upGrad

X-Education : Lead Scoring Case Study

Detection of Hot Leads to concentrate more of marketing efforts on them, improving conversion rates for X Education

Team Members: Aman Soni , Sushil Kumar and Shivani Siddhu

Table of Contents

Background of X Education Company

- An education company named X Education sells online courses to industry professionals. Professionals who are interested in the courses land on their website and browse for courses.
- Courses are marketed 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 people fill up a form providing their email address or phone number, they are classified to be a lead. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%.

Problem Statement & Objective of Case Study

Problem Statement:

- X Education gets a lot of leads however lead conversion rate is very poor at around 30% .
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads.
- Sales team want to know the potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Objective of the Study:

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- Company requires us to build a model where we need 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 a lower lead score have a lower conversion chance.
- CEO has given a ballpark of the target lead conversion rate to be around 80%

Analysis Approach Used

- Reading & Understanding Data
- Data Quality Checks
- Exploratory Data Analysis
- Model Building
- Model Evaluation
- Making Predictions
- Observation

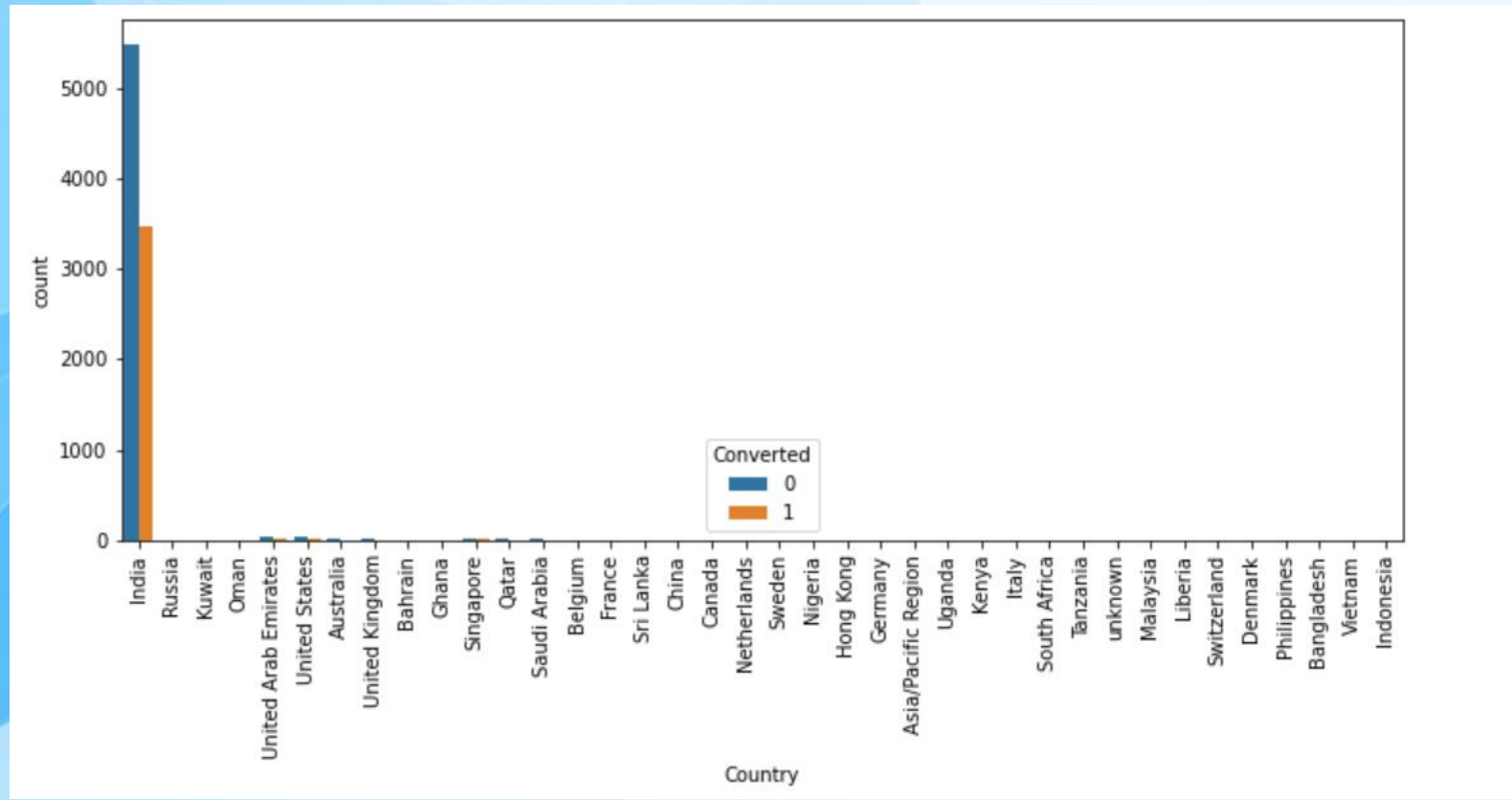
Data Cleaning

- Performed Data quality checks to find Duplicates & NULL value checks to impute/remove records.
- Column with more than 45% null value were dropped.
- Treated the missing values by imputing the favorable aggregate function like (Mean, Median, and Mode).
- Identified & Dropped highly skewed columns.
- Dropped columns that were of no significance.
- Detected Outliers & Dropped the top and bottom 1% datapoints to treat outliers.
- Low frequency values were grouped together to “Others”.
- Fixed Invalid values & Standardized Data in columns by checking casing styles (lead source has Google, google) etc.

