Lead Scoring Case Study

Submitted by:

Vivek Venugopal & Thousif Ahmed

Goals of the Case Study

X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Steps followed(Method)

Data Cleaning

- Dealt with null values – through dropping as well as imputing.
- Dealt with skewed data by grouping categories (binning). E.g., India in 'Country' variable.
- Checked for outliers and capped when required.

EDA

- Categorical variables analysis with respect to target variable ('Converted')
- Univariate and segmented analysis
- Mapping and creation of dummy variables for categorical variables
- Splitting the data into training and testing sets for logistic model building
- Feature scaling using Standard scalar

Model building

- REF based feature selection – 20 variables
- Dropping variables with p-value > 0.05
- VIF analysis
- ROC for ideal cut-off
- Final model

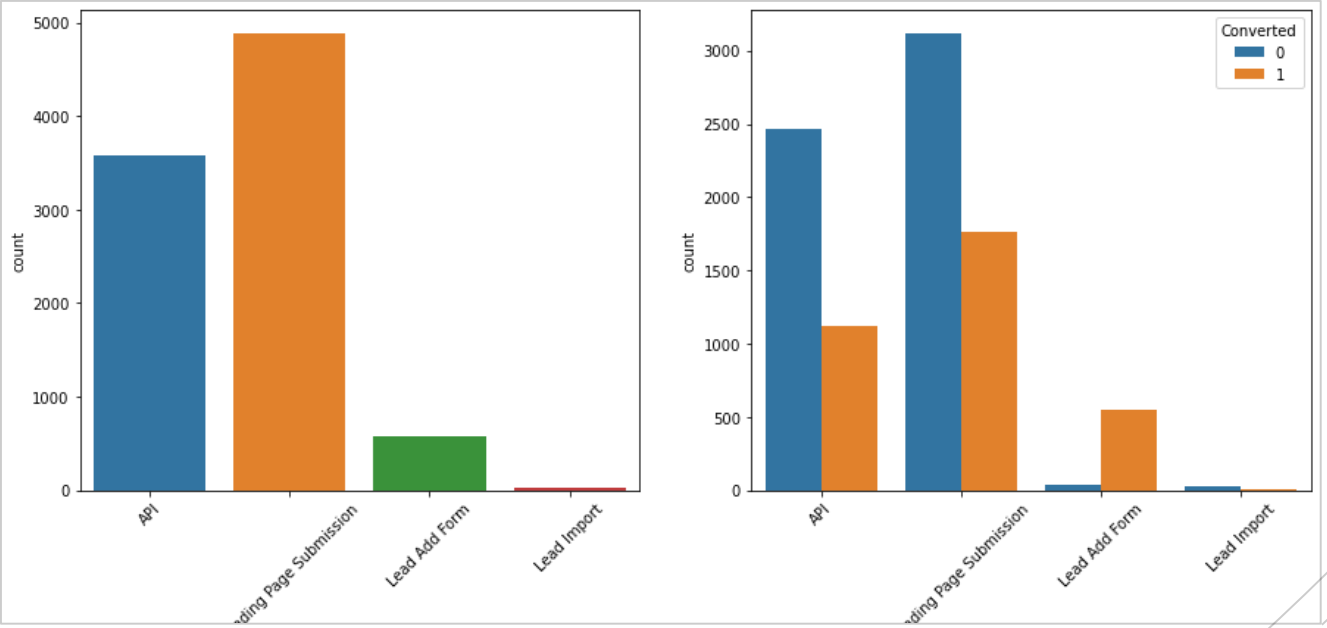
EDA insights

- Lead origin variable

EDA insights

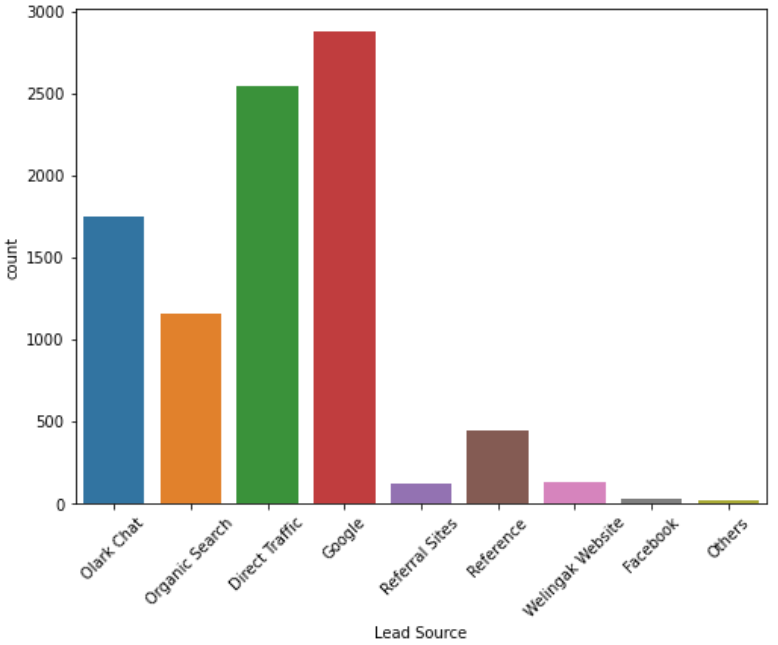
Inference:

