LEAD SCORING CASE STUDY

Submitted By : Pallavi Jadhav
Vamsiram Pandlluru
Parthib Ray

Batch DS – C60 PROGRAM : upGrad & IIITB | Data Science Program –September 2023

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. 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%.

Business Goals

- Education company wants to identify ‘Hot Leads’ so that sales team can focus more on potential leads rather than making calls to everyone.

- Company want to build a model wherein we need to assign a lead score to each of the leads such that 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%

Solution Strategy

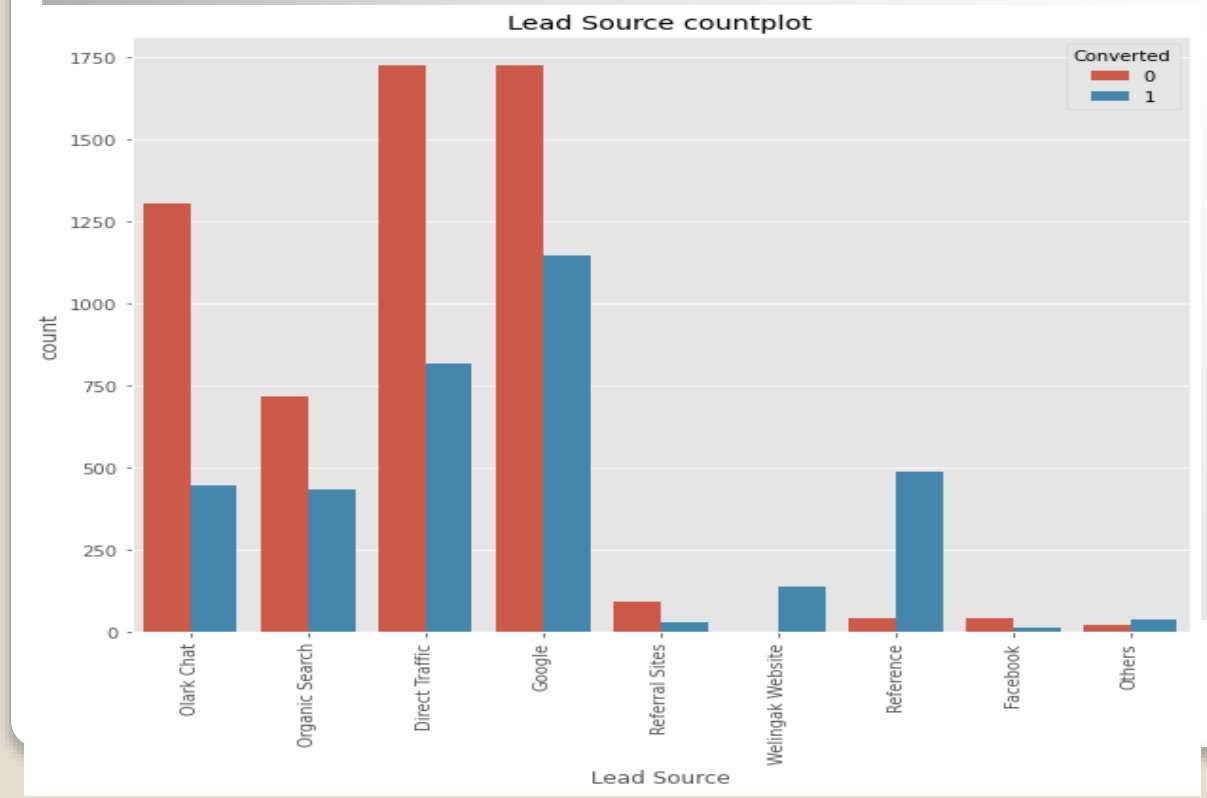
- Importing and understanding the data
- Cleaning the data and preparing it for Analysis
- Exploratory Data Analysis
- Splitting data into training set and test set
- Scaling
- Building logistic regression model with the best features
- Evaluating the model on training set
- Finding the optimal cut-off to get the best accuracy, sensitivity and specificity.
- Evaluating the model on test set
- Compute lead score and check if sensitivity is around 80% on test and train sets.

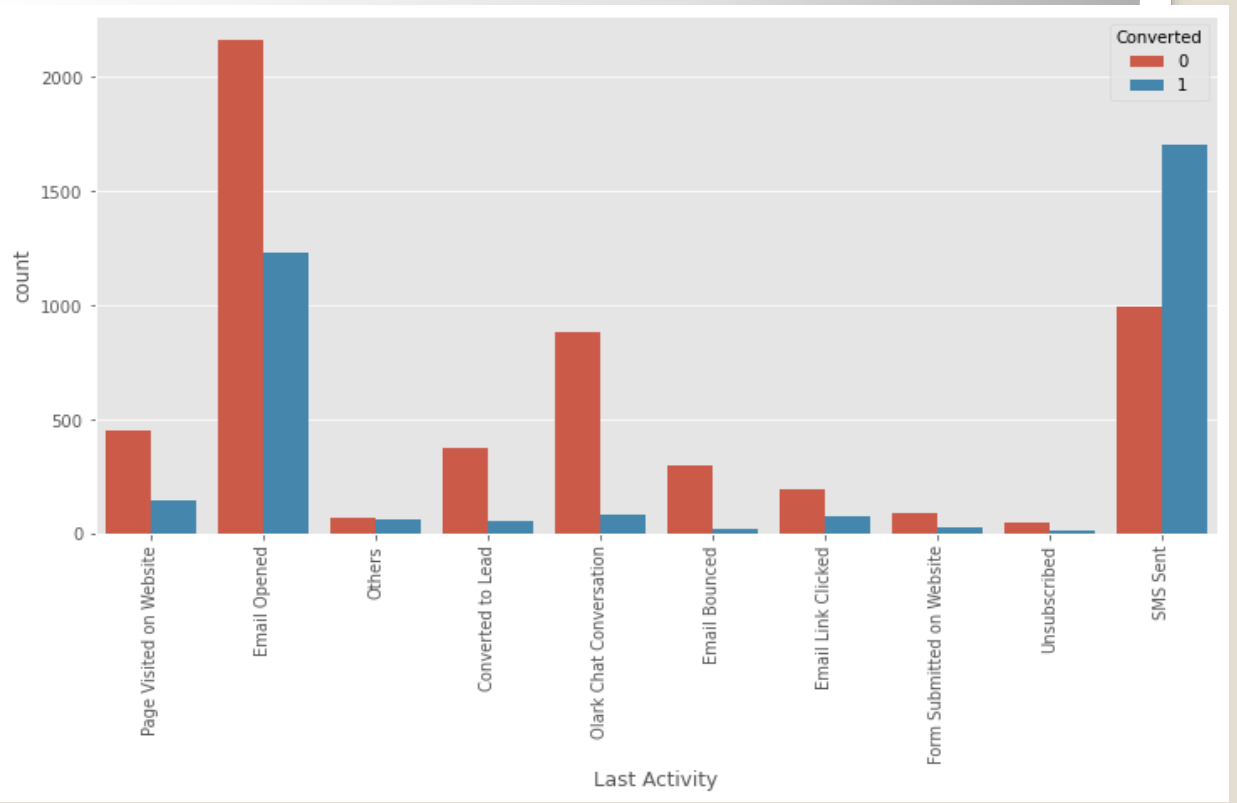
Findings of Exploratory Data Analysis

- We have around 39% conversion rate in the dataset

- Conversion rate is higher for Reference and Welingak Website
- improving conversion rate for Direct Traffic, Google, Organic Search and Olark Chat