Exploratory Data Analysis

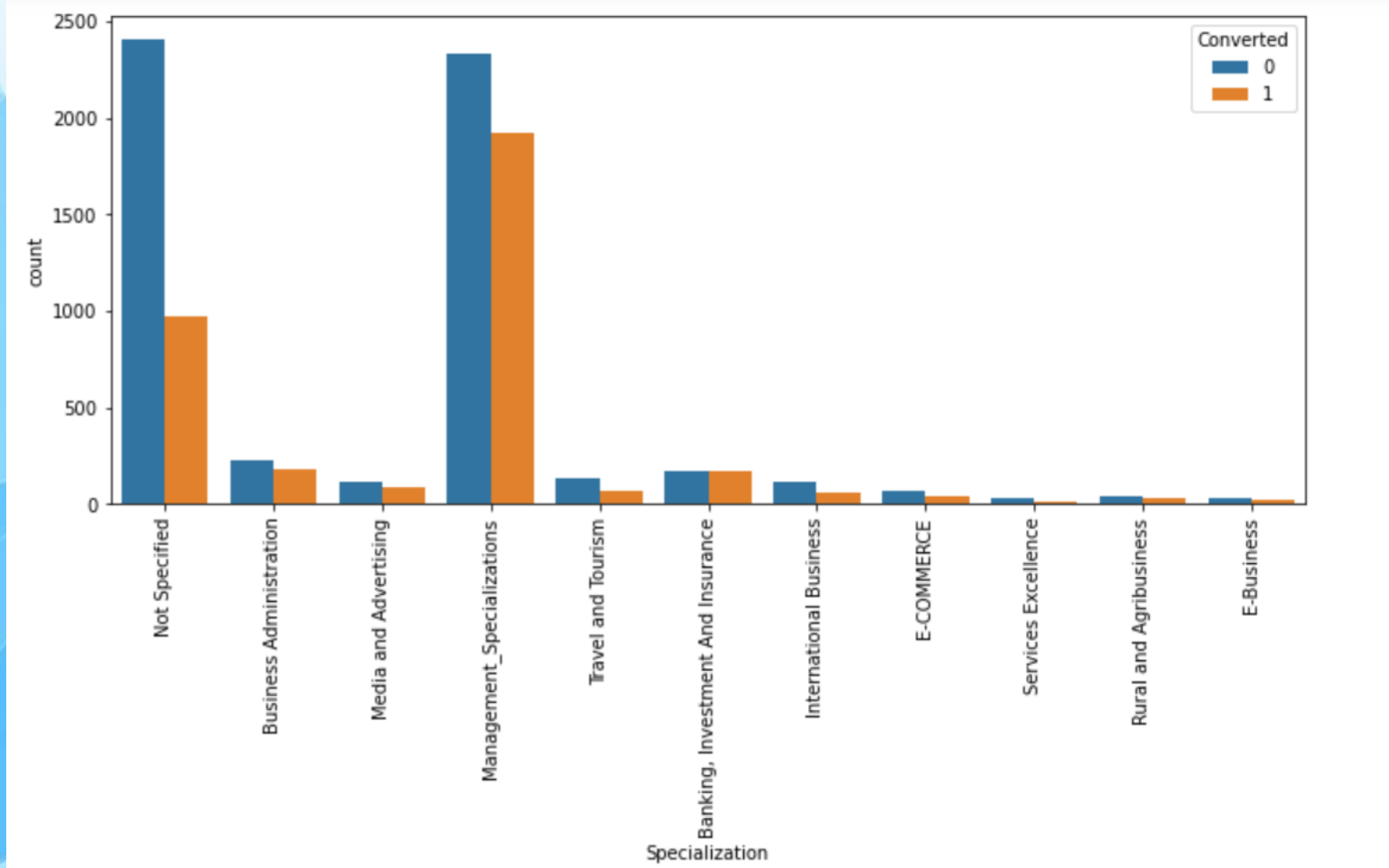
- **Country Column**

India is country from which most people are checking for the course.



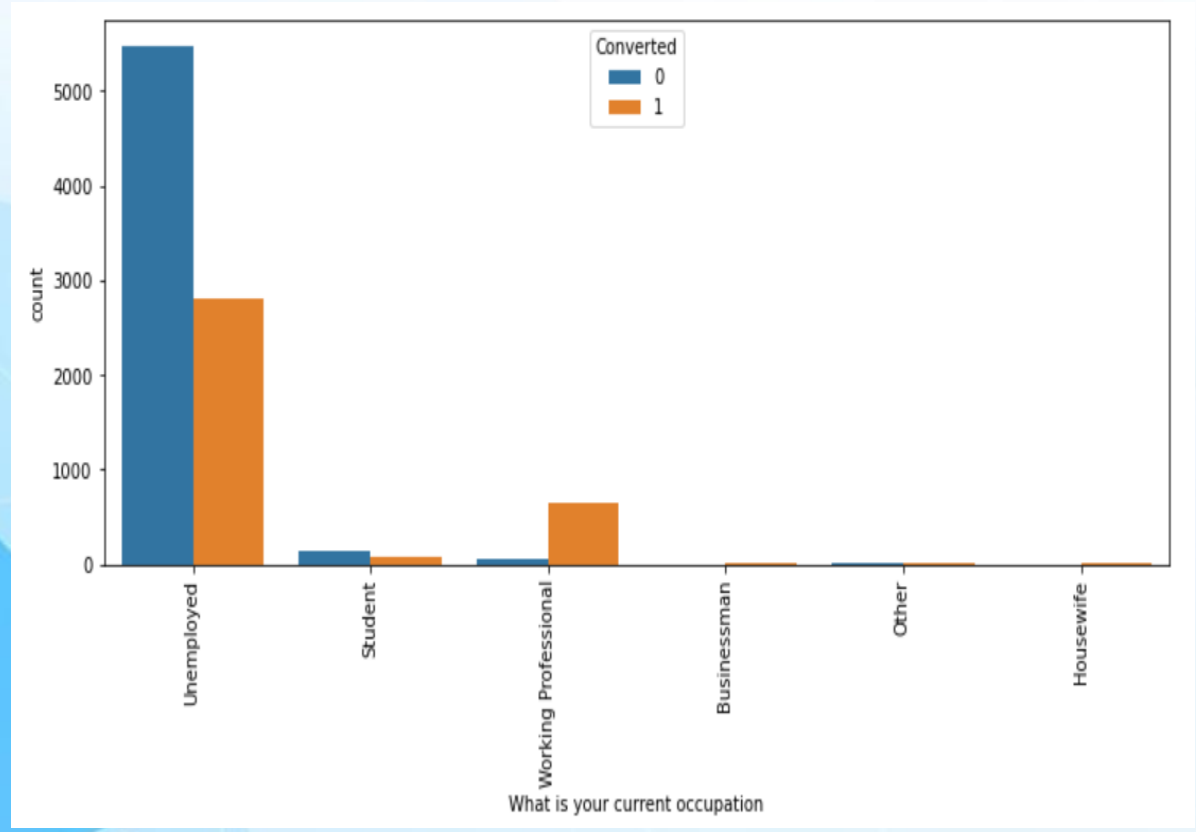
- **Management Specialization**

This is a very significant column. People specializing in management have more chance of converting into “HOT Leads”.



