LEAD SCORE CASE STUDY

SUBMITTED BY

Ankita Sethi
Chaitali Toke

Problem Statement

- X Education sells online courses to industry professionals. 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.

Goal of case study

Goals of the Case Study There are quite a few goals for this 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 your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.**

Steps Involved

a) Read and understand data

b) Clean the data

c) Prepare the data for modelling.

d) Model Building

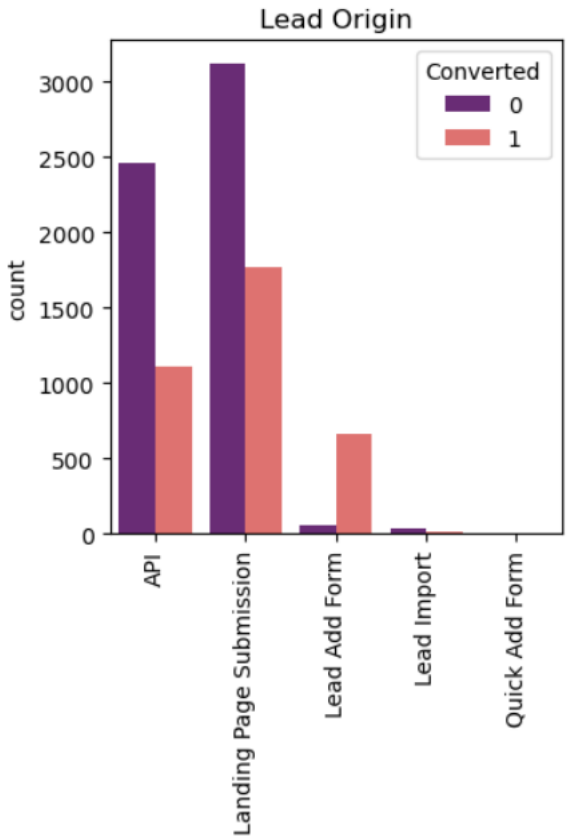
e) Model Evaluation

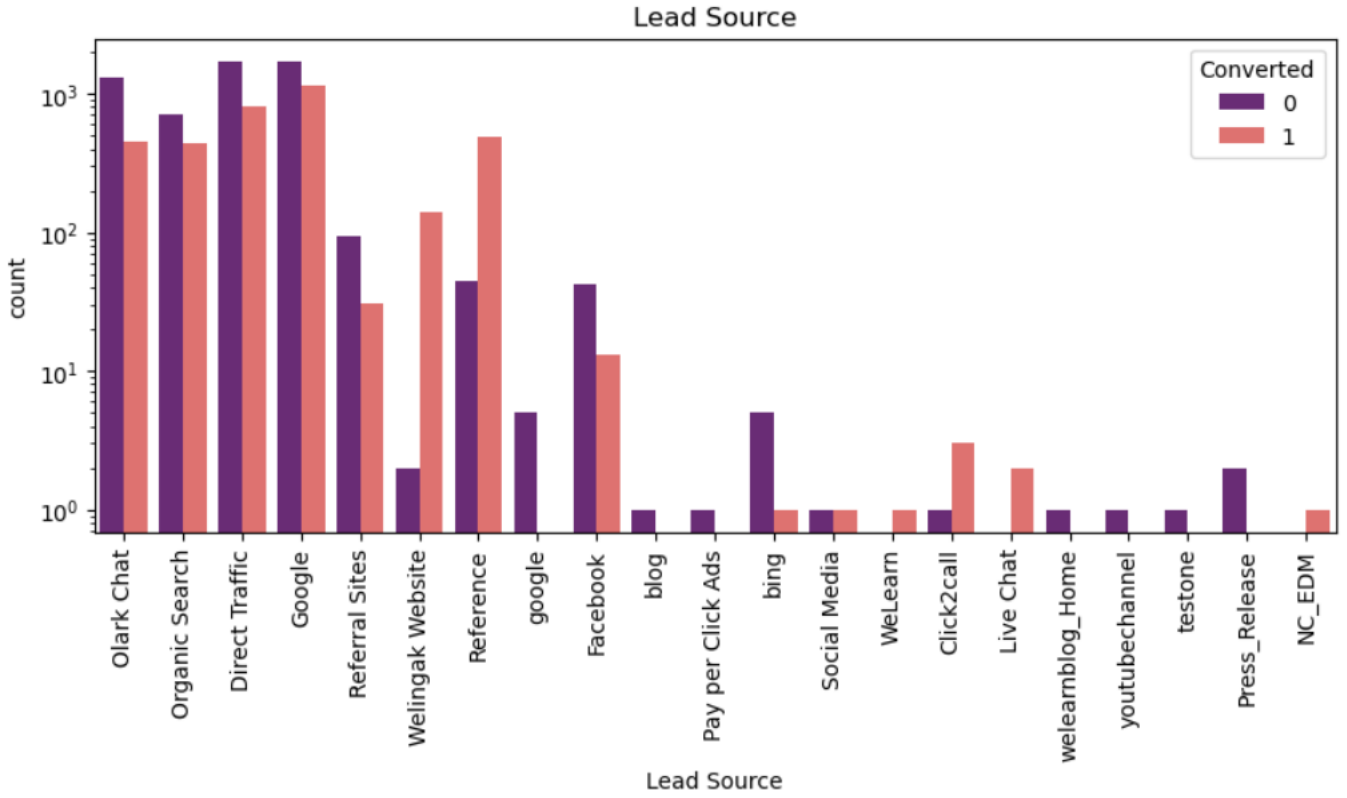
f) Making Predictions on test set

g) Conclusion and summary

- Initial number of columns and rows are respectively 9240 ,37.
- Columns having select values were replaced by nan.
- Data cleaning was done by dropping columns having more than 40 % of missing values.
- EDA was done on both categorical variables and numerical variables .
- Outliers were removed.
- As per the value counts of categorical variables we can observe few columns in which only one value is given .These columns are Search , Magazine , Do not call , Newspaper Article ,X Education Forums ,Newspaper , Digital Advertisement ,Through Recommendations, Receive More Updates About Our Courses ,Update me on Supply Chain Content ,Get updates on DM Content, I agree to pay the amount through cheque.so they are dropped.
- Columns such as country, city , tags , what matters most in choosing the course were dropped.

