LEAD SCORE CASE STUDY

Panchali Kar

Problem Statement

An education company named X Education sells online courses to industry professionals.

The company markets its courses on several websites and search engines like Google.

The company is generating lot of leads but only 30% out of them is getting converted. The company wants a higher lead conversion rate by focusing it's resources on leads who are actually going to be converted.

BUISNESS GOAL

The goal is to build a ML model to find out the potential leads that can be converted to paying customer.

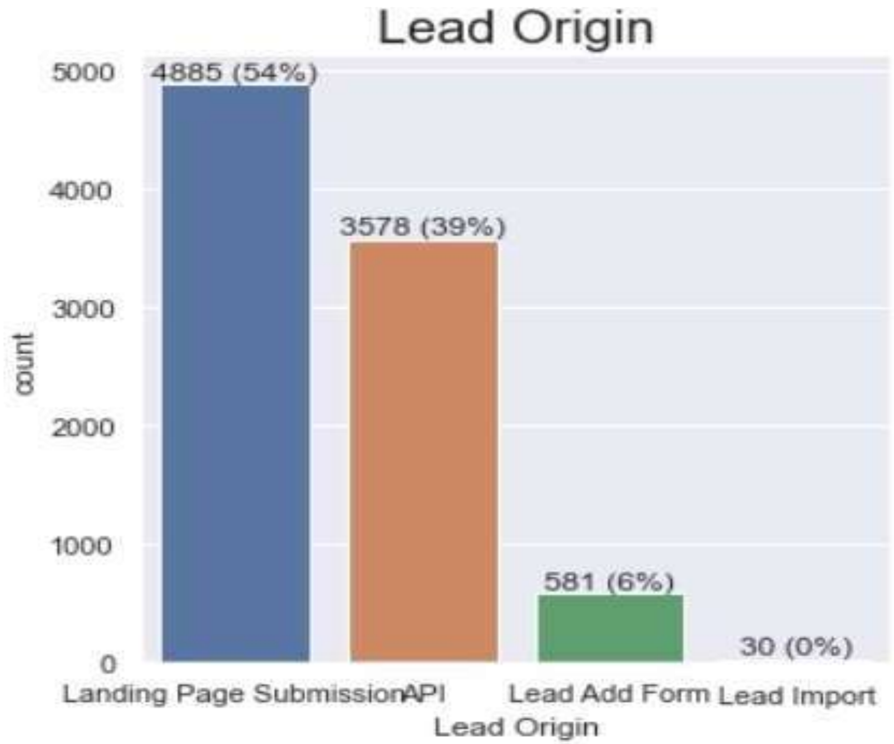
Lead score to be assigned between 0 to 100 to each of the leads which can be used by the Company to target the potential leads.