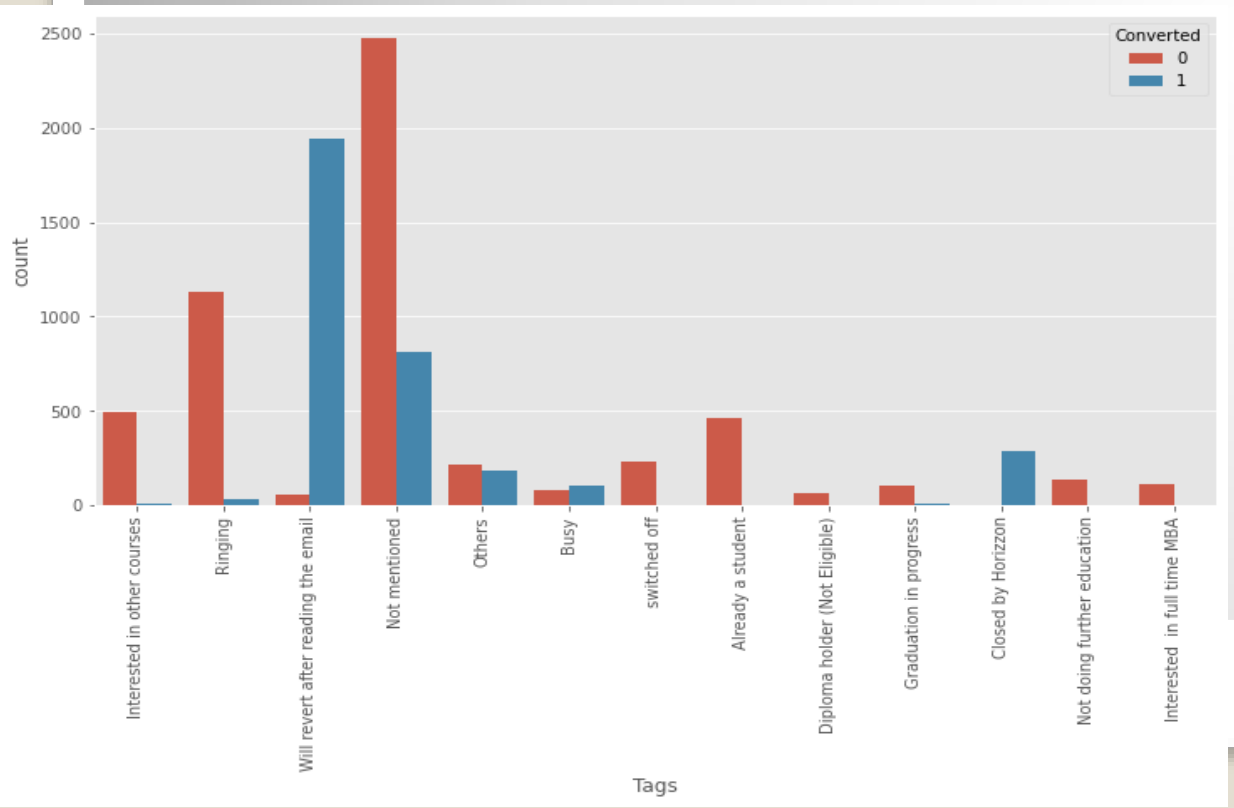
- Conversion rate is higher where Last Activity is 'SMS Sent'.
- Conversion rate need to be increased for all other customers majorly for Email opened,

