Lead Score Case Study

Group Members
1. Preetam Kumar Singh
2. Pranali Desale
3. Atharv Joshi

Problem Statement :

- X Education sells online courses to industry professionals. This case study is done to analyze and find ways to get more industry professionals to join their courses.
- 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.
- The company needs a model to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- 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.

Business Objective:

- To help X Education select most promising leads, i.e. the leads that are most likely to convert into paying customers.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Methodology :

- **Reading and Understanding the Data :**
  - Importing and Observing the past data provided by the Company
- **Data Cleaning :**
  - Few columns have ‘Select’ value which means that the lead did not choose any option. So, we replaced it with Null values.
  - Removing duplicate data and other redundancies
  - We drop the columns having 45% or more Null values.
  - We removed some of the variables because they are redundant and do not provide any useful information.
- **EDA :**
  - We perform EDA on the data to check various categorical variables and the presence of outliers in numerical variables.
  - Univariate and Bivariate analysis

Methodology :

❑**Data Preparation :**

- We convert some binary variables to 0/1.
- We create Dummy Variables for categorical variables and dropped repeated variables.

❑**Model Building :**

- We used RFE to select the features. Then, we remove the variables depending on their p-value and VIF value.
- Variables having p-value<0.05 and VIF<5 were significant.

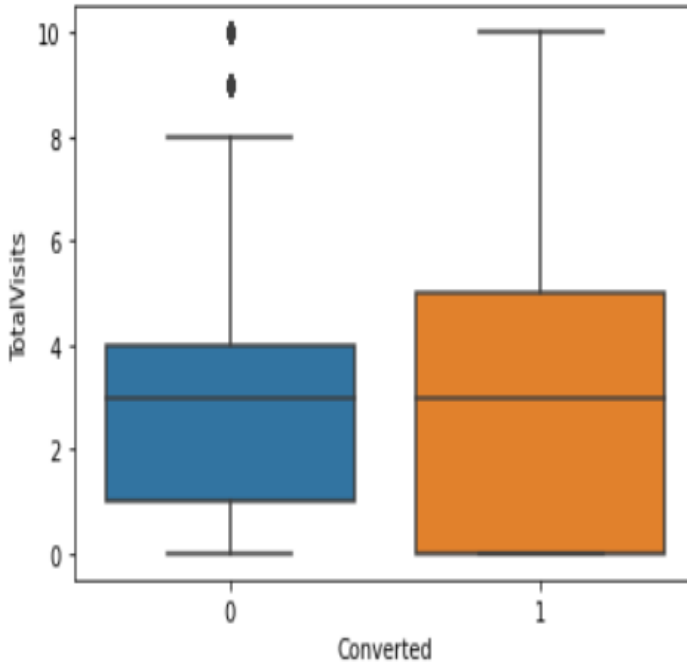
❑**Model Evaluation :**

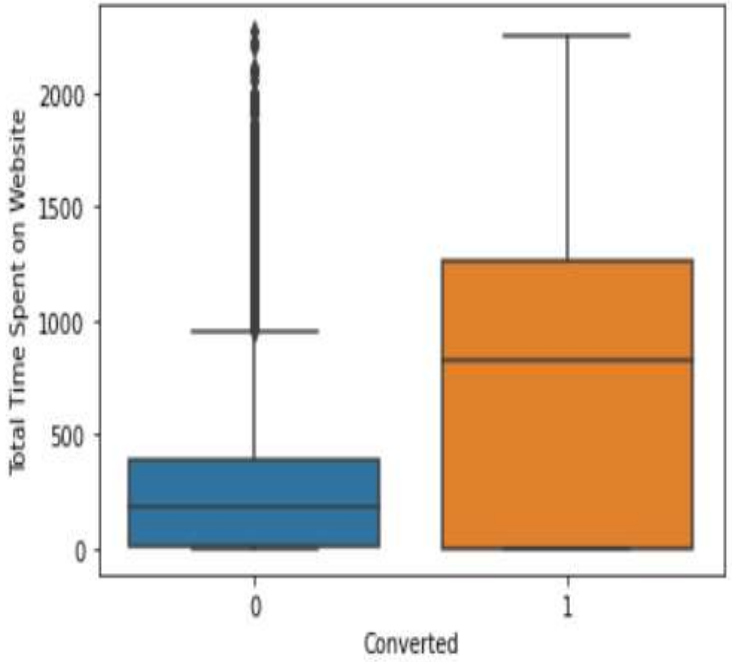
- Evaluating model based on various evaluation metrics
- Finding the optimal probability threshold