The model should be built on an lead conversion rate of around 80%

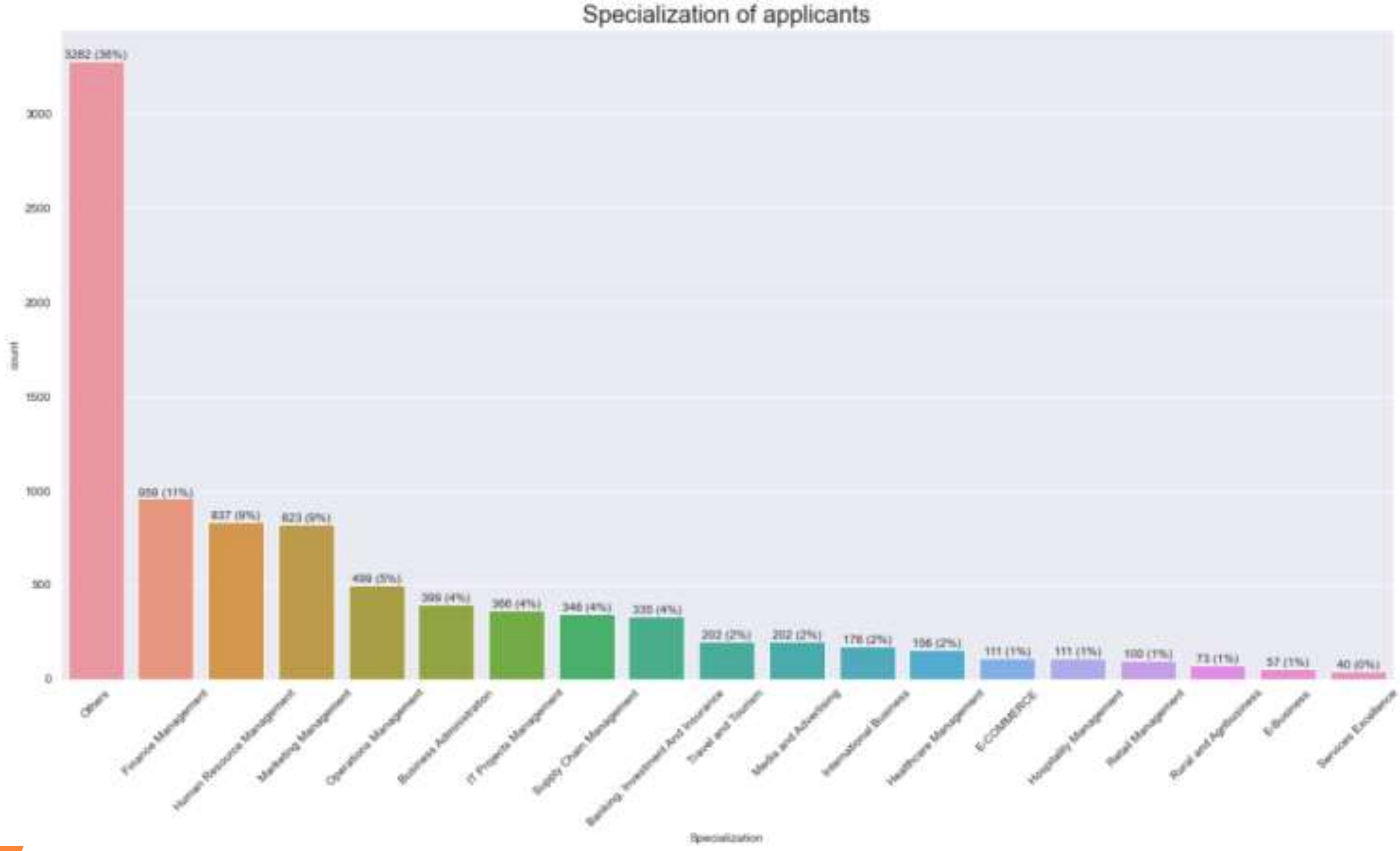
PROCESS FOLLOWED

1. Reading the data
2. Cleaning of data for further analysis
3. Imputing missing values of the dataset
4. Performing Exploratory data analysis to find out the most important attributes that will help us in further analysis.
5. Dividing data into train-test set
6. Scaling of continuous variables
7. Building of logistic regression model
8. Assign lead score to each of the leads
9. Test the model on Train-set
10. Evaluate model accuracy, sensitivity, specificity and other parameters
11. Test the model on Test-set
12. Evaluate different parameters of the model.

