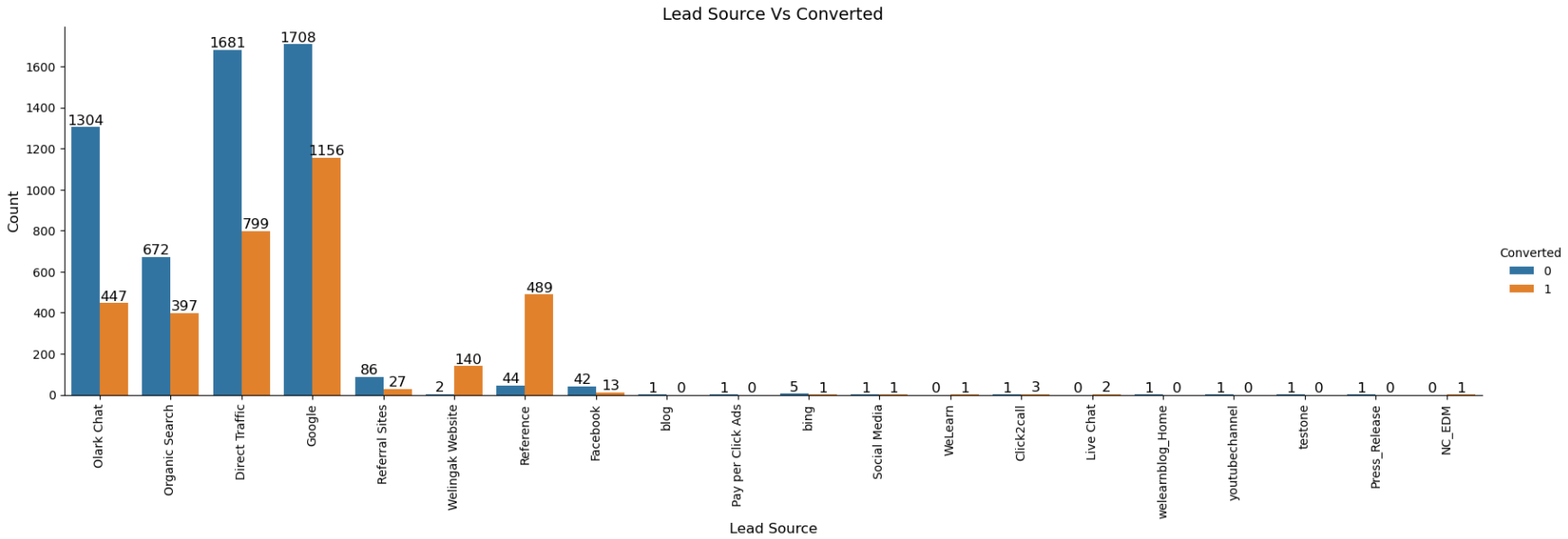
Exploratory Data Analysis



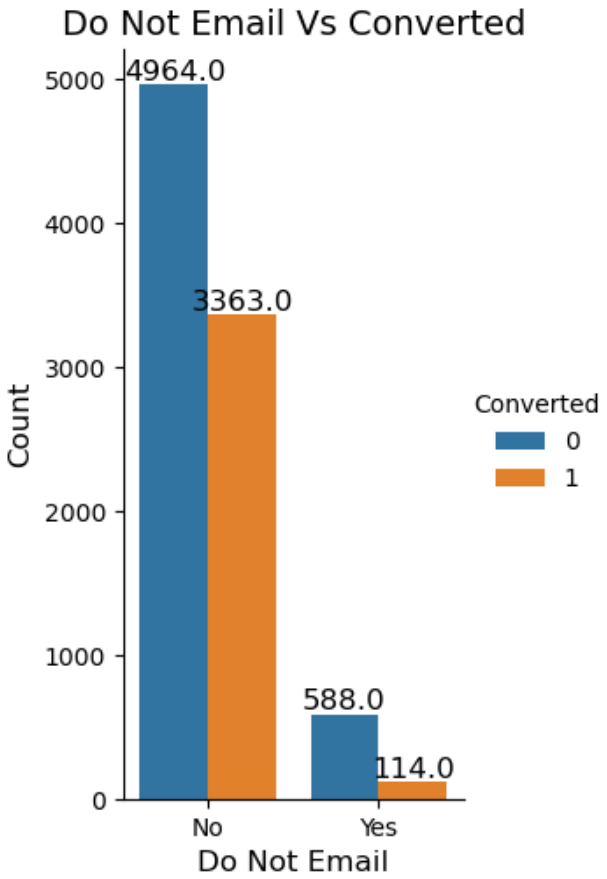
The overall conversion rate is around 39%



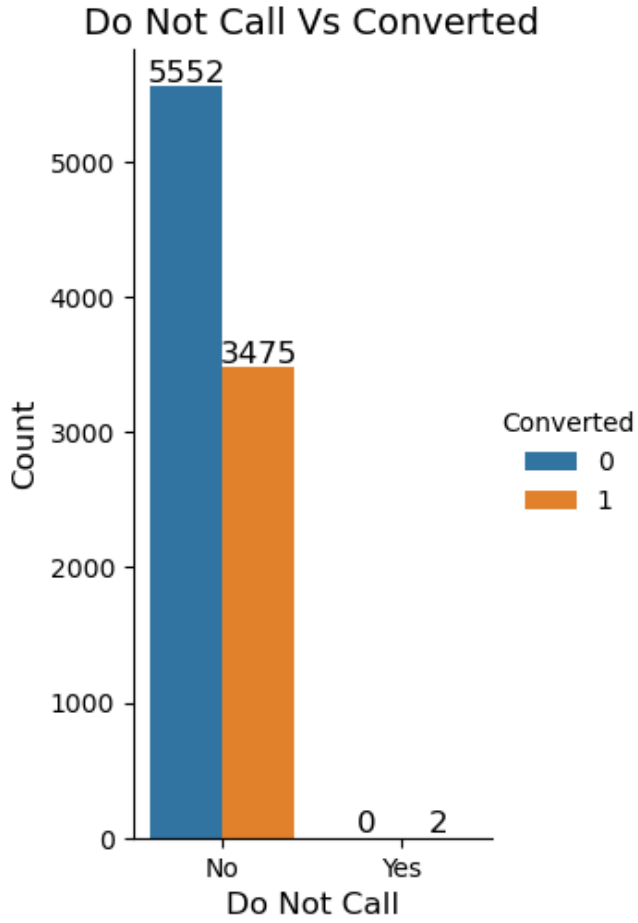
So, we can see that the maximum conversion happened from Landing Page Submission. Also, there was only one request from Quick Add Form which got converted.



