Contents

- Problem Statement
- Business Objective
- Our Approach
- EDA
- Categorical Variable vs. Target Variable
- Model Building
- ROC Curve
- Sensitivity, Accuracy and Specificity
- Conclusion

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- They acquire 100 leads in a day, only about 30 of them are converted. The typical lead conversion rate at X education is around 30%.
- 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.

Business Objective

- X Education wants to know most promising Leads.

- For that they want to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- Deployment of the model for the future use.

- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Our Approach

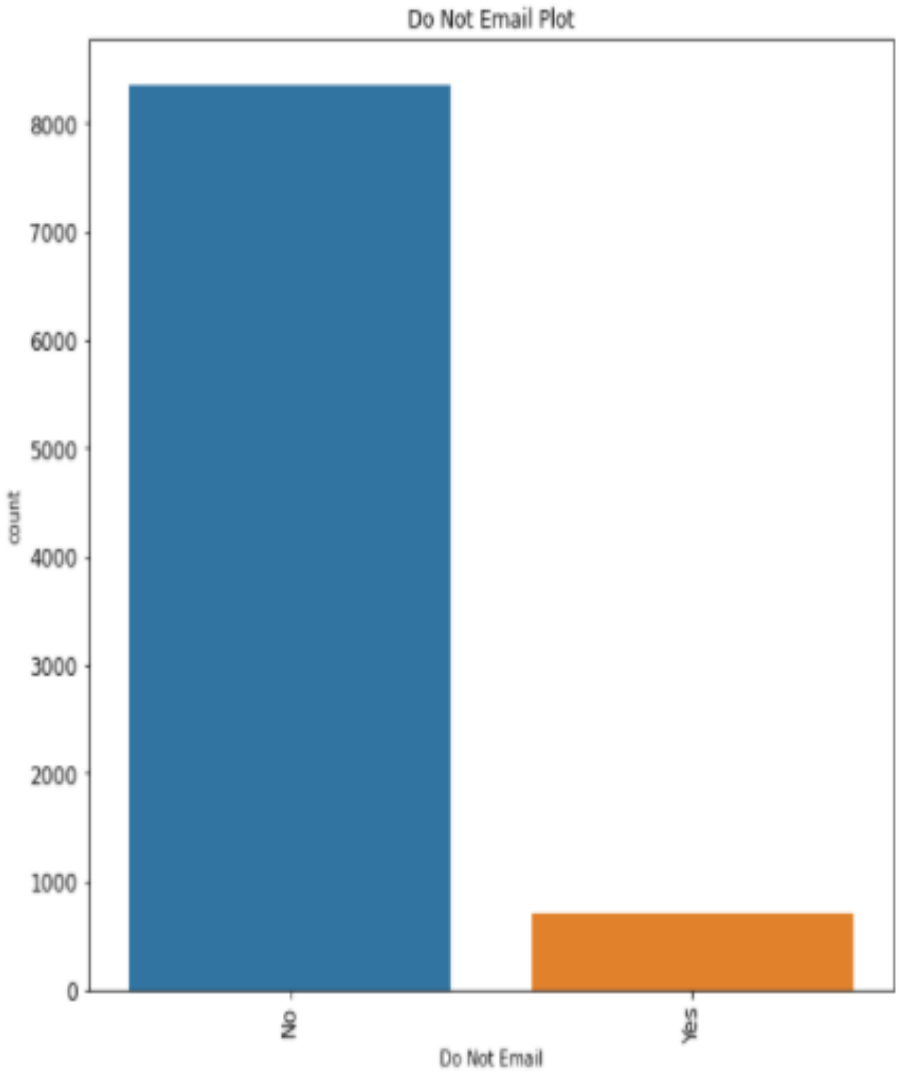
Data Sourcing, Cleaning and Preparation

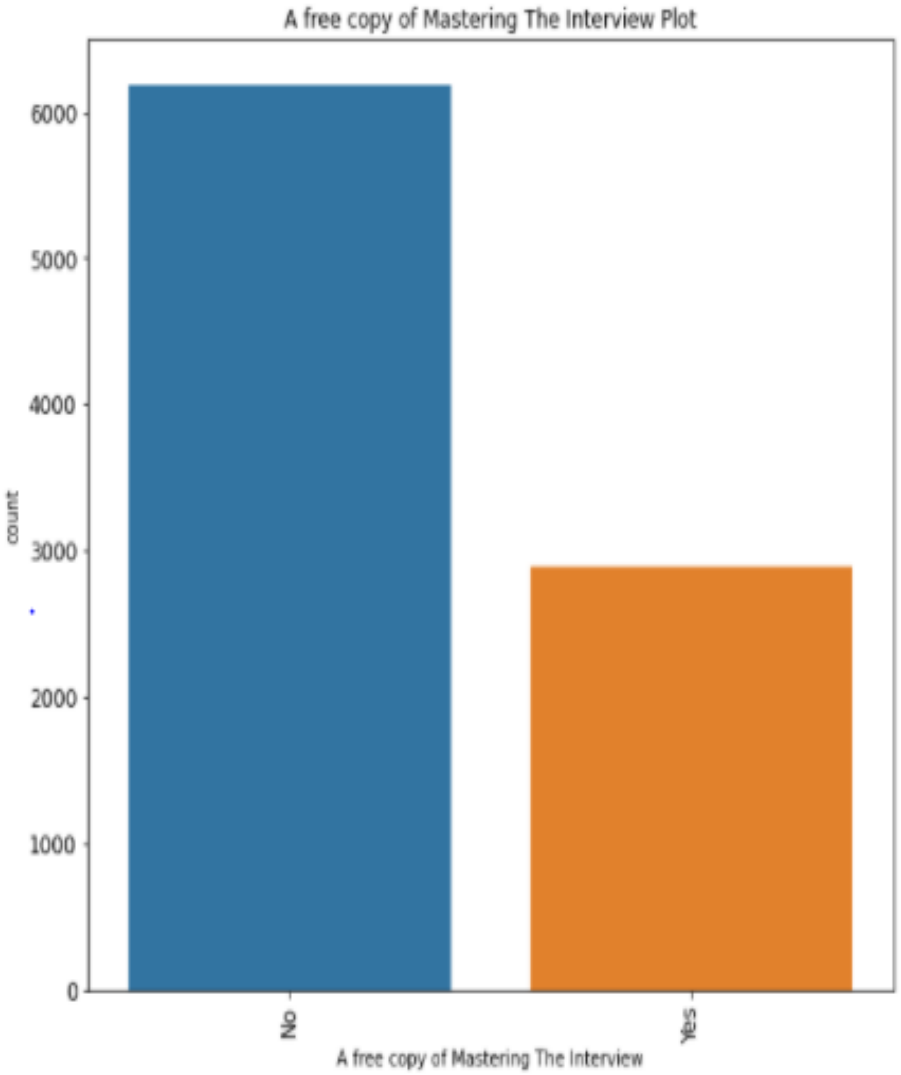
- Read the data from the source in a Pandas Data Frame and convert the data into clean format suitable for analysis.
- Removing the duplicate data if present.
- Identifying the categorical and numerical variables for further analysis.
- Drop columns if it contains large amount of missing values which is not useful for our analysis.
- Check and handle outliers in the if present.
- Imputation of values if necessary.