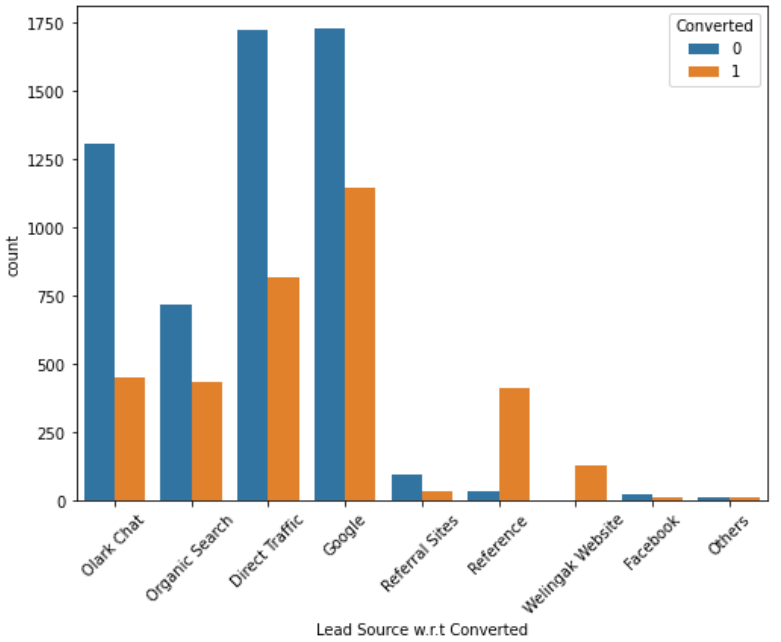
- Lead Add Form has highest conversion rate(90%) but number of leads are low.
- API and Landing Page Submission have lower conversion rate compared to Lead Add Form but number of leads originated from them are higher.
- Lead Import are very less in count.



- Lead Source

EDA insights





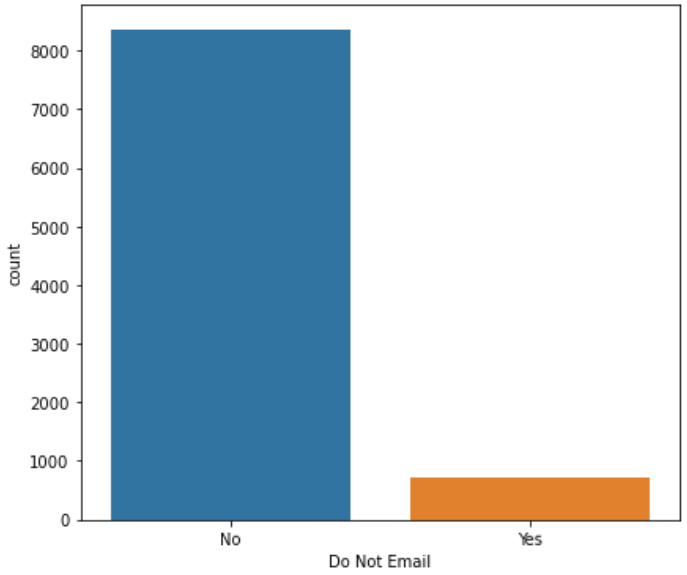
Inference :

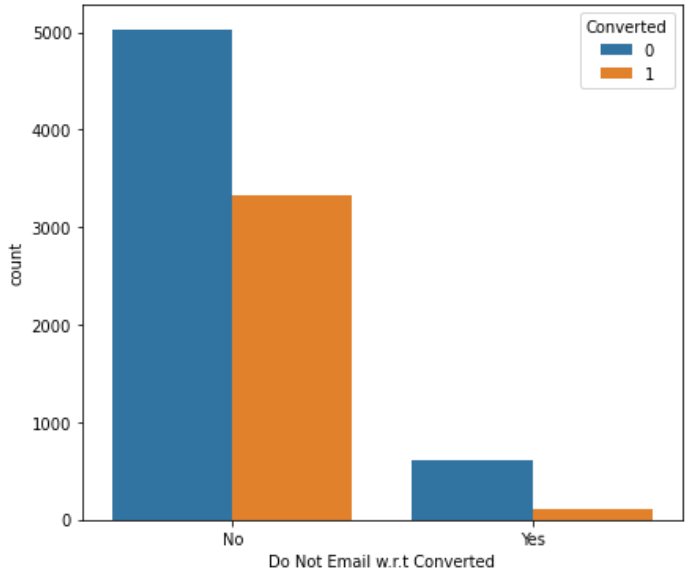
•The top four Lead Source's are Google, Direc Traffic, Olark Chat and Organic Search

•Reference and Welingak Website have higher conversion rate than other Lead Sources

