Lead Scoring Model

Problem Statement

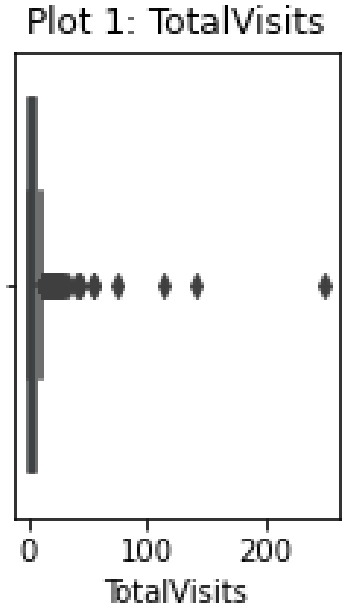
1.Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will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

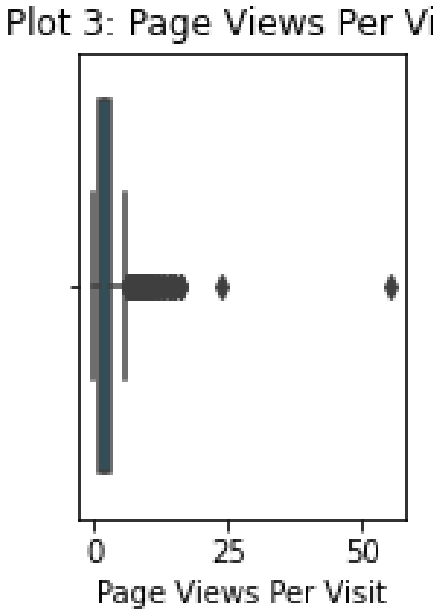
2.There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

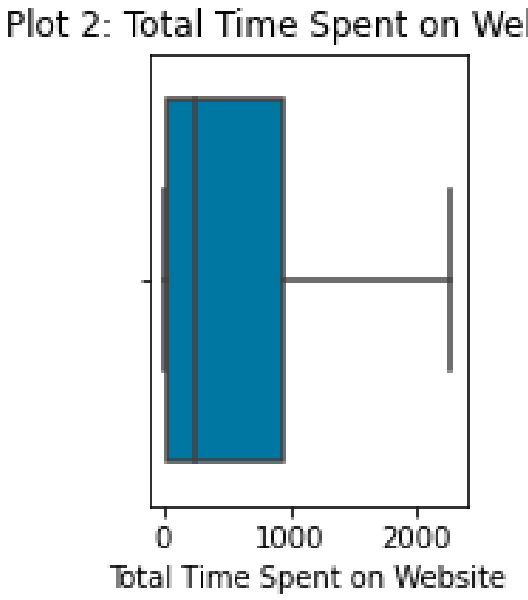
Approach

1. Remove columns with a missing value percentage greater than 30%.
2. Drop categorical variables like 'Do Not Call', 'Newspaper Article', 'X Education Forums', 'Search', 'Newspaper', 'Digital Advertisement', and 'Through Recommendations' as they are nearly constant, with over 99% of data points having the same category.
3. For the remaining categorical variables, fill missing values with the mode (most frequent value), and for numerical variables, fill missing values with the median.
4. Drop 'TotalVisits' and 'Page Views Per Visit' numerical variables due to suspected outlier values, especially since they have low correlation with the target variable.
5. Normalize the remaining numerical variables using standard scaling, and encode categorical variables using dummy variables.