Data Visualization

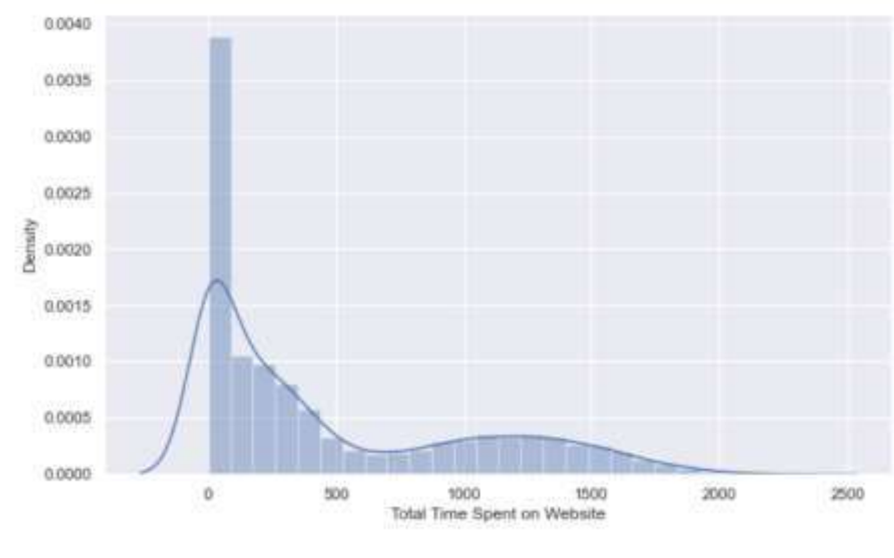


Maximum leads are generated by landing page submission

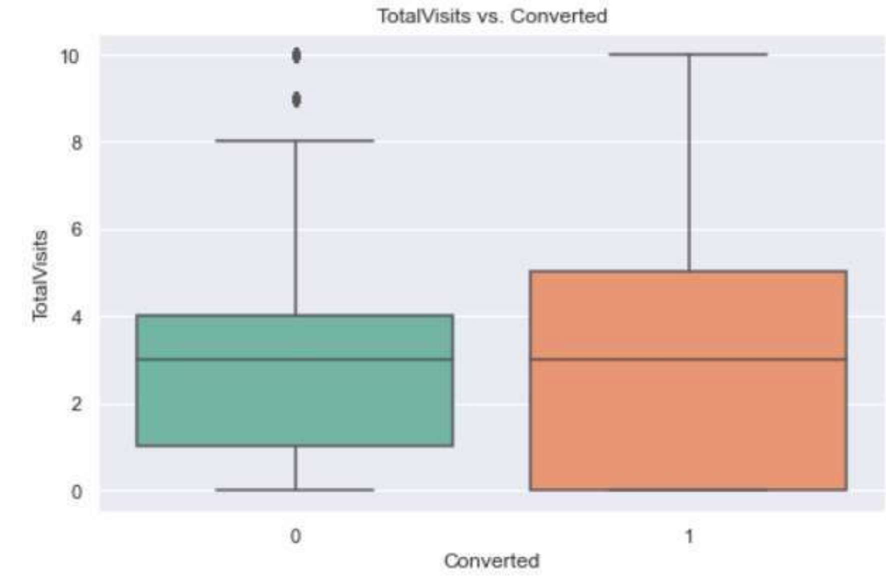


- Majority of the leads didn't provided their specialization while filling up the form.

- People from finance domain form the majority of the applicant pool, followed by HR and marketing