EDA insights

- Do not Email



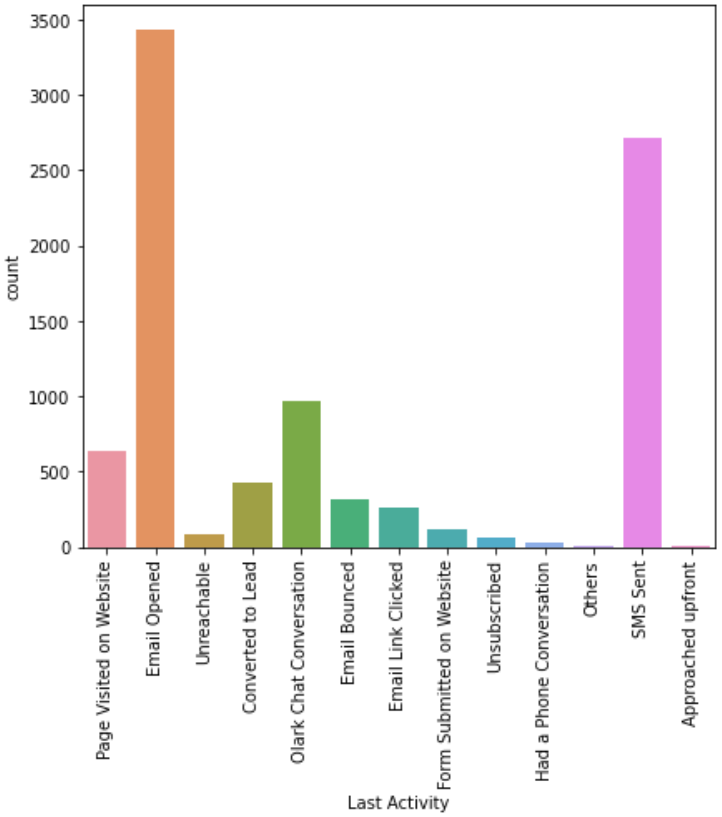


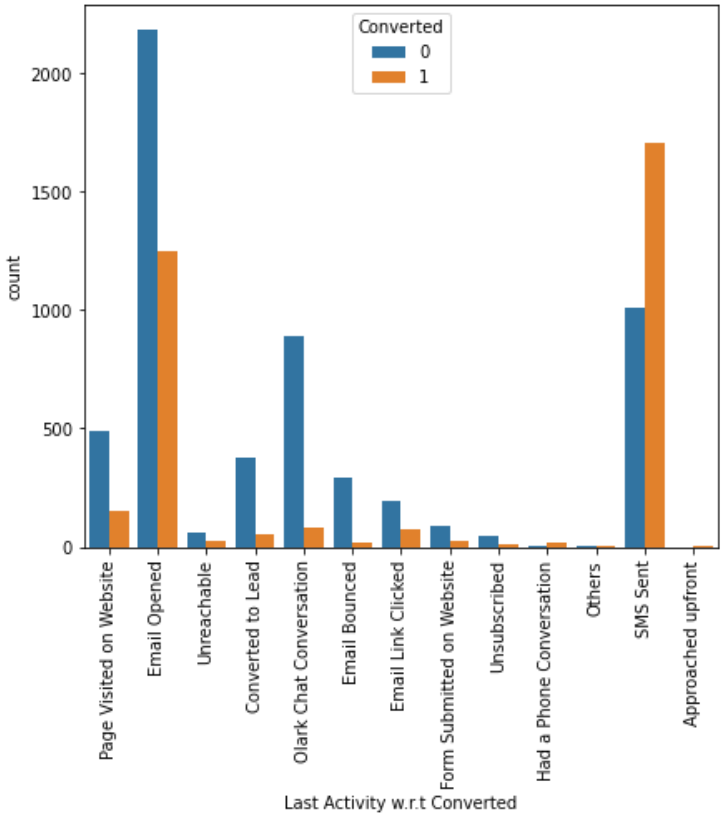
Inference :

•There is 40% conversion in people who have opted for email

- Last Activity

EDA insights



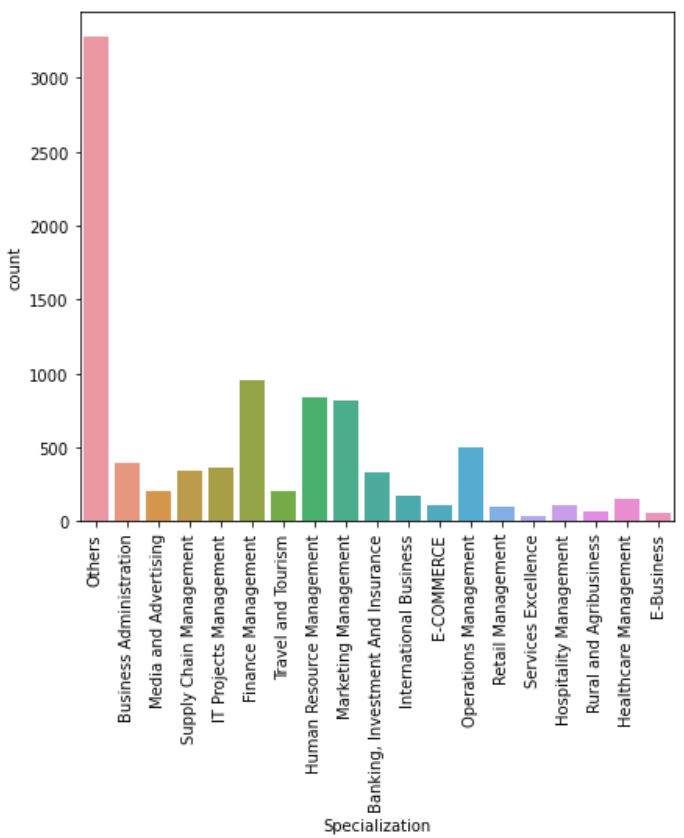


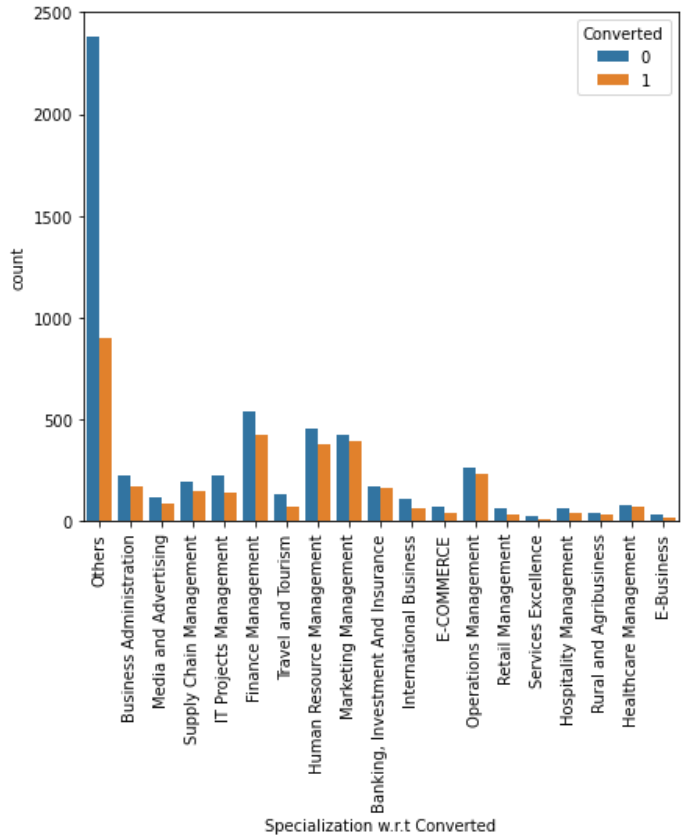
Inference :

•SMS Sent and Email Opened activities have higher conversion rate. The sales team can talk to these people directly and try to increase the conversion rate by explaining the advantages of the courses and difference in the content offered by them v/s company X competitors or provide a discount if possible to such customers