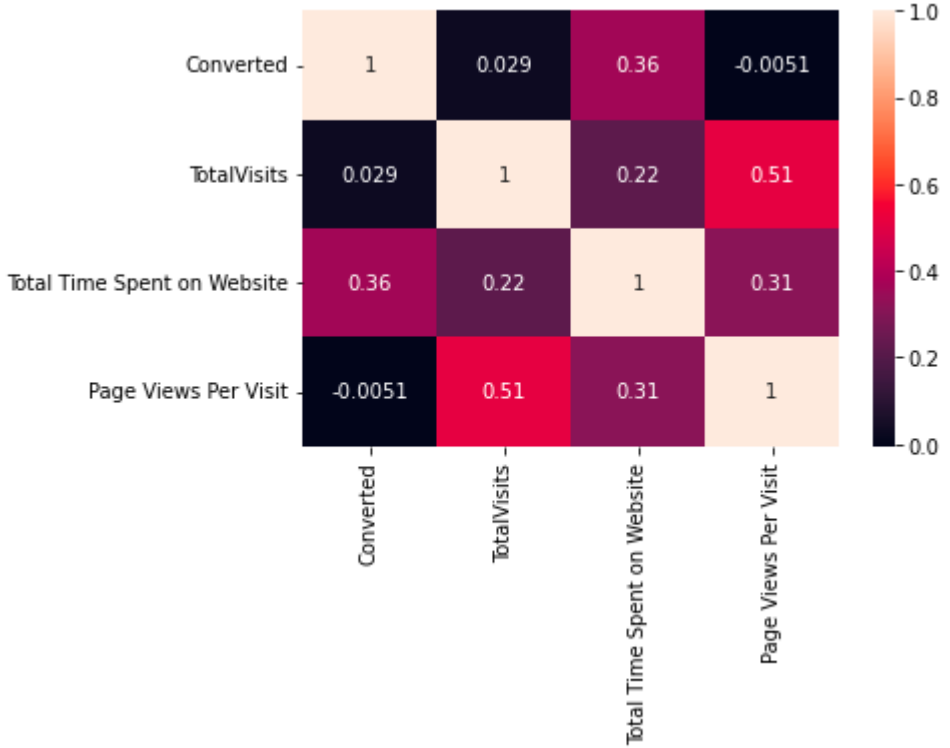
Outliers







Correlation



Data Analysis (EDA)

1. API and Landing Page submissions are the primary sources of leads, while Lead Add Form leads have a higher conversion rate.
2. Lead Import and Quick Add Form contribute very few leads, with Quick Add Form showing zero conversion rates.

3. Direct Traffic and Olark Chat bring in a significant number of leads but have low conversion rates. Google, on the other hand, contributes to a good number of leads with a moderate conversion rate.

4. The reference as a lead source shows the highest conversion rate.

5. Activities like having a Phone Conversation and sending SMS seem to generate leads with a higher chance of conversion.

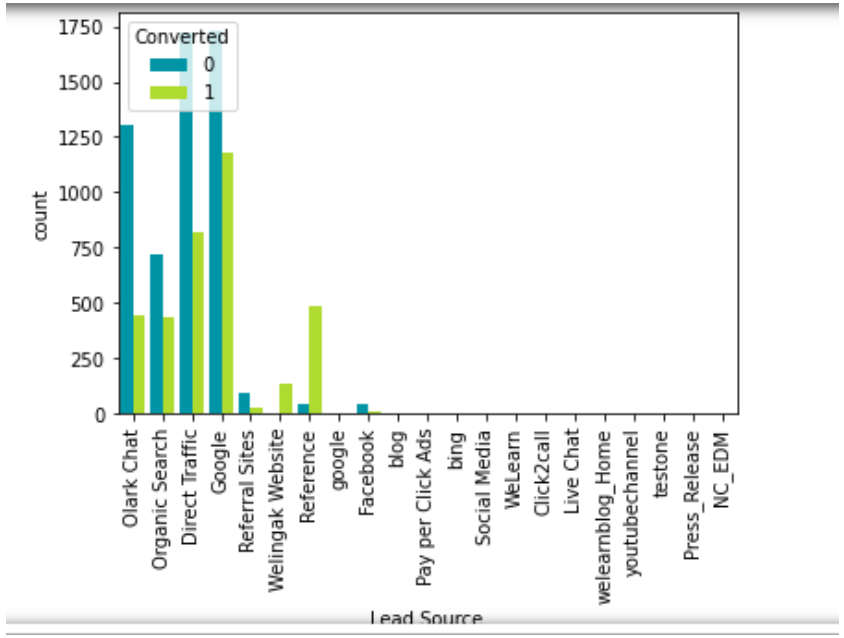
6. Unemployed individuals constitute a majority of leads but have a lower conversion rate, around half of the leads.