Our Approach cont.

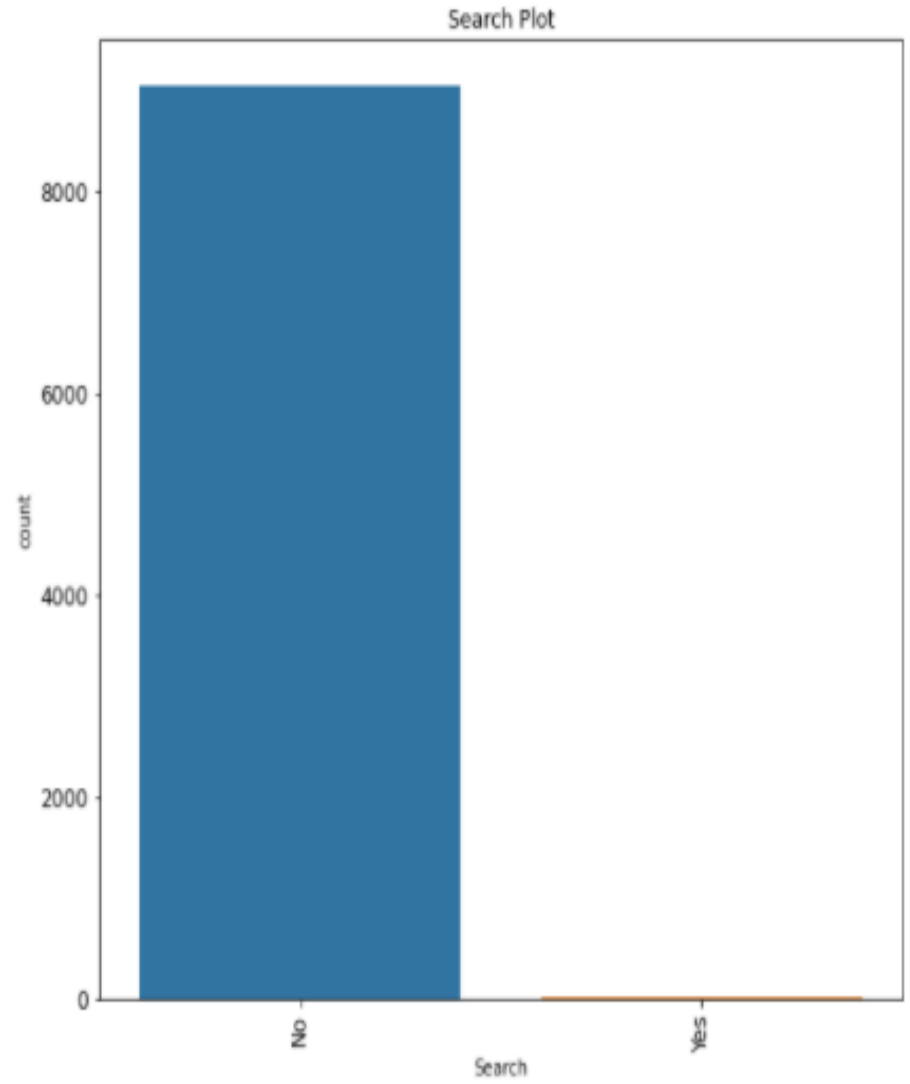
- We dropped the columns having null values greater than 40%.
- Few columns have values as ‘select’ which means the user hasn’t given any selection so, we converted those values to nan for ease in our Analysis.
- There were few columns not answered by the customers and they were having values as nan. Instead of dropping them we categorized them as ‘Not Answered’.
- This is done in order to prevent us from losing too many data, which might have impacted our analysis.

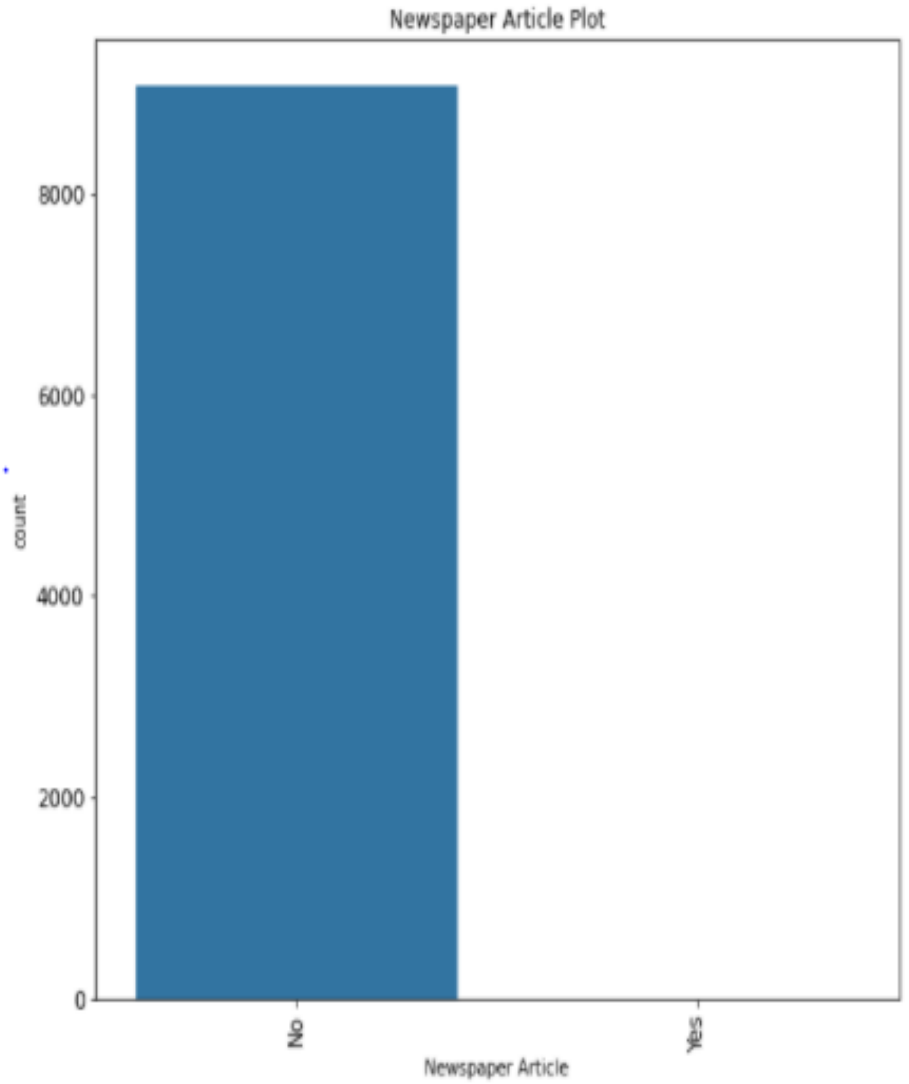
Exploratory Data Analysis (EDA)