- Specialization

EDA insights





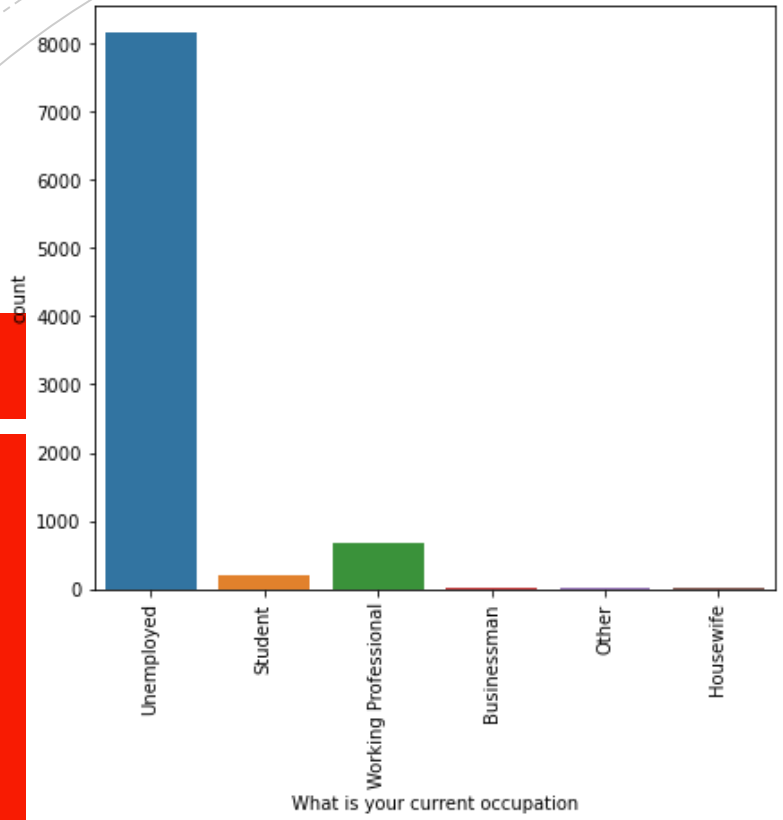
Inference :

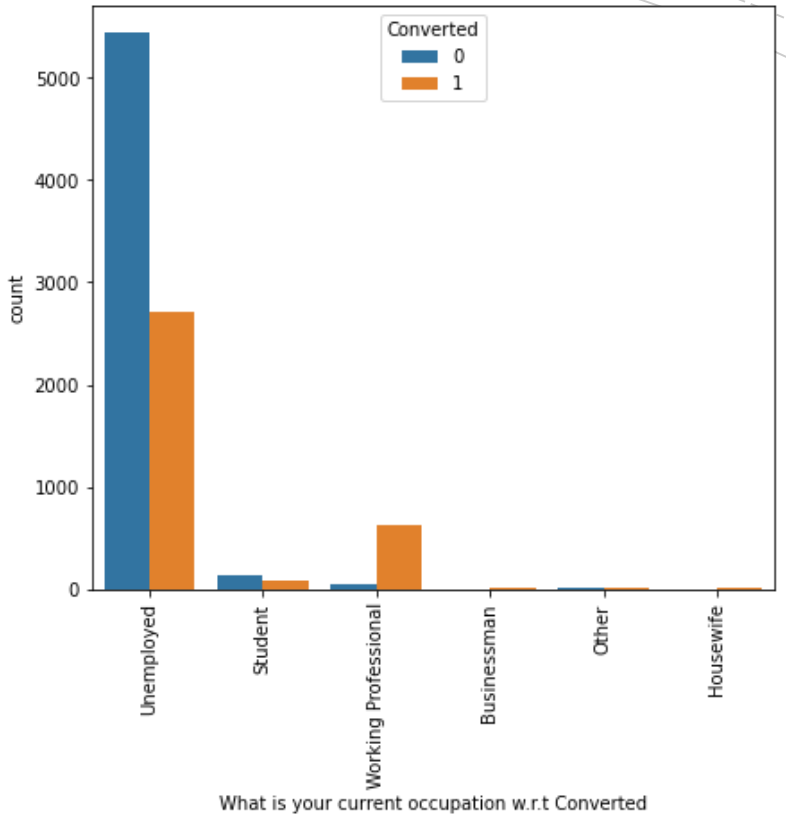
•Many data points have not specified their Specialization ('Others category was replaced by np.NaN')

•Higher conversion rates are observed in Human Resource Management, Finance Management, Banking, Investment And Insurance, Business Administration and few others

- What is your current occupation

EDA insights





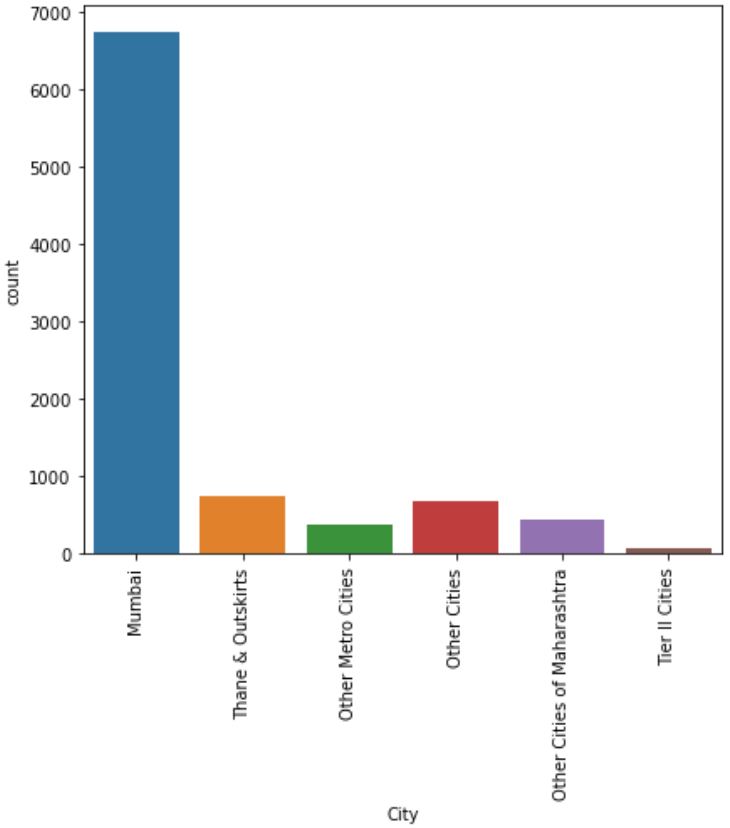
Inference :

