
Lead
Scoring

PURPOSE

- The Purpose is to optimize the lead scoring mechanism based on their fit, demographics, behaviours, buying tendency etc. By implementing explicit & Implicit lead scoring modelling with lead point system.

PROBLEM STATEMENT

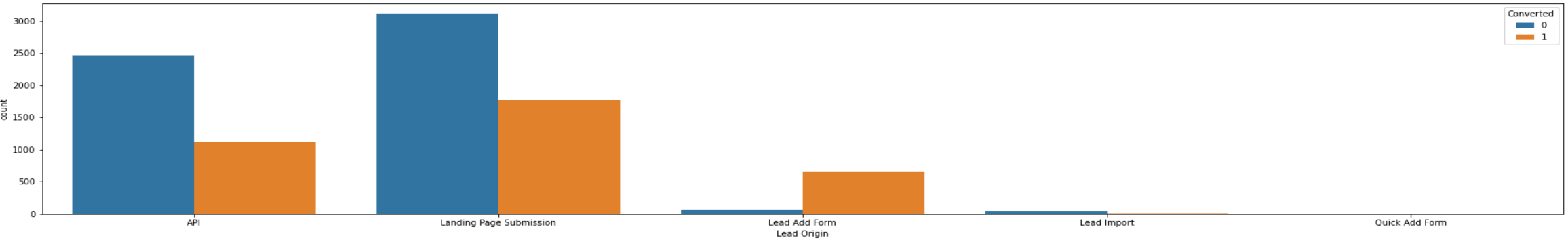
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

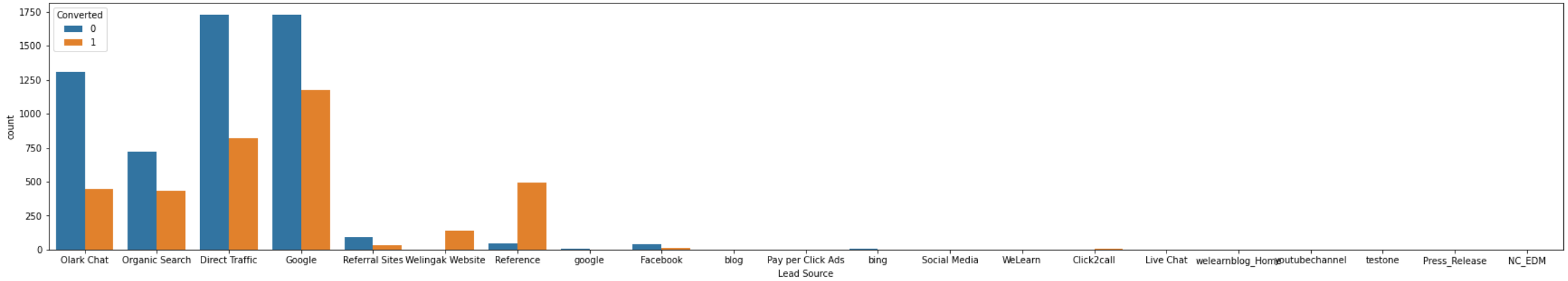
Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If 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.