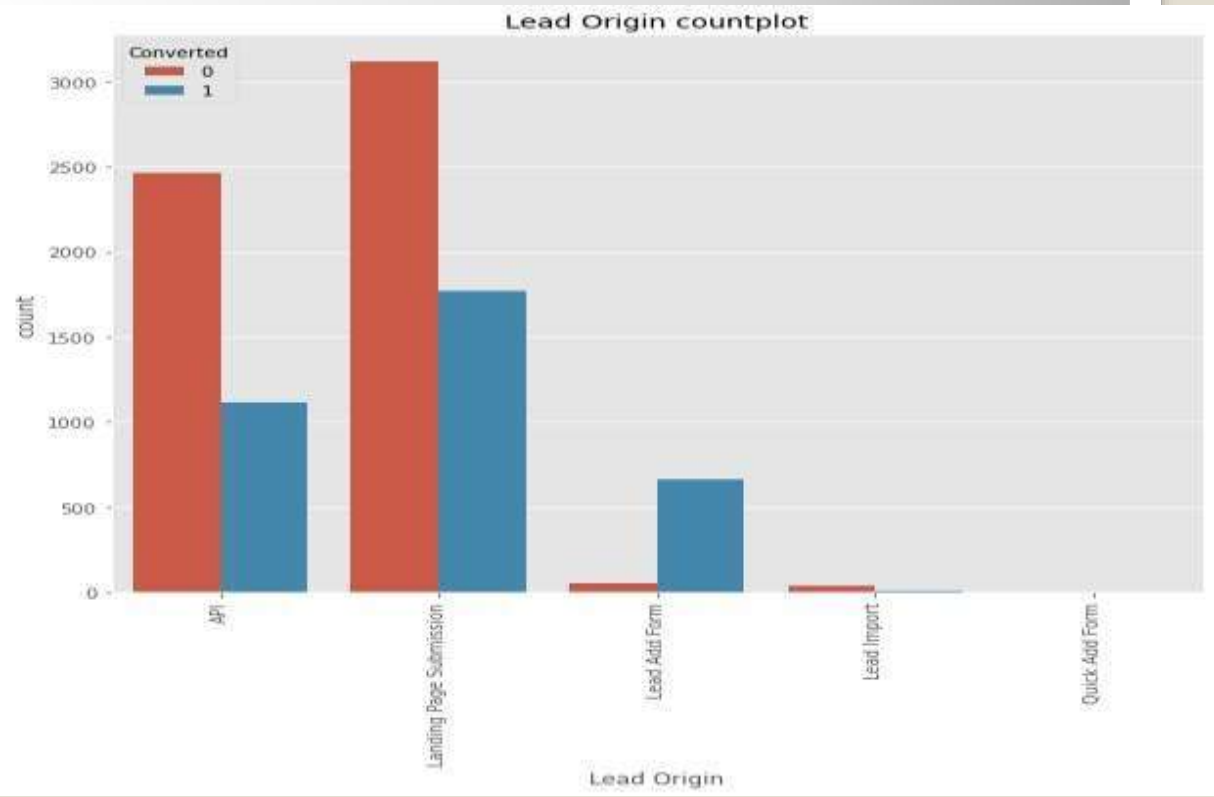




Findings of Exploratory Data Analysis

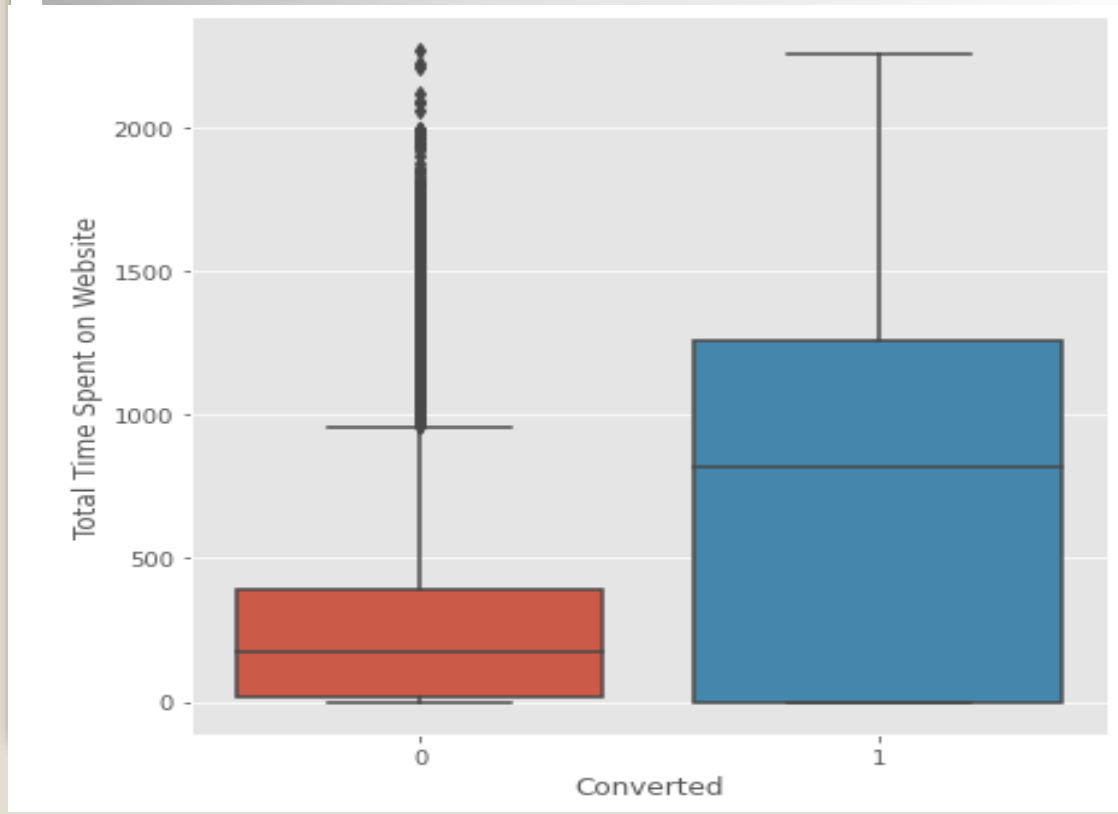
- Conversion rate is high for `Will revert after reading the email and Closed by Horizzon tags
- Conversion rate is very low for rows having Interested in other courses, Interested in full time MBA and Ringing tags
- API and Landing Page Submission brought more number of leads and conversions are also more. However, conversions for these two Lead Origin can be improved
- Lead Add Form has very high conversion rate but count of leads is less.



