Lead Scoring Case Study

Submitted By- Durgesh Kumar

PROBLEM STATEMENT

- X Education in the Ed-Tech Industry sells online courses to industry professionals.
- It’s lead conversion is very poor.
- The company wants to assign Lead Score to each lead using Machine Learning Model.

Assumptions

- Unique value variables like ‘Prospect ID’ and Single value variables are dropped as they do not provide any significant information.
- Variables with high missing values and data imbalances are also not considered

APPROACH FOLLOWED

1. Data Cleaning and Data Manipulation
2. Exploratory Data Analysis
3. Data Pre-processing
4. Model Building - Logistic Regression Model
5. Model Evaluation
6. Predictions
7. Conclusions and Recommendations

Notable EDA Conclusions

```
In [6]: df.head()
```

Out[6]:

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content	Lead Profile	City	Asymmetrique Activity Index	Asymmetriqu Profile Inde
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	Select	Select	02.Medium	02.Mediu
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No	Select	Select	02.Medium	02.Mediu
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	...	No	Potential Lead	Mumbai	02.Medium	01.Hig
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	...	No	Select	Mumbai	02.Medium	01.Hig
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	...	No	Select	Mumbai	02.Medium	01.Hig

5 rows × 37 columns

Notable EDA Conclusions

```
#check for columns with one unique value, count and freq is same
df.describe(include = 'object')
```

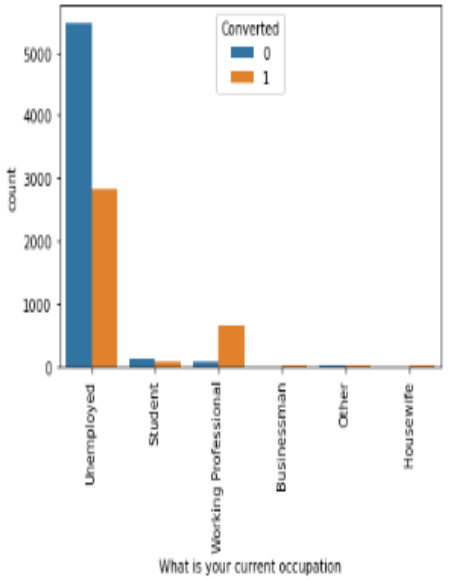
	Lead Origin	Lead Source	Do Not Email	Do Not Call	Last Activity	Country	Specialization	How did you hear about X Education	What is your current occupation	What matters most to you in choosing a course	...	Digital Advertisement	Through Recommendations
count	9074	9074	9074	9074	9074	9074	9074	9074	9074	9074	...	9074	9074
unique	4	21	2	2	17	38	11	9	6	3	...	2	2
top	Landing Page Submission	Google	No	No	Email Opened	India	Management_Specializations	Online Search	Unemployed	Better Career Prospects	...	No	No
freq	4885	2868	8358	9072	3432	8787	4197	7894	8159	9072	...	9070	9067

4 rows × 25 columns

NOTE: Following columns have only one unique value:

- 'I agree to pay the amount through cheque',
- 'Get updates on DM Content',
- 'Update me on Supply Chain Content',
- 'Receive More Updates About Our Courses',
- 'Magazine'

These columns are of no use as they have only one category of response from customer and can be dropped:



- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in terms of Absolute numbers.

Dropping columns of no use for modeling

NOTE: Columns such as:

- 'Last Notable Activity'
- 'City'
- 'Country'
- 'How did you hear about X Education'
- 'Lead Profile'
- 'Converted'
- 'What matters most to you in choosing a course'

Above columns do not add any value to the model. Dropping these columns will remove unnecessary data from the dataframe.