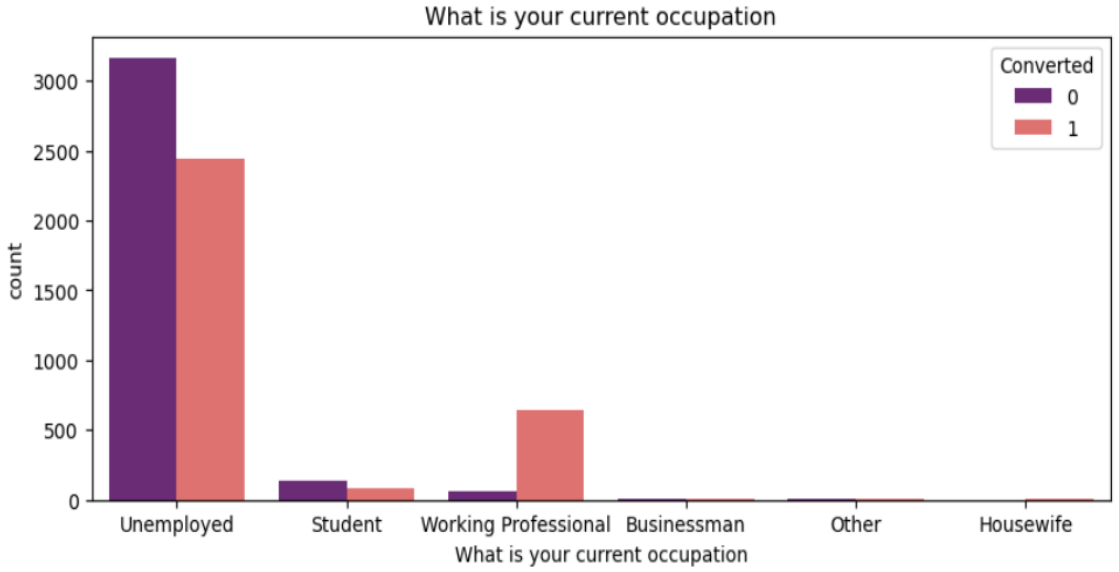
EDA

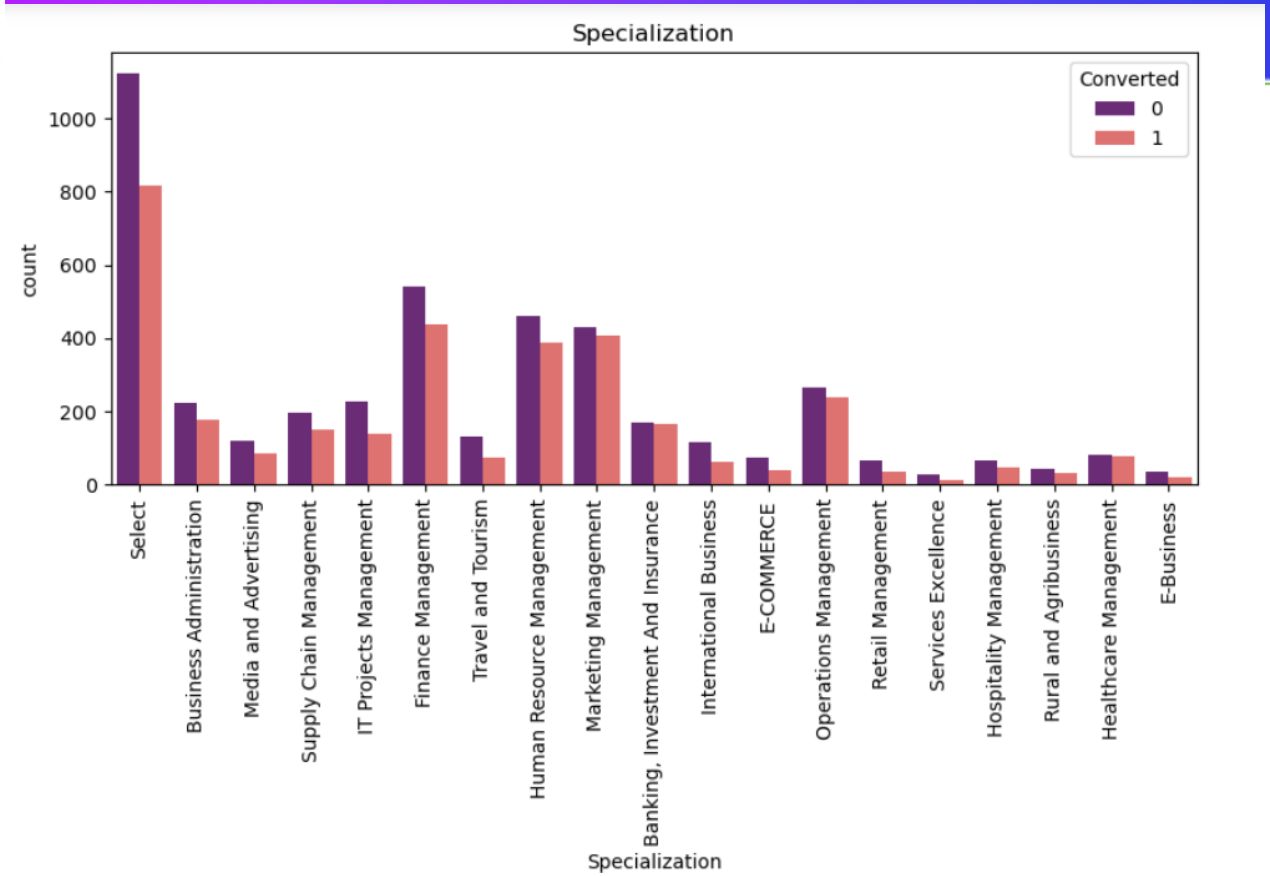


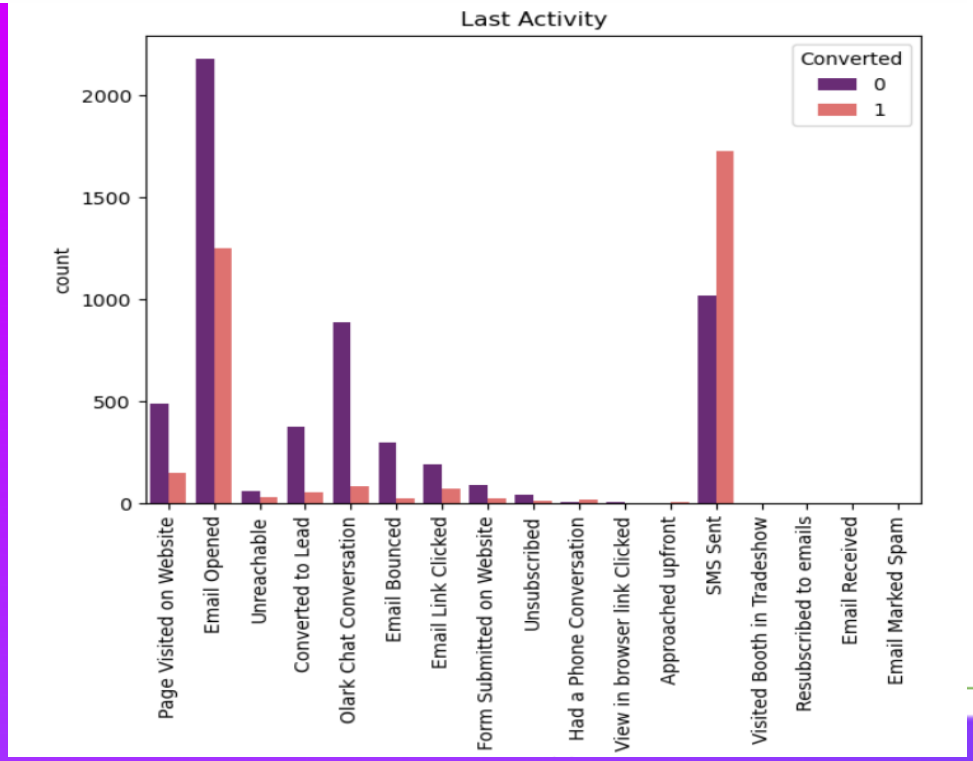


1.Lead Origin - More than 50 % of all leads orginated from "Landing Page Submission" and they only have more conversion rate followed by API.

2.Lead Source - More conversion rate can be seen in a) Olark b) Organic c) Direct d) Google e) Reference f) Welingak website