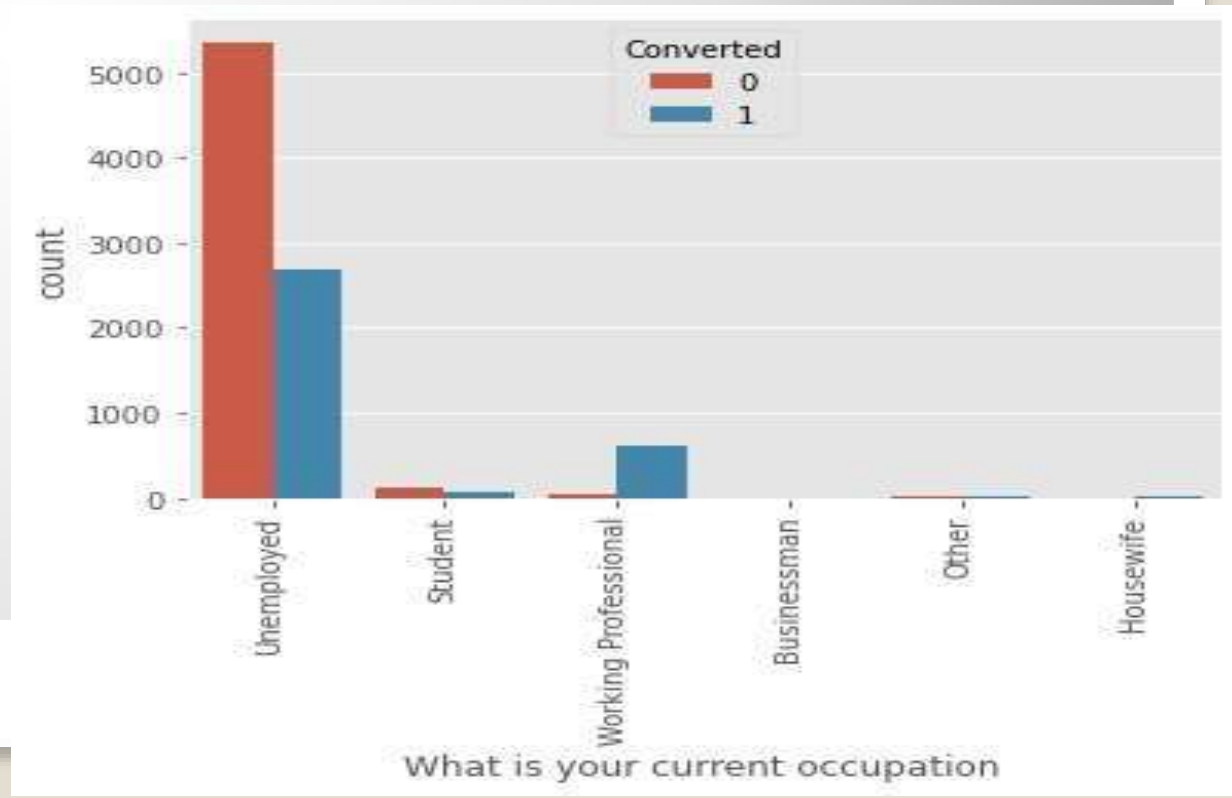


Findings of Exploratory Data Analysis

- Converted Leads have spent more time on Website

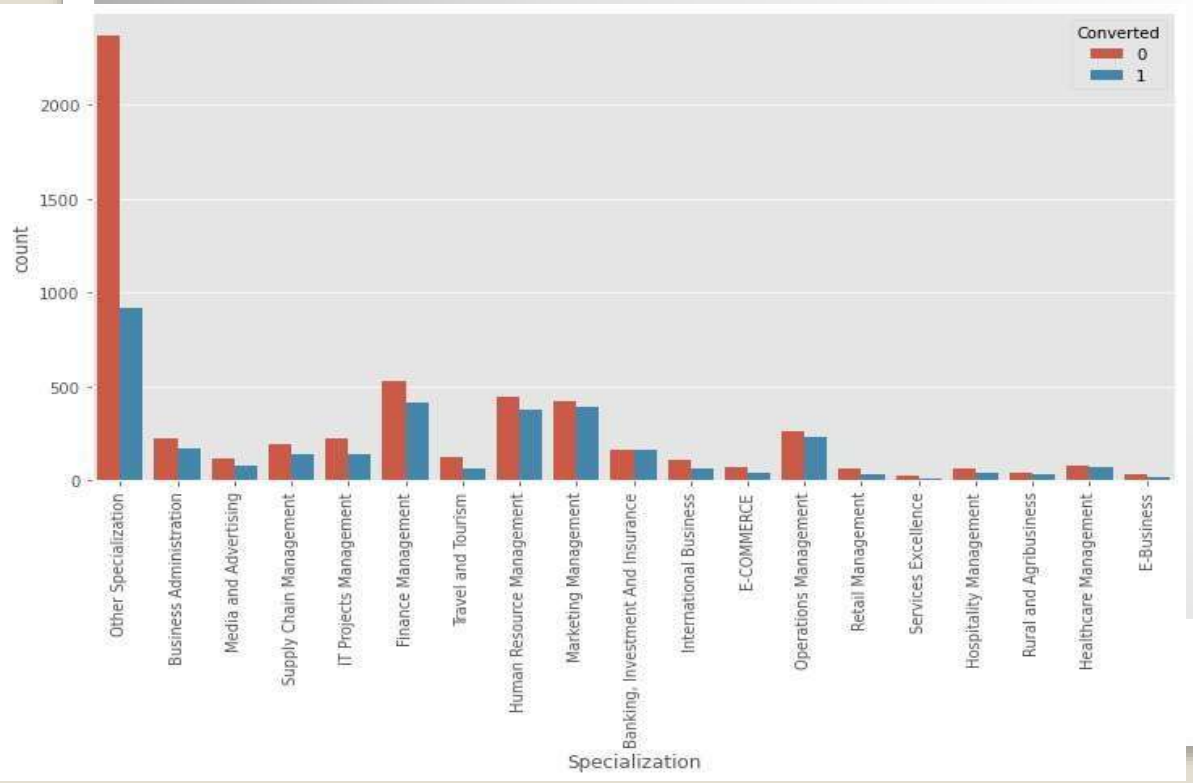


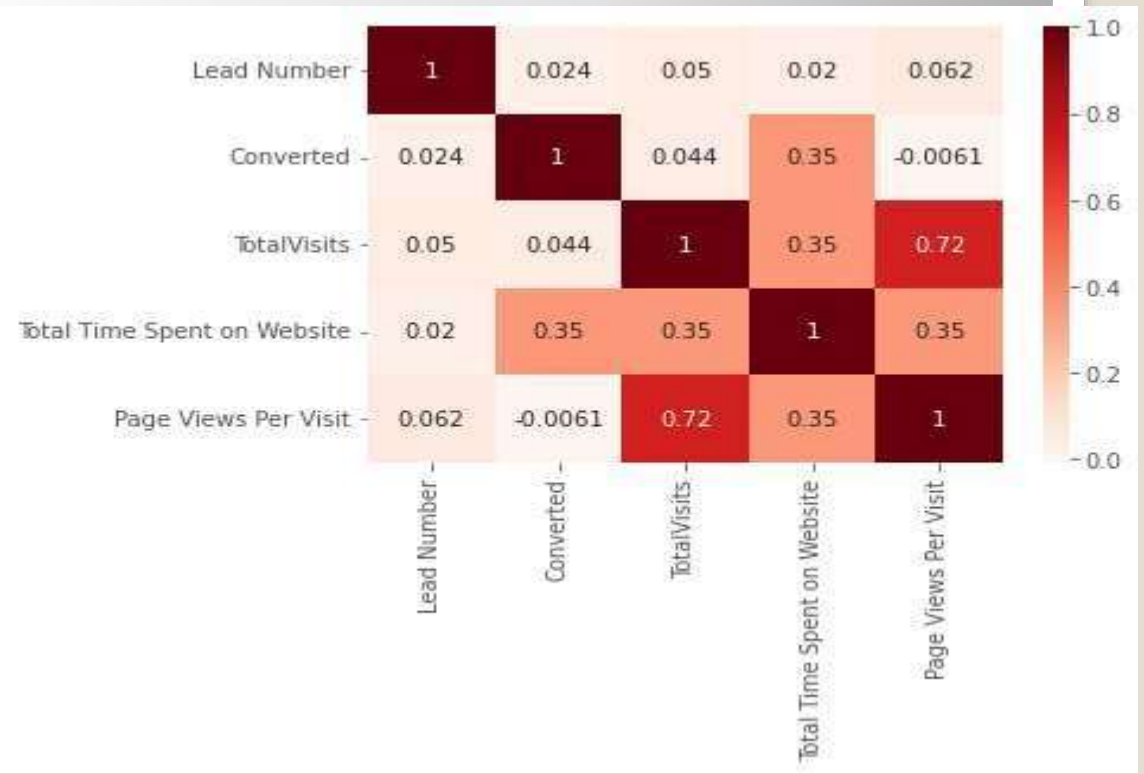
- Conversion rate is high for Working Professional.
- Company received more leads from Unemployed people in terms of absolute numbers



Findings of Exploratory Data Analysis

- Customers with Management, Business Administration & Media and Advertising specialization have higher conversion rate.
- Heatmap shows that `Page Views Per Visit` is highly correlated with `TotalVisits`.