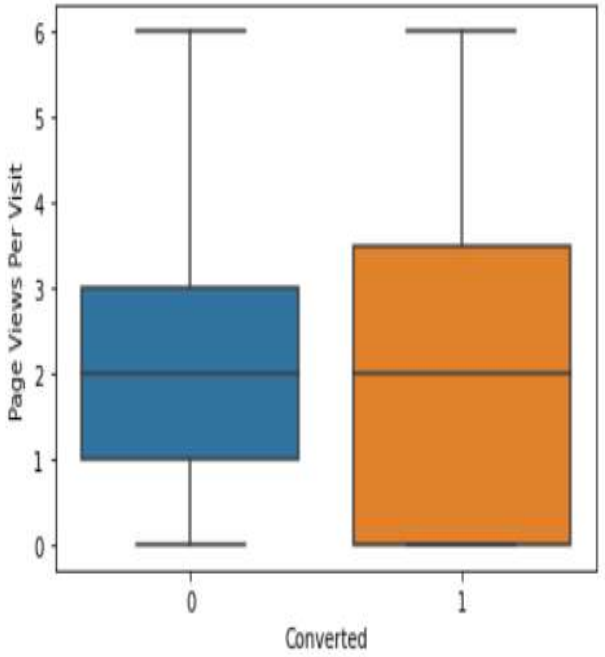
❑**Making Predictions on the Test Set:**

- Predictions on the Test set was done. Precision and Recall was calculated as 89% and 66% respectively.