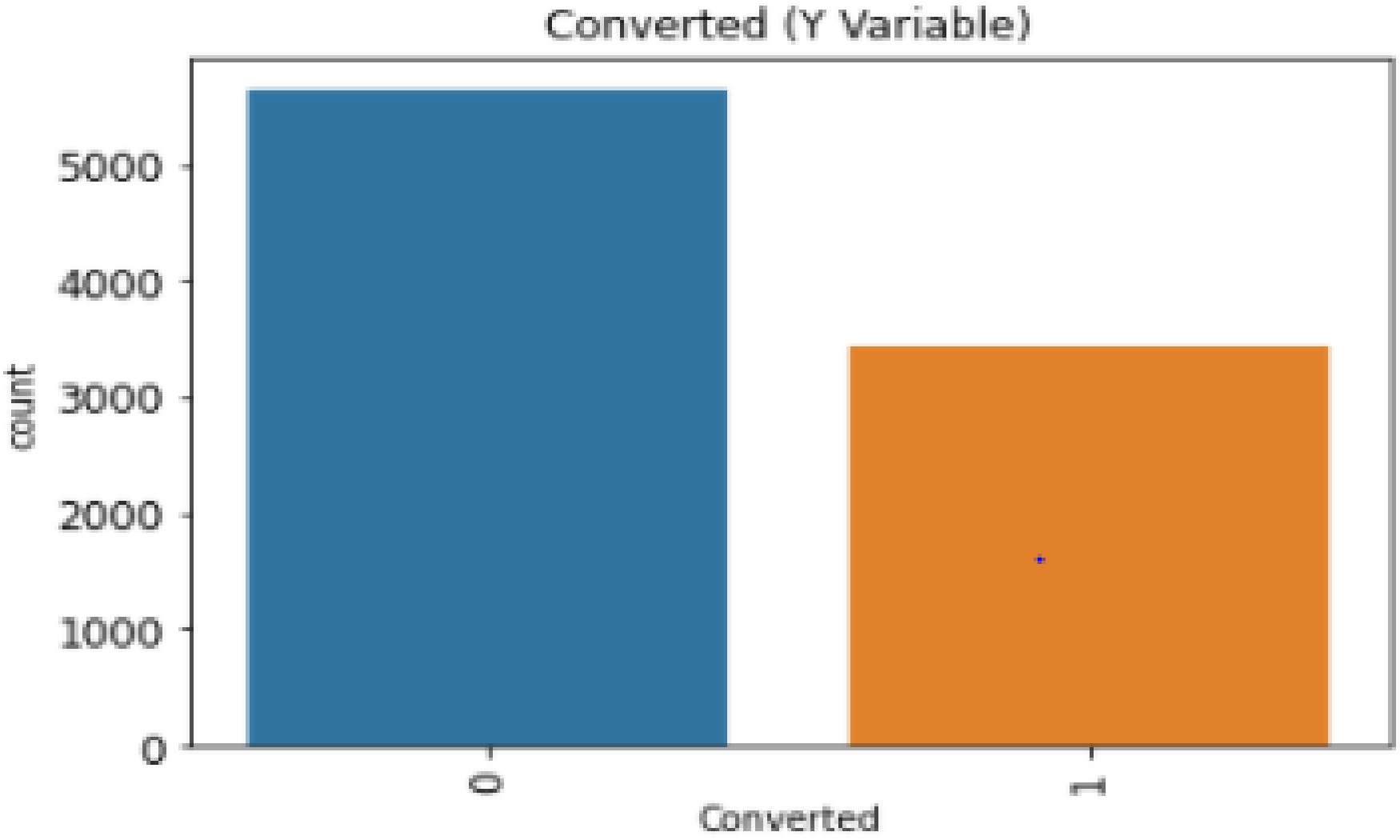


Exploratory Data Analysis (EDA)