Variables impacting the conversion rate

]:

Generalized Linear Model Regression Results

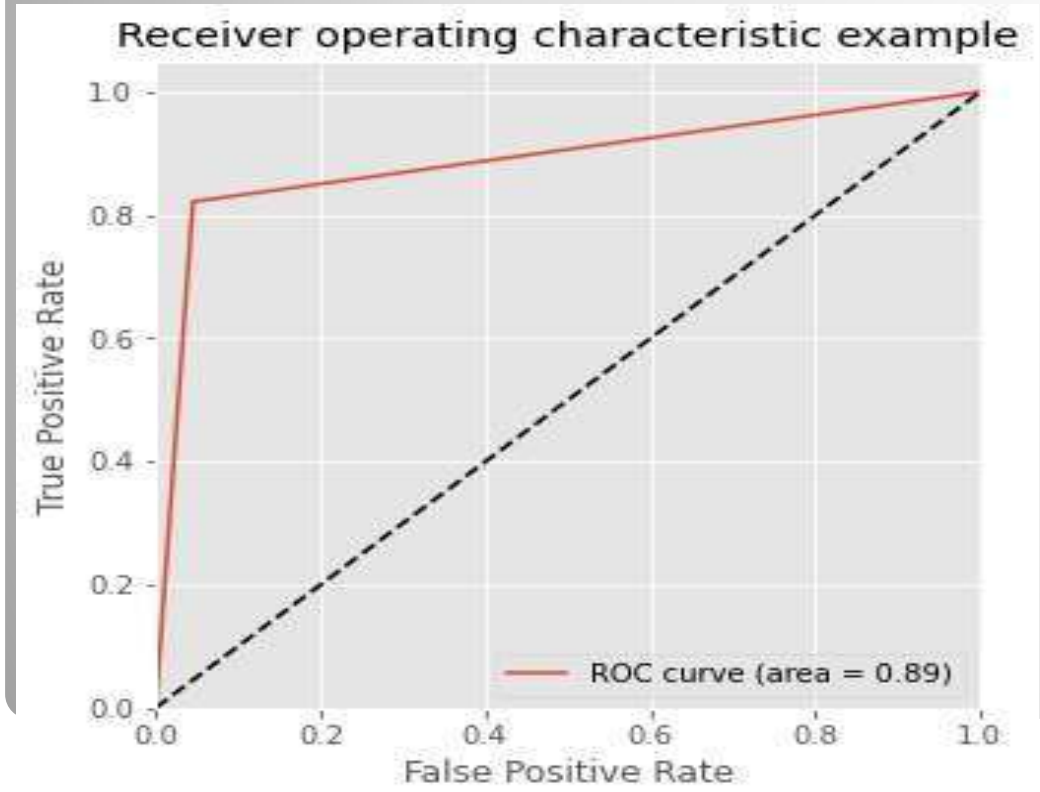
Dep. Variable:	Converted	**No. Observations:**	6267
Model:	GLM	**Df Residuals:**	6253
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1615.3
Date:	Sat, 25 Feb 2023	**Deviance:**	3230.6
Time:	22:44:24	**Pearson chi2:**	8.15e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.5566
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3346	0.059	-22.478	0.000	-1.451	-1.218
Do Not Email	-1.3085	0.206	-6.353	0.000	-1.712	-0.905
Lead Origin_Lead Add Form	1.4489	0.371	3.903	0.000	0.721	2.177
Lead Source_Welingak Website	3.6895	1.080	3.416	0.001	1.573	5.806
Last Activity_Olark Chat Conversation	-1.5026	0.197	-7.608	0.000	-1.890	-1.116
Tags_Already a student	-3.5273	0.712	-4.955	0.000	-4.923	-2.132
Tags_Closed by Horizzon	6.3094	1.008	6.261	0.000	4.334	8.285
Tags_Interested in full time MBA	-1.7154	0.593	-2.893	0.004	-2.878	-0.553
Tags_Interested in other courses	-2.3057	0.367	-6.290	0.000	-3.024	-1.587
Tags_Not doing further education	-2.8022	1.009	-2.777	0.005	-4.780	-0.825
Tags_Ringing	-3.4037	0.228	-14.910	0.000	-3.851	-2.956
Tags_Will revert after reading the email	4.2725	0.169	25.316	0.000	3.942	4.603
Tags_switched off	-3.9086	0.594	-6.581	0.000	-5.073	-2.745
Last Notable Activity_SMS Sent	2.2163	0.111	19.921	0.000	1.998	2.434