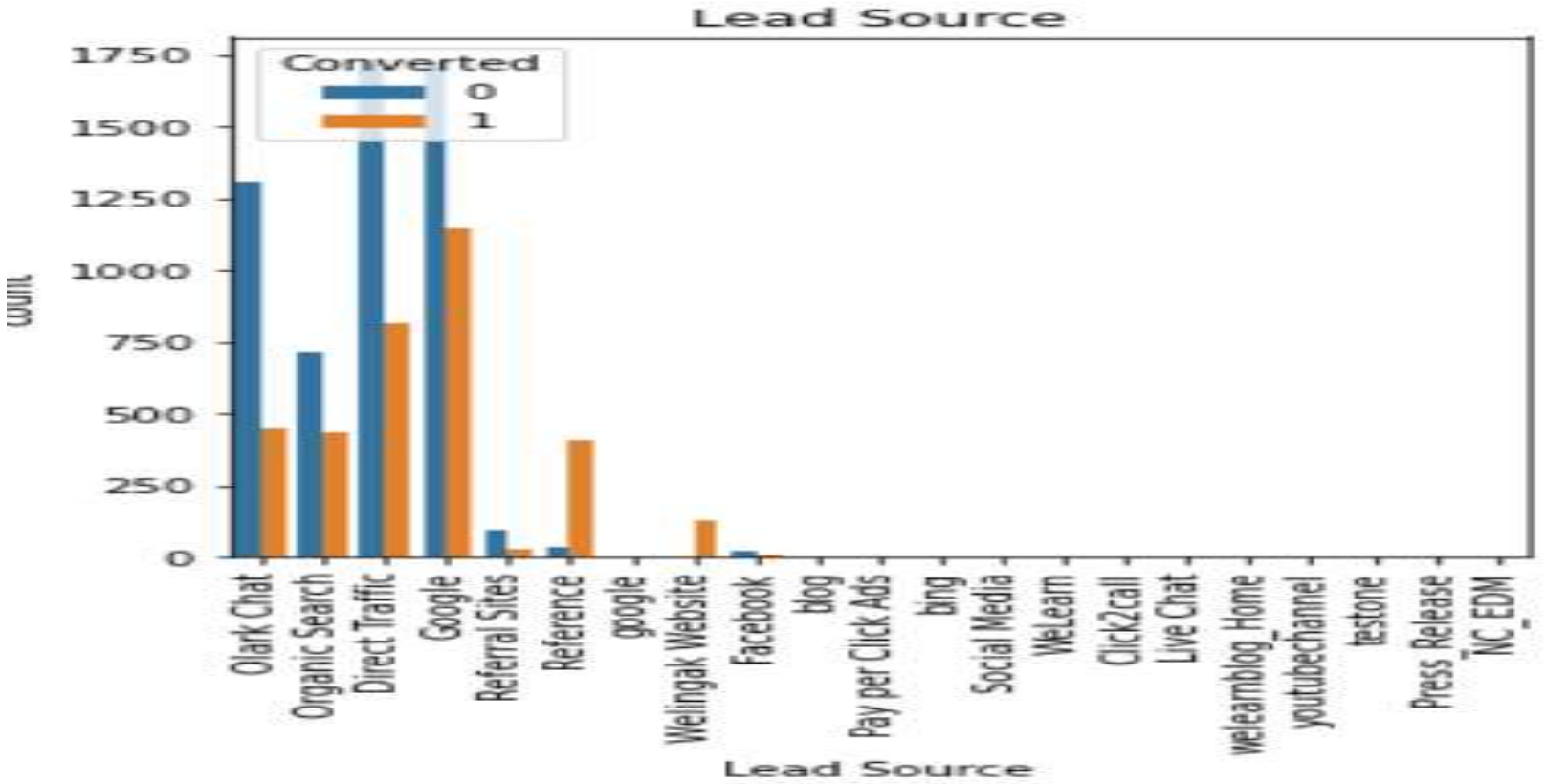
Data Visualization :