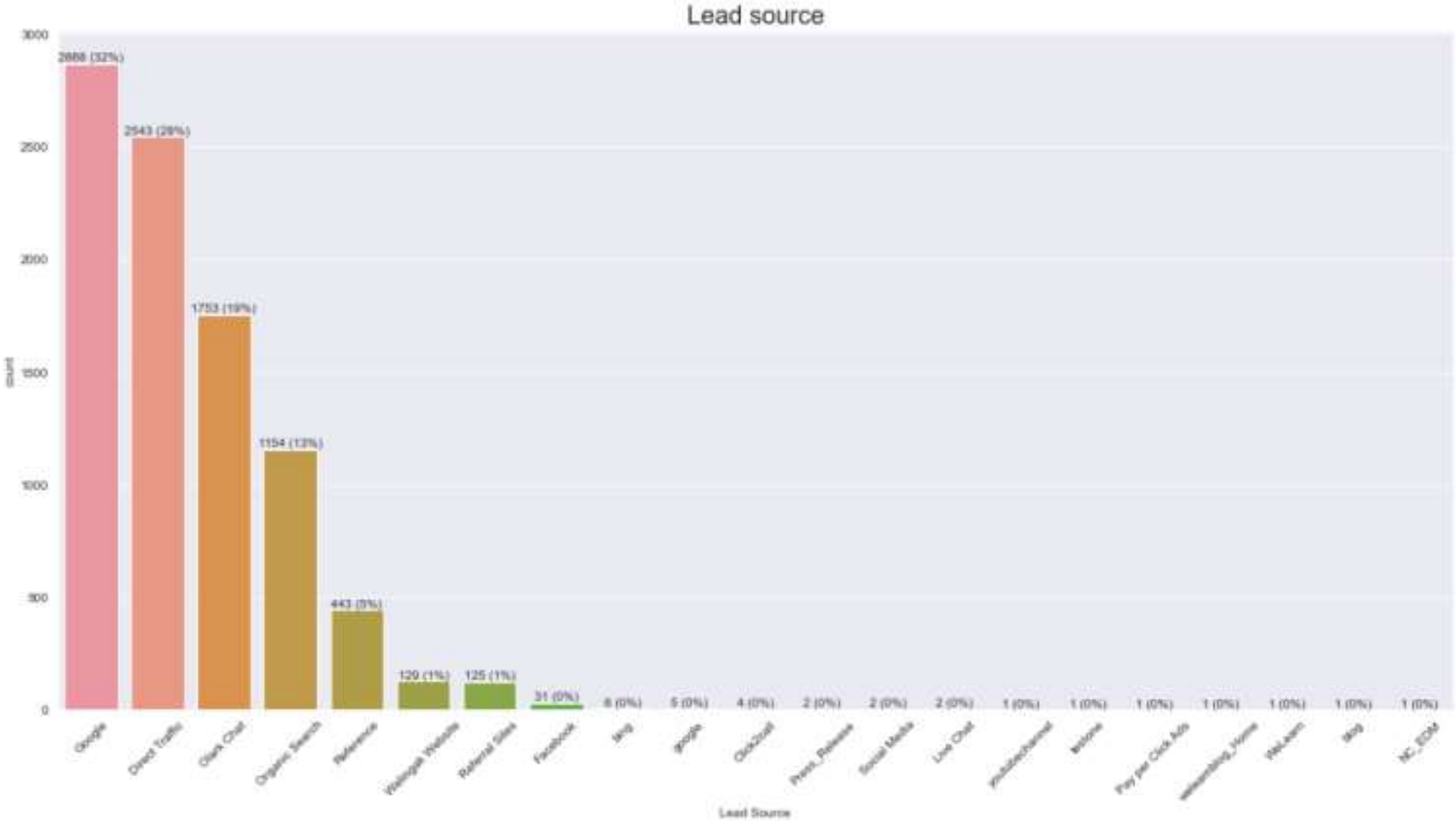


Most leads spend very less time on the website

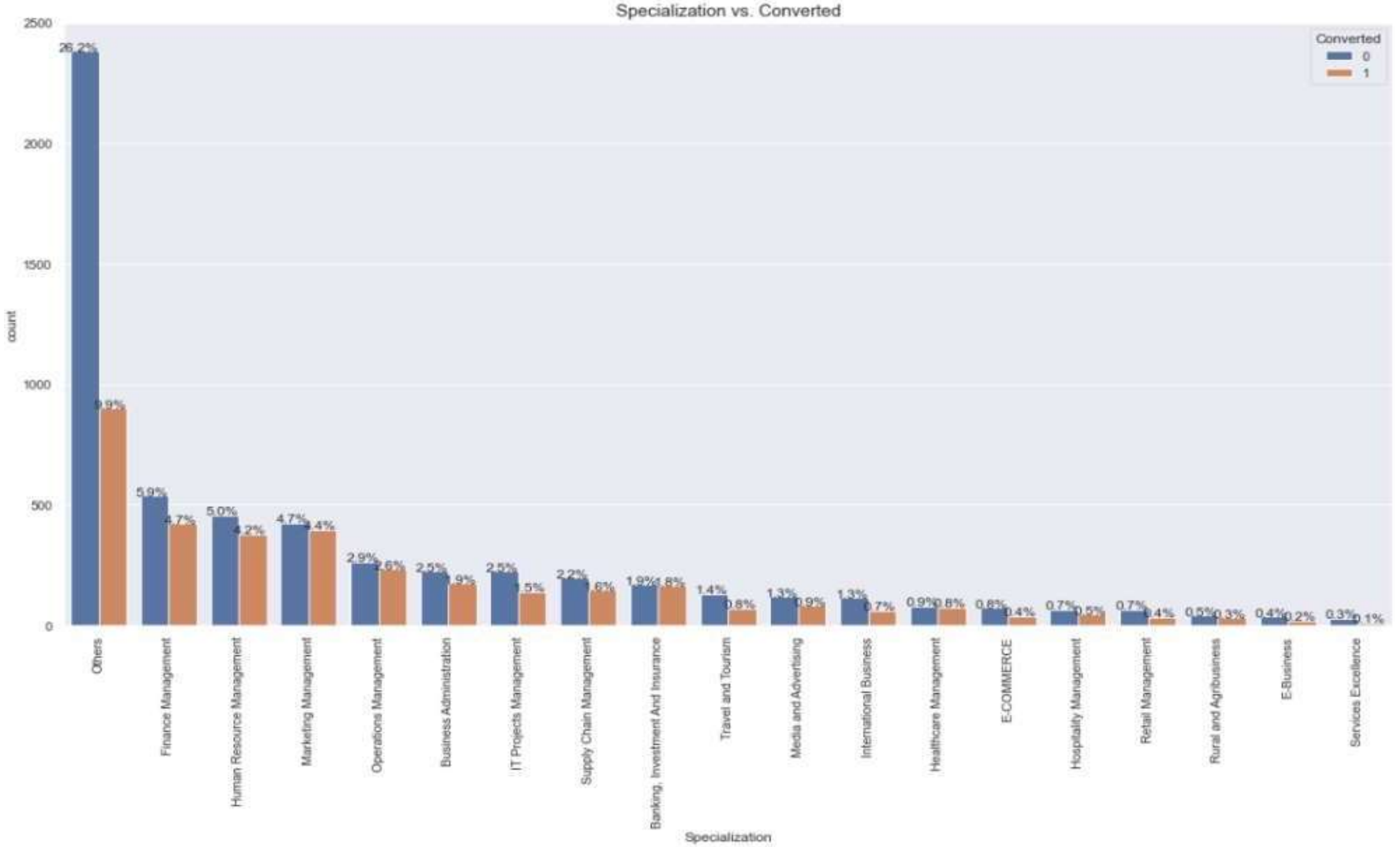


- The median value for both