- **What is your current Occupation**

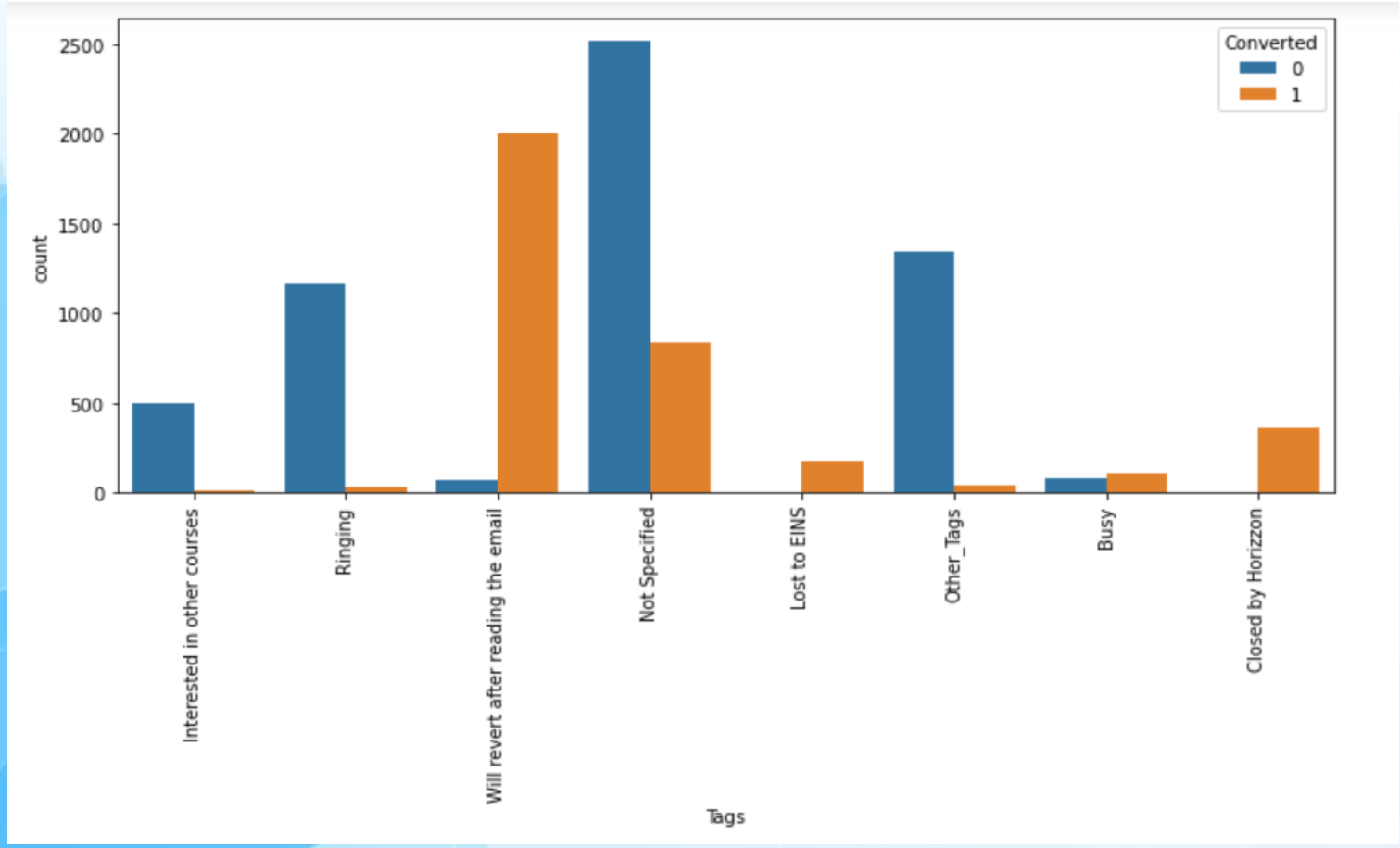
The data look skewed in favor of 'Unemployed' leads but we found a trend:- leads who are 'working professionals' are more likely to get converted.



- **Tags**

Leads tagged 'Will revert after reading mail' have highest chances of being converted followed by 'Lost To EINS', 'Closed By Horizon and 'Busy' and efforts should be made to generate more leads from these tags.

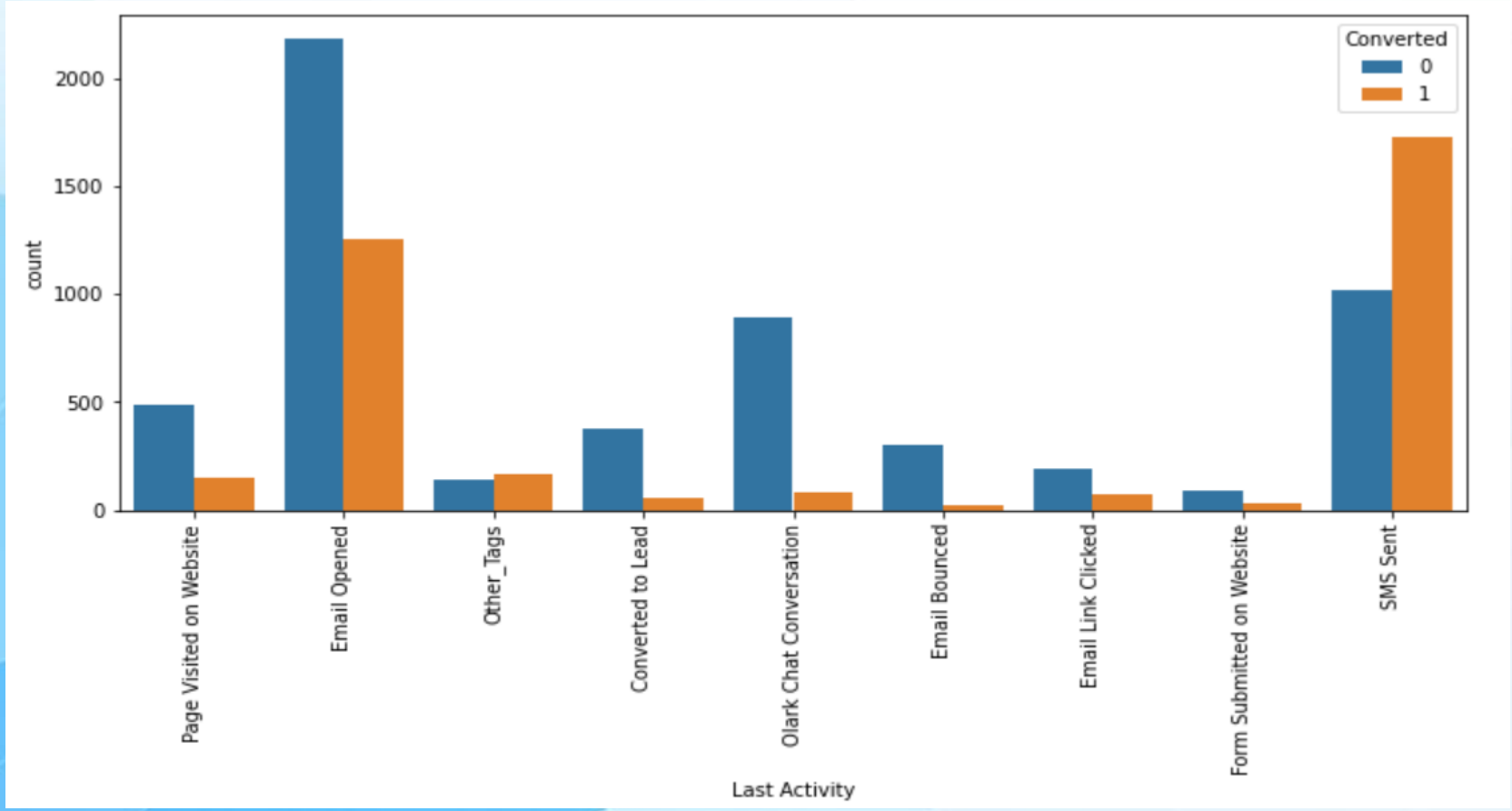
'Ringing' and 'Not Specified' tagged leads are more in number and hence efforts should be made to maximize conversion from these tags.