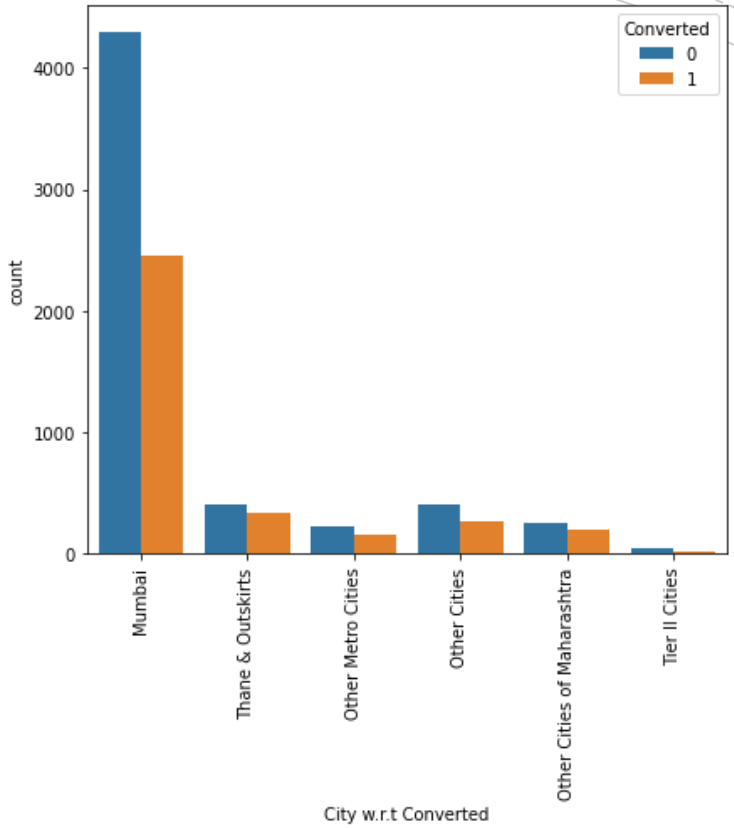
•Unemployed category is the biggest contributor to conversion rate

•But Working Professional seem to have higher conversion rate

- City

EDA insights



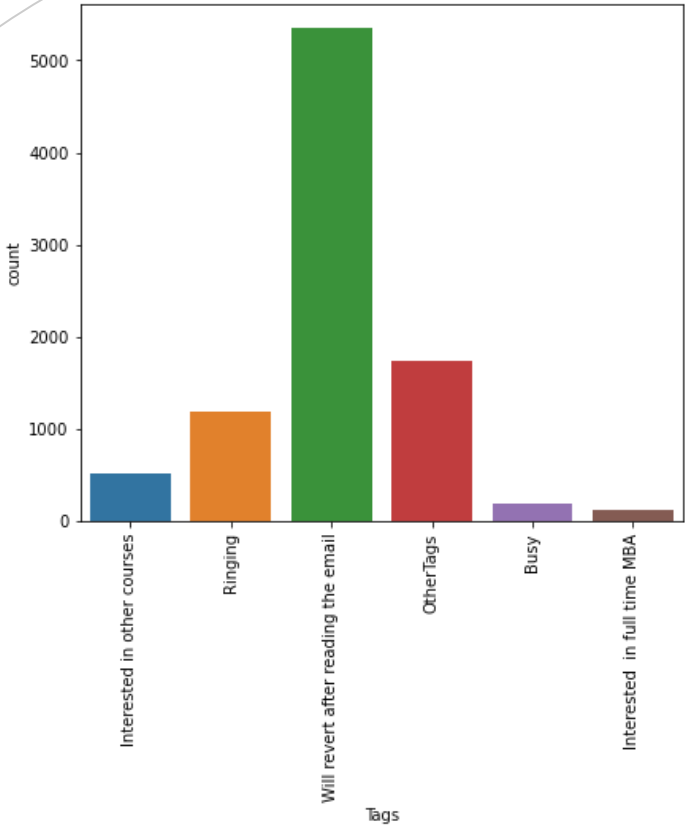


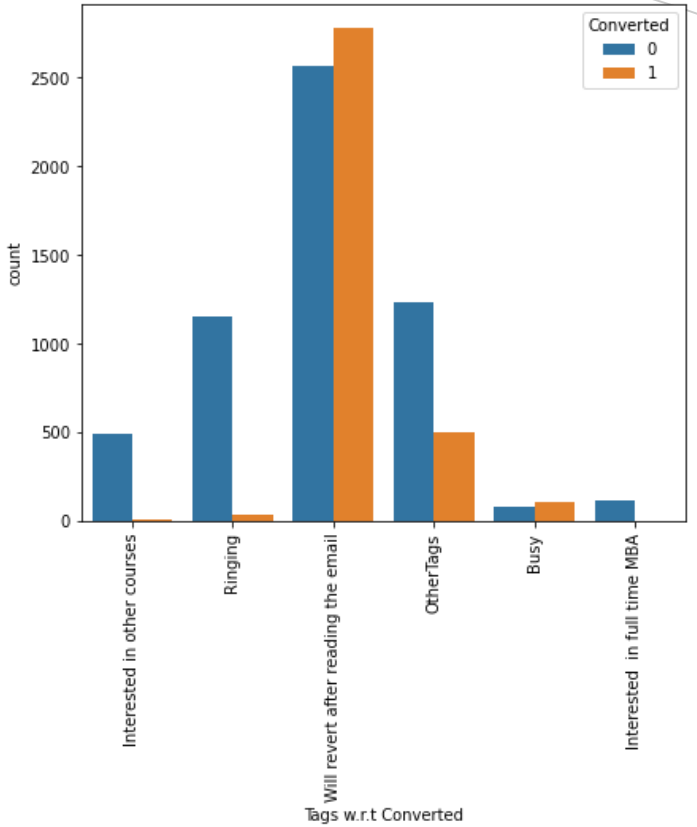
Inference :

•The sales employee can focus on people from Maharasthra to improvise the conversion rate as Mumbai and other cities of Maharashtra have higher conversion rate and also represent majority of data population