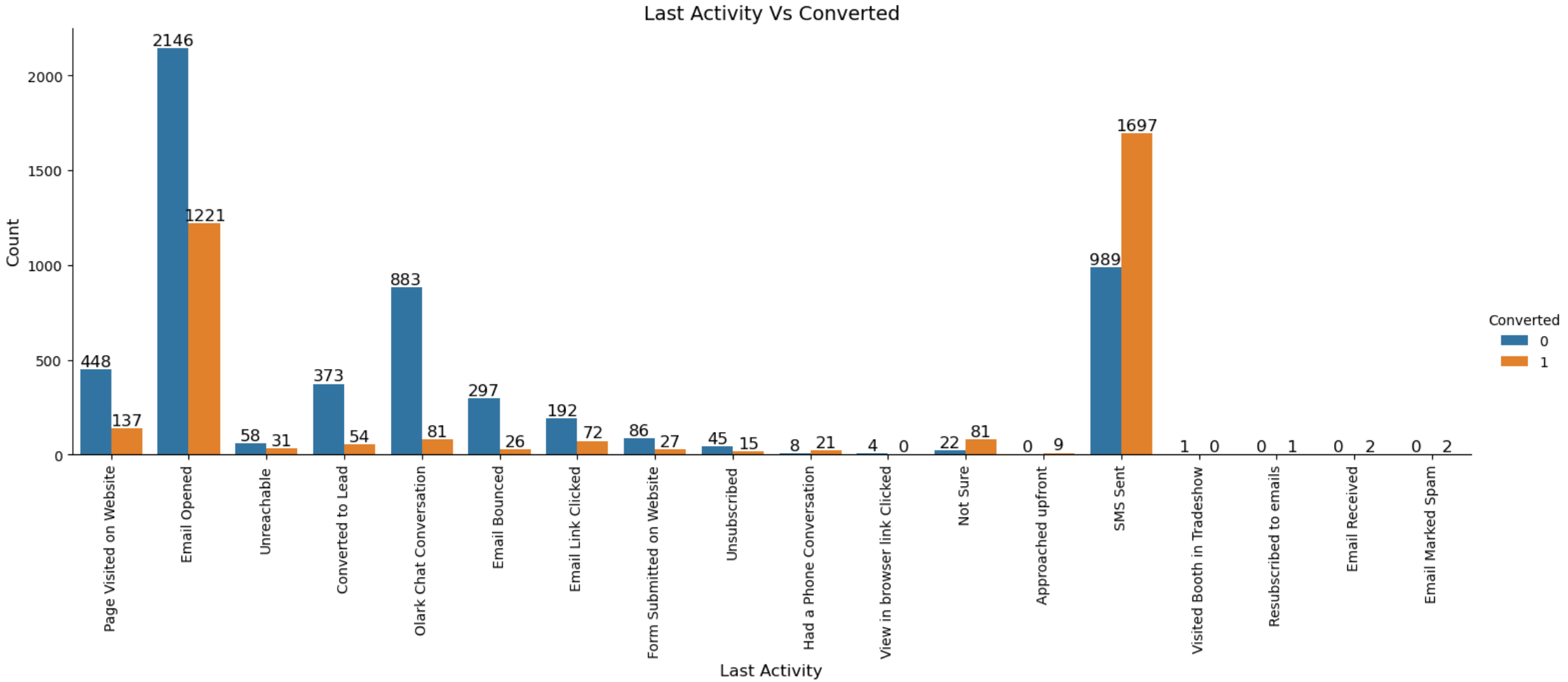
So, we can see that major conversion in the lead source is from Google.