APPROACH

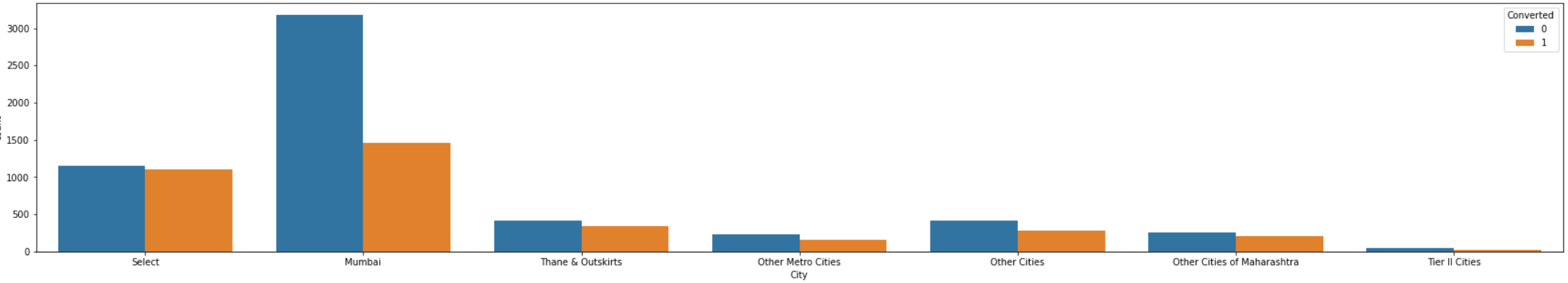
- Source the data for Analysis
- Reading & Understanding the data
- Data Cleaning
- EDA
- Feature Scaling
- Splitting the data into Train and Test dataset
- Model Building
- Model Evaluation – Specificity and Sensitivity or precision recall
- Making Prediction on the test set



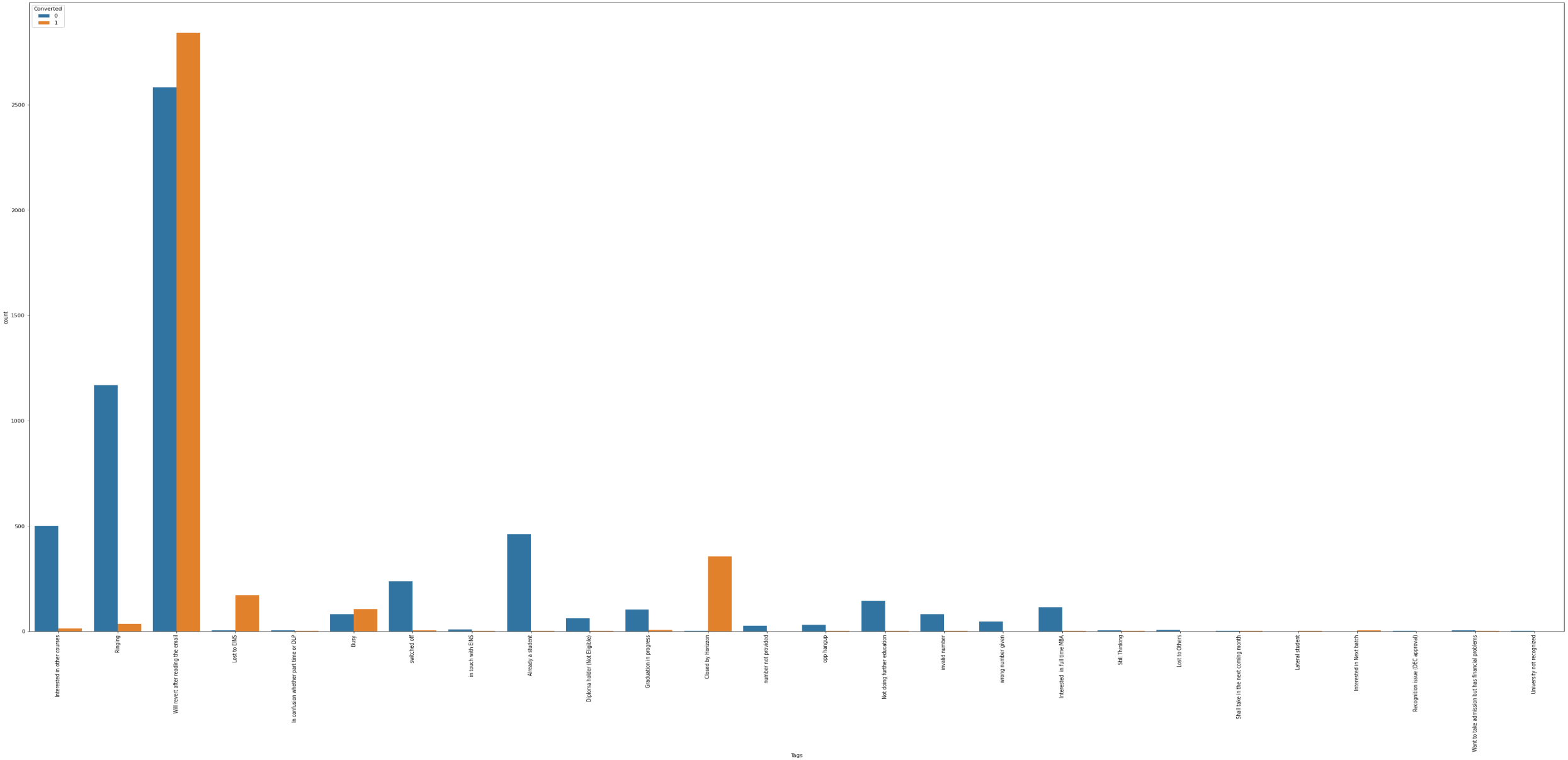
- Customer identification and conversion, both are higher in Landing Page Submission.