converted and not-converted are same.

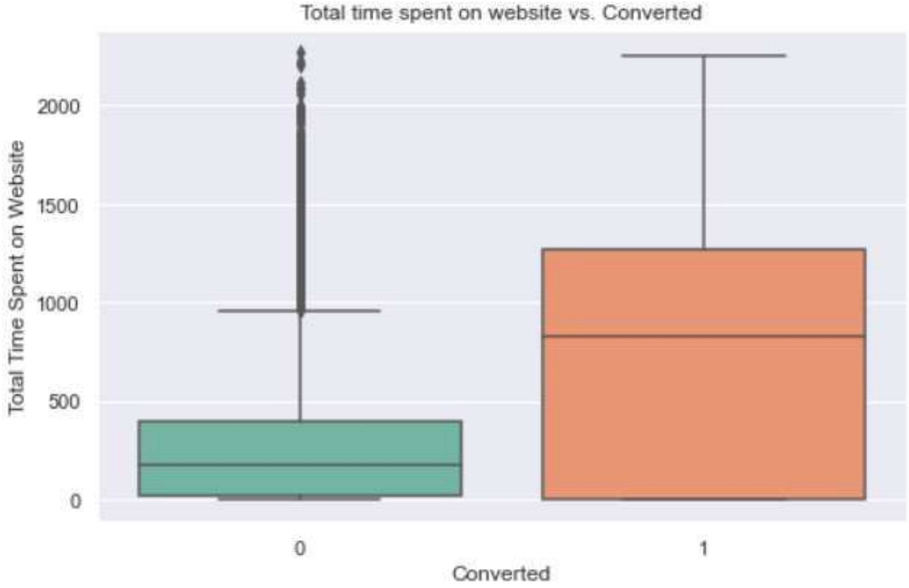
- Higher the visit on website greater is the chance of that lead to get converted



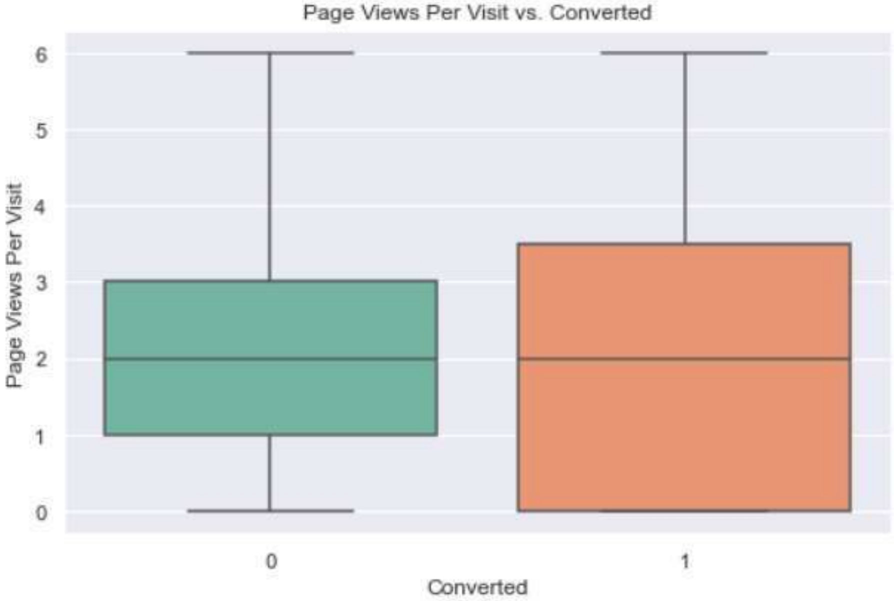
- Maximum leads comes from Google but the no-conversion rate is very high.
- Leads coming through someone‘s reference has the high chances of getting converted



- Leads who have not mentioned their specialization while filling up the form has the high chances of not enrolling for any course.
- Leads coming from marketing, operations, BFSI, Healthcare has the high chance of getting converted.