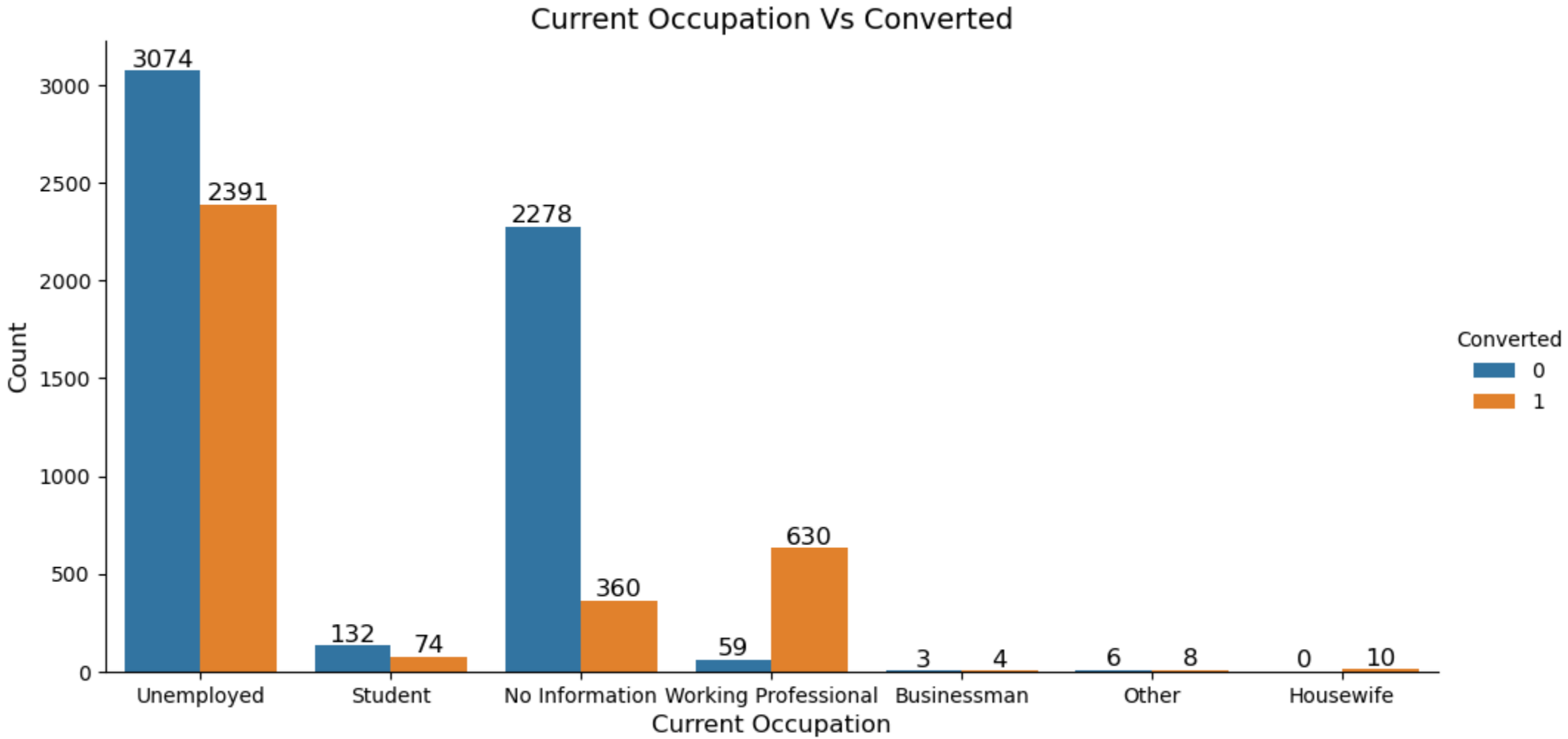
So, from the above graph we can see that major conversions have happened from the emails that have been sent.



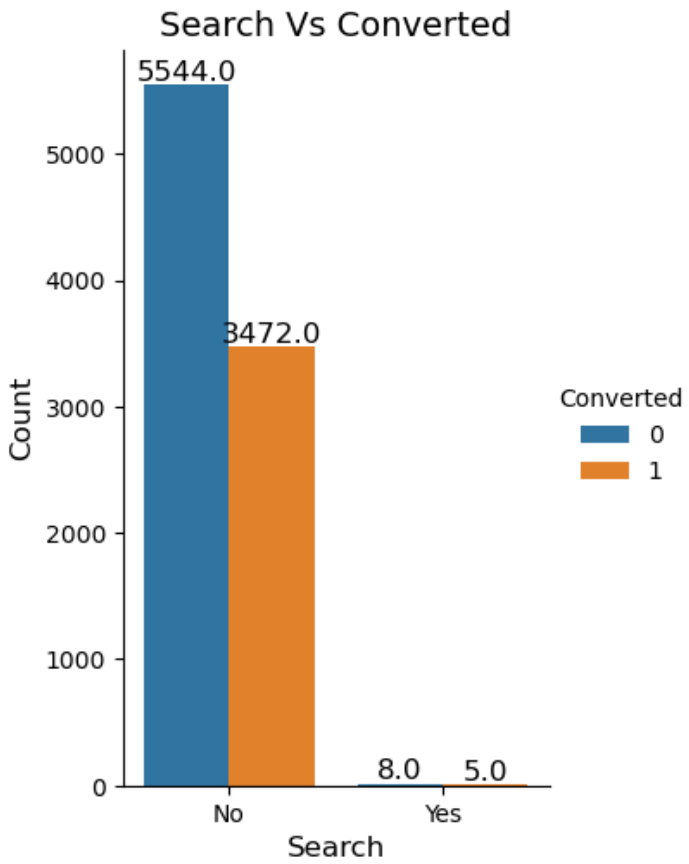
So from the above graph we can see that major conversions happened when the calls were made. But 2 leads still got converted even though they had opted for 'Do Not Call'.



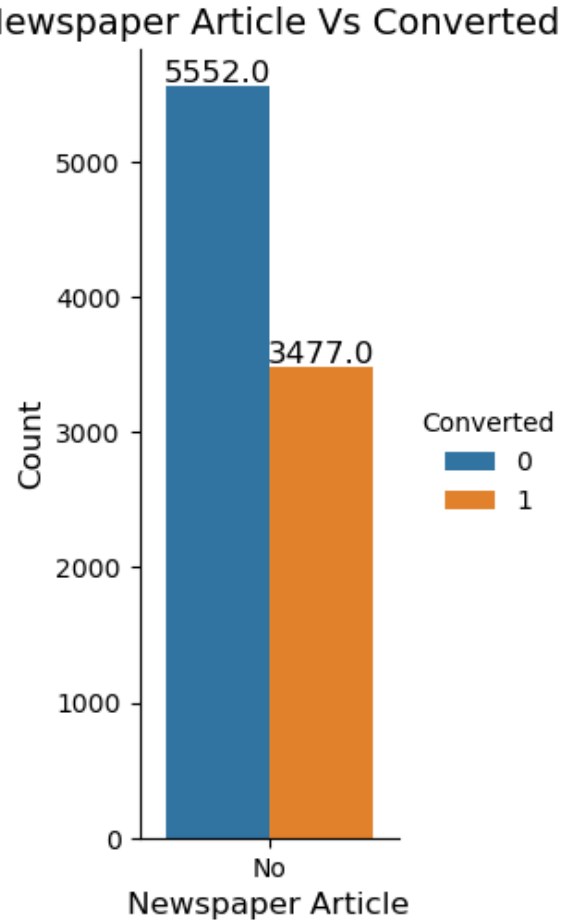
So from the above graph, we can see that Last Activity value 'SMS Sent' had more conversion.