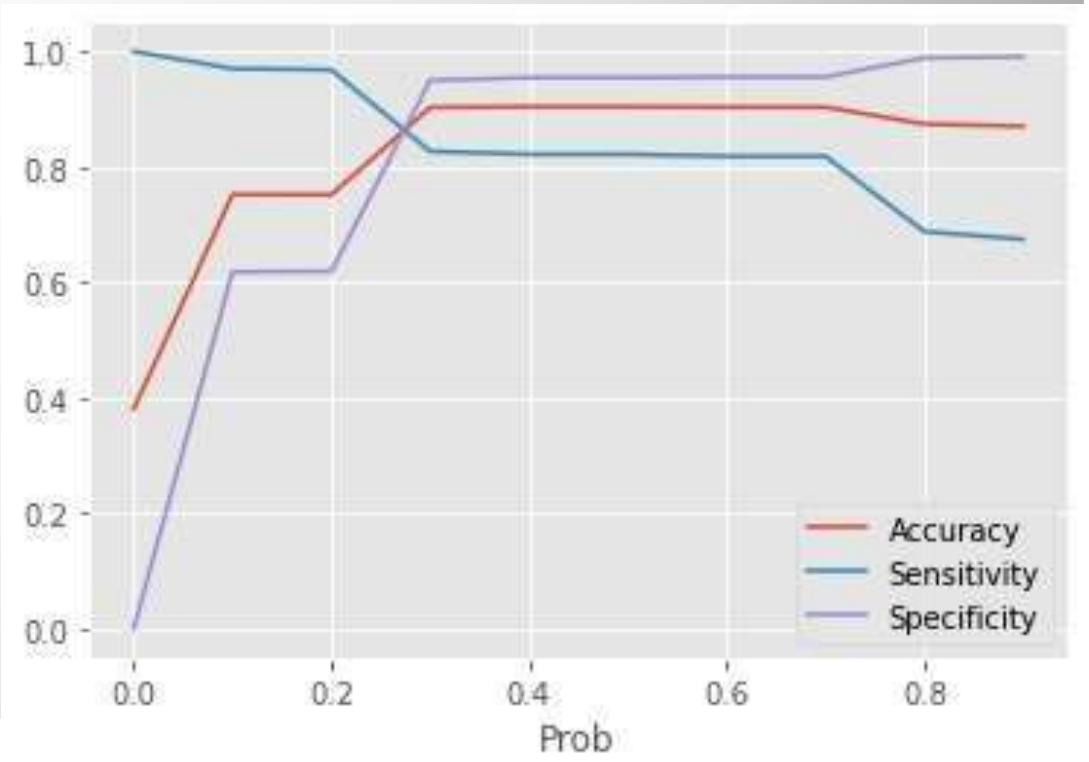
- According to the summary obtained, we can find the
- most impacting variables:
  - Tags_Closed by Horizzon
  - Tags_Will revert after reading the email
  - Lead Source_Welingak Website
  - Last Notable Activity_SMS Sent
  - Occupation_Working Professional

Model Evaluation on Train Data

- In the ROC plot it can be seen that the curve is going close to Y axis and near the value of 1. Also the area under the curve is very high. so the model is reliable.



- From the accuracy, sensitivity and specificity curve, we can see that optimal cut-off point is 0.3



Comparing results with test data

TRAIN DATA SET	TEST DATA SET
- Accuracy = 90.29%	- Accuracy = 90.32%
- Sensitivity/Recall = 82.68%	- Sensitivity/Recall = 81.98%
- Specificity = 94.97%	- Specificity = 95.34%
- Precision = 91.00%	- Precision = 91.39%
- F1-Score = 86.64%	- F1-Score = 86.43%
- ROC AUC Score = 0.8883	- ROC AUC Score = 0.8866

Conclusion

- If the company focus on features influencing conversion rate, it would improve the conversion rate from the existing 30%.

- Top 3 variables increasing conversion rate are
  1. Lead Source_Welingak Website
  2. Tags_Will revert after reading the email
  3. Tags_Closed by Horizzon

- Model Evaluation parameters are almost same for train and test dataset, hence we can conclude that model is performing well and can be used .