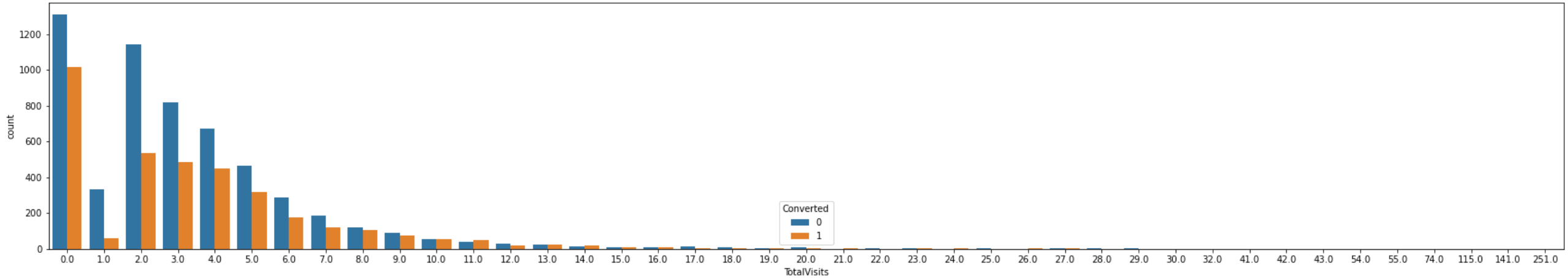
- The Lead Generation and Conversion, both are higher in Google than all other sources.



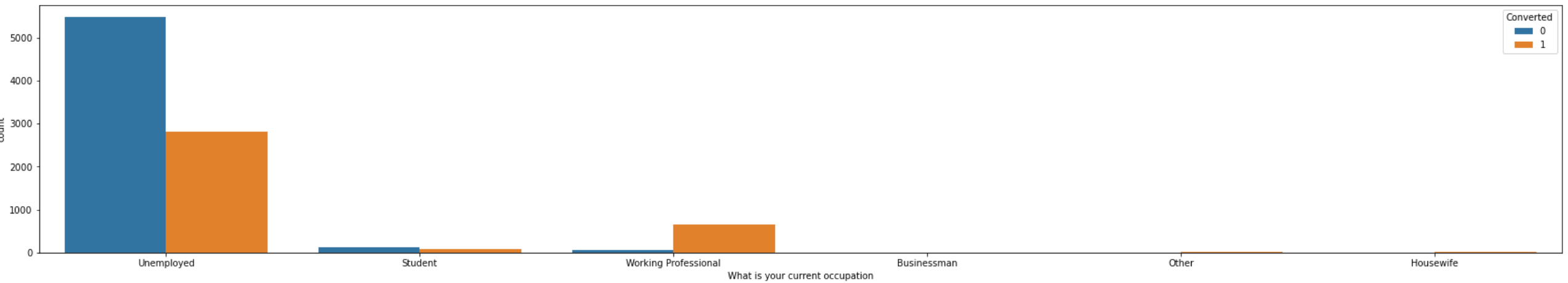
- Mumbai has the highest number of leads and conversion out of all.