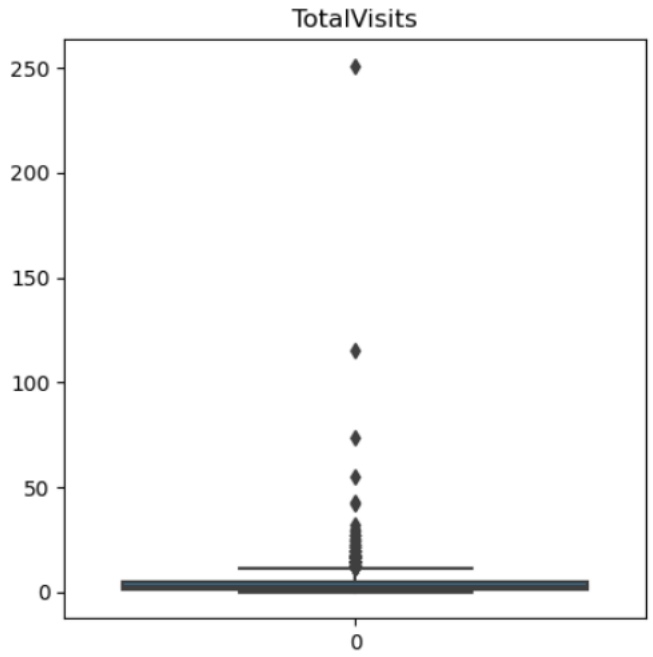


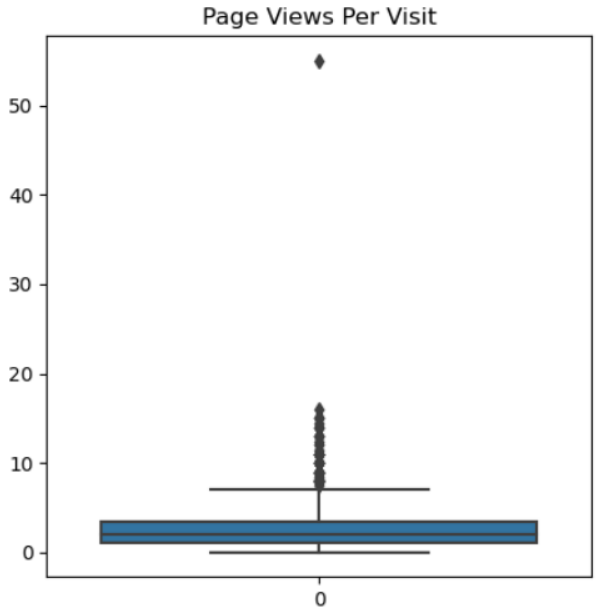




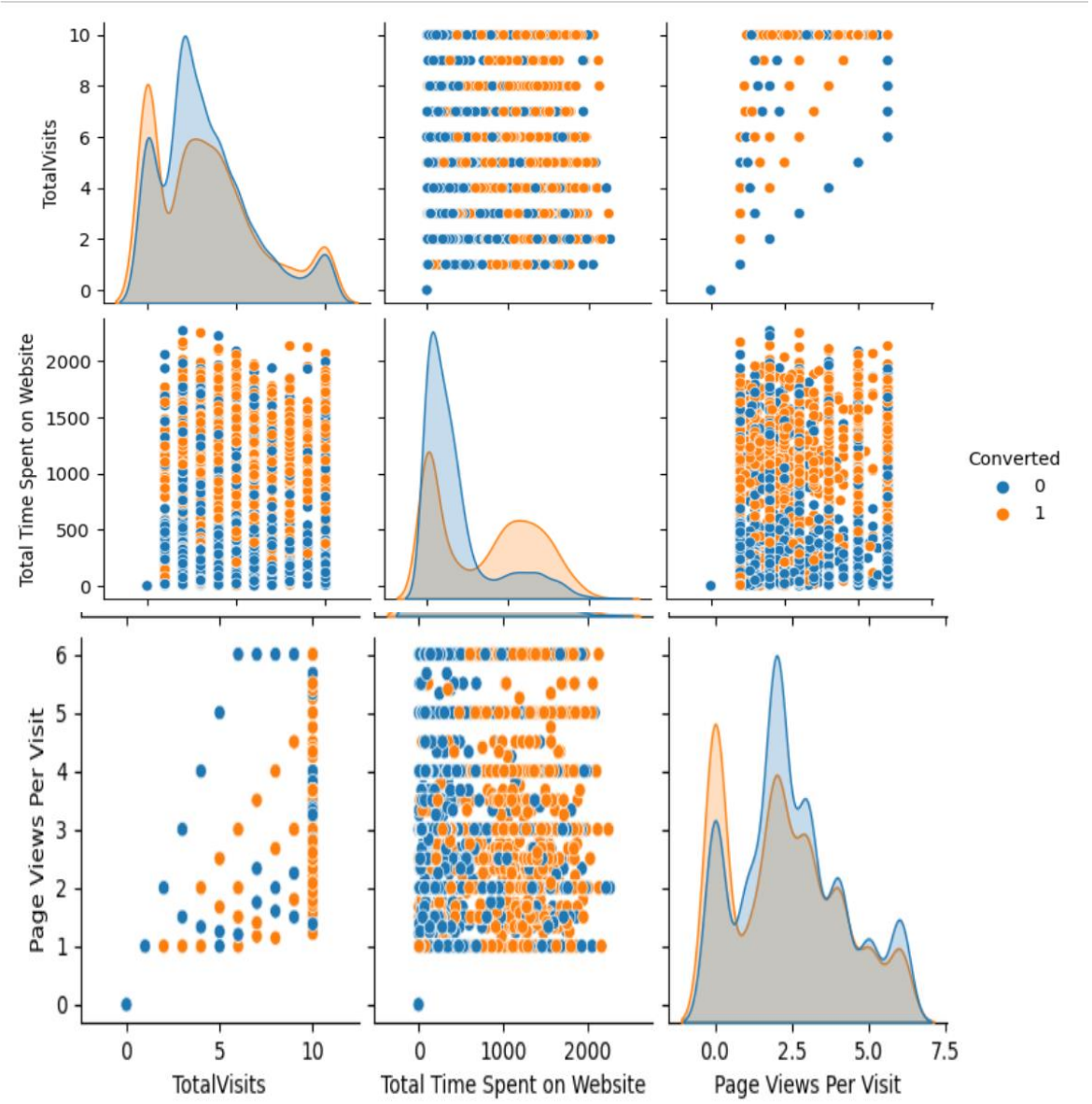
1.Specialization – Focus on a) Finance b) HR c) Management Marketing d) Operation management
2.What is your current occupation a) Working professionals have huge conversion rate.
3.Last Activity - SMS sent has high conversion rate.