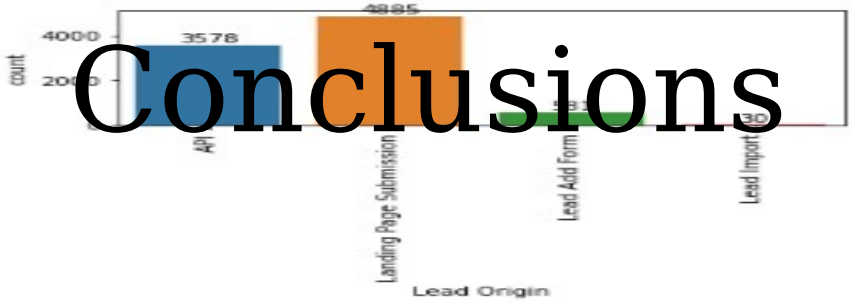
Dropping columns of no use for modeling

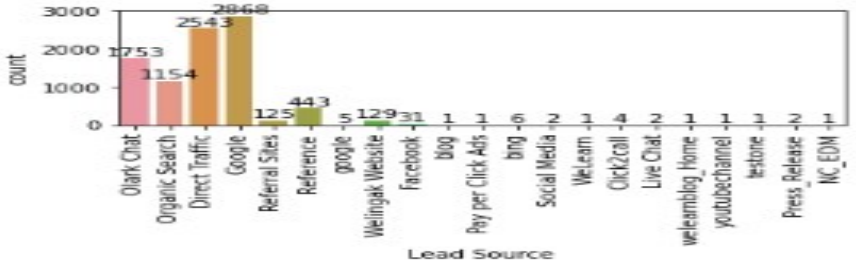
NOTE: Columns such as:

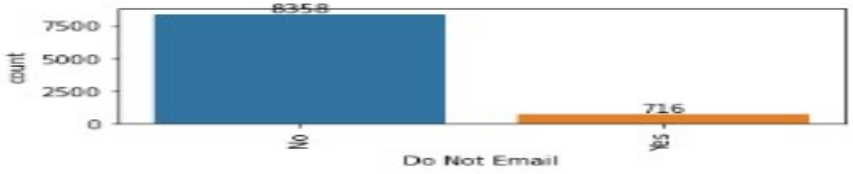
- 'Last Notable Activity'
- 'City'
- 'Country'
- 'How did you hear about X Education'
- 'Lead Profile'
- 'Converted'
- 'What matters most to you in choosing a course'

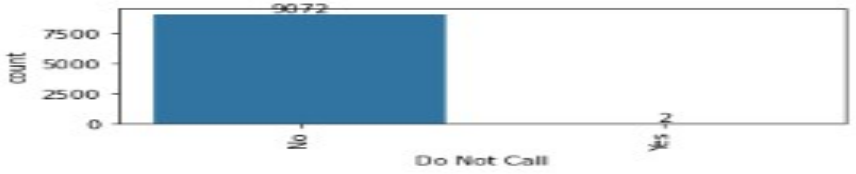
Above columns do not add any value to the model. Dropping these columns will remove unnecessary data from the dataframe.

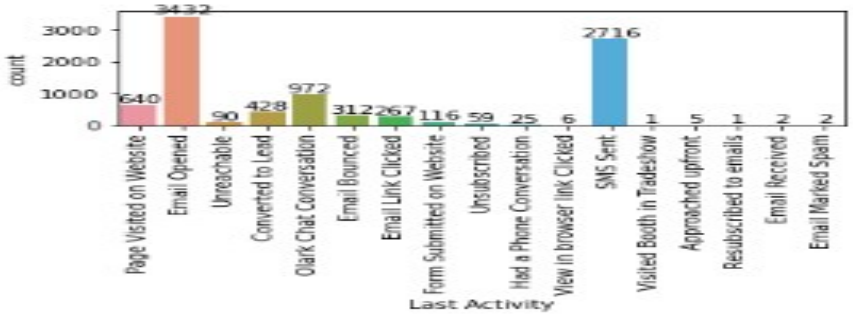
Notable EDA Conclusions

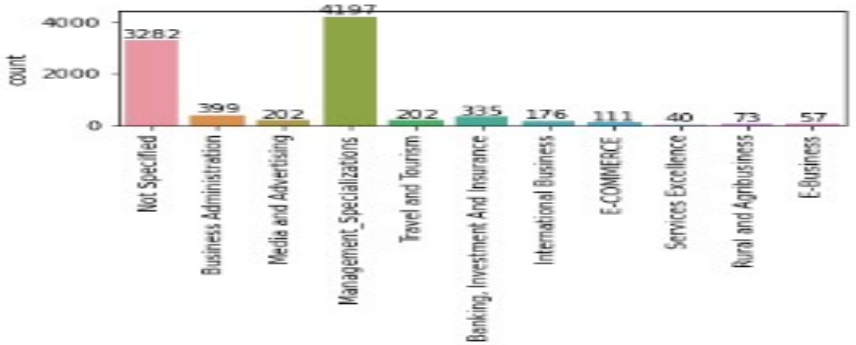


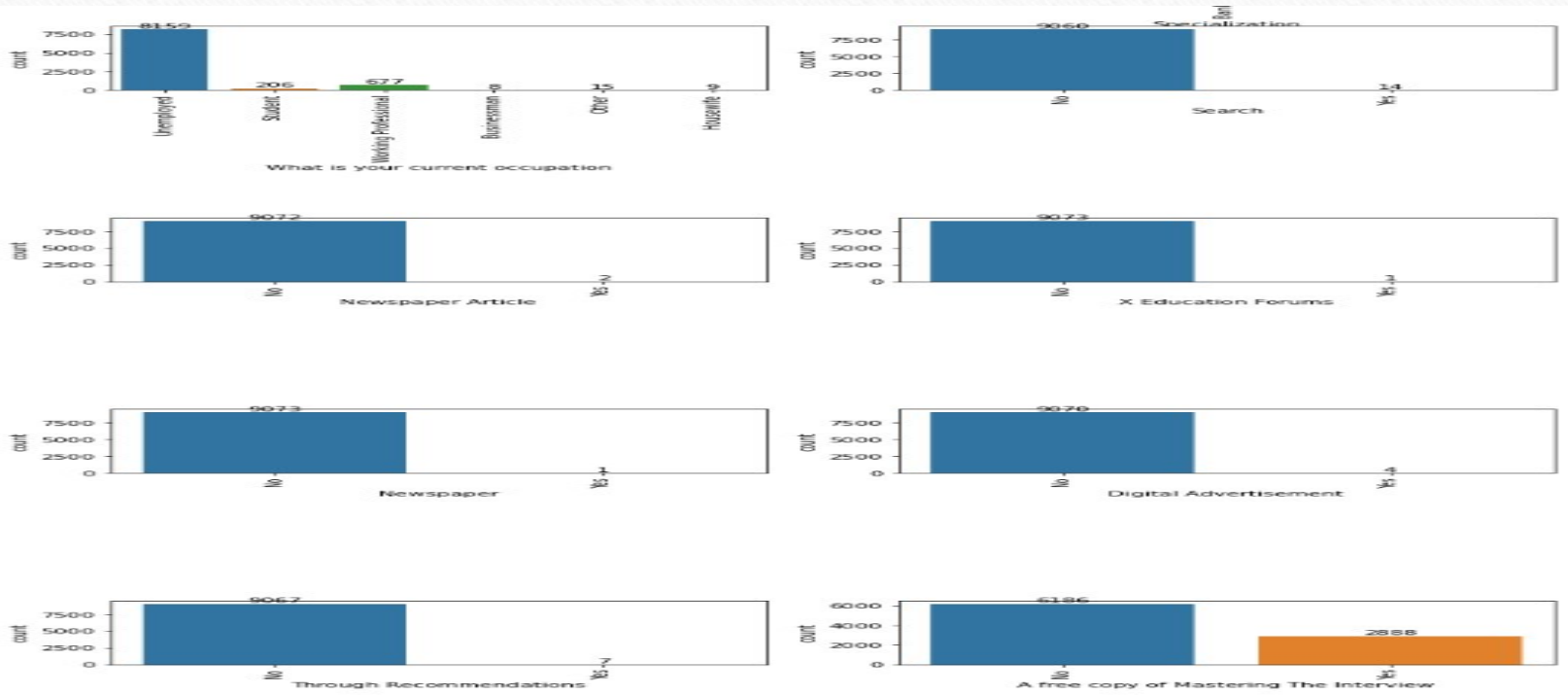








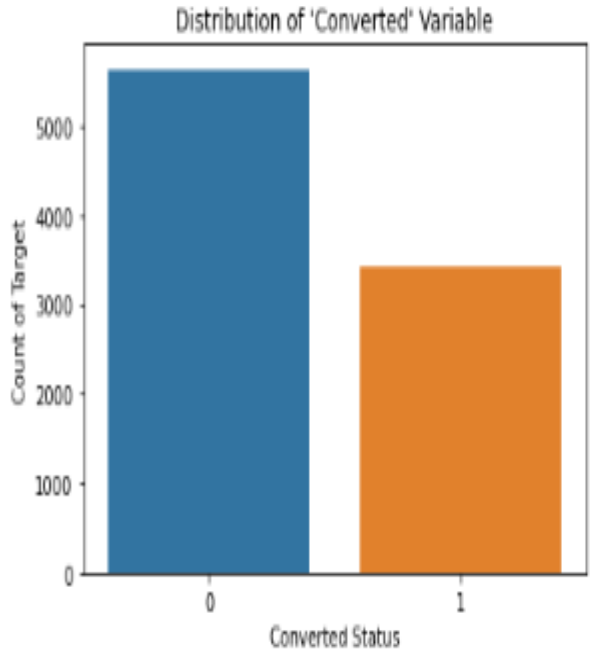




NOTE: Following columns have data which is highly skewed :

- 'Do Not Call',
- 'Search',
- 'Newspaper Article',
- 'X Education Forums',
- 'Newspaper',
- 'Digital Advertisement',
- 'Through Recommendations'.