- **Last Activity**

Leads having last activity as 'SMS Sent' have the most conversion rate.

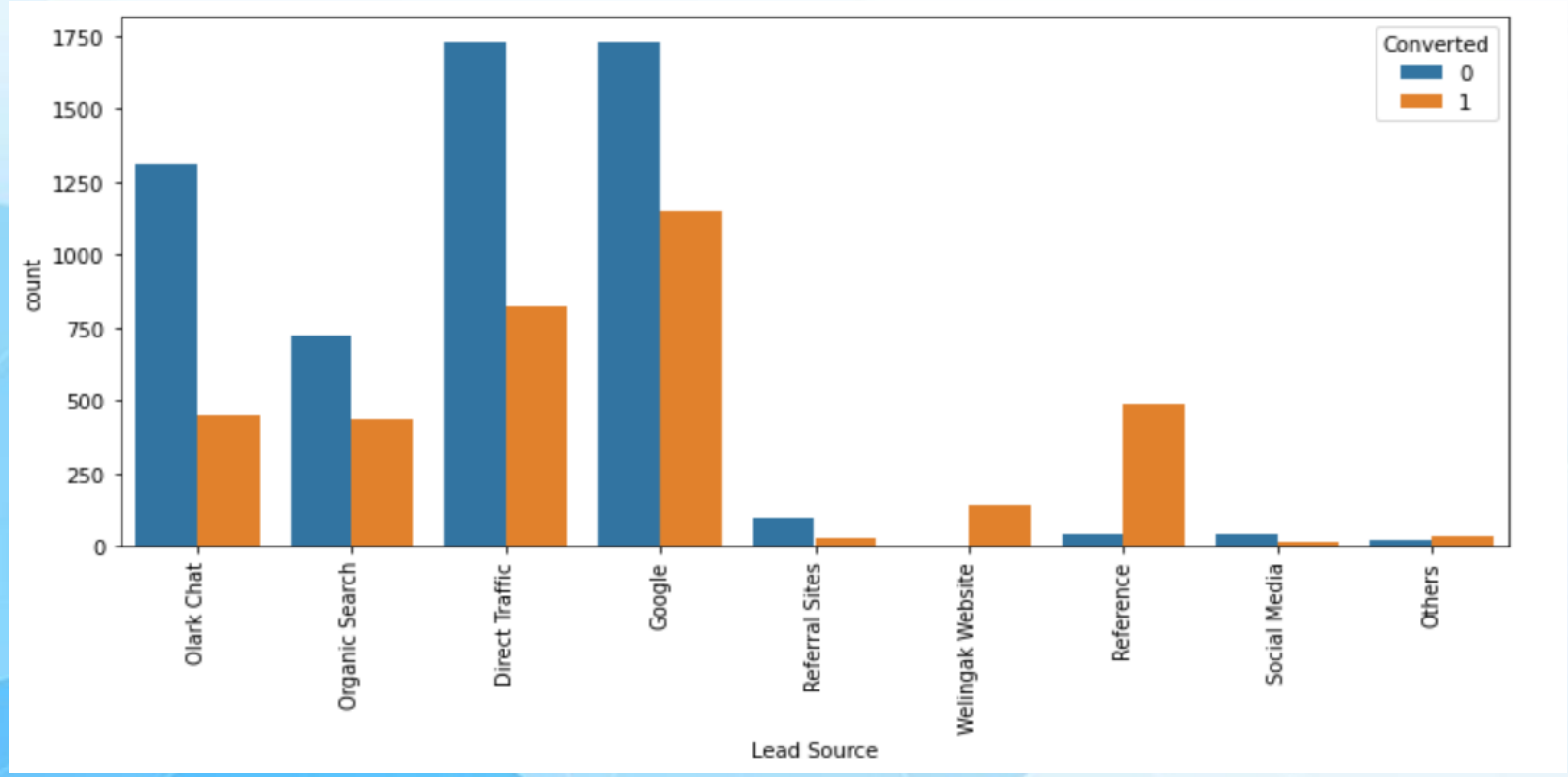
'Email Opened' brings maximum no. of leads and has second most conversion as well.



- **Lead Source**

'Google' and 'Direct Traffic' generate most number of leads and lead conversions while some like 'Welingak Website', 'Reference' and 'Others' are having maximum conversion of leads.

To improve overall lead conversion rate, focus should be on improving lead conversion of 'direct traffic' and 'google leads' and efforts should be put to generate more leads from 'reference' and 'welingak website' as they are conversion rate is very strong.