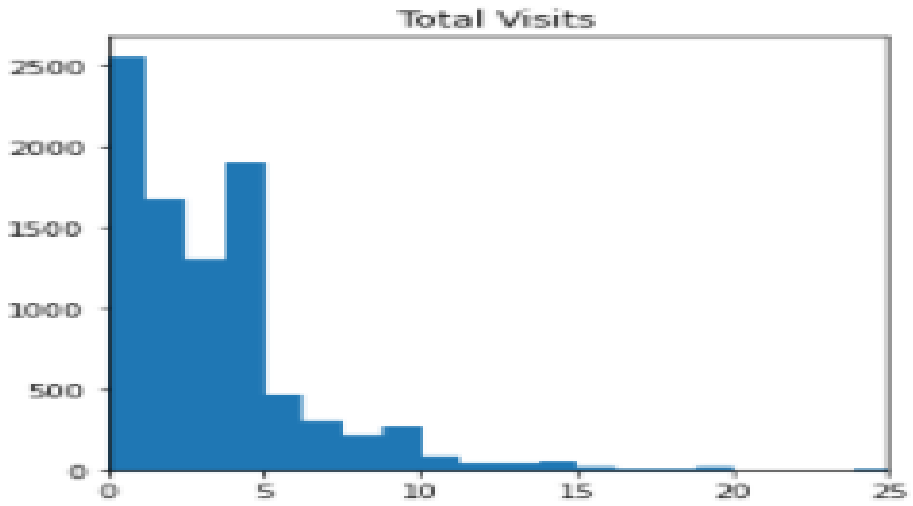


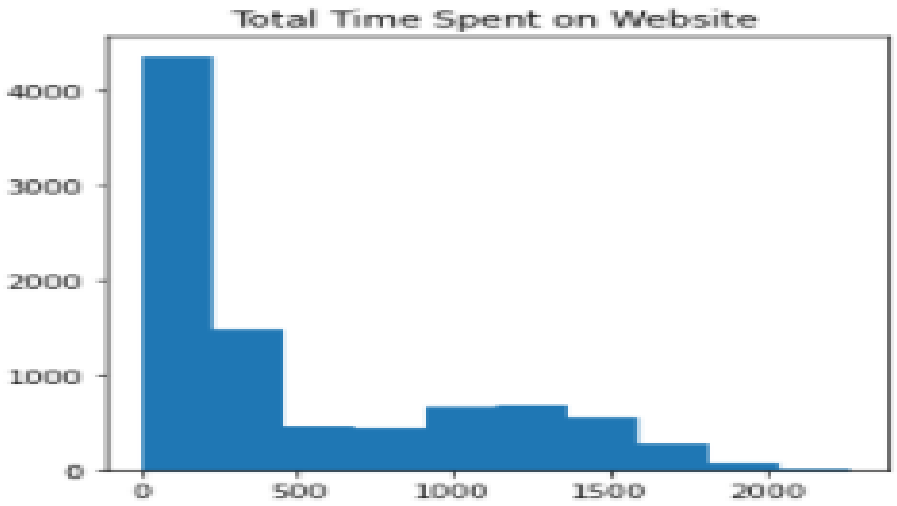


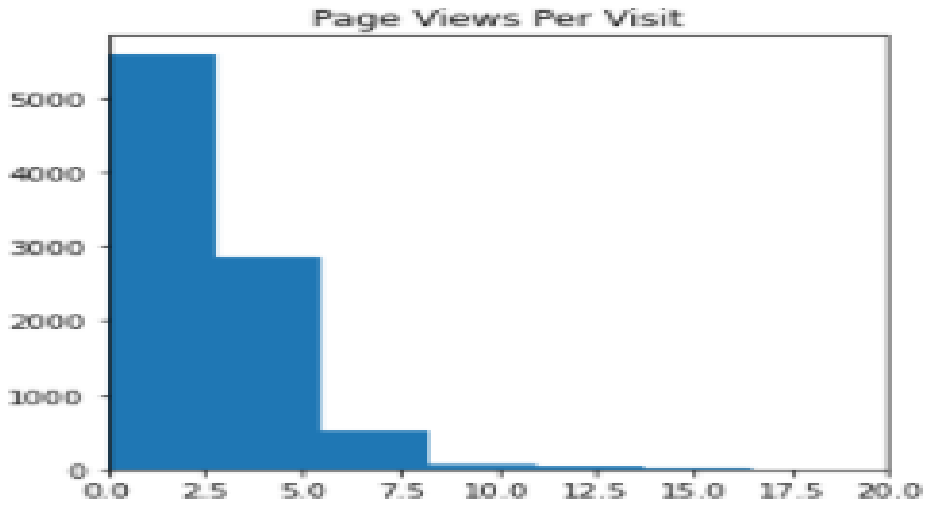
Exploratory Data Analysis (EDA)



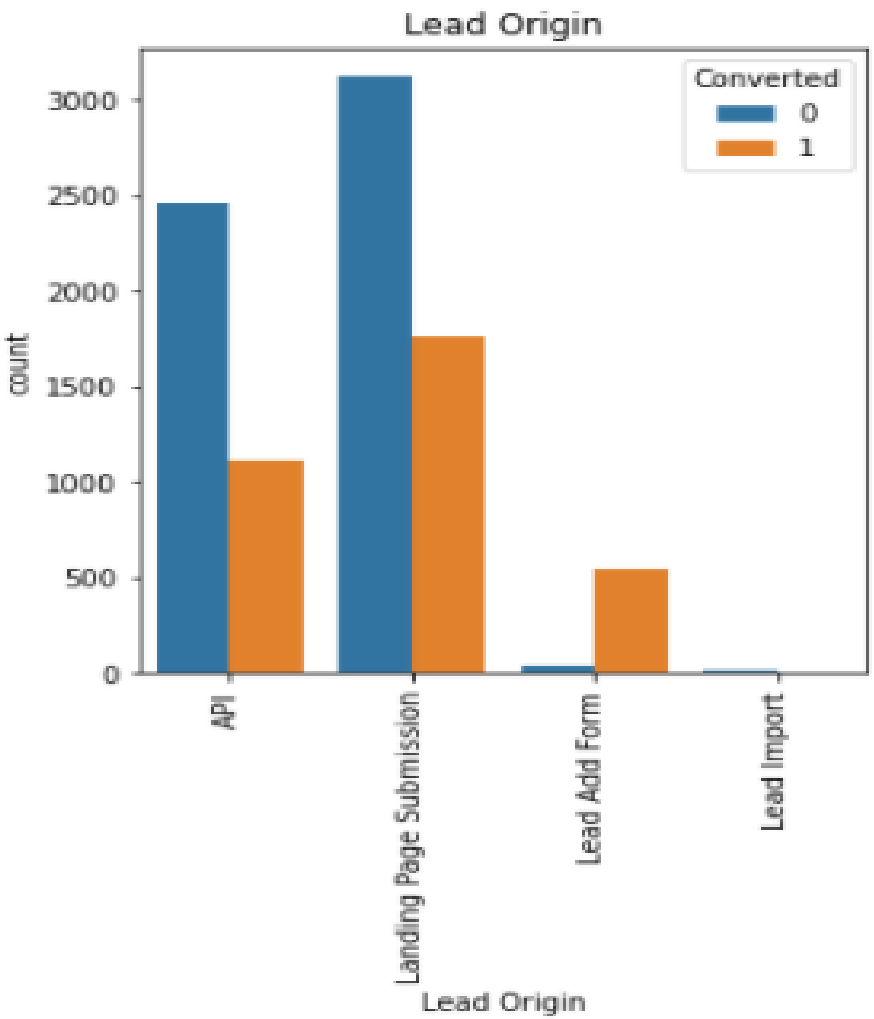
Exploratory Data Analysis (EDA)