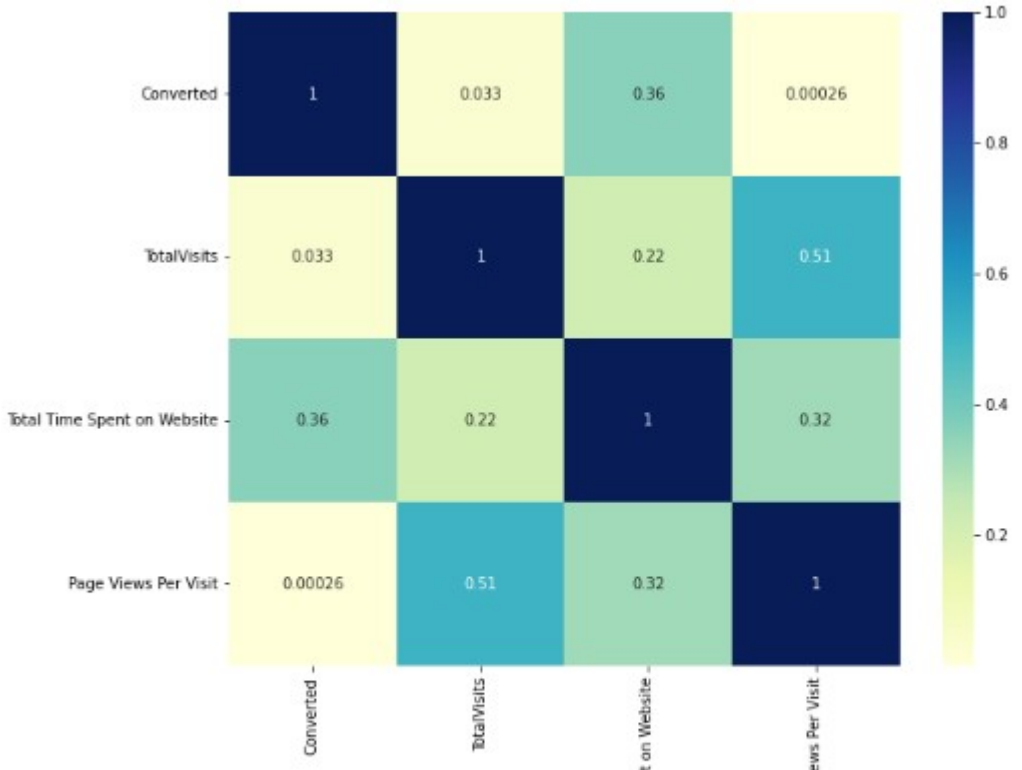
Hence these columns will be dropped as they will not add any value to the model. Morever, Skewed variables can affect the performance of logistic regression models, as they can lead to biased or inaccurate parameter estimates.




```
# Finding out conversion rate
Converted = (sum(df['Converted'])/len(df['Converted'].index))*100
Converted
```