OUTLIERS

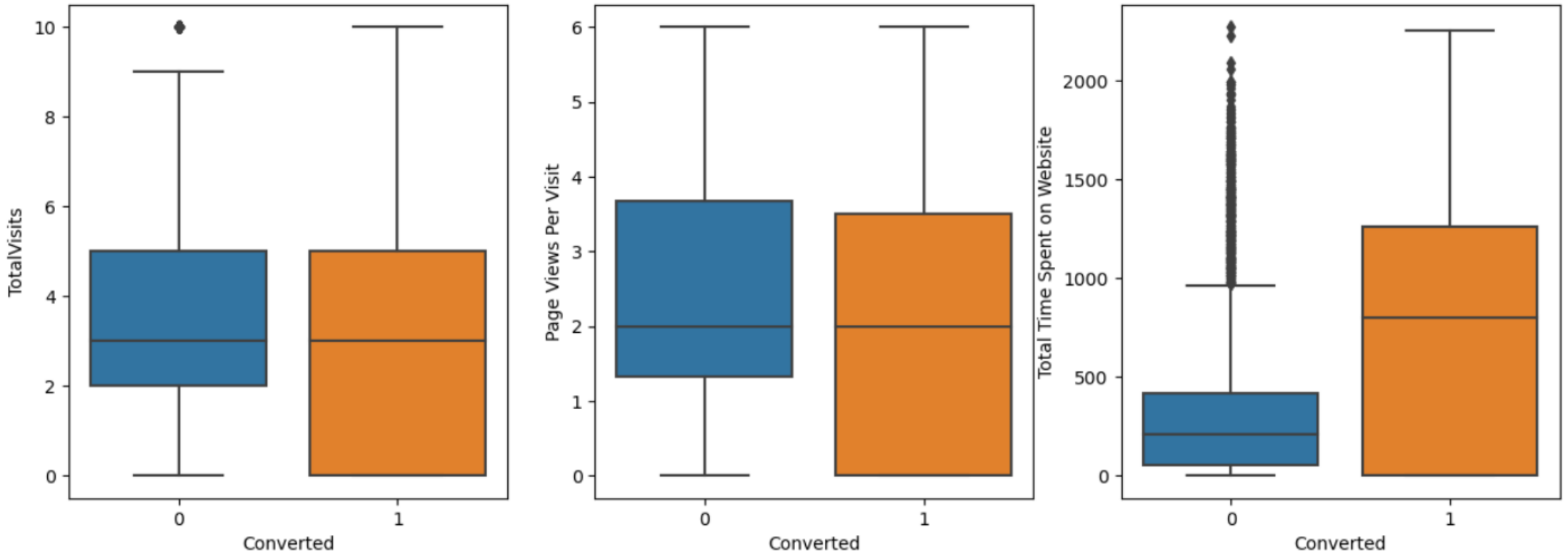




We can see outliers in Total Visits and page per view visit hence they were capped.



- Graph (pair plots were plotted for numerical variables after removing outliers.
- The pair plot on the left shows the spread of distribution across the converted and non - converted classes
- Dummy variables were created.