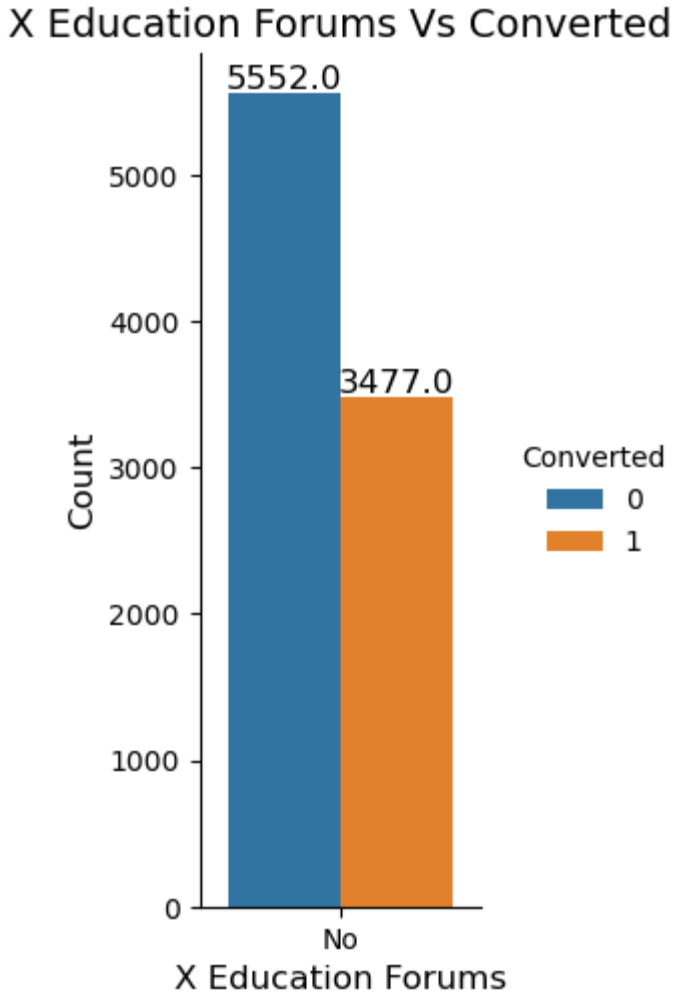
From the above graph we can say that people who are unemployed have converted more. Out of 7 businessman, 4 got converted. All the 10 out of 10 housewives got converted.



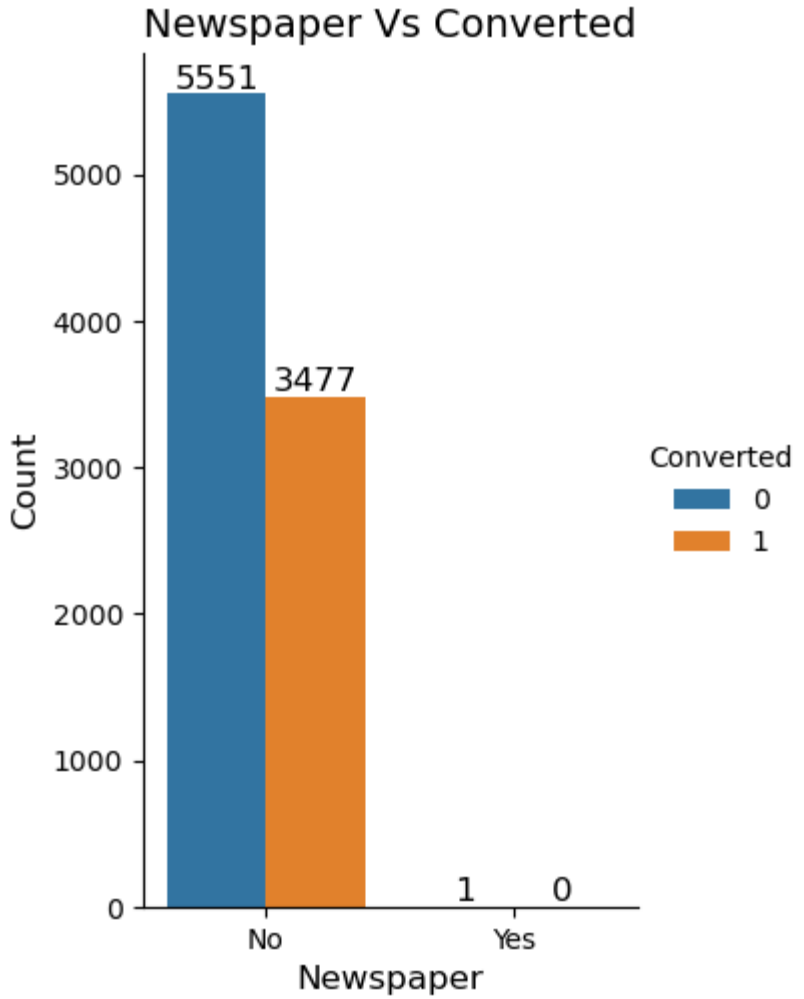
Conversion rate is high for leads who are not through search.