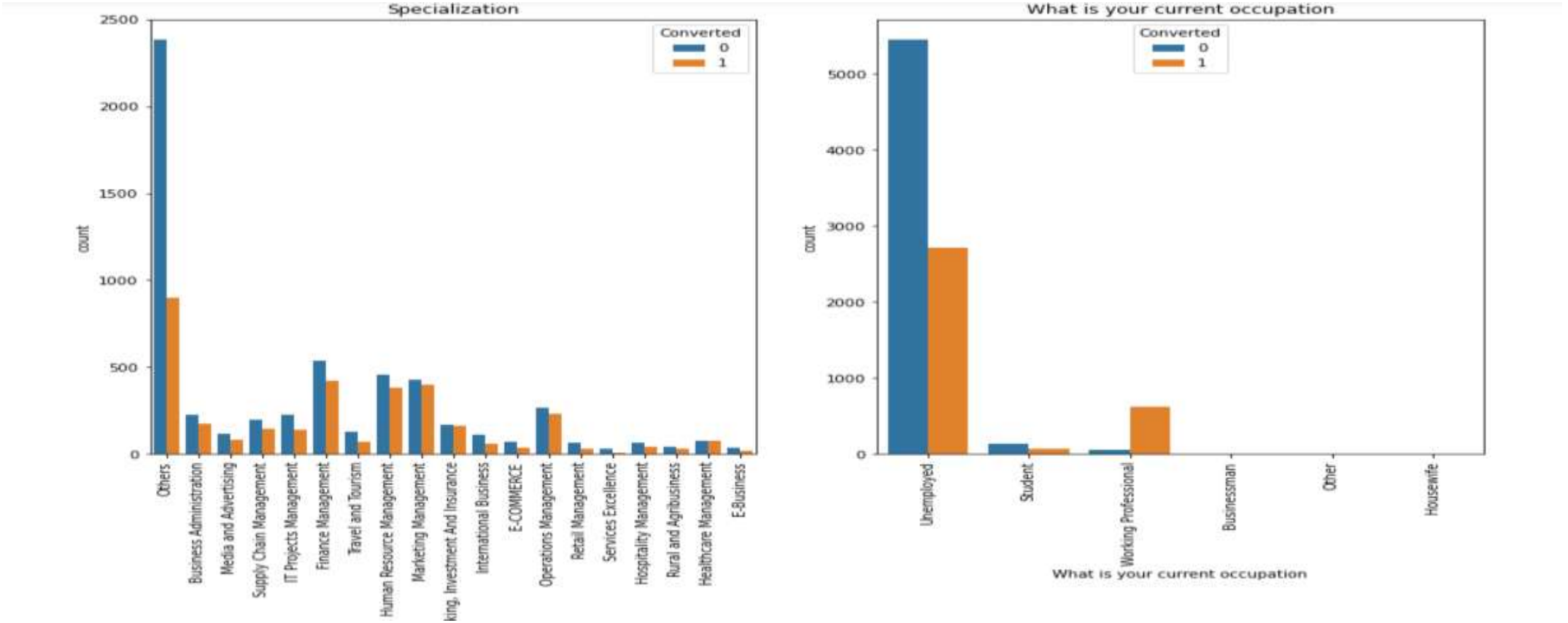


Lead Source :