- Tags

EDA insights





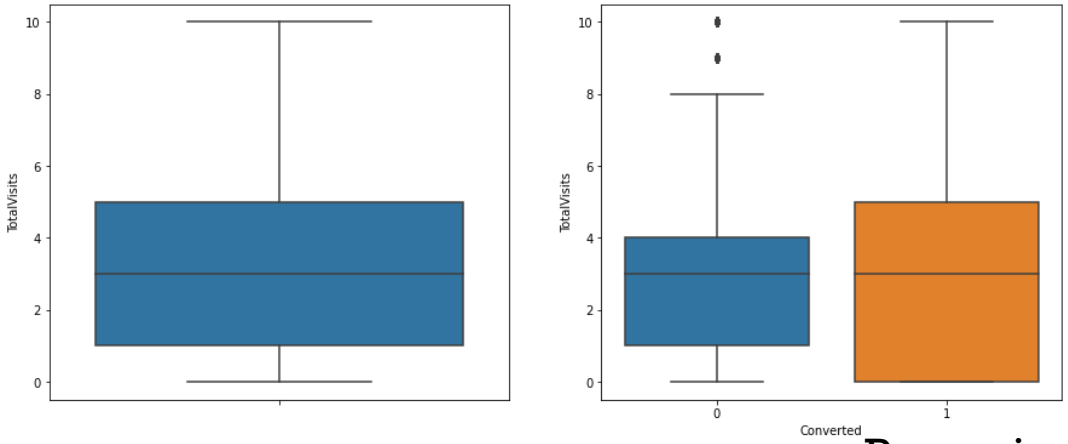
Inference :

•'Will Revert After Reading Email' has higher conversion rate and higher representation.

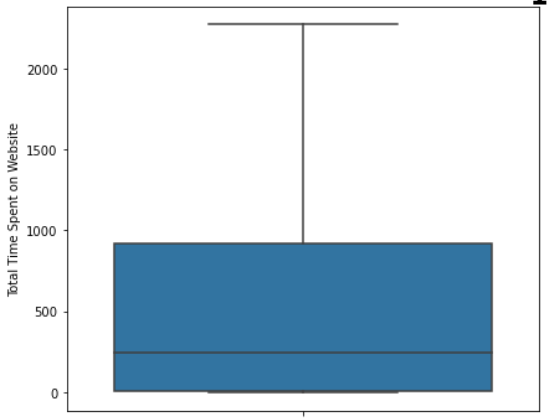
•'Busy' and 'Closed By Horizon' have next higher conversion rate apart form the above.