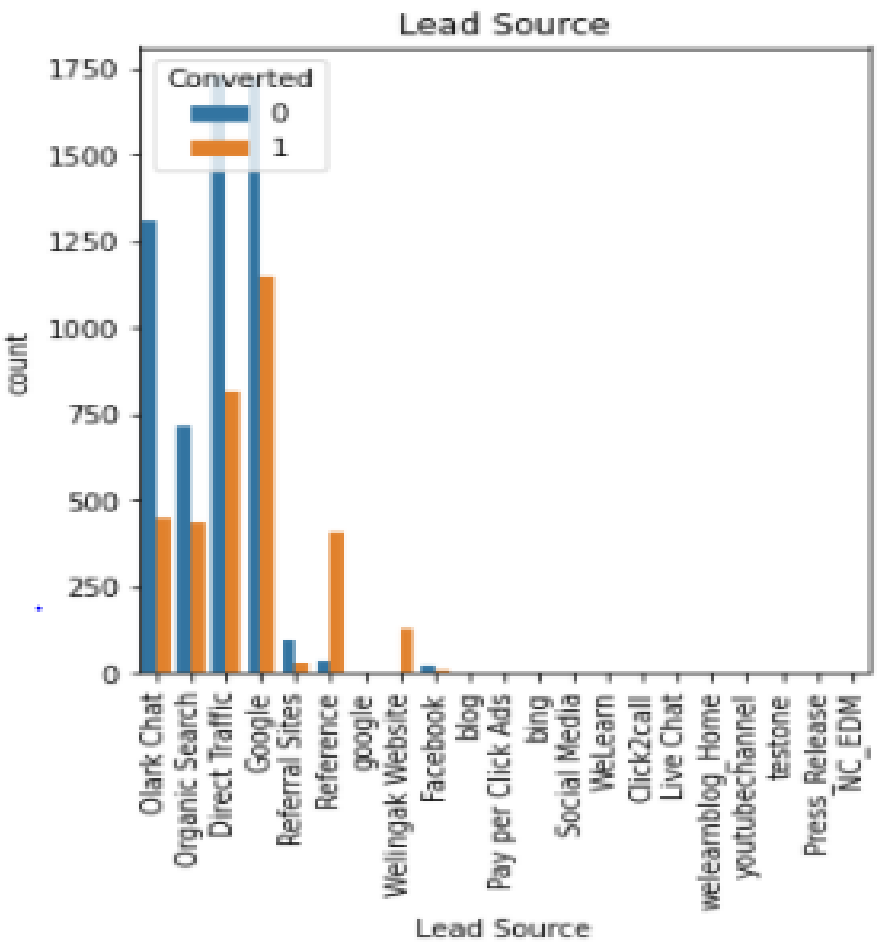




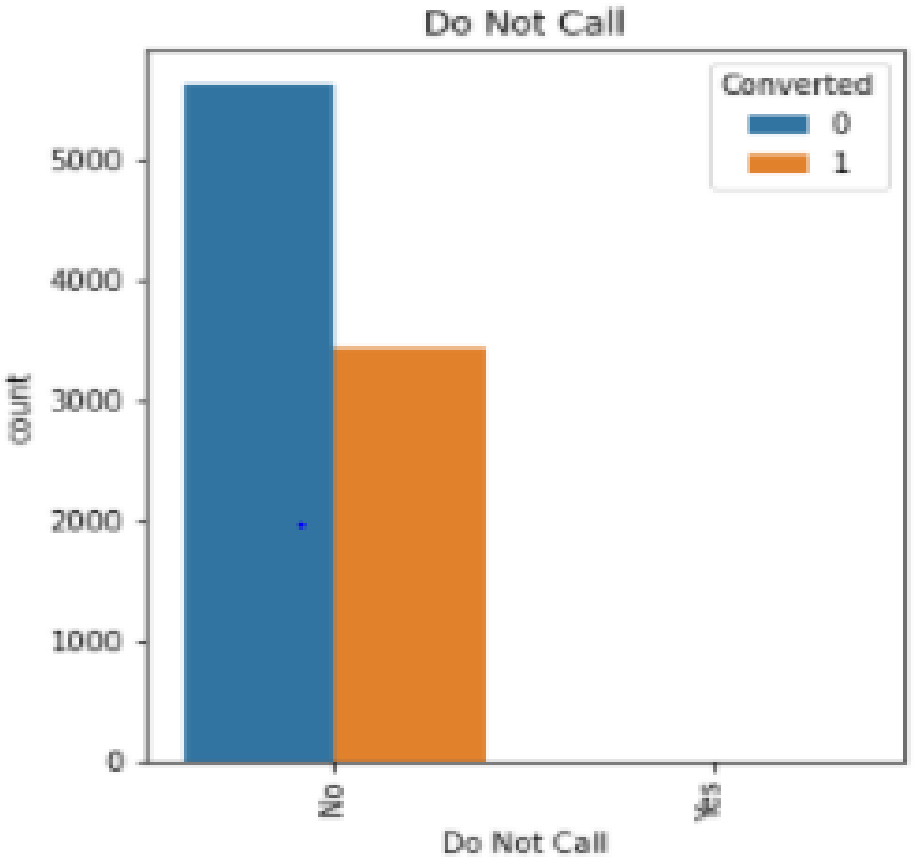


Categorical Variable Vs. Target Variable

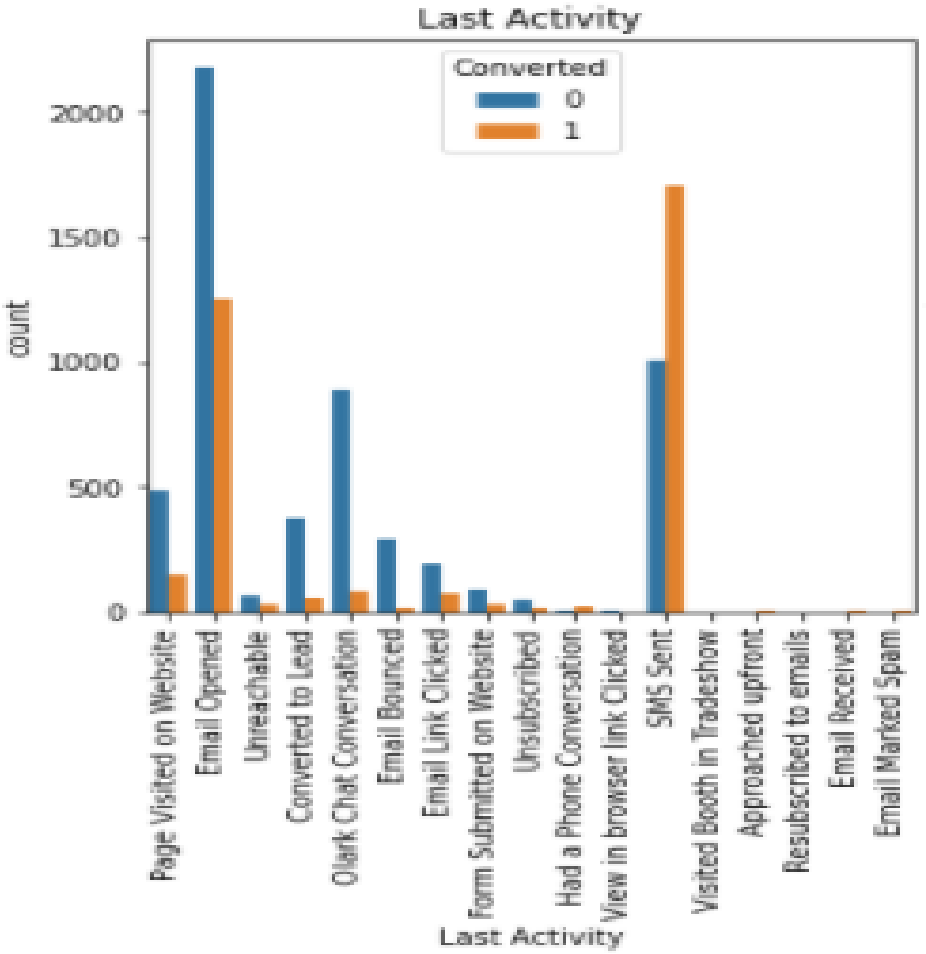