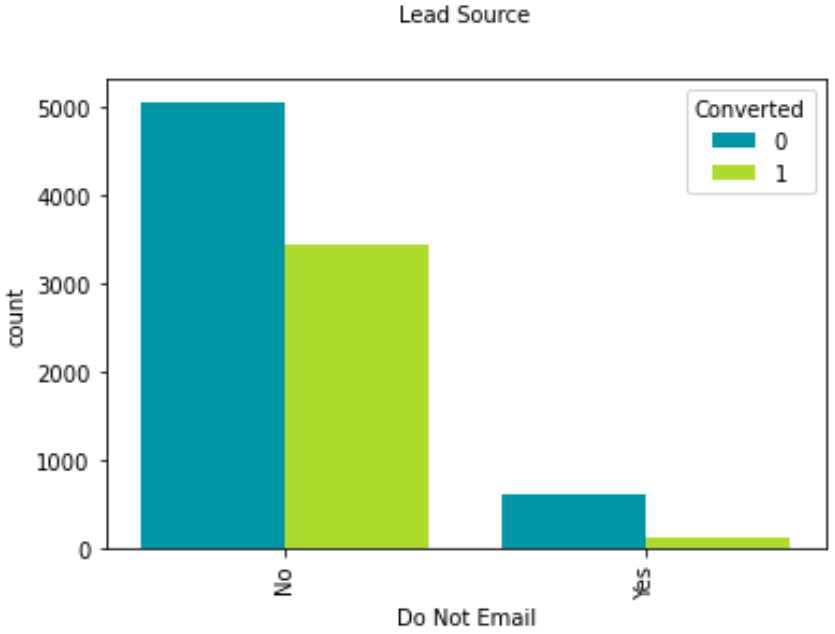
7. Businessmen and Working Professionals contribute to higher conversions.

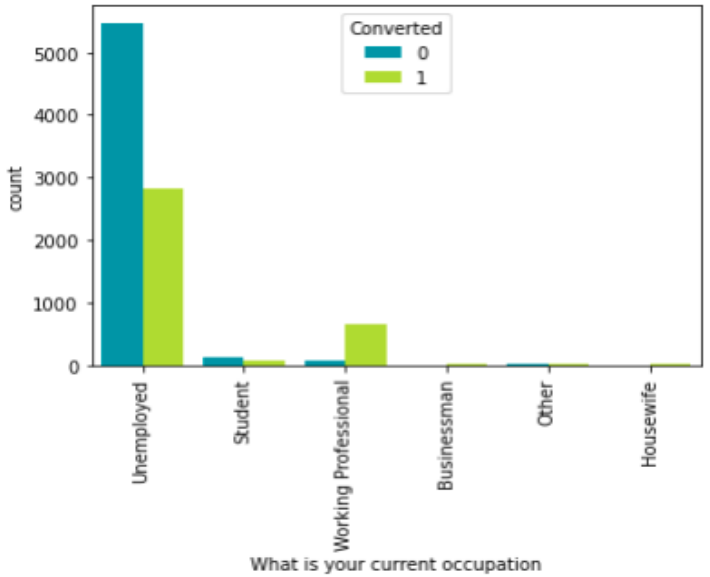
8. Although Housewives generate fewer leads, the leads they generate tend to have a higher conversion rate.

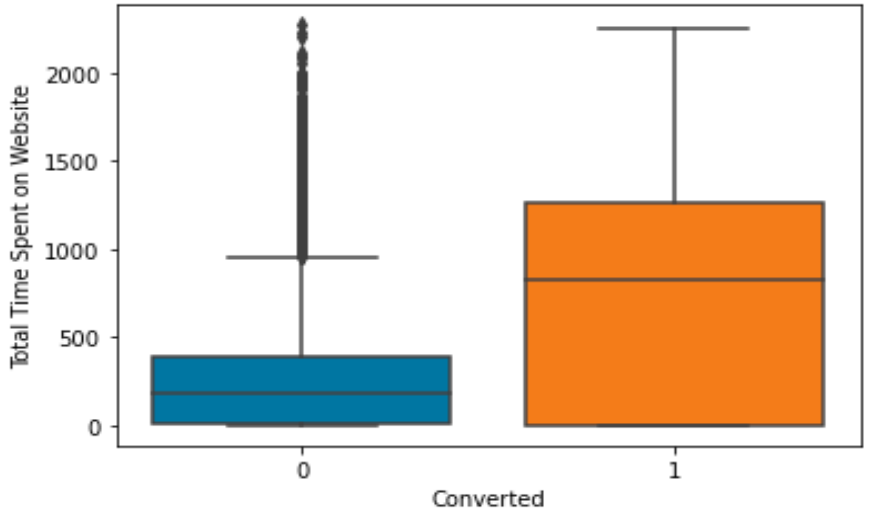
9. Leads spending more time on the website are more likely to convert.

Plot for Overview









Performance of Model

	variable	vif
11	Last Notable Activity_Modified	2.053099
6	Last Activity_Olark Chat Conversation	2.038582
5	Last Activity_Email Bounced	1.878718
3	Do Not Email_Yes	1.844504
2	Lead Source_Olark Chat	1.676651
12	Last Notable Activity_Olark Chat Conversation	1.337010
4	Last Activity_Converted to Lead	1.250938
0	Total Time Spent on Website	1.213874
1	Lead Origin_Lead Add Form	1.162495
7	Last Activity_Page Visited on Website	1.117900
8	What is your current occupation_Working Profes...	1.116013
10	Last Notable Activity_Email Opened	1.098823
9	Last Notable Activity_Email Link Clicked	1.017879

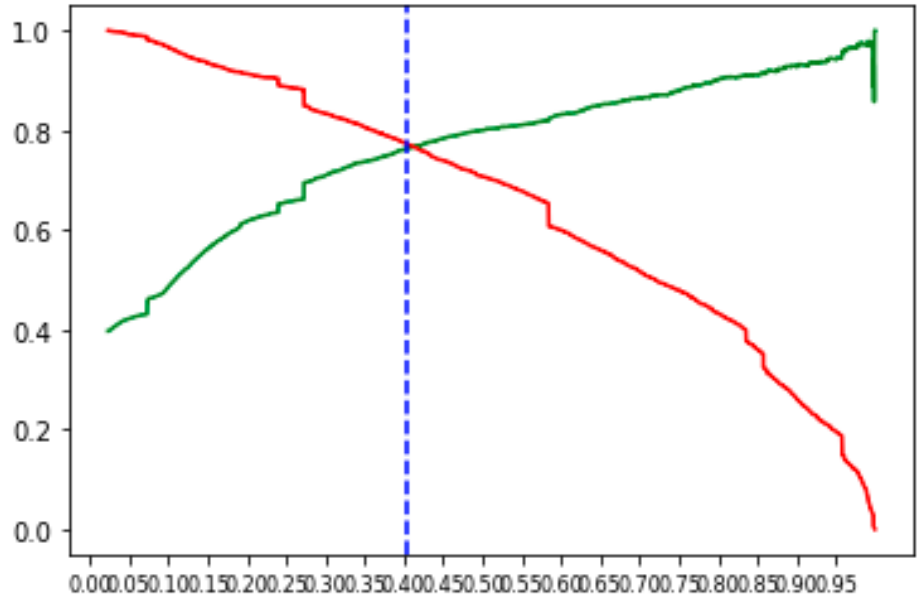
Generalized Linear Model Regression Results

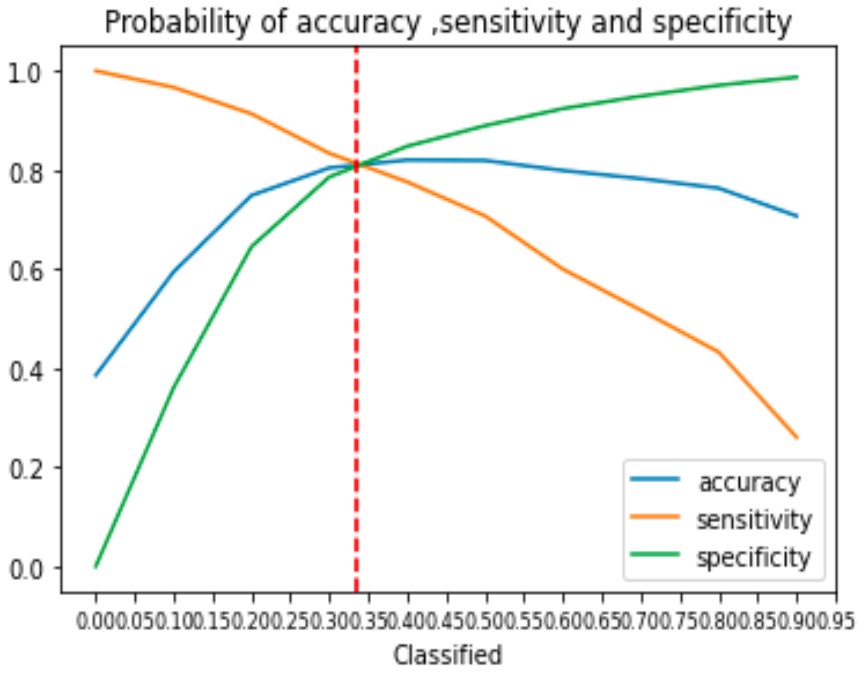
Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6454
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2690.7
Date:	Fri, 12 Apr 2024	Deviance:	5381.3
Time:	11:54:46	Pearson chi2:	8.07e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3944
Covariance Type:	nonrobust		