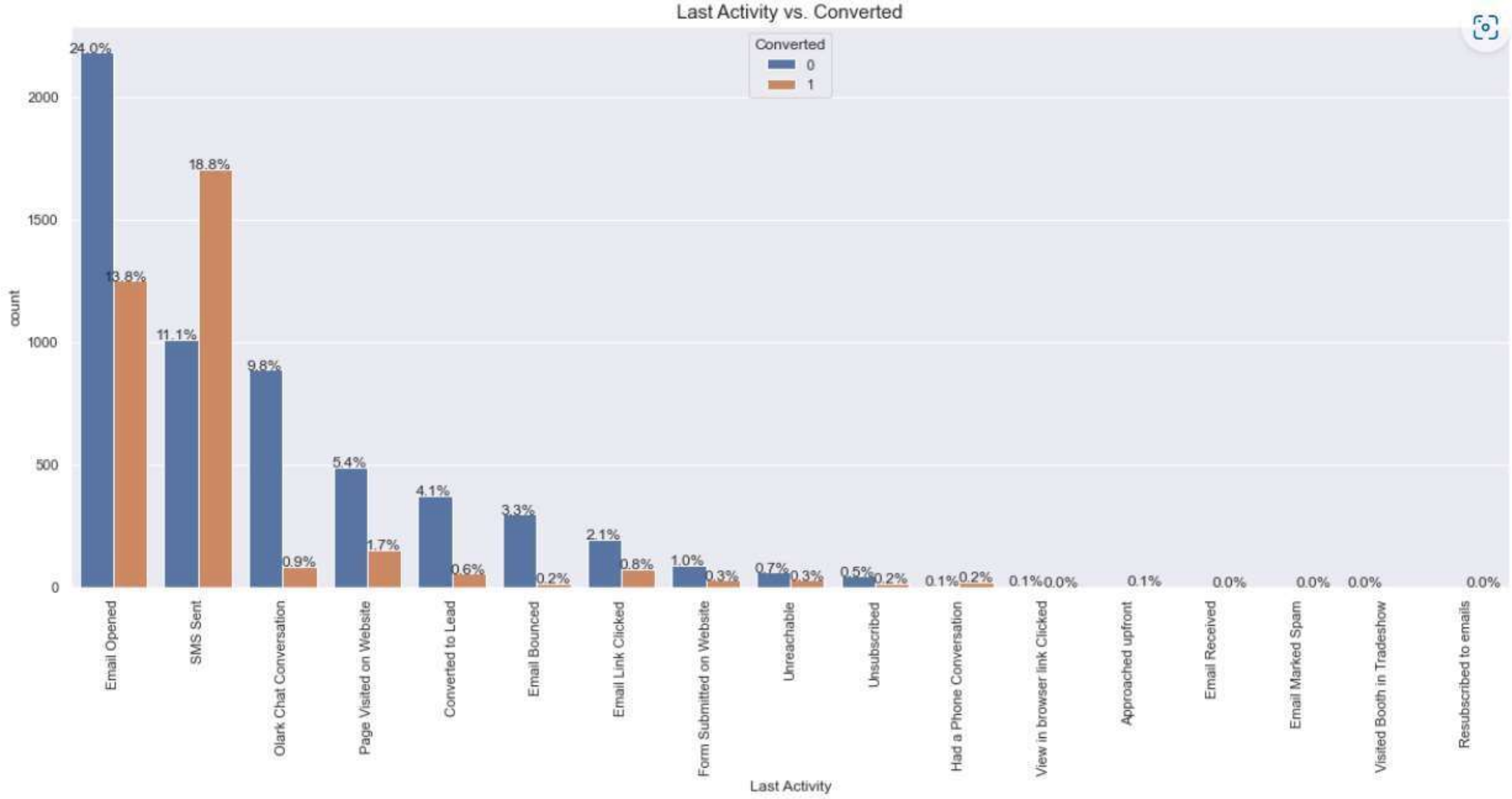


Higher the time spent on website greater is the chance of a lead to get converted



Median value of converted and not-converted are same

Nothing conclusive can be drawn from this



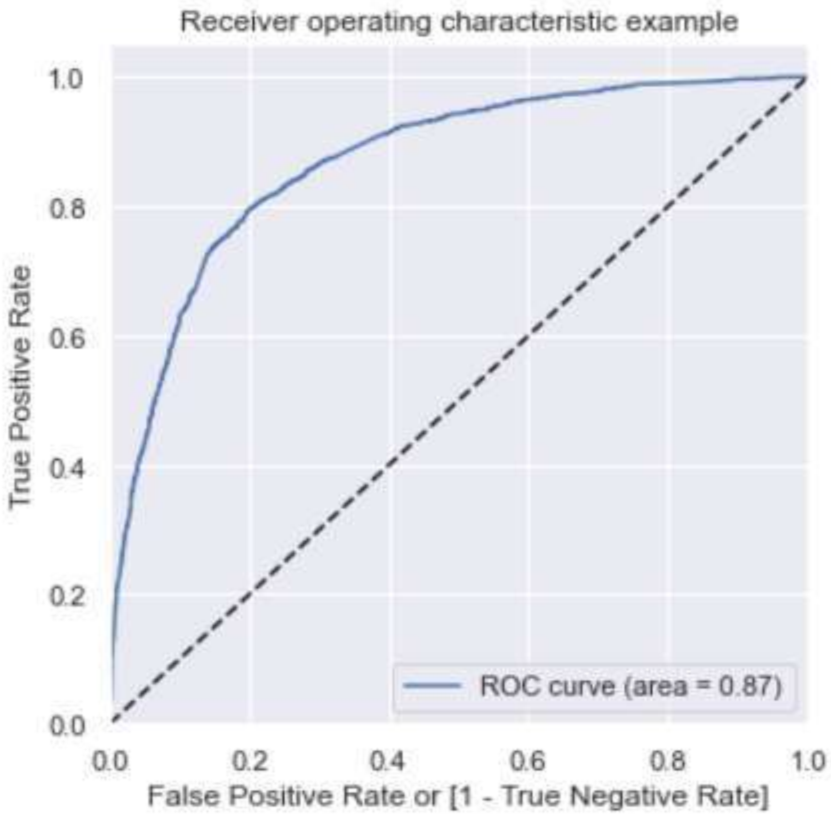
- Conversion rate is higher for leads having 'SMS sent' as their last activity.
- Most of the leads have 'Email Opened' as their last activity