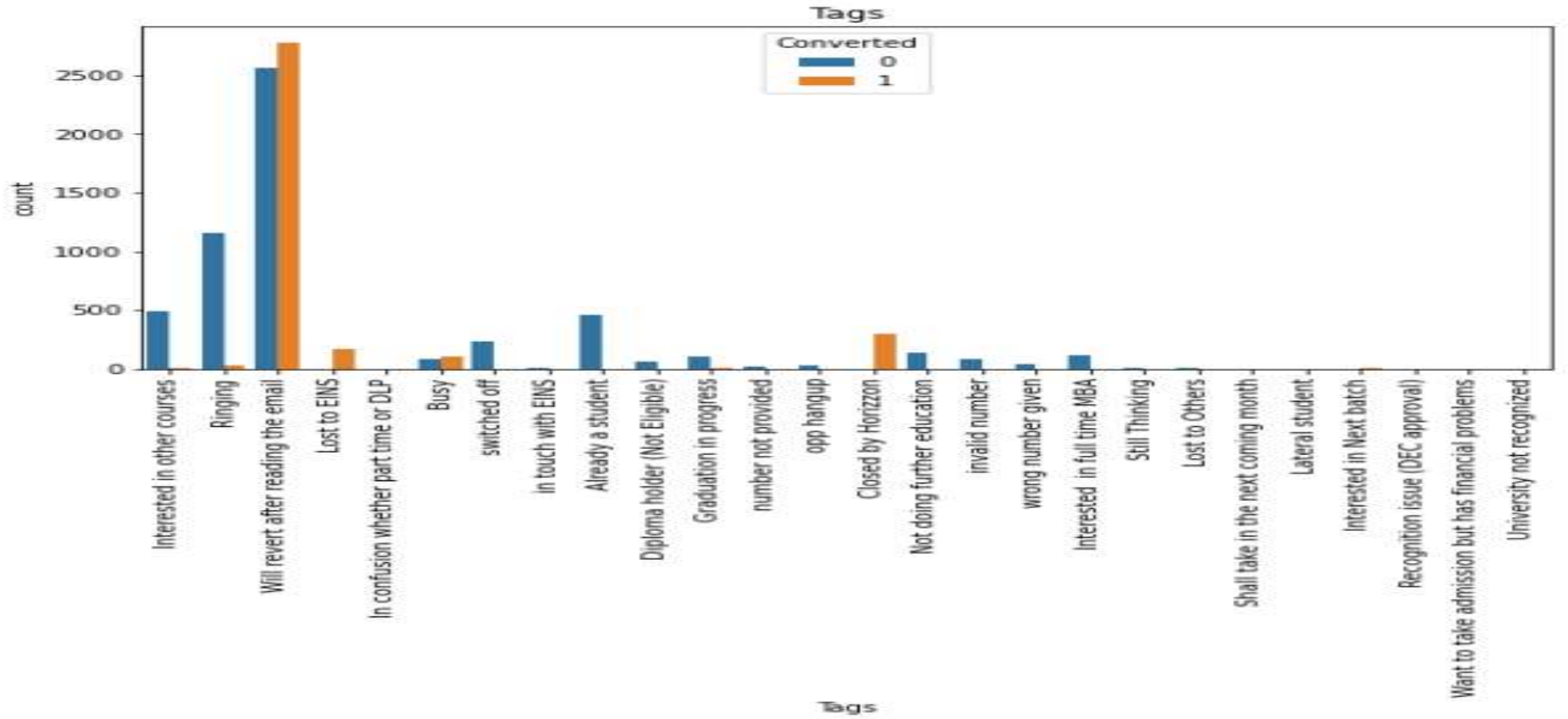


- Very high conversion rates for lead sources ‘Reference’ and ‘Welingak Website’.
- Most leads are generated through ‘Direct Traffic’ and ‘Google’.

Specialization and What is your current Occupation :