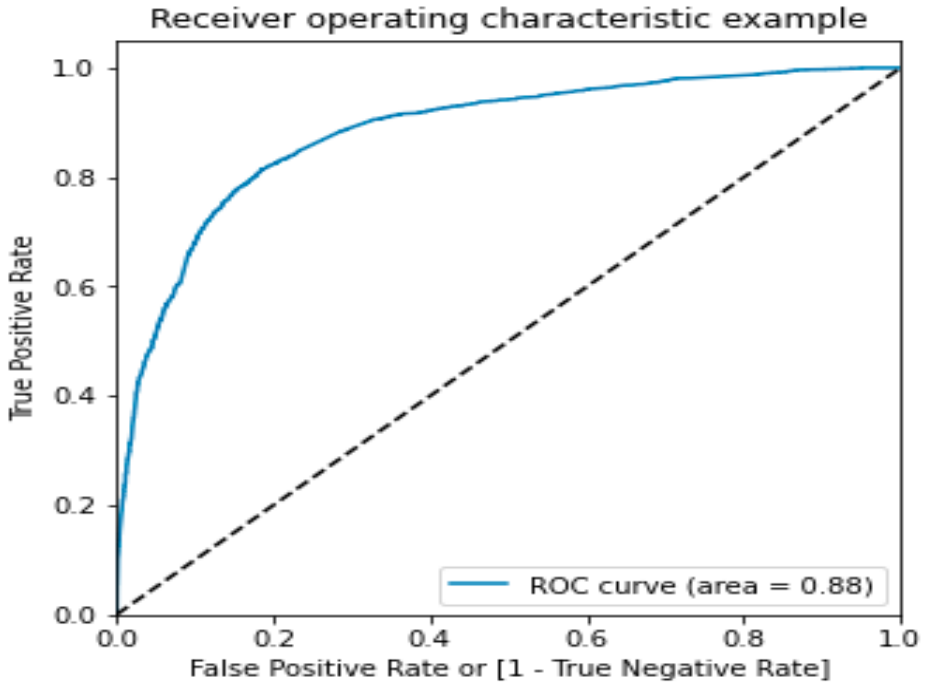
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0792	0.068	1.171	0.241	-0.053	0.212
Total Time Spent on Website	1.1539	0.041	28.459	0.000	1.074	1.233
Lead Origin_Lead Add Form	4.0715	0.195	20.912	0.000	3.690	4.453
Lead Source_Olark Chat	1.2939	0.103	12.577	0.000	1.092	1.496
Do Not Email_Yes	-1.2895	0.189	-6.814	0.000	-1.660	-0.919
Last Activity_Converted to Lead	-0.9888	0.218	-4.546	0.000	-1.415	-0.563
Last Activity_Email Bounced	-1.0947	0.344	-3.187	0.001	-1.768	-0.421
Last Activity_Olark Chat Conversation	-1.3905	0.185	-7.516	0.000	-1.753	-1.028
Last Activity_Page Visited on Website	-1.2526	0.156	-8.009	0.000	-1.559	-0.946
What is your current occupation_Working Professional	2.8560	0.194	14.702	0.000	2.475	3.237
Last Notable Activity_Email Link Clicked	-1.7813	0.247	-7.212	0.000	-2.265	-1.297
Last Notable Activity_Email Opened	-1.3212	0.086	-15.344	0.000	-1.490	-1.152
Last Notable Activity_Modified	-1.4927	0.098	-15.270	0.000	-1.684	-1.301
Last Notable Activity_Olark Chat Conversation	-1.4772	0.373	-3.959	0.000	-2.208	-0.746