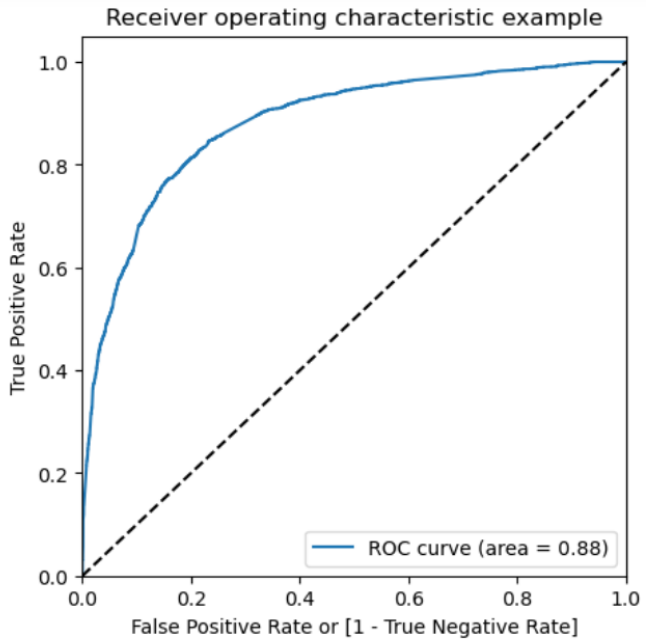


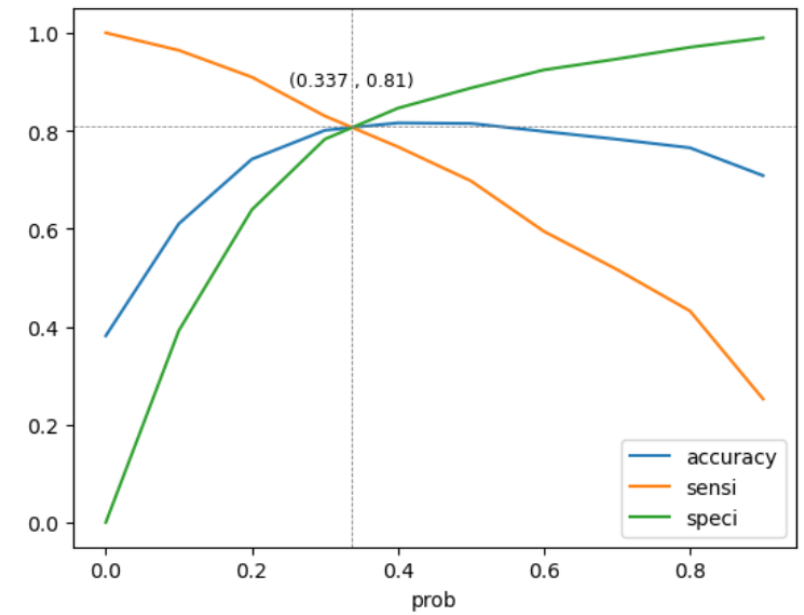
OBSERVATION - Past leads who spend more time on website are successfully converted than those who spend less as seen in the boxplot.

Building model

	Features	VIF
1	TotalVisits	2.02
2	Total Time Spent on Website	1.97
9	Last Notable Activity_Modified	1.50
6	Last Activity_SMS Sent	1.44
4	Lead Source_Olark Chat	1.25
3	Lead Origin_Lead Add Form	1.17
7	What is your current occupation_Working Profes...	1.16
5	Last Activity_Others	1.13
0	Do Not Email	1.11
10	Last Notable Activity_Olark Chat Conversation	1.07
8	Last Notable Activity_Had a Phone Conversation	1.06

- Train test split was done by taking 70:30 ratio.
- Scaling was done using Minmax scaler.
- Model was built using RFE approach by selecting top 15 features and based on p-values and ViF some variables were dropped.
- After dropping variables whose p values are more than 0.05 and Vif more than 5 we get top 11 features.