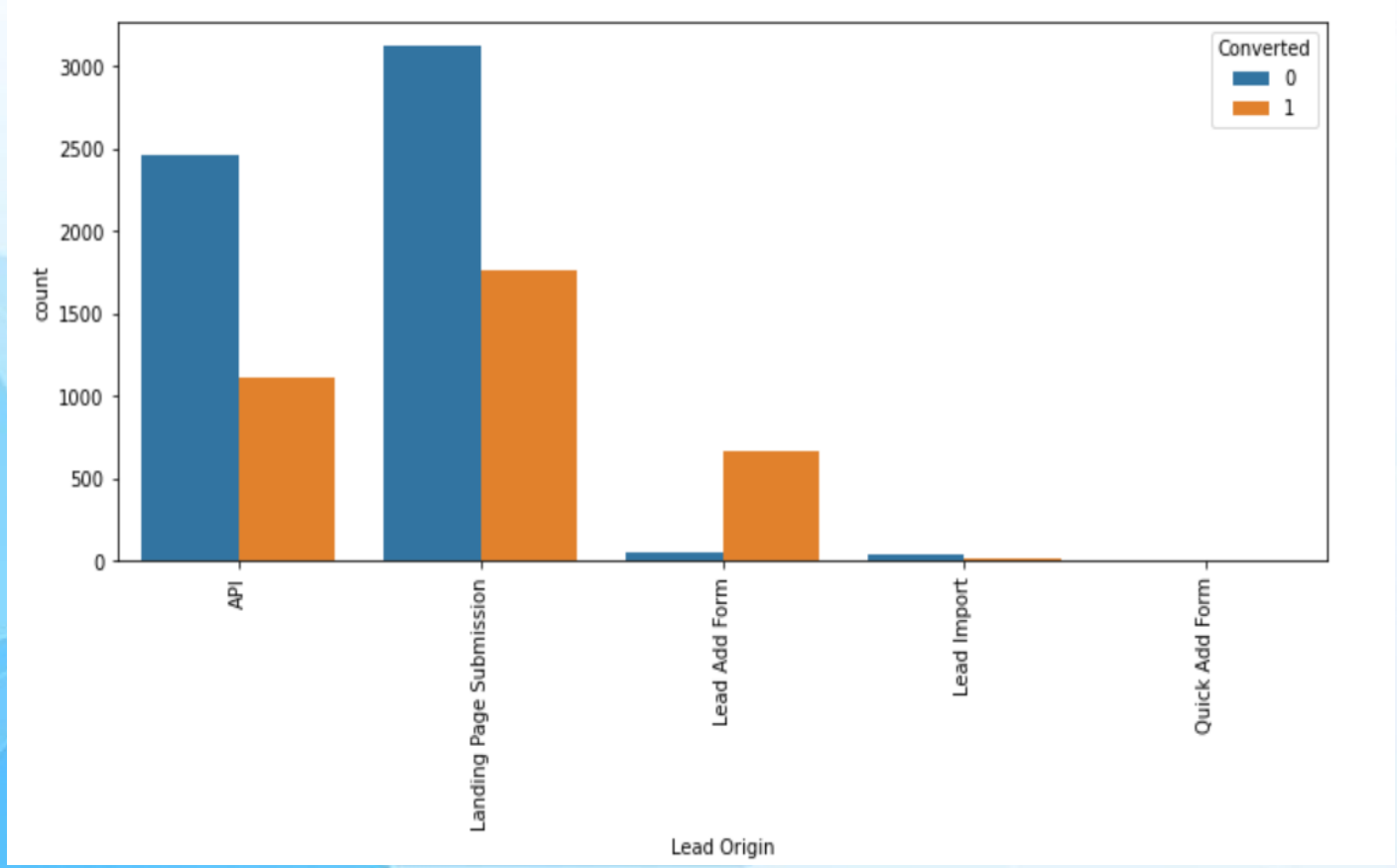
- **Lead Origin**

This is very significant column as we can see that 'Lead Add Form' is a very good origin of leads due to its very strong conversion rate.

Lead Import and Quick Add Form get very few leads.

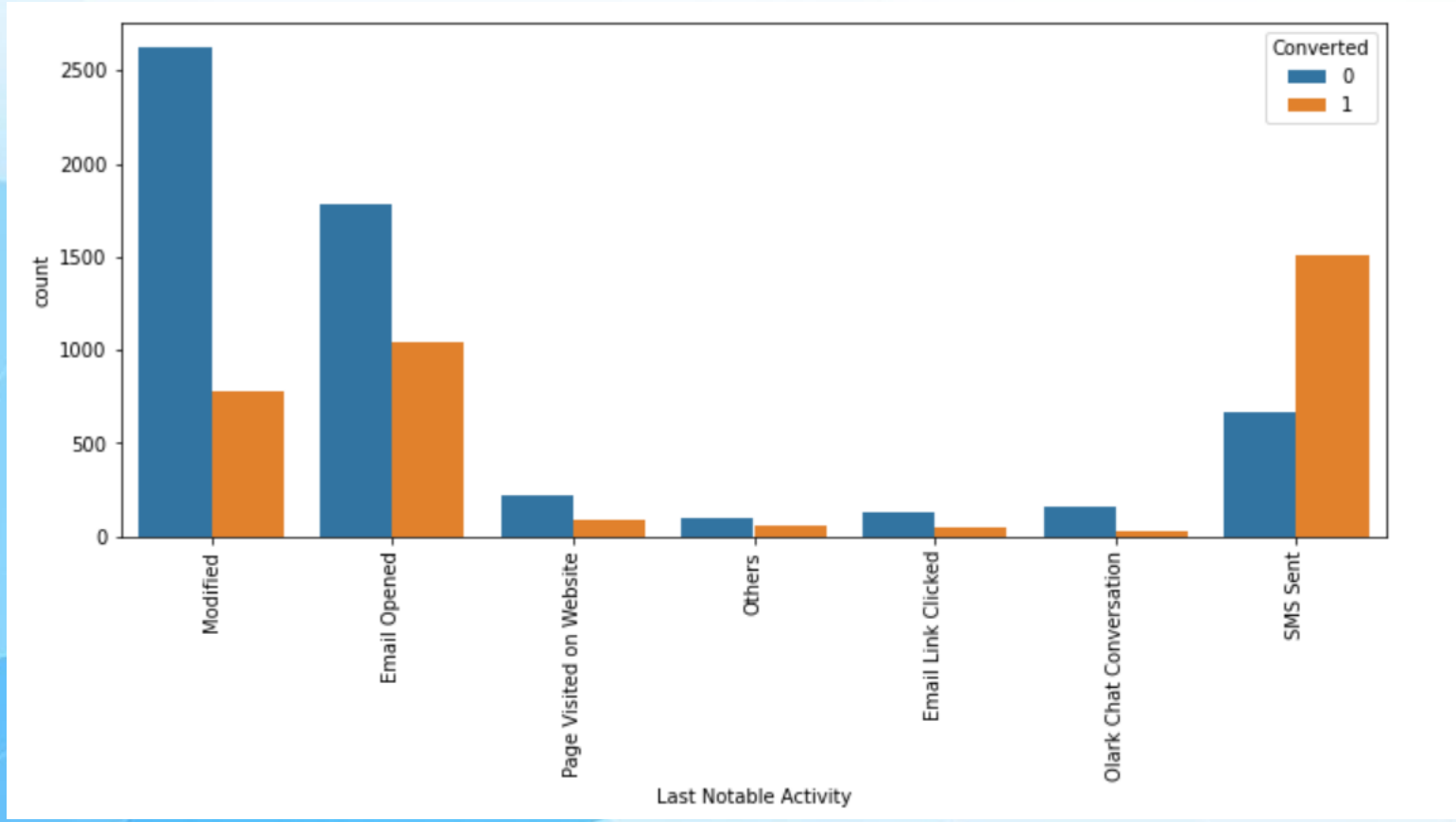
'Landing Page Submissions' and 'API' bring a higher amount of leads and see more lead conversion as well.

To improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



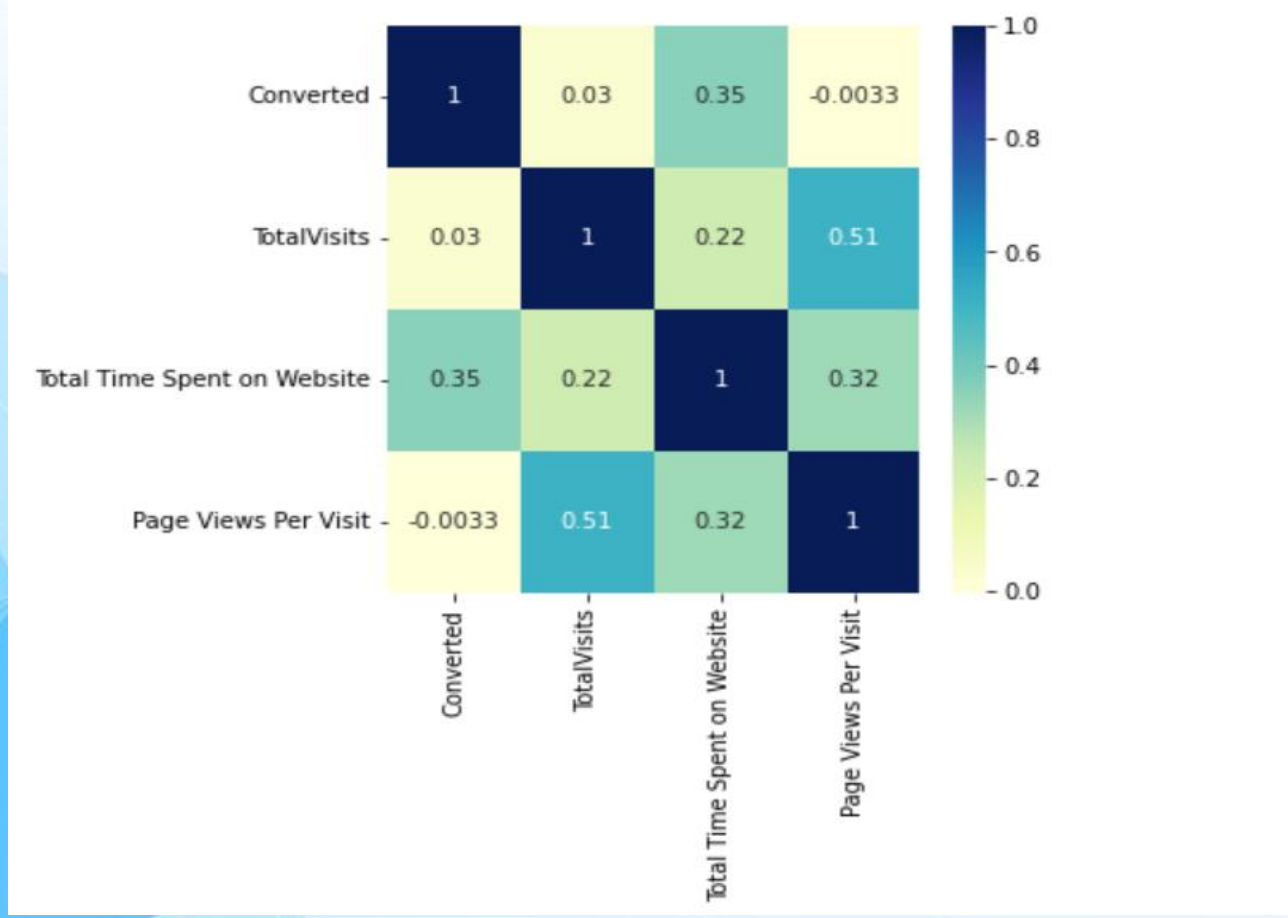
- **Last Notable Activity**

leads having last notable activity as 'SMS Sent' have a very high conversion rate.

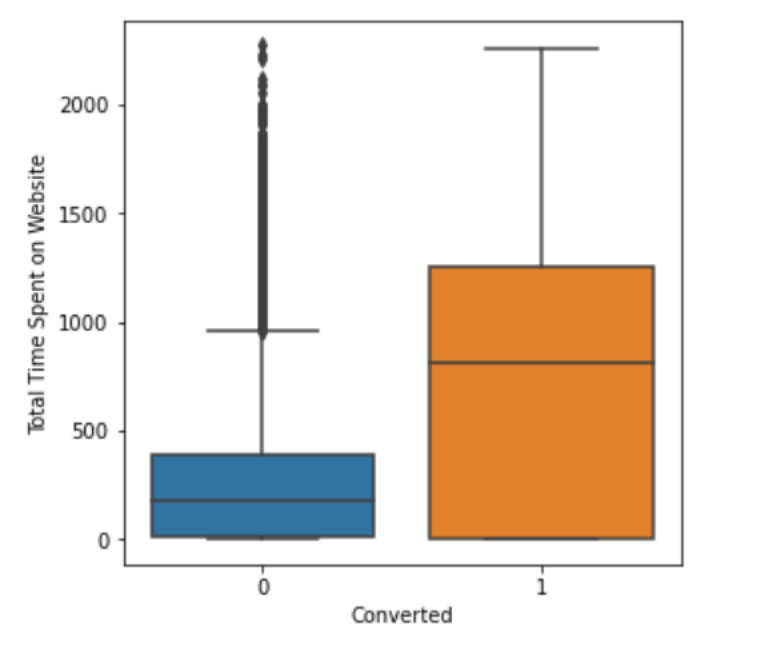


- **Correlation Matrix**

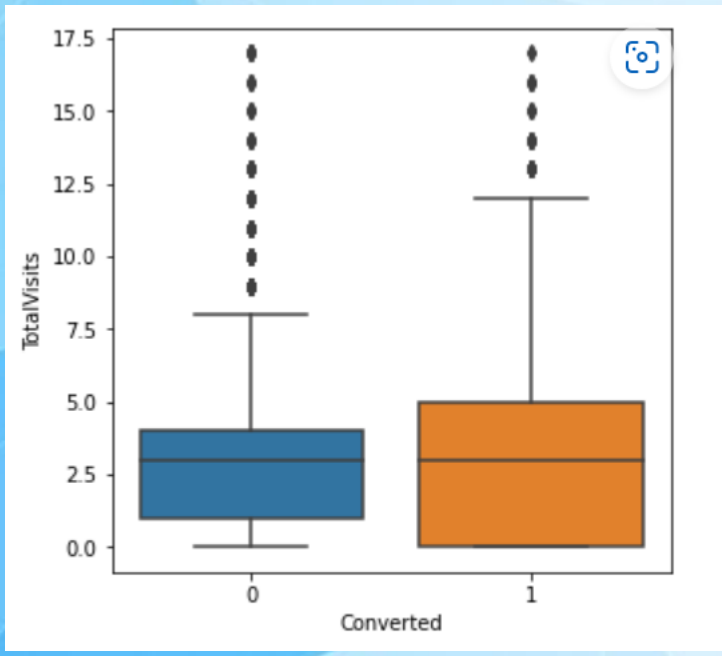
'Total Visits' and 'Page Views Per Visit' have the most correlation with each other