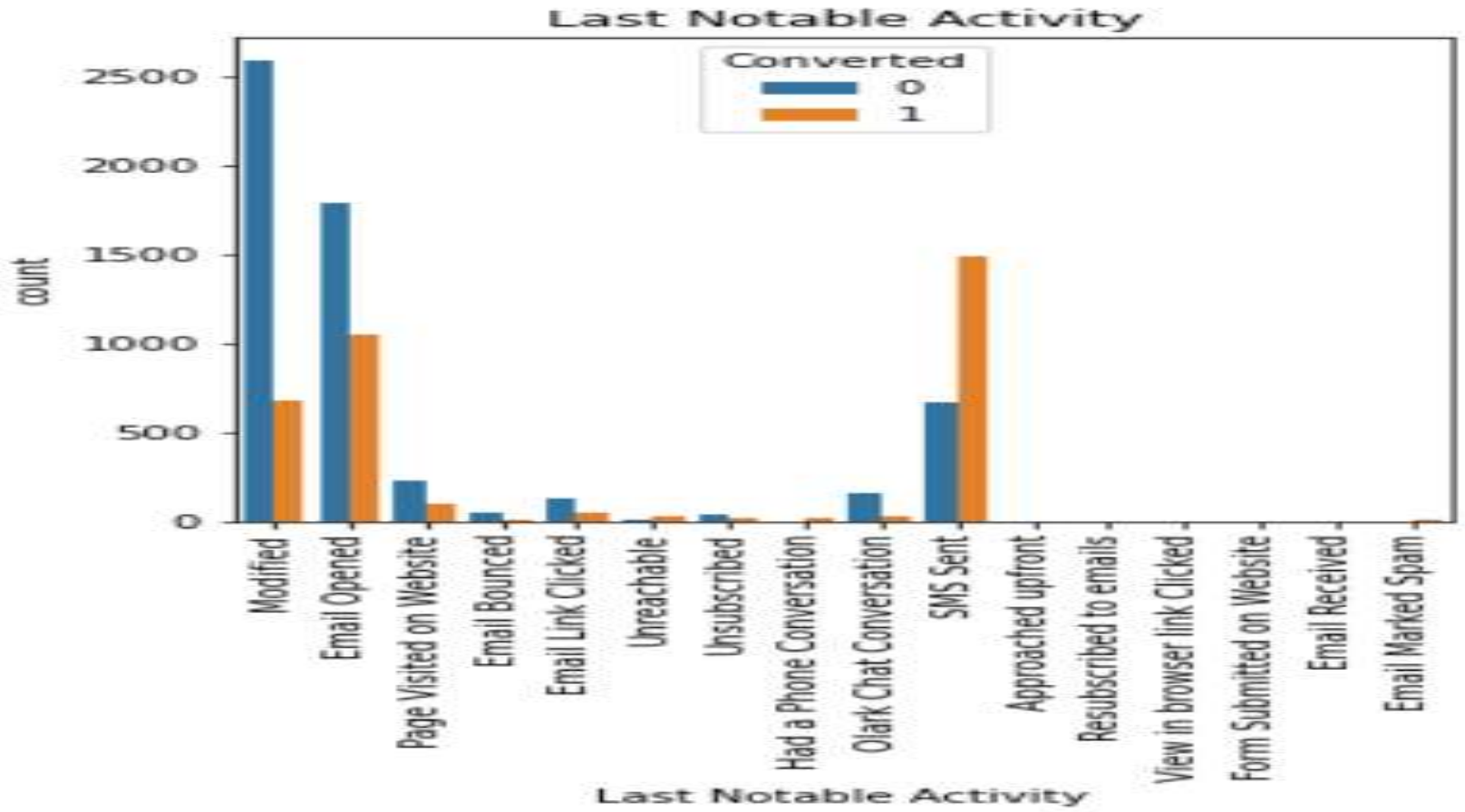


Tags :



High conversion rates for tags 'Will revert after reading the email', 'Closed by Horizon', 'Lost to EINS', and 'Busy'.

Last Notable Activity :



Highest conversion rate is for the last notable activity ‘SMS Sent’.

Model Evaluation :

Final Model Evaluation

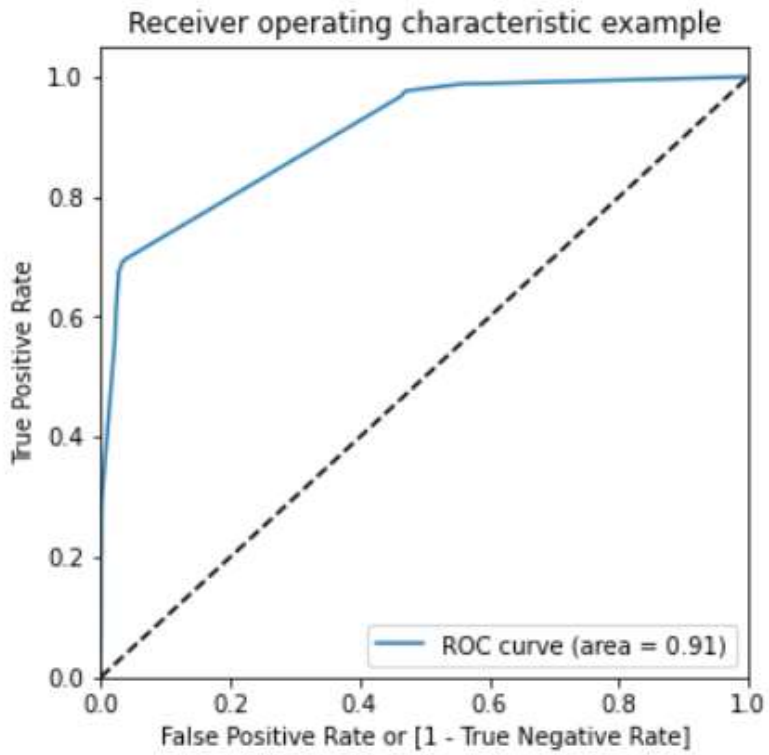
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6341
Model Family:	Binomial	**Df Model:**	9
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2146.3
Date:	Mon, 27 Feb 2023	**Deviance:**	4292.6
Time:	19:51:19	**Pearson chi2:**	1.03e+04
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.4817
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4992	0.209	-11.979	0.000	-2.908	-2.090
Lead Origin_Lead Add Form	2.6691	0.242	11.042	0.000	2.195	3.143
Last Activity_Had a Phone Conversation	3.0385	0.883	3.441	0.001	1.308	4.769
Last Activity_Unsubscribed	1.0723	0.505	2.122	0.034	0.082	2.063
What is your current occupation_Unemployed	-2.6575	0.179	-14.874	0.000	-3.008	-2.307
Tags_Busy	3.9204	0.271	14.491	0.000	3.390	4.451
Tags_Closed by Horizzon	9.1126	0.731	12.463	0.000	7.680	10.546
Tags_Lost to EINS	8.9879	0.733	12.267	0.000	7.552	10.424
Tags_Will revert after reading the email	4.3170	0.159	27.194	0.000	4.006	4.628
Last Notable Activity_SMS Sent	2.7264	0.105	25.867	0.000	2.520	2.933