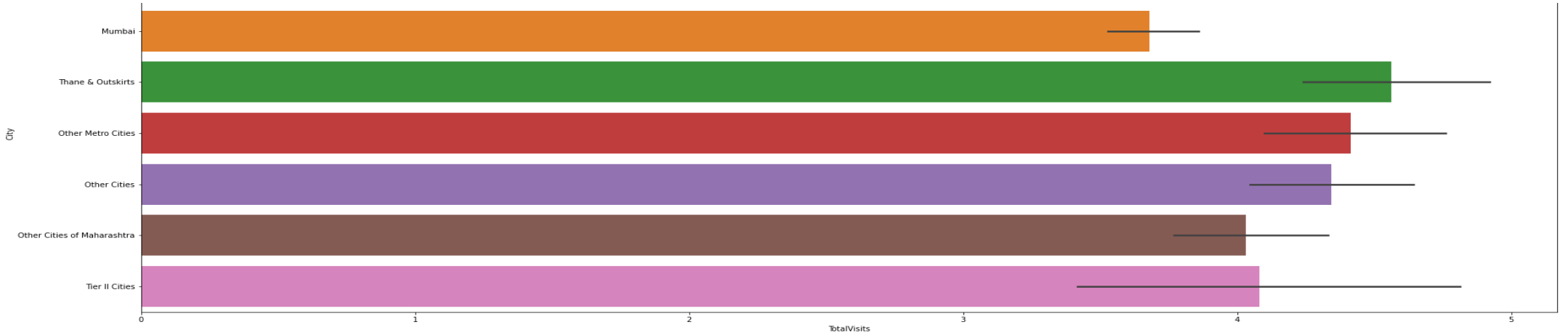
- As per the graph above, maximum conversion is through reverting emails.



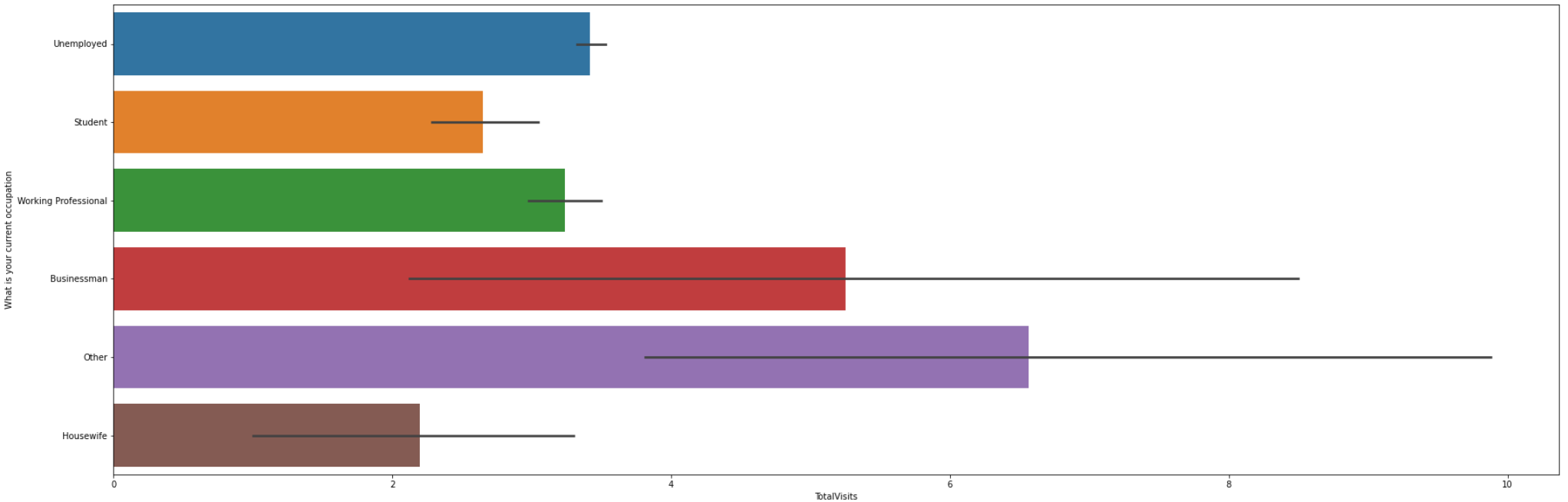
- In this analysis the maximum no. visit is 0 that means the people are getting information through 'other' sources. i.e. Newspaper Articles, Word of mouth etc.