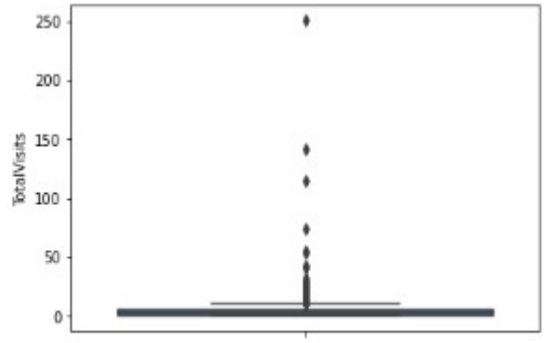
```
37.85541106458012
```

Currently, lead Conversion rate is 37.85% only

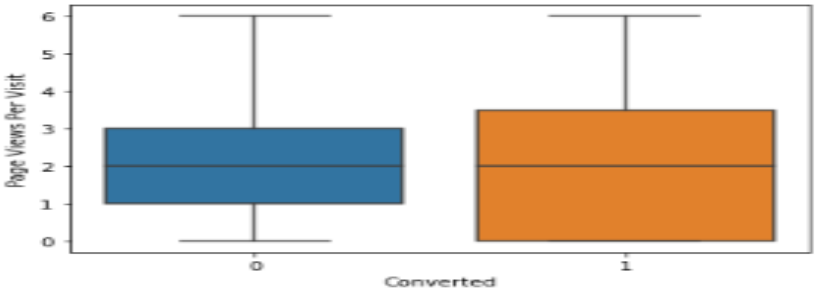


The top three variables that contribute most towards the probability of a lead getting converted are

- Total Time Spent on Website
- Lead Source_Welingak Website
- Lead Source_Reference

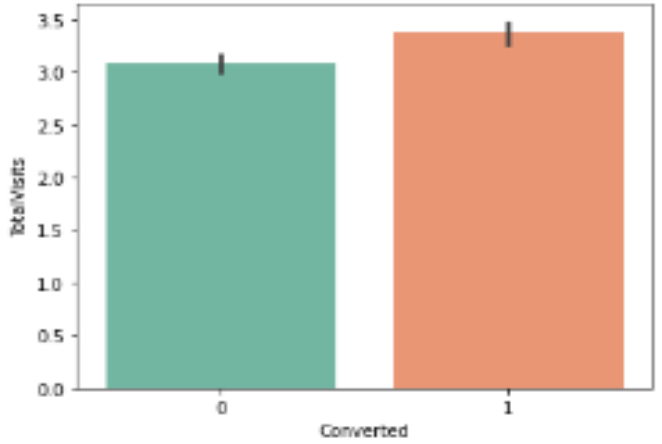


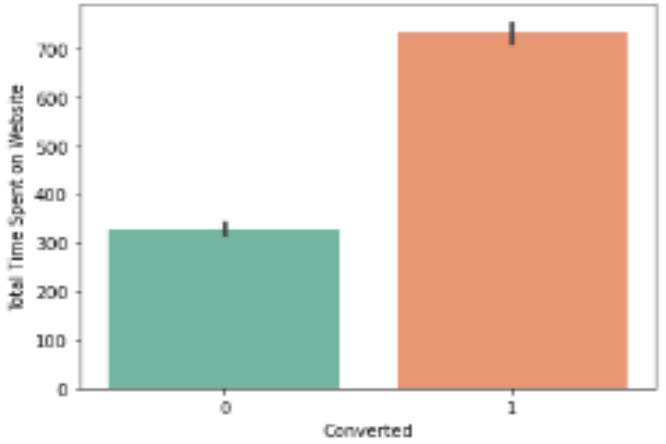
Presence of outliers can be seen clearly

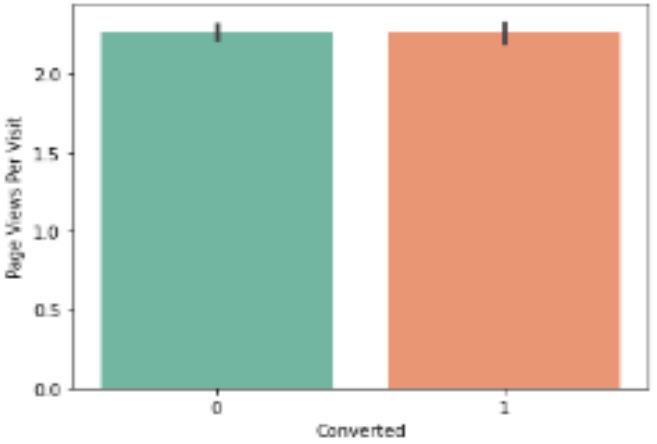


Inference

- Median for converted and not converted leads is almost same.
- Nothing conclusive can be said on the basis of Page Views Per Visit.