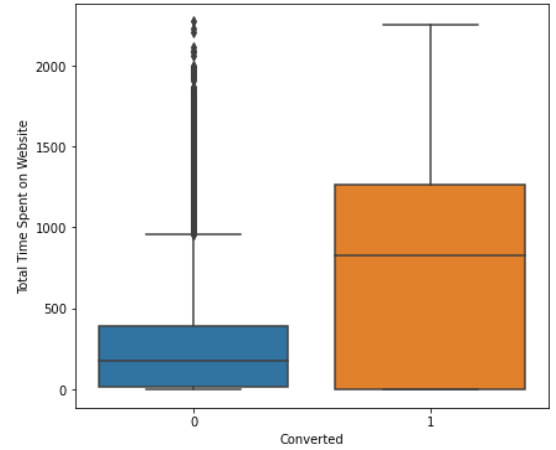
EDA insights

Total Visits

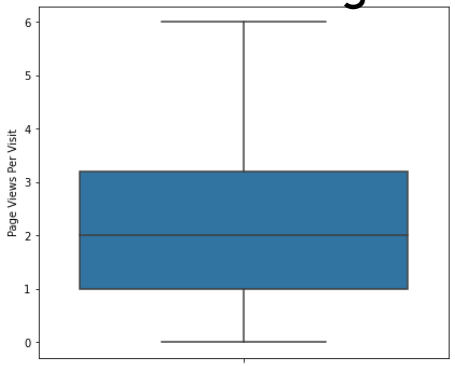


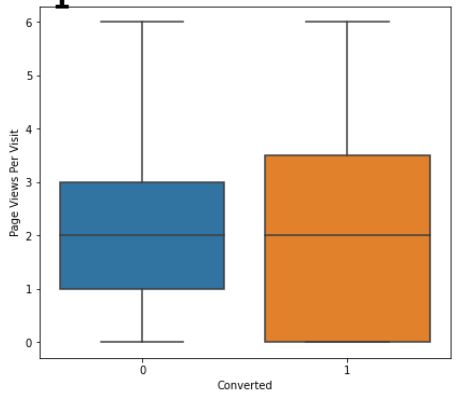
Total time spent on website





Page views per visit



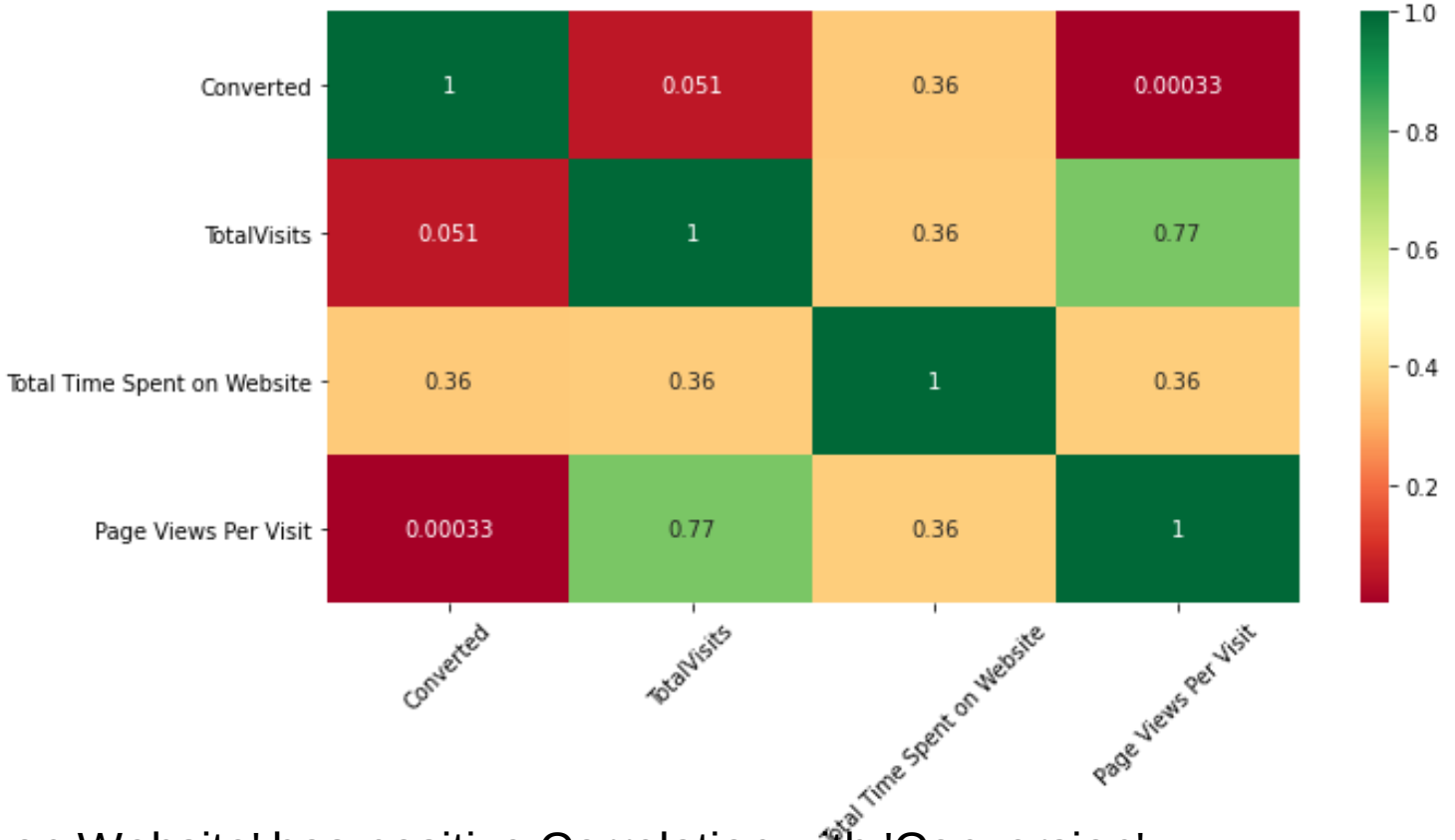


Inference

- We cannot infer much from Total Visits.
- We can observe that people spending more time on website have are likely to be converted
- Nothing can be inferred from Page Views Per Visit

EDA insights

- Checking correlations of numerical columns with ‘Converted’



Inference :

- We can observe that 'Total Time Spent on Website' has positive Correlation with 'Conversion'
- There is almost no correlation in 'Page Views Per Visit' and 'TotalVisits' with 'Conversion'
- 'Page Views Per Visit' and 'Total Time Spend on Website' have highest correlation with 'Conversion'
- There seems to be some correlation between 'total visits' and 'page views per visit'

Model evaluation and feature scaling

Logistic regression model

- Confusion matrix of the initial model

Predicted →	Converted	Not converted
Actual ↓		
Converted	3270	365
Not converted	579	708

Overall accuracy: 0.89

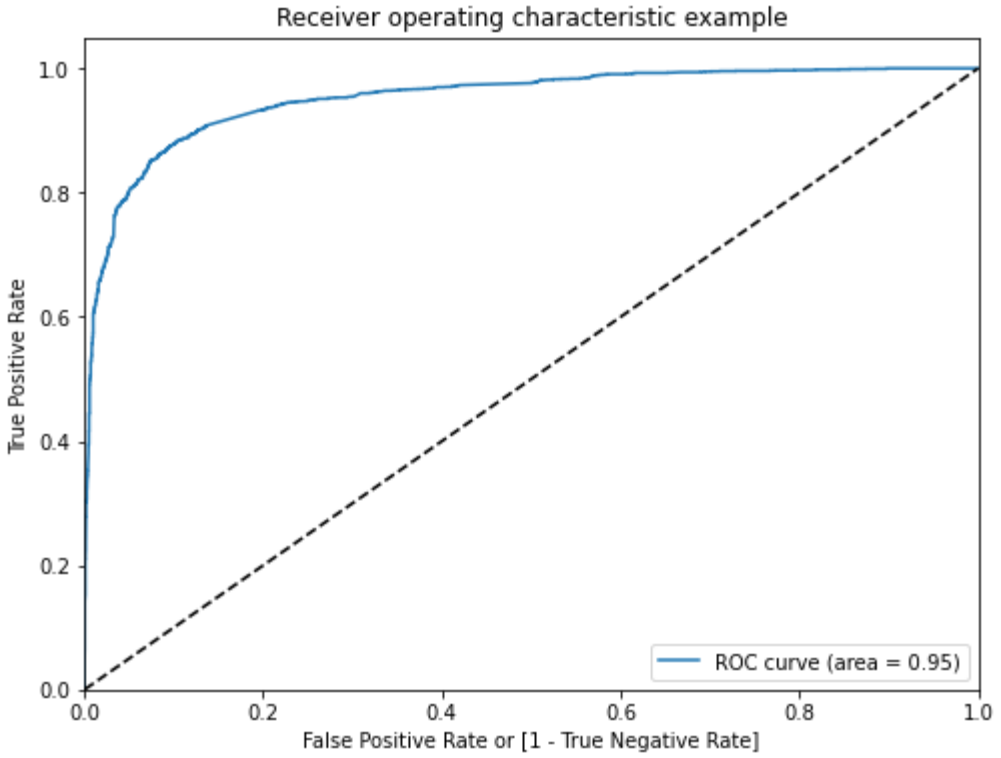
Logistic regression model

- VIF

	Features	VIF
13	Lead Quality_Not Sure	3.62
11	Tags_Will revert after reading the email	3.45
12	Lead Quality_Might be	1.97
2	Lead Source_Olark Chat	1.82
10	Tags_Ringing	1.65
15	Last Notable Activity_SMS Sent	1.63
6	Last Activity_Olark Chat Conversation	1.44
1	Total Time Spent on Website	1.35
7	What is your current occupation_Working Profes...	1.32
3	Lead Source_Reference	1.24
9	Tags_Interested in other courses	1.21
14	Lead Quality_Worst	1.14
0	Do Not Email	1.13
5	Last Activity_Converted to Lead	1.10
4	Lead Source_Welingak Website	1.07
8	Tags_Interested in full time MBA	1.05