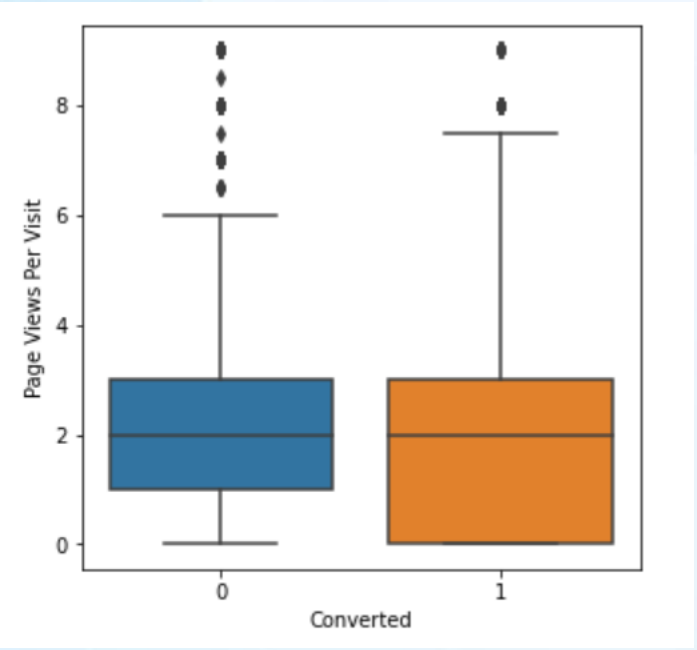
- Converted vs Total Time Spent
- Converted vs Total Visits
- Converted vs Page Views Per Visit



Converted vs Total Time Spent



Converted vs Total Visits



Converted vs Page Views Per Visit

Data Preparation

- Binary level categorical columns were already mapped to 1 / 0 in previous step.
- Created dummy features (one-hot encoded) for categorical variables : Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation .
- Splitting Train & Test Sets : 70:30 % ratio was chosen for the split.
- Feature scaling : Standardization method was used to scale the features
- Checking the correlations : Predictor variables which were highly correlated with each other were dropped.

Model Building

Below steps were used for Model Building :-

- We build the model manually starting with the 15 best variables as selected by RFE.
- Checked which model is a better fit by dropping variables that are not required depending upon their VIF values/p-values.
- Drop variables either having VIF > 5 or having p-values > 0.05.
- VIF parameter(i.e. multicollinearity) must be < 5 while p-values that will determine significance of the variables must be < 0.05.
- Used the stats models to view a detailed summary of different parameters and make decisions.
- Keep dropping/adding variables to find the best model and then test it against the test dataset.

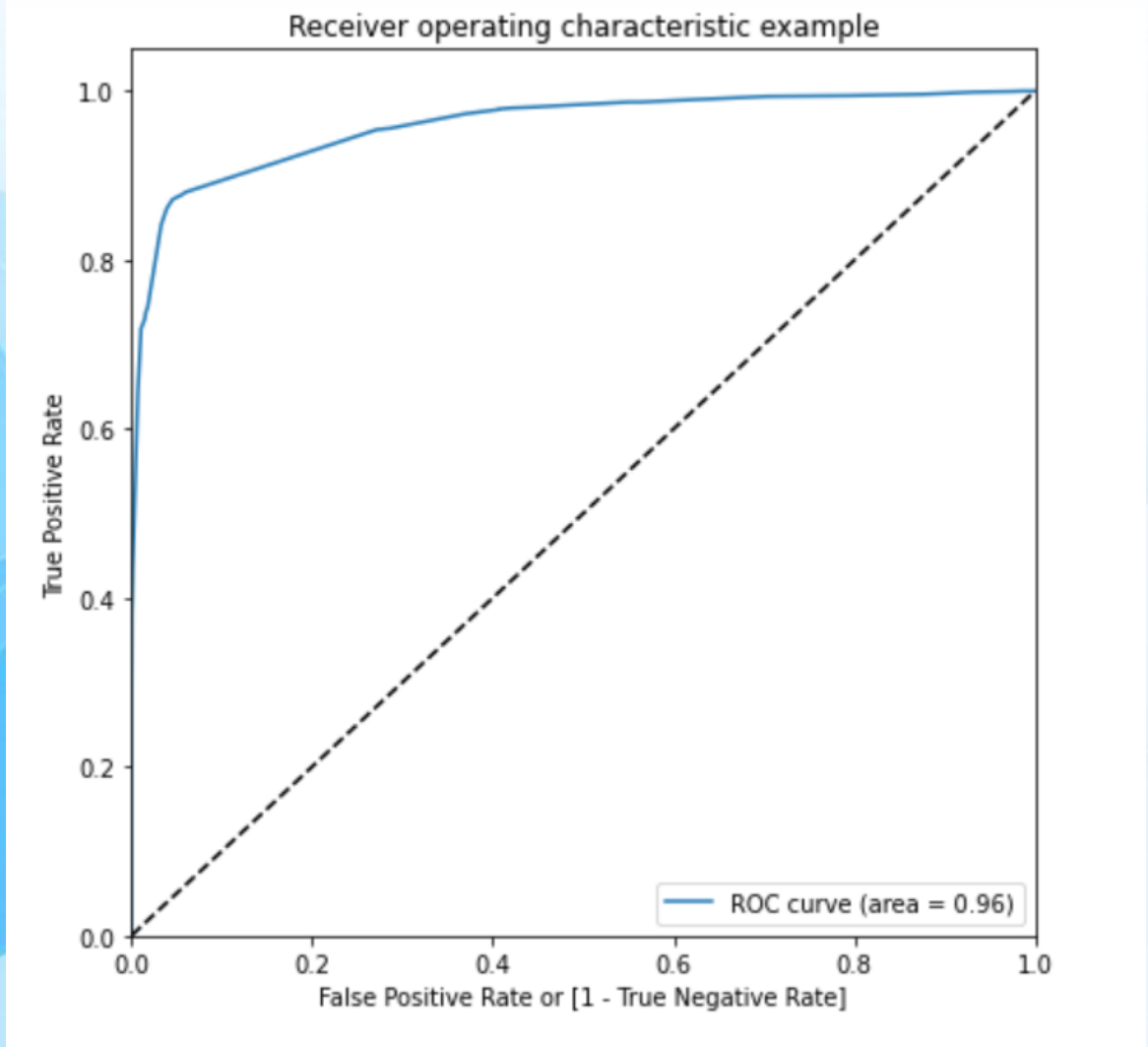
- Model 5 looks stable after four iteration with: significant p-values within the threshold (p-values < 0.05) and No sign of multicollinearity with VIFs less than 5
- Hence, res5 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6267
Model:	GLM	**Df Residuals:**	6255
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1418.6
Date:	Tue, 21 Mar 2023	**Deviance:**	2837.1
Time:	14:05:41	**Pearson chi2:**	1.01e+04
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.5835
Covariance Type:	nonrobust		