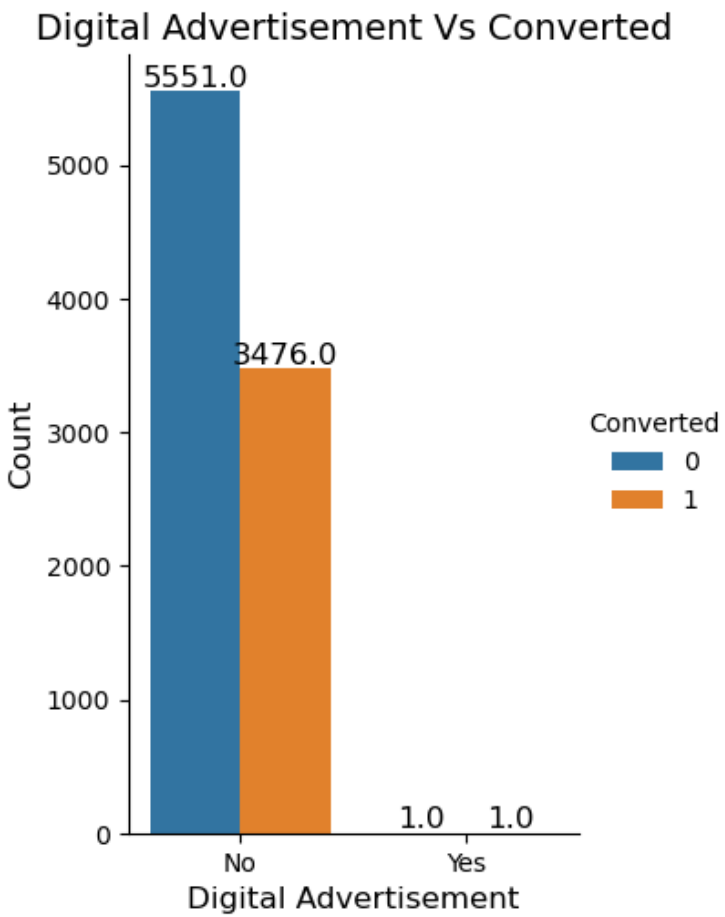
The Newspaper Article has only one value 'No', so we can drop this column.



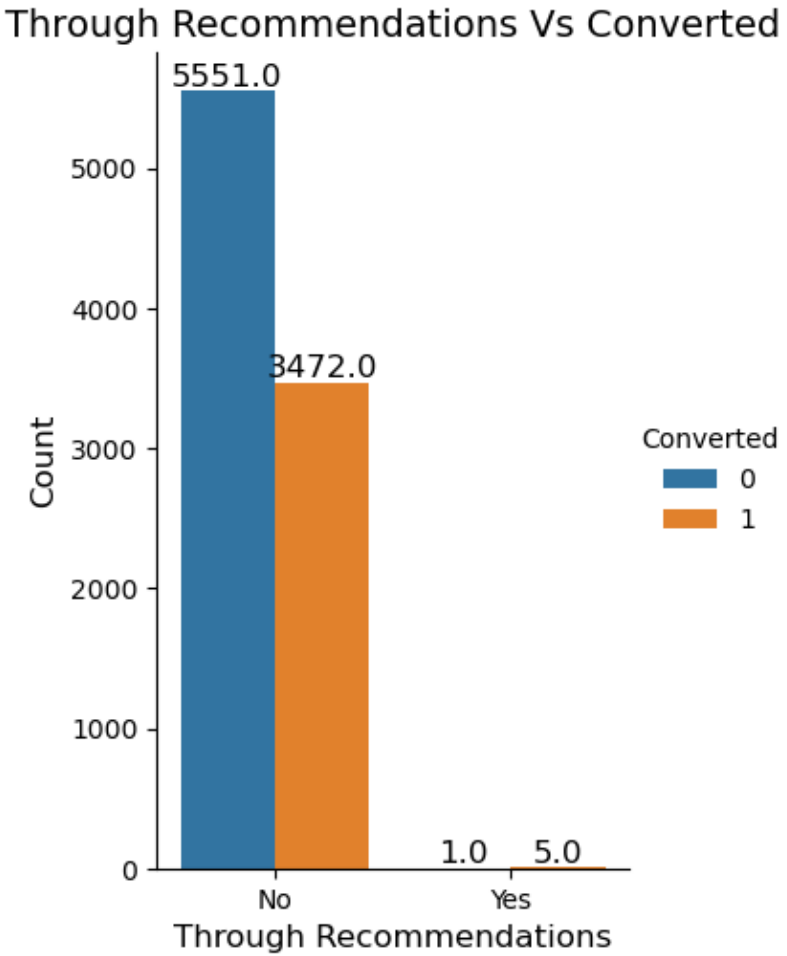
The X Education Forums column has only one value 'No', so we can drop this column.



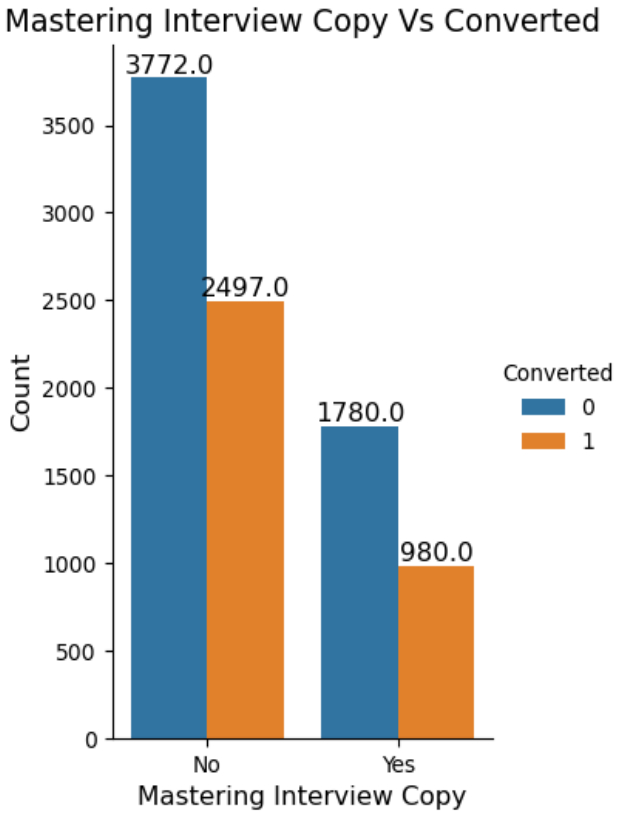
The Newspaper column has only one row with "Yes" as the value and that too did not get converted, therefore we can drop this column.