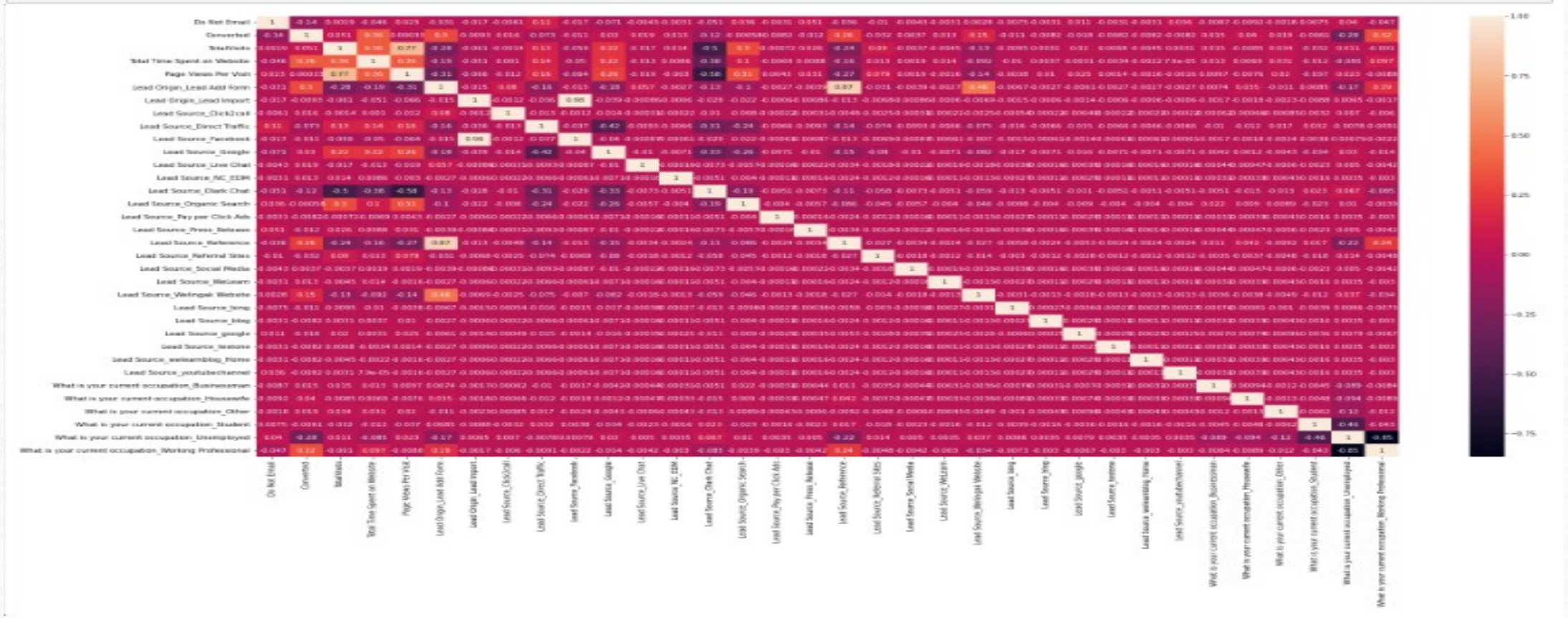




Inference

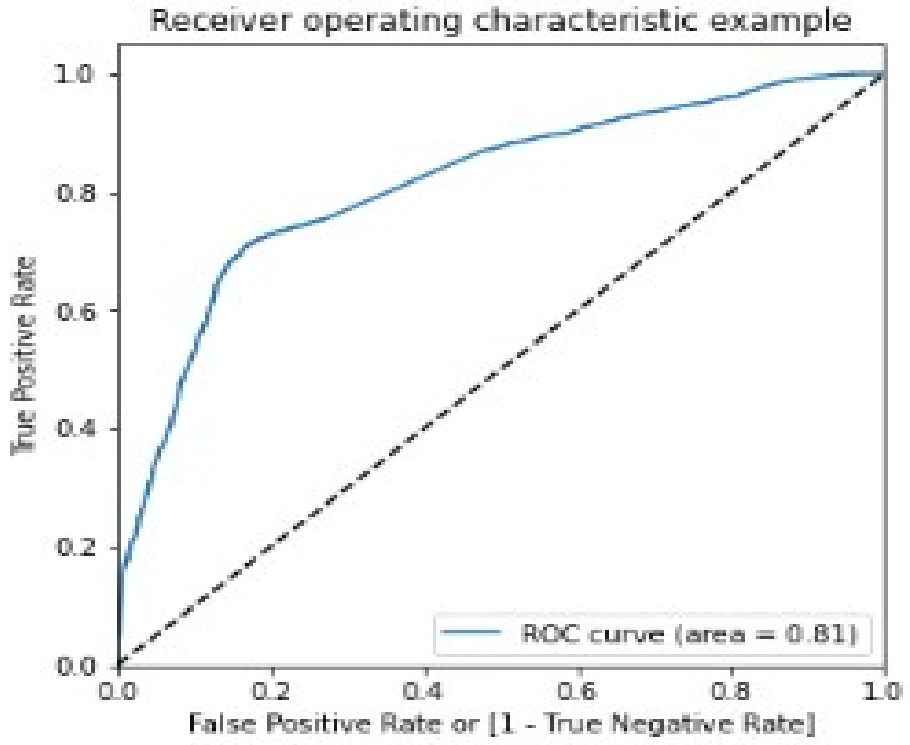
The conversion rate is high for Total Visits, Total Time Spent on Website and Page Views Per Visit

NOTABLE EDA CONCLUSIONS (CONT.)



Dropping highly correlated dummy variables

RESULTS

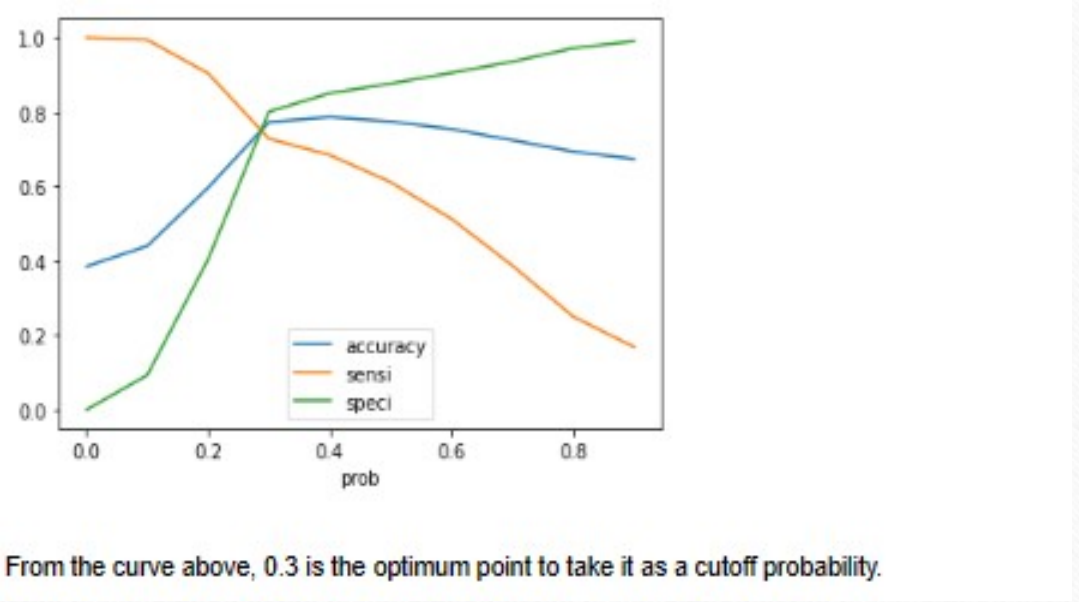


- ROC Curve
- 96% of the area is under ROC curve.
- Classification Probability of lead conversion by the model is very high

The ROC Curve should be a value close to 1. We are getting a good value of 0.81 indicating a good predictive model.

NOTABLE EDA CONCLUSIONS (CONT.)

- 0.3 is the optimum point to take it as a cut-off probability.
- Accuracy : 77.45%
- Sensitivity :71.58%
- Specificity : 80.79%



RECOMMENDATIONS

Follow ups through calls and emails with high conversion probability leads is suggested.

- Focus more on customers who spend a lot of time on the company's website as their conversion rate is high as per EDA.
- Providing special offers to customers who are highly interested and are seen visiting back to the website.
- Leads who have Tags such as 'Ringing', 'Switched Off', 'Invalid Number' can be avoided as the probability of them converting is very low.

Thank You !