Train Data Set

1.Model Accuracy value: 81.91%
2.Model Sensitivity value: 70.73%
3.Model Specificity value: 88.93%
4.Model Precision value: 80.04%
5.Model Recall value: 70.73%
6.Model True Positive Rate (TPR): 70.73%
7.Model False Positive Rate (FPR): 11.07%
8.Model Positive Prediction Value: 80.04%
9.Model Negative Prediction value: 82.88%

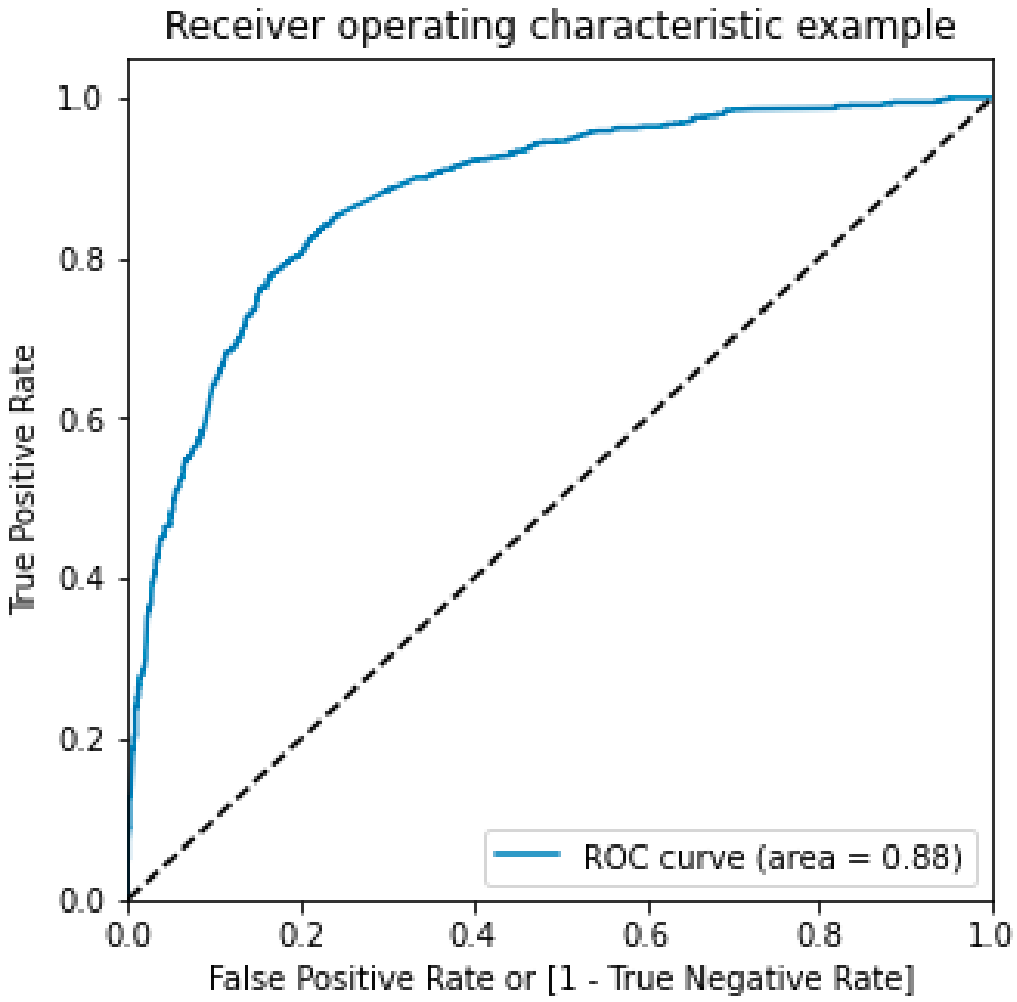






Test Data Set

1.Model Accuracy value: 81.1%
2.Model Sensitivity value: 76.48%
3.Model Specificity value: 83.99%
4.Model Precision value: 74.93%
5.Model Recall value: 76.48%
6.Model True Positive Rate (TPR): 76.48%
7.Model False Positive Rate (FPR): 16.01%
8.Model Positive Prediction Value: 74.93%
9.Model Negative Prediction value: 85.09%



Conclusion

1. Working professionals are more likely to convert, while unemployed leads, despite being numerous, have lower conversion rates.

2. Leads who spend more time on the website tend to convert at higher rates, indicating the importance of focusing on these leads.

3. The Lead Add Form appears to be the most effective lead source.

Thanks and Regards
Pradum Kumar Chaubey