Building Model

Steps involved are as follows:

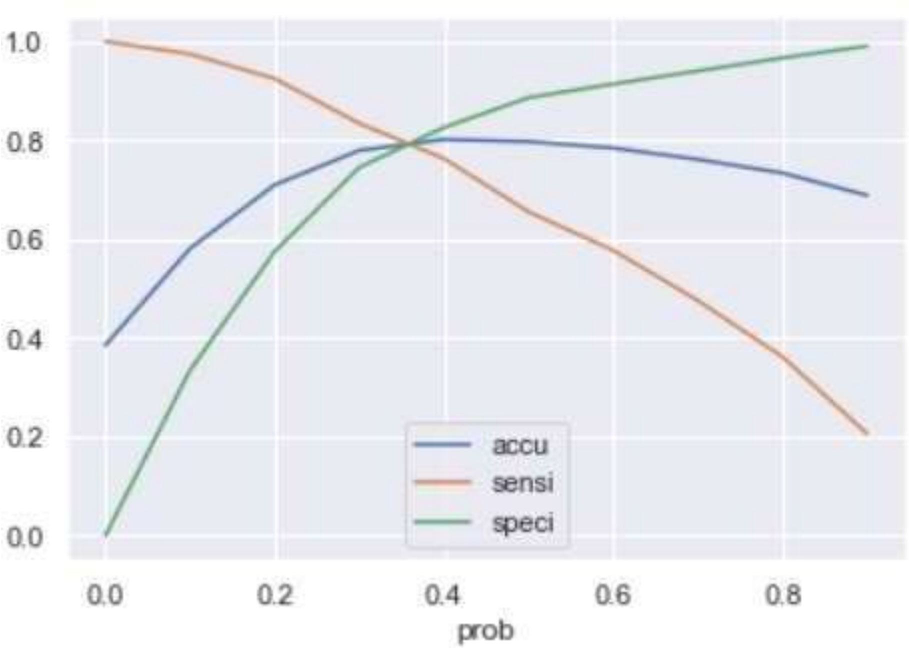
1. Split data into Train-Test set
2. Scale the continuous variables
3. Build the model
4. RFE to eliminate redundant variables
5. Build the next model
6. Drop variables with high p-values
7. Continue step 5 and 6 until p values are lower
8. Check for VIF score
9. Remove variables with high VIF score
10. Predict model on train set
11. Evaluate different parameters of Train set
12. Predict on Test set
13. Evaluate different parameters of Test set

ROC Curve



Area under the curve is higher (0.87) which is good. Therefore, we can conclude that our Model is a good model.

Optimal Probability Cut-Off



From the above graph we can see that 0.36 is the optimum probability cut-off point

Accuracy, Sensitivity and Specificity (Train) & Confusion Matrix

- accuracy: 79.78%
- sensitivity: 79.76%
- specificity: 79.79%

Confusion Matrix	Actual positive	Actual negative
Predicted posittive	True Positives **3116**	False positives **789**
Predicted negative	False negatives **489**	True negatives **1951**

Accuracy, Sensitivity and Specificity (Test) & Confusion Matrix

- accuracy: 79.51%
- sensitivity: 78.76%
- specificity: 79.93%

Confusion Matrix	Actual positive	Actual negative
Predicted posittive	True Positives **1386**	False positives **348**
Predicted negative	False negatives **210**	True negatives **779**

CONCLUSION

- SMS messages have high impact on the leads.
- Leads who are apending more time on the website has greater chances of getting converted
- Marketing team should look out for 'landing page submission' as the source of more number of leads
- Leads coming from Marketing, Operation, BFSI and Healthcare has the maximum chances of getting converted
- Leads who are coming to website from someone's reference arre most likely to get converted
- The final logistic regression model shows an accuracy of 79.51% (~80% targeted accuracy)
- The model shows sensitivity of 78.76% and specificity of 79.93%
- The area under curve is 0.87 which is good.
- The model finds out the hot-leads correctly which proves that the model is performing well.