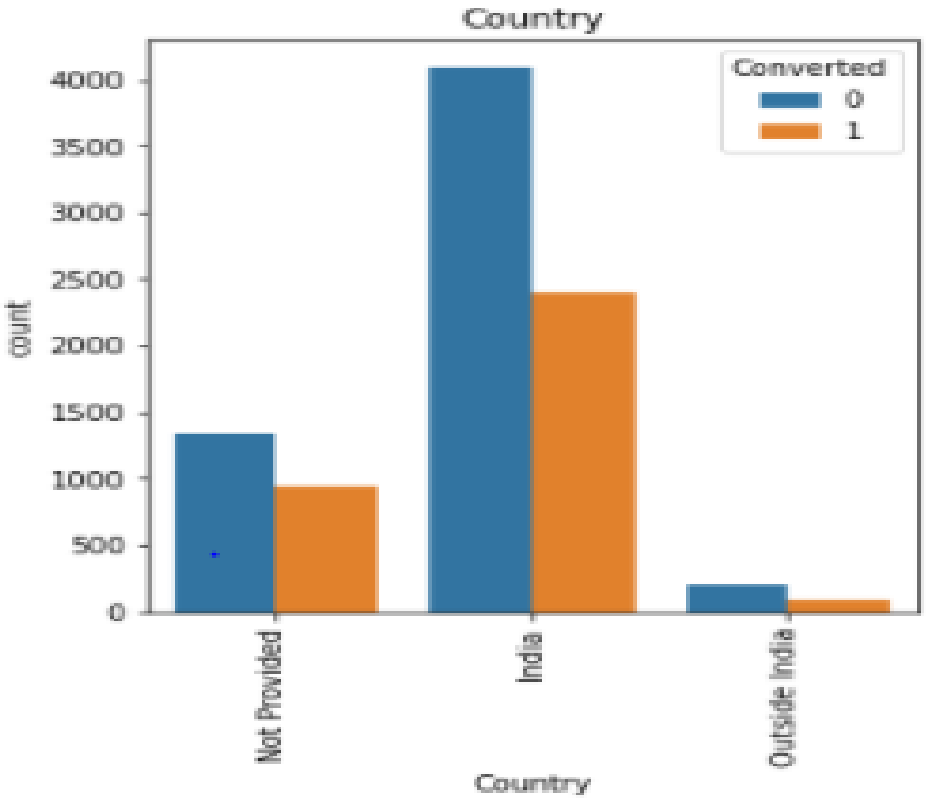


Categorical Variable Vs. Target Variable

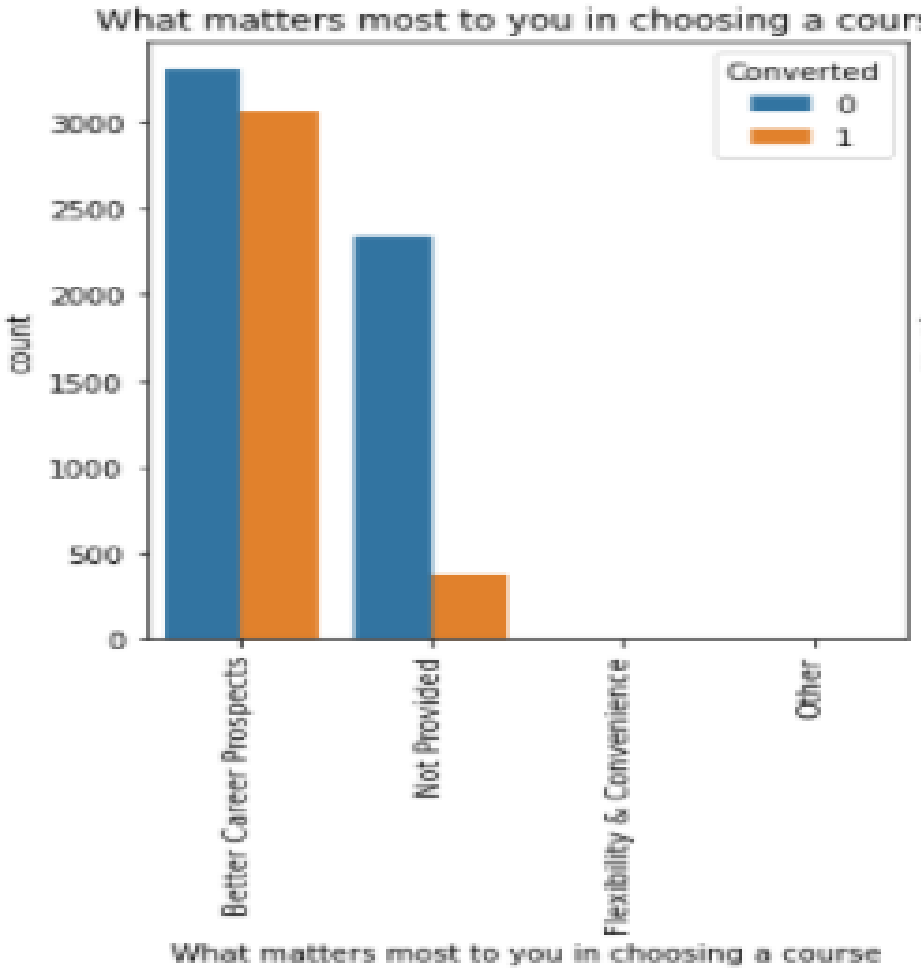


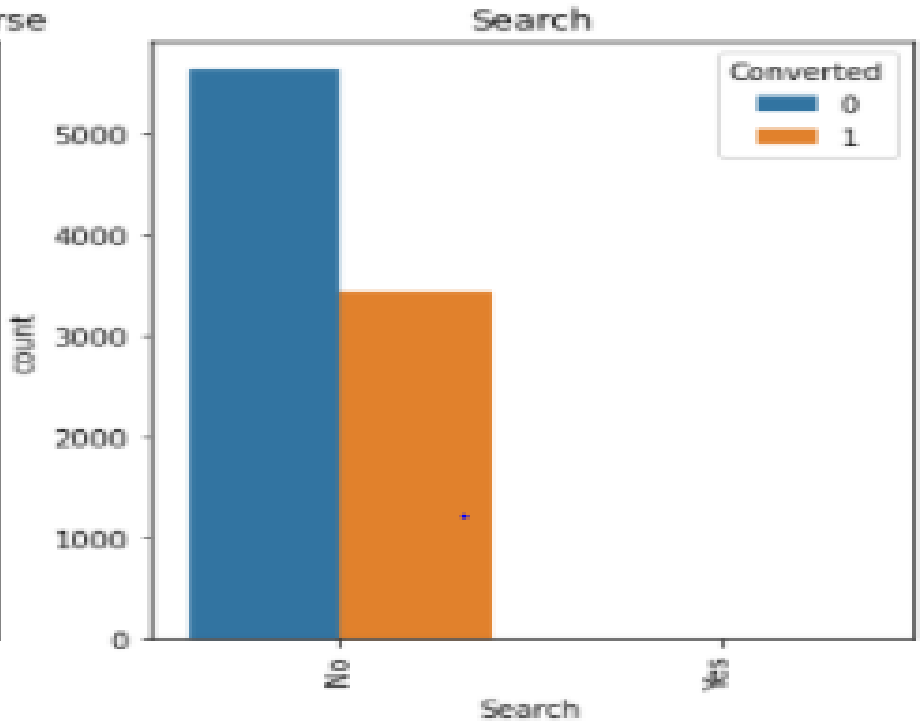
Categorical Variable Vs. Target Variable