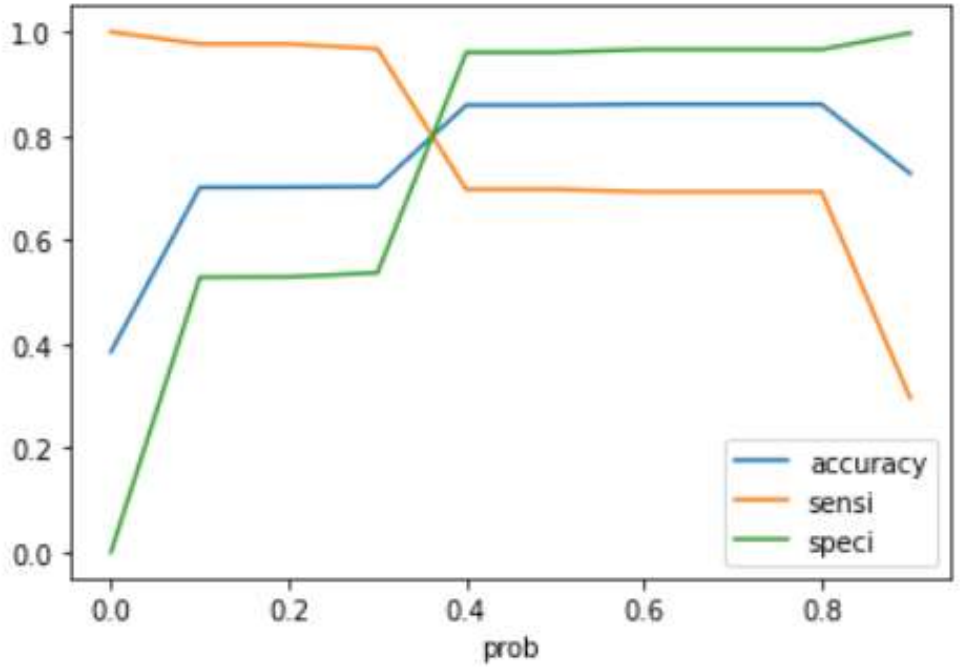
The p- values are significant for each other.

ROC Curve :



The area under ROC curve is 0.91.

Finding Optimal Cut-off point :



Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values Optimal cutoff = 0.35

Inference :

After running the model on the Test Set we have:

* Accuracy: 84.75%

* Sensitivity: 66%

* Specificity: 95.5%

Results :

Comparing the values obtain for the Train and test set :

Train Data:

* Accuracy: 85.9%
* Sensitivity: 69.7%
* Specificity: 96%

Test Data:

* Accuracy: 84.75%
* Sensitivity: 66%
* Specificity: 95.5%

Conclusion :

* The Accuracy, Sensitivity, Specificity of the Test set are very close to their respective values of the Train set.

* The Optimal cut-off is considered on the basis of Sensitivity and Specificity.

* We have got 714 Leads who have a high chance of getting converted.

* The overall Accuracy of the model is very good.