- Graph above proves that, we have maximum conversion from unemployed people.



- We can see that website has been most visited by people in Thane & Outskirts region.



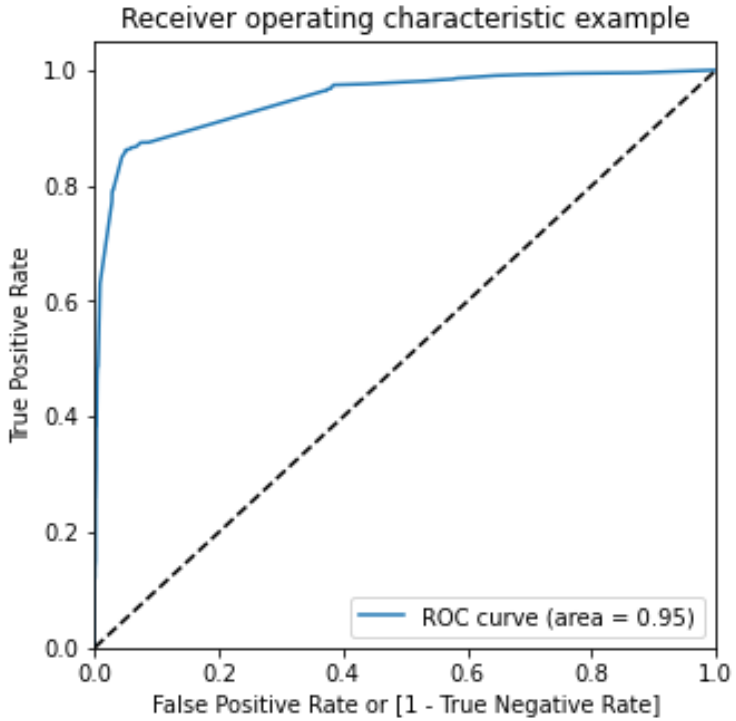
- Analysis above shows that both people visiting the website belong to ‘Other’ occupations.

- Some variables are highly co-related to each other. So, we have dropped the variables avoiding multi-collinearity.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.0511	0.195	-15.664	0.000	-3.433	-2.669
Lead Source_Welingak Website	2.6363	0.750	3.514	0.000	1.166	4.107
Last Activity_Email Bounced	-1.4365	0.298	-4.827	0.000	-2.020	-0.853
Last Activity_Olark Chat Conversation	-1.5875	0.183	-8.668	0.000	-1.946	-1.229
What is your current occupation_Working Professional	2.4233	0.266	9.104	0.000	1.902	2.945
Tags_Busy	3.9968	0.310	12.896	0.000	3.389	4.604
Tags_Closed by Horizzon	10.3251	1.034	9.981	0.000	8.298	12.353
Tags_Lost to EINS	10.1628	0.759	13.386	0.000	8.675	11.651
Tags_Ringing	-0.7443	0.302	-2.465	0.014	-1.336	-0.152
Tags_Will revert after reading the email	5.0259	0.221	22.781	0.000	4.593	5.458
Tags_switched off	-1.3086	0.588	-2.227	0.026	-2.460	-0.157
Lead Profile_Select	-3.3067	0.141	-23.454	0.000	-3.583	-3.030
Lead Profile_Student of SomeSchool	-3.8246	1.091	-3.507	0.000	-5.962	-1.687
City_Select	1.8540	0.121	15.308	0.000	1.617	2.091
Last Notable Activity_SMS Sent	2.7955	0.120	23.280	0.000	2.560	3.031