- Optimal cutoff = 0.337

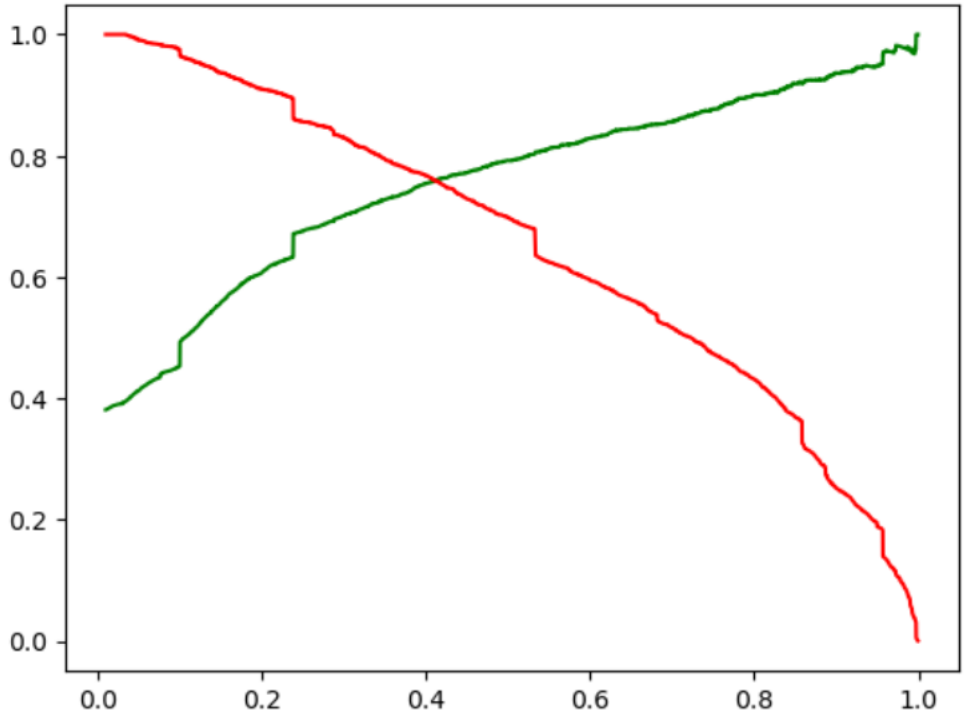
Training Set

```
Sensitivity : 80.37 %
Specificity : 80.93 %
Accuracy =    80.72 %
```

Test Set

```
Sensitivity =  80.46 %
Specificity =  81.40 %
Accuracy    =  81.02 %
```

Precision and Recall



Optimal cutoff =0.41

Training set

```
Precision = 72.2040072859745
Recall =  80.37307380373075
```

Test set

```
Precision = 73.84744341994971
Recall =  80.45662100456622
```

Top features along with their coefficients

Total Time Spent on Website	4.545584
Lead Origin_Lead Add Form	4.268686
What is your current occupation_Working Professional	2.808038
Last Notable Activity_Had a Phone Conversation	2.663266
Lead Source_Olark Chat	1.302173
Last Activity_SMS Sent	1.296274
Last Activity_Others	0.678950
TotalVisits	0.669318
Last Notable Activity_Modified	-1.039097
Do Not Email	-1.282595
Last Notable Activity_Olark Chat Conversation	-1.342796
const	-2.469008

Conclusion

Top three variables –

Top three features are with their coefficents

a) Total Time Spent on Website - 4.545584
b) Lead Origin_Lead Add Form - 4.268686
c) What is your current occupation_Working Professional - 2.808038

Recommendations

- We need to focus on employed or working professionals as they have higher chances of enrolling themselves.
- We need to give briefing to leads as in what are the oppurtunities they will be getting after doing this particular course and it will lead to their professional growth.
- Use broadcast messages, emails to reach out to the maximum audience.
- Do not focus on students as they are already involved in some courses.
- Focus on features with positive coefficients for targeted marketing stratergies.
- More budget can be spend on Olark chat as more leads are coming from their in terms of advertising.
- Incentives or discounts for providing refrence that convert to lead.
- The company should make calls who spend more time on website.
- They should not make more calls to the leads who chose the option of Do not email as yes.
- More the number of visits to the website more are the chances of the lead conversion.

THANK YOU !