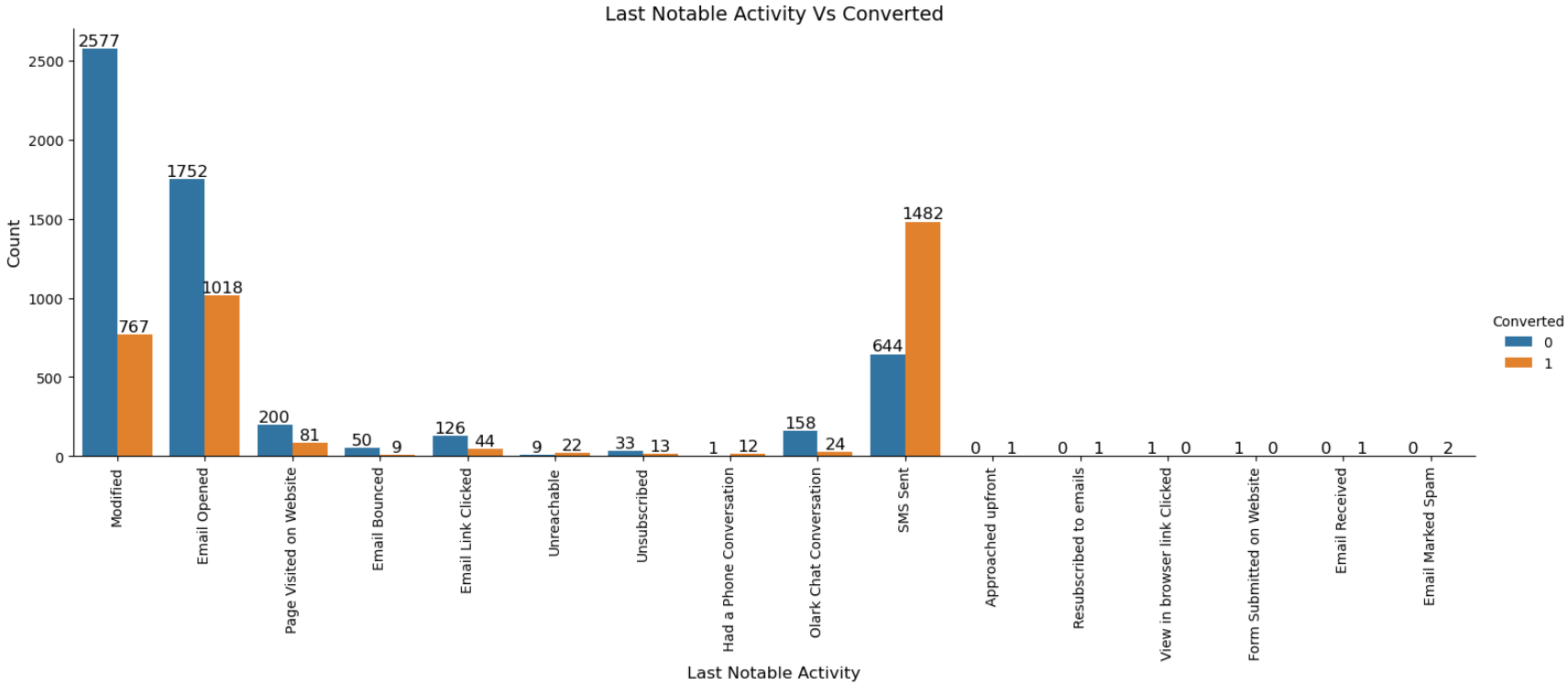
From the above graph we can see that there are two leads from Digital Marketing, out of which one got converted.



We can that six leads were from Recommendations out of which 5 got converted.