	Features	VIF
9	Tags_switched off	1.11
5	Tags_Closed by Horizzon	1.10
4	Tags_Busy	1.08
0	Lead Source_Welingak Website	1.07
1	Last Activity_Email Bounced	1.06
6	Tags_Lost to EINS	1.05
11	Lead Profile_Student of SomeSchool	1.04
3	What is your current occupation_Working Profes...	0.61
2	Last Activity_Olark Chat Conversation	0.27
8	Tags_Will revert after reading the email	0.16
10	Lead Profile_Select	0.15
7	Tags_Ringing	0.14
12	City_Select	0.04
13	Last Notable Activity_SMS Sent	0.00

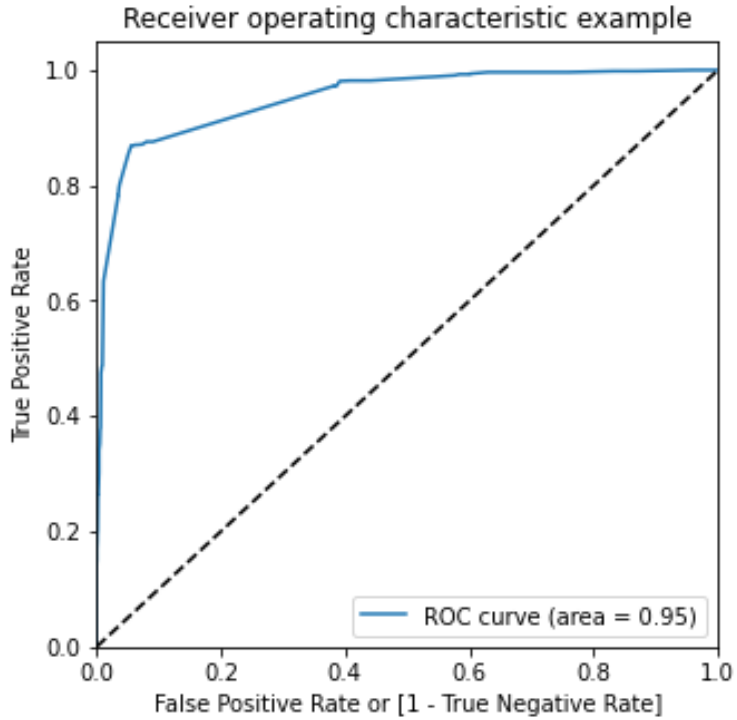
- We have dropped 5 features for “P Value” to be less than 0.05.
- In VIF, all the features are less than 5.




```
0.9160482374768089
```

```
sensitivity: 0.8600973236009732
specificity: 0.9505247376311844
false +ve rate: 0.0494752623688155
+ve predictive value: 0.9146183699870634
-ve predictive value: 0.9168474331164136
```

- ROC Curve for Train Set
- Accuracy of Train Set is 0.916
- Overall Performance of Train Set




```
0.9145021645021645
```

```
sensitivity: 0.869406392694064
specificity: 0.9439475253428742
false +ve rate: 0.05605247465712582
+ve predictive value: 0.9101338432122371
-ve predictive value: 0.9171494785631518
```

- ROC Curve for test Set
- Accuracy of Train Set is 0.9145, which almost equal to Train Set Accuracy
- Overall Performance of Train Set

CONCLUSION

- While we have checked both sensitivity-specificity as well as Precision and recall metrics, we have considered that optimal cutoff based on sensitivity & specificity for calculating the final prediction.
- Accuracy, Sensitivity & specificity values of the test set are around 91%, 86%, 94% which are approximately closer to values calculated using Trained dataset.
- Lead score calculated for the conversion rate final model on Train & Test dataset is 95%.
- Hence, Overall Model seems to be good.


Thank you