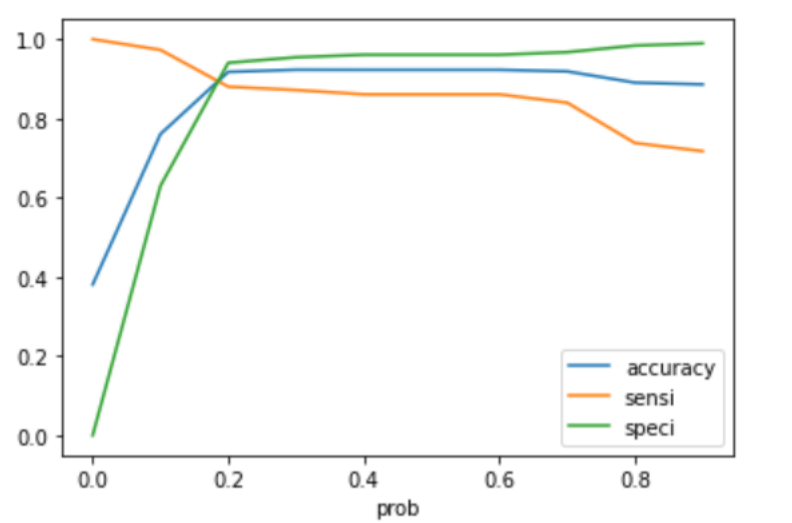
	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.7108	0.179	-20.683	0.000	-4.062	-3.359
Lead Source_Direct Traffic	-0.6530	0.120	-5.453	0.000	-0.888	-0.418
Lead Source_Welingak Website	5.0258	1.019	4.930	0.000	3.028	7.024
Last Activity_Email Bounced	-1.4331	0.420	-3.409	0.001	-2.257	-0.609
Last Activity_Olark Chat Conversation	-1.7220	0.224	-7.687	0.000	-2.161	-1.283
Tags_Busy	2.9903	0.267	11.200	0.000	2.467	3.514
Tags_Closed by Horizzon	9.2285	1.020	9.051	0.000	7.230	11.227
Tags_Lost to EINS	7.8543	0.619	12.696	0.000	6.642	9.067
Tags_Not Specified	2.2801	0.179	12.743	0.000	1.929	2.631
Tags_Ringing	-1.0360	0.272	-3.805	0.000	-1.570	-0.502
Tags_Will revert after reading the email	6.9142	0.236	29.265	0.000	6.451	7.377
Last Notable Activity_SMS Sent	2.4994	0.118	21.140	0.000	2.268	2.731

- ROC Curve : We have got a value of 0.96. ROC curve should be a value closer to 1 for a good model. so our value is good



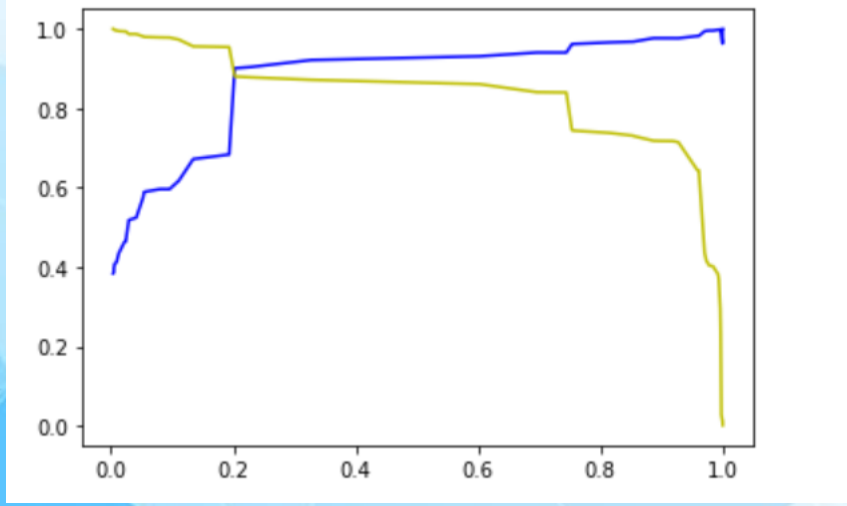
Model Evaluation – Train Data

Confusion Matrix & Evaluation matrix with 0.5 as cut off



True Positive	2053
True Negative	3730
False Positive	152
False Negative	332
Confusion Matrix	[[3730 152] [332 2053]]
Specificity	0.9608449252962391
Sensitivity	0.8607966457023061
Positive Predicted Value	0.9310657596371882
Negative Predicted Value	0.9182668636139832
False Positive Rate	0.03915507470376095

Confusion Matrix & Evaluation matrix with 0.2 as cut off



True Positive	2099
True Negative	3651
False Positive	231
False Negative	286
Confusion Matrix	[[3651, 231], [286, 2099]]
Specificity	0.9404945904173106
Sensitivity	0.880083857442348
Positive Predicted Value	0.9008583690987124
Negative Predicted Value	0.9273558547117094
False Positive Rate	0.059505409582689336

Predictions On Test Dataset by Model Built From Training Dataset

True Positive	893
True Negative	1582
False Positive	94
False Negative	117
Confusion Matrix	[[1582, 94], [117, 893]]
Specificity	0.9439140811455847
Sensitivity	0.8841584158415842
Precision	0.9047619047619048
Recall	0.8841584158415842

Train & Test Dataset Model Conclusion

Conclusion On Training Dataset

- The model seems to be performing very well. ROC curve has a value of 0.96 which is good.

- Statistics of the model:

```
. Accuracy:    91.75%
. Sensitivity: 88%
. Specificity: 94.04%
```

Conclusion On Test Dataset

After running the trained model on the test dataset, we got below values :-

```
. Accuracy:    92.14%
. Sensitivity: 88.40%
. Specificity: 94.39%
```

Comparing the values obtained by our Train and Test dataset:

1. Train Dataset
   - Accuracy: 91.75%
   - Sensitivity: 88%
   - Specificity: 94.04%

2. Test Dataset
   - Accuracy: 92.14%
   - Sensitivity: 88.40%
   - Specificity: 94.39%

Recommendation

Below are the factors that are responsible for HOT Leads :-

- The total time spend on the Website
- Total number of visits.
- When the lead source was: - Olark Chat
- When the last activity was: - SMS - Olark chat conversation
- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Optimize communication channels based on lead engagement impact.
- Engage working professionals with tailored messaging.
- More budget/spend can be done on **Welingak Website** in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

Thank You