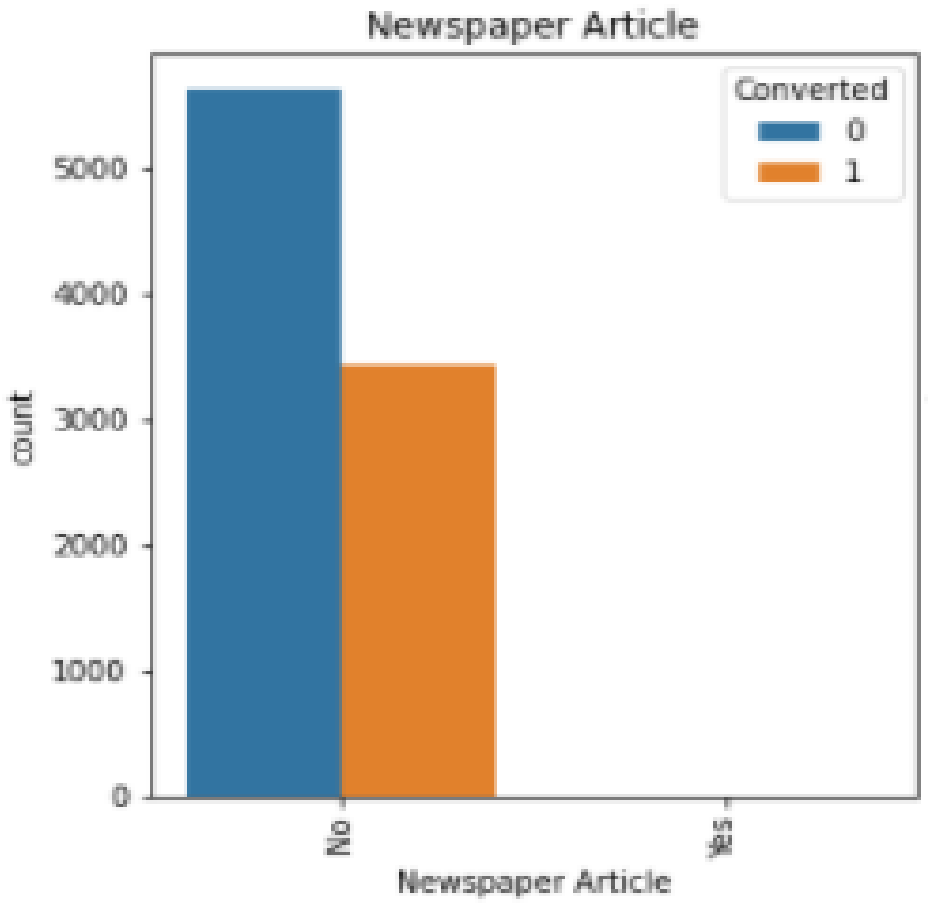


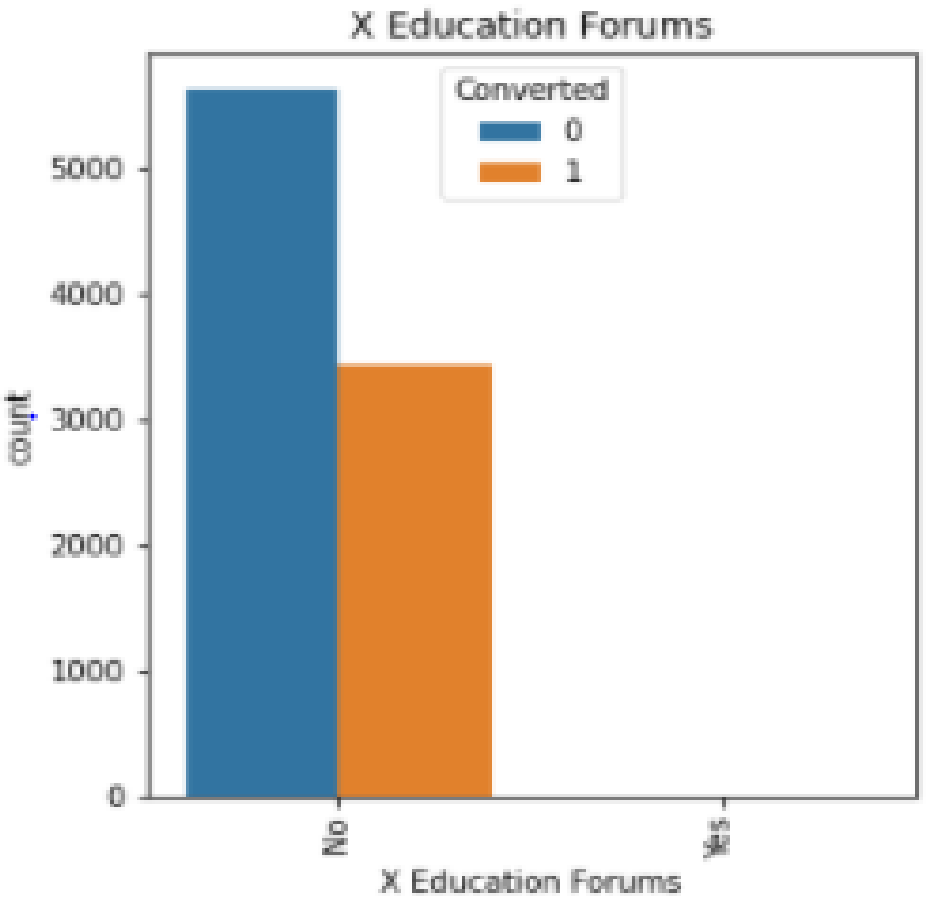
Categorical Variable Vs. Target Variable





Categorical Variable Vs. Target Variable





Model Building

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6335
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2540.3
Date:	Mon, 14 Jun 2021	Deviance:	5080.6
Time:	01:10:46	Pearson chi2:	6.30e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