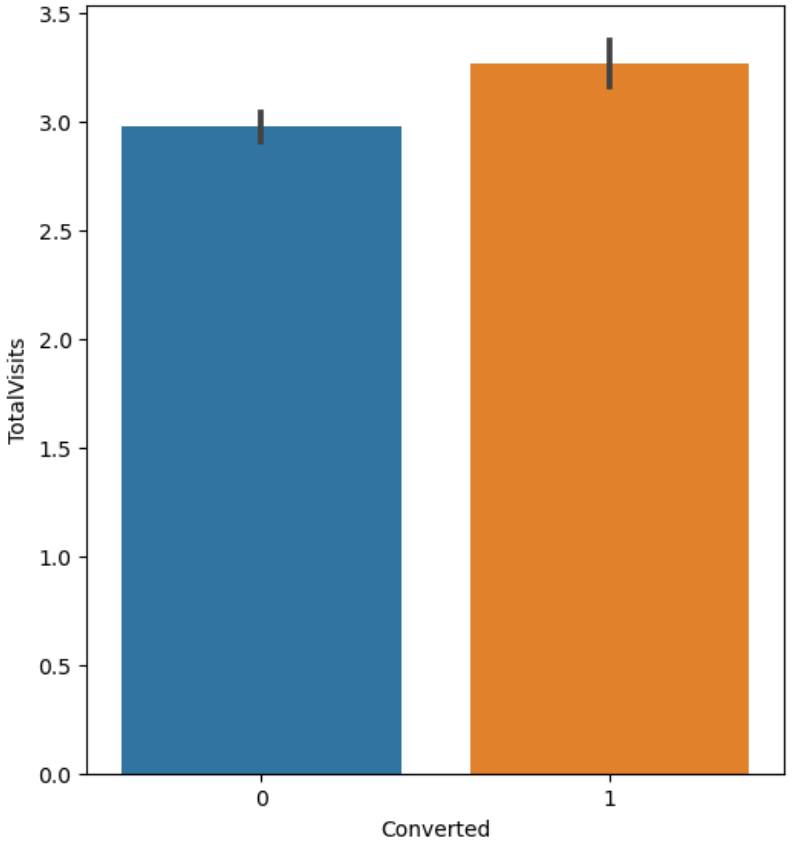


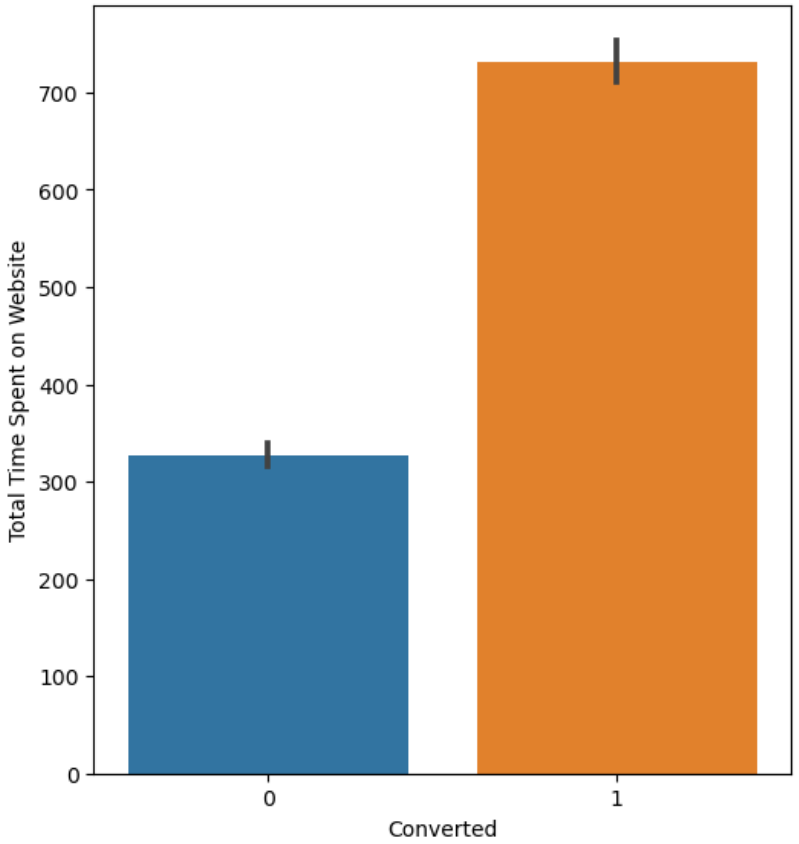
There is a high conversion rate for those leads who do not want a free copy of Mastering Interviews.

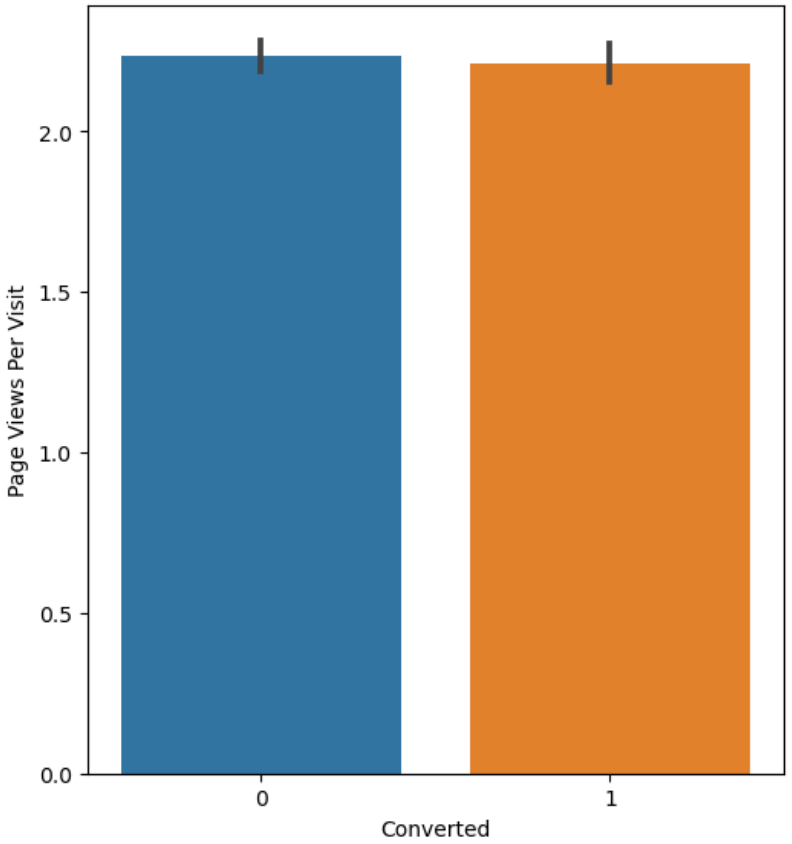


From the above graph we can see that the conversion rate is high for "SMS Sent".