Logistic regression model

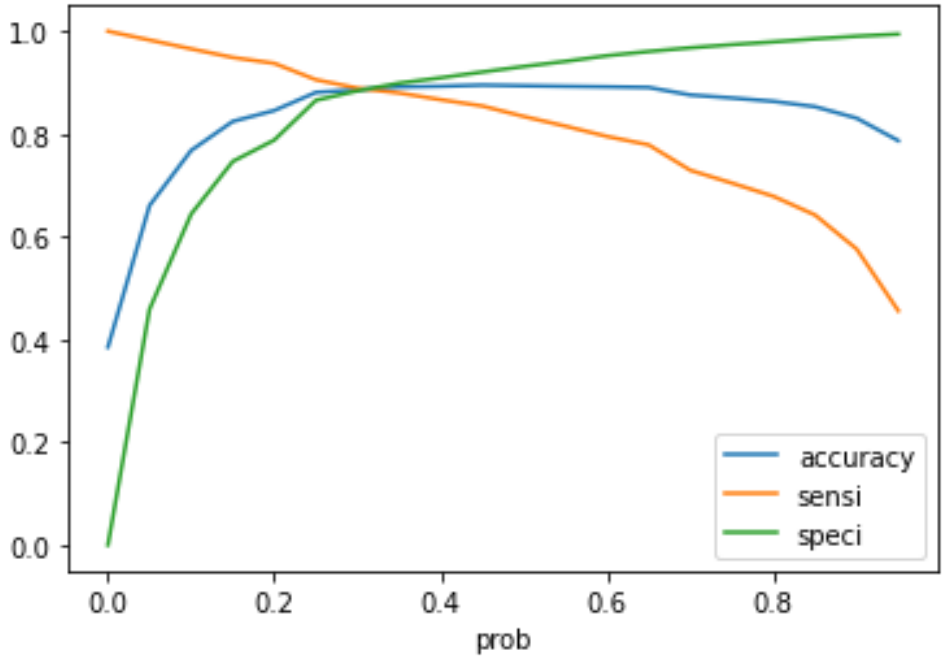
- ROC



•If the curve is closer to the top left of the graph the model is performing good

Logistic regression model

- Optimal cut-off using threshold



•We will be choosing the cut-off-threshold as 0.3 to strike a good balance among the following metrices : accuracy, sensitivity and specificity

Logistic regression model

- Final model

Confusion matrix

Predicted →	Converted	Not converted
Actual ↓		
Converted	3452	453
Not converted	271	2175

Training Metrices

Accuracy:0.886002204377263 4,
Sensitivity:0.886002204377263 4,
Specificity:0.886002204377263 4

Logistic regression model

- Model Prediction

Confusion matrix

Predicted →	Converted	Not converted
Actual ↓		
Converted	1522	212
Not converted	129	860

Training Metrices

Accuracy:0.874770473742196,
Sensitivity:0.874770473742196,
Specificity:0.874770473742196

As both the metrices for training and testing data from the model are good we can conclude our final model is good

Final model results summary

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6334
Model Family:	Binomial	**Df Model:**	16
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1692.3
Date:	Mon, 11 Oct 2021	**Deviance:**	3384.6
Time:	16:02:25	**Pearson chi2:**	1.01e+04
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.2487	0.174	7.164	0.000	0.907	1.590
Do Not Email	-1.8095	0.210	-8.599	0.000	-2.222	-1.397
Total Time Spent on Website	1.1359	0.053	21.621	0.000	1.033	1.239
Lead Source_Olark Chat	1.2522	0.129	9.685	0.000	0.999	1.506
Lead Source_Reference	3.2806	0.348	9.439	0.000	2.599	3.962
Lead Source_Welingak Website	6.0962	0.739	8.247	0.000	4.647	7.545
Last Activity_Converted to Lead	-1.1326	0.271	-4.175	0.000	-1.664	-0.601
Last Activity_Olark Chat Conversation	-1.4354	0.184	-7.822	0.000	-1.795	-1.076
What is your current occupation_Working Professional	1.5271	0.272	5.611	0.000	0.994	2.061
Tags_Interested in full time MBA	-3.1553	0.890	-3.546	0.000	-4.899	-1.411
Tags_Interested in other courses	-3.1221	0.398	-7.854	0.000	-3.901	-2.343
Tags_Ringing	-3.9020	0.278	-14.021	0.000	-4.448	-3.357
Tags_Will revert after reading the email	0.8364	0.112	7.487	0.000	0.617	1.055
Lead Quality_Might be	-1.4047	0.197	-7.131	0.000	-1.791	-1.019
Lead Quality_Not Sure	-3.4582	0.172	-20.119	0.000	-3.795	-3.121
Lead Quality_Worst	-5.1577	0.415	-12.431	0.000	-5.971	-4.344
Last Notable Activity_SMS Sent	1.9559	0.111	17.622	0.000	1.738	2.173

Conclusion

- The top three variables in your model which contribute most towards the probability of a lead getting converted:
  - Lead Source (Welingak Website)
  - Last Notable Activity (SMS Sent)
  - What is your current occupation (Working Professional)
- The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?
  - Lead Source_Welingak Website
  - Lead Source_Reference
  - Last Notable Activity_SMS Sent
- Phone calls must be targeted to prospective people who meet the following criteria:
  - Target working professionals
  - People visiting (or leads from) the Welingak Website.
  - Follow up with people to whom SMS has been sent.
  - People who spend time on website
  - Phone leads from references, and Olark chat.
  - Follow up with people with whom the company has already had a telephone conversation.
- During non-peak times, the company should focus on:
  - Focus on improving the quality of chat on Olark as this has a negative effect on conversion.
  - Develop strategies for starting new courses as some candidates are not converted because they are interested in other courses.