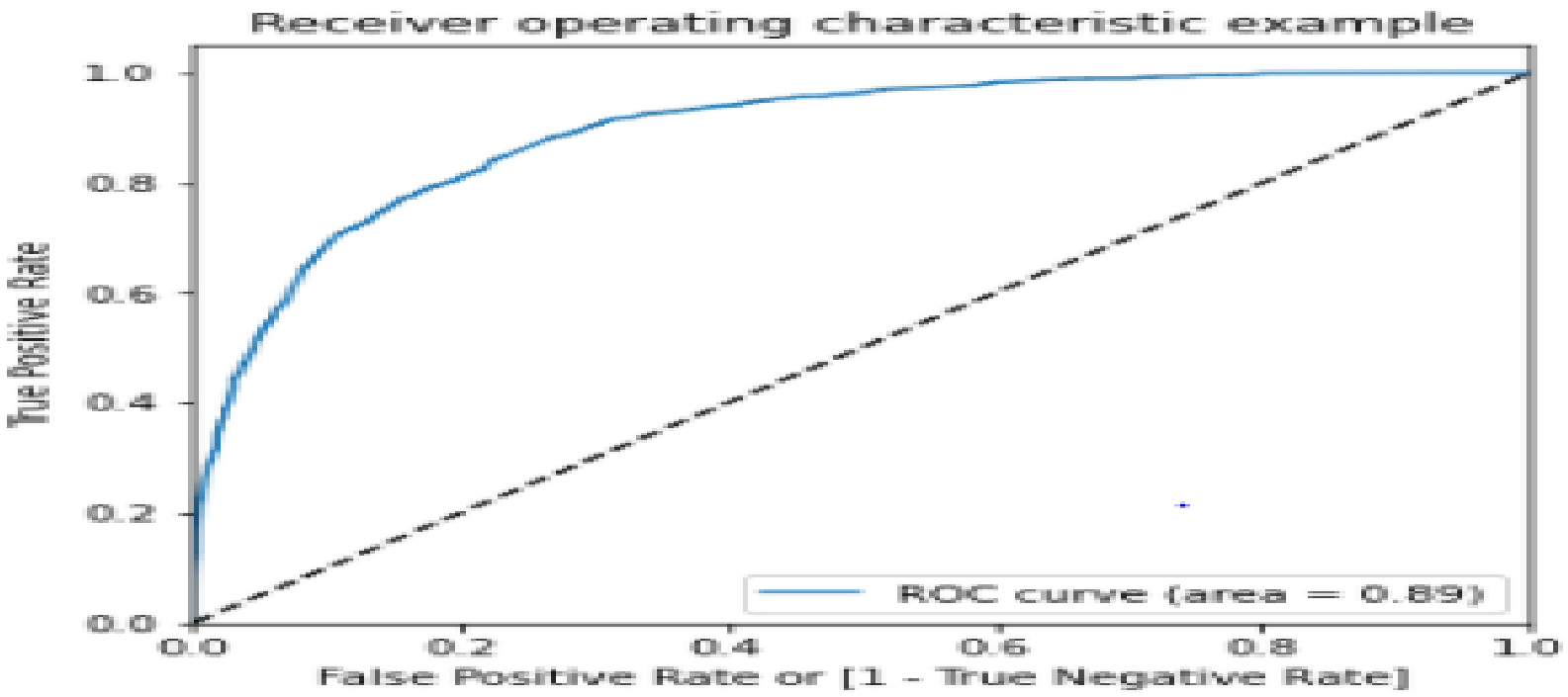
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2370	0.140	-8.823	0.000	-1.512	-0.962
TotalVisits	5.5976	1.982	2.824	0.005	1.712	9.483
Total Time Spent on Website	4.5526	0.168	27.165	0.000	4.224	4.881
Lead Origin_Landing Page Submission	-1.0243	0.128	-7.990	0.000	-1.276	-0.773
Lead Source_Olark Chat	1.3194	0.128	10.296	0.000	1.068	1.571
Lead Source_Reference	3.1776	0.243	13.096	0.000	2.702	3.653
Lead Source_Welingak Website	5.5753	0.730	7.639	0.000	4.145	7.006
Last Activity_Converted to Lead	-1.1867	0.225	-5.264	0.000	-1.629	-0.745
Last Activity_Olark Chat Conversation	-1.3723	0.169	-8.135	0.000	-1.703	-1.042
Last Activity_SMS Sent	1.2836	0.076	16.813	0.000	1.134	1.433
Specialization_Not Provided	-0.9095	0.126	-7.208	0.000	-1.157	-0.662
What is your current occupation_Not Provided	-1.2395	0.089	-13.849	0.000	-1.415	-1.064
What is your current occupation_Working Professional	2.3955	0.192	12.463	0.000	2.019	2.772
Last Notable Activity_Had a Phone Conversation	3.4665	1.135	3.053	0.002	1.241	5.692
Last Notable Activity_Unreachable	1.9558	0.499	3.920	0.000	0.978	2.934
Do Not Email_Yes	-1.6526	0.174	-9.477	0.000	-1.994	-1.311

	Features	VIF
2	Lead Origin_Landing Page Submission	2.61
9	Specialization_Not Provided	2.53
1	Total Time Spent on Website	2.03
3	Lead Source_Olark Chat	2.00
8	Last Activity_SMS Sent	1.61
10	What is your current occupation_Not Provided	1.60
0	TotalVisits	1.59
7	Last Activity_Olark Chat Conversation	1.47
11	What is your current occupation_Working Profes...	1.21
4	Lead Source_Reference	1.17
14	Do Not Email_Yes	1.12
6	Last Activity_Converted to Lead	1.11
5	Lead Source_Welingak Website	1.09
13	Last Notable Activity_Unreachable	1.01
12	Last Notable Activity_Had a Phone Conversation	1.00

Model Building Cont.

- We build our Model with considering 15 variables.
- VIF values have been considered to be less than 5.
- P-value have been considered to be less than 0.05

ROC Curve



The ROC curve bending toward the left side of the border, hence our model is having great accuracy. The area under the curve is 87% of the total area.

Sensitivity, Accuracy and Specificity

```
# Checking the overall accuracy

metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.Predicted)
```

```
0.8182963312864115
```

```
# Calculating the sensitivity

TP/(TP+FN)
```

```
0.7837285363859362
```

```
# Calculating the specificity

TN/(TN+FP)
```

```
0.829449423815621
```

Accuracy, Sensitivity and Specificity calculated on Train Data Set.

Sensitivity, Accuracy and Specificity cont.

```
# Check the overall accuracy

metrics.accuracy_score(y_pred_final['Converted'], y_pred_final.final_predicted)
```

```
0.8189496878442893
```

```
# Calculating the sensitivity

TP/(TP+FN)
```

```
0.78159757330637
```

```
# Calculating the specificity

TN/(TN+FP)
```

```
0.8356401384083045
```

Accuracy, Sensitivity and Specificity calculated on Test Data Set.

Conclusion

- The values we have got for Accuracy, sensitivity and specificity are in acceptable range.
- The sensitivity score is inversely proportional to specificity .
- The ROC curve is bending toward the left side of the border, hence our model is having great accuracy.
- We have high recall score than precision score which we were exactly looking for.
- The variables which matters the most in terms of finding the genuine applicants are:

    The total time spend on the Website

    Total number of visits

    Lead Source_Welingak Website


Thank You