Checking the conversions for all numeric values





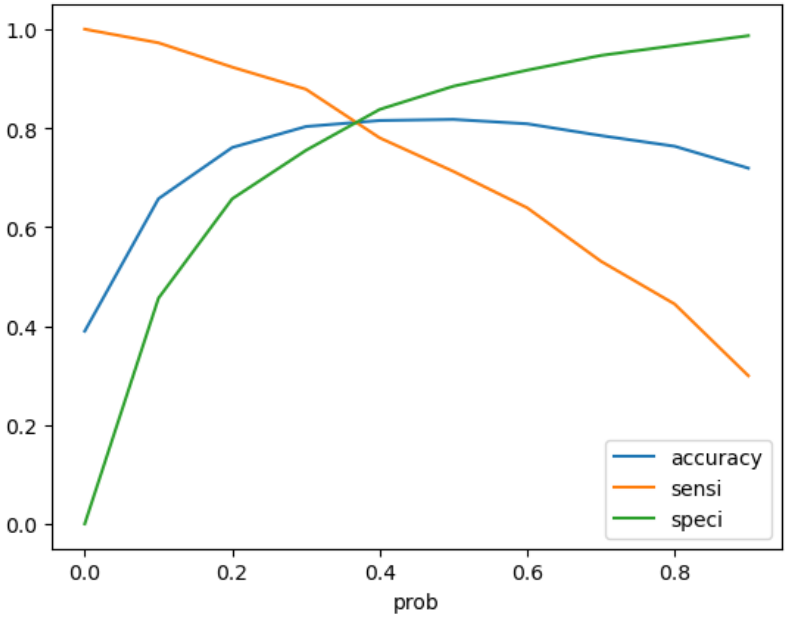


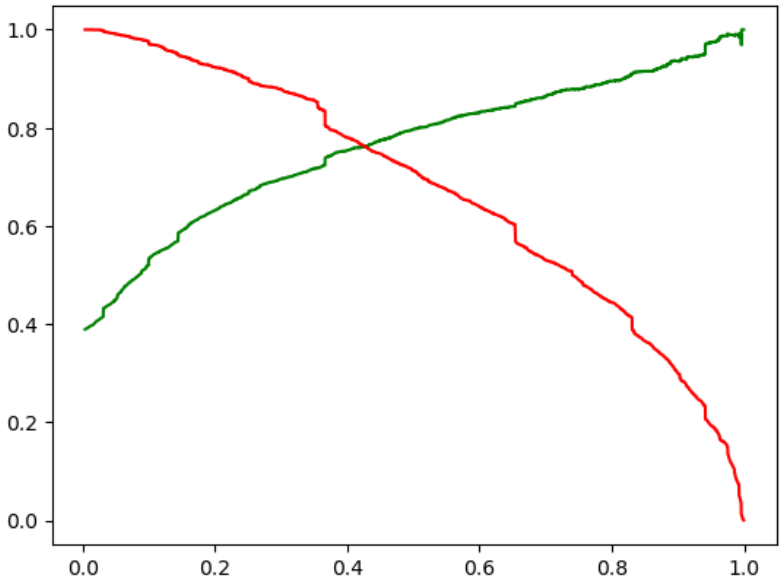
So, we can say that the conversion rates are high for Total Visits, Total time spent on website and Page views per visit.

MODEL BUILDING

- Splitting into train and test set
- Scale variables in train set
- Build the initial model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate the variables based on high p-values
- Check VIF value for all the existing columns
- Predict using train set
- Evaluate accuracy and other metrics
- Predict using test set
- Precision and recall analysis on the test predictions

MODEL EVALUATION (TRAIN)





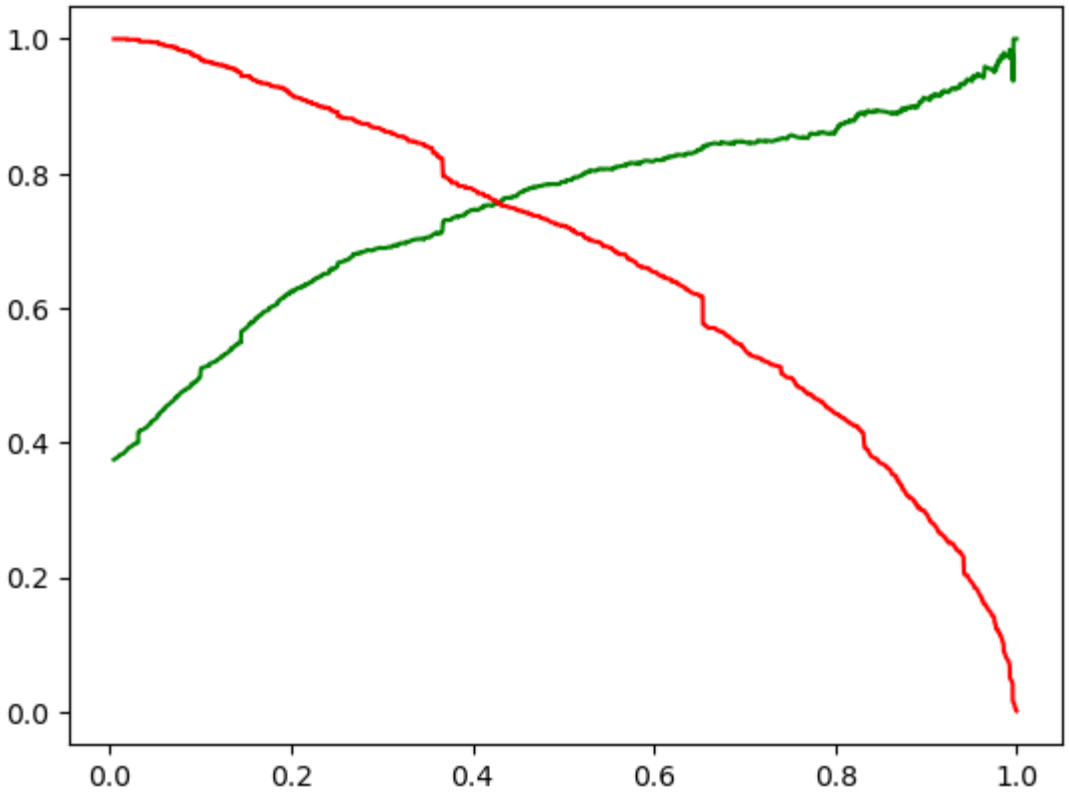
Confusion Matrix :-

[3167, 691],
[487, 1975]

Accuracy : 81.36 %
Sensitivity : 80.2%
Specificity : 82.08 %

Precision : 79.7 %
Recall : 71.2%

MODEL EVALUATION (TEST)



Confusion Matrix :-

[1397, 297],
[209, 806]

Accuracy : 81.32 %
Sensitivity : 79.4 %
Specificity : 82.46 %

Precision : 73.07 %
Recall : 79.40 %

CONCLUSION

EDA :

- People spending higher than average time are promising leads, so targeting them and approaching them can be helpful in conversions
- SMS can have a high impact on lead conversion
- Landing page submission can help find out more leads
- Unemployed people have converted more

Logistic Regression Model :

- The model shows about 81% accuracy
- The threshold has been selected from accuracy, sensitivity, specificity measures and precision, recall curves
- The model shows about 79% sensitivity and about 82% specificity
- The model finds correct promising leads and the leads that have less chances of getting converted
- Overall this model proves to be accurate