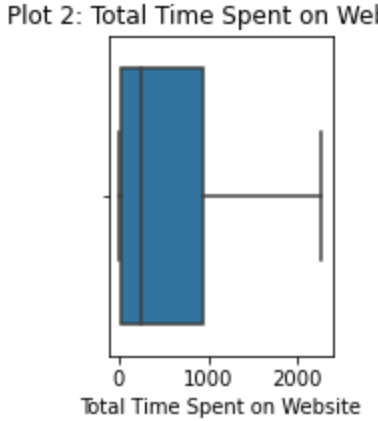
Lead Scoring Case study

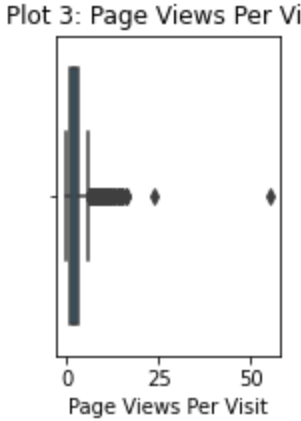
Problem Statement

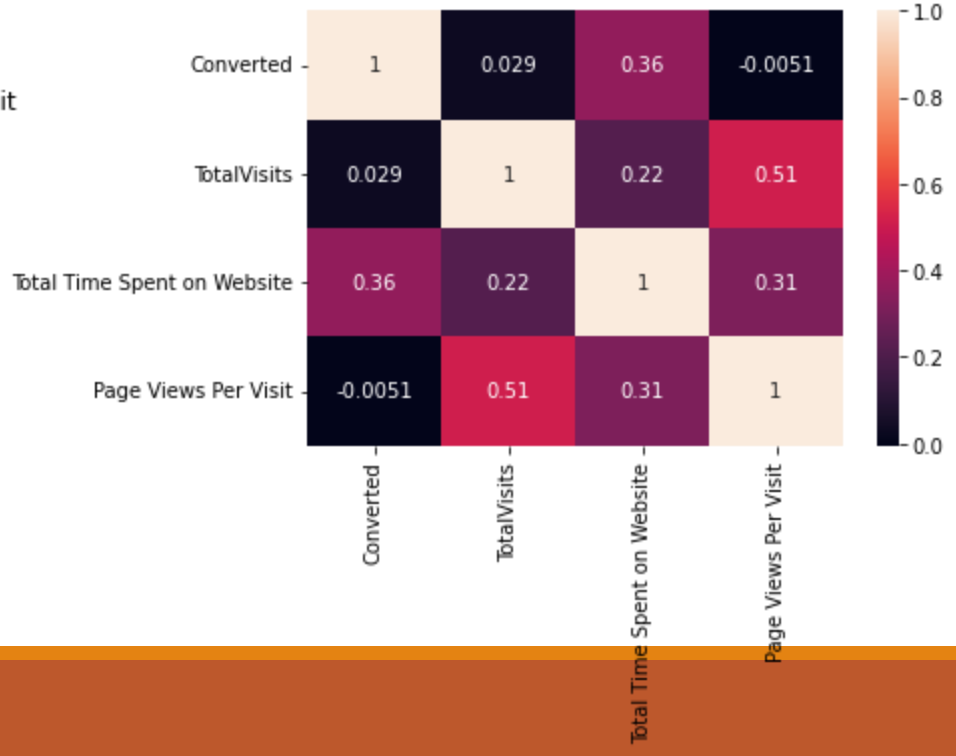
To build a Lead Scoring(logistic regression) model to assign a lead score to each of the leads which can be used by X Education to target potential leads which would help the company to get more customers.

Pre-Process Approach

- First, columns with more than 30% of missing value percentage is removed.
- Categorical variables 'Do Not Call', 'Newspaper Article', 'X Education Forums', 'Search', 'Newspaper', 'Digital Advertisement' and 'Through Recommendations' had constants with more than 99% of data points having same category and hence dropped.
- For remaining categorical variables, the missing values were imputed with Mode and numerical with Median.
- Numerical Variables 'TotalVisits' and 'Page Views Per Visit' had outlier values and has less correlation with the target and these were dropped.
- The remaining numerical variables were normalized using standard scalar and categorical variables were encoded using dummy variables.



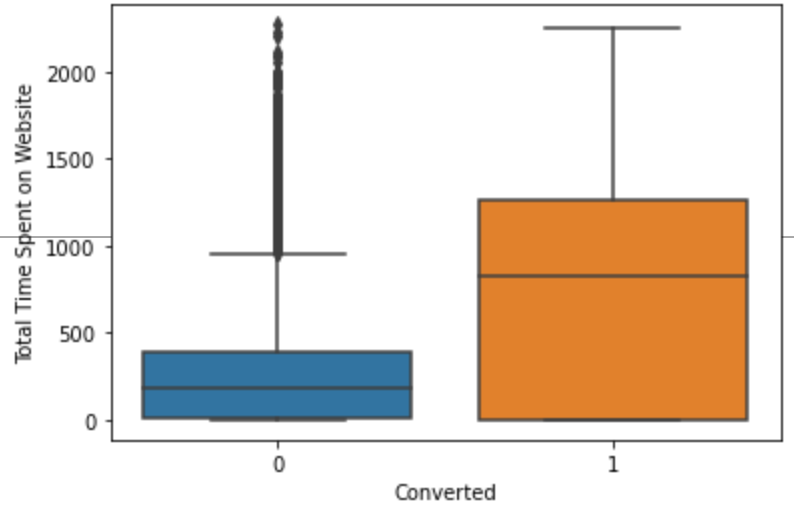


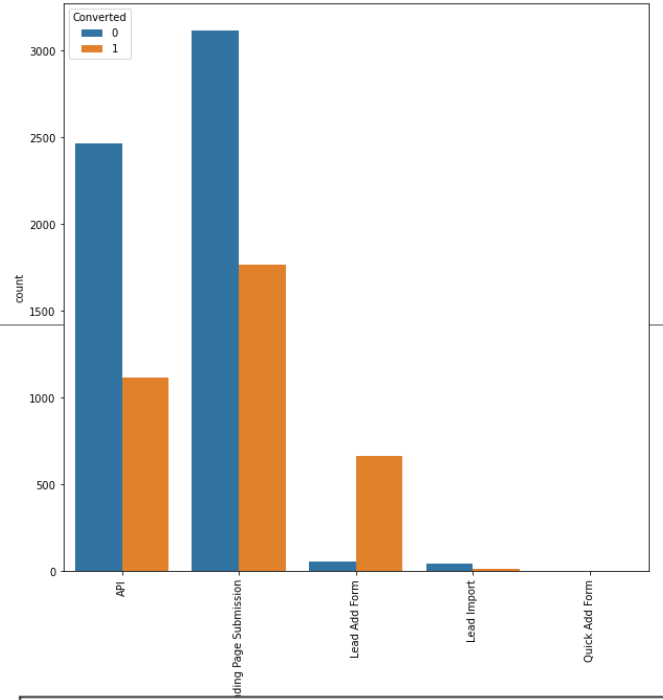


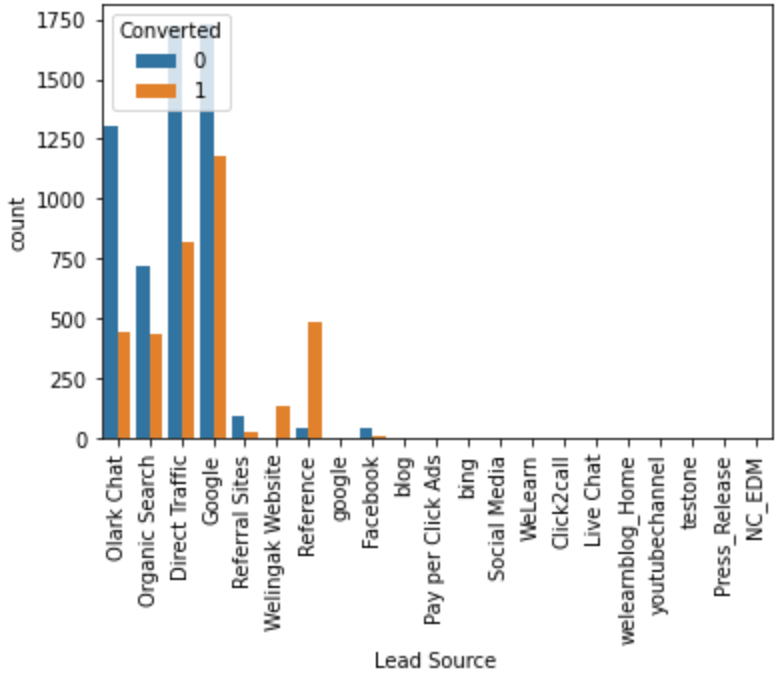
EDA

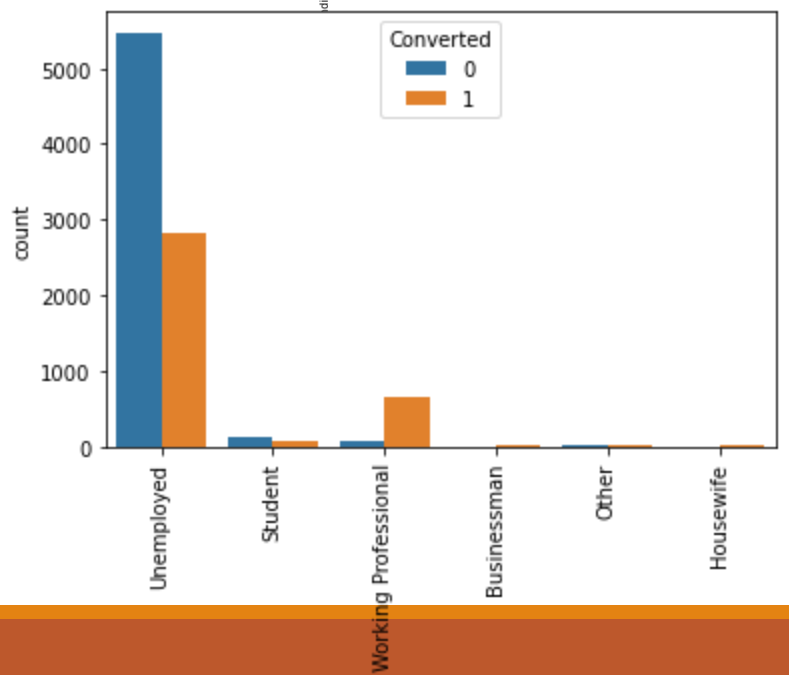
- API and Landing Page submission brings most of the leads whereas the conversion rate is higher for Lead Add Form.
- Lead Import and Quick Add Form brings very few leads and Quick Add Form is having a zero conversion rate.
- In lead source, Direct Traffic and Olark Chat brings in huge number of leads but suffers very low conversion whereas Google has good lead inputs and decent conversion.
- In lead source, reference shows the highest conversion rate
- In Last Activity, Had a Phone Conversation and SMS sent seems to generate hot leads having good conversion rate.
- Unemployed people seems to be making up for most of the leads but with low conversion of about half.
- Businessman and Working Professional contribute for higher conversions.
- Housewives are having less lead generation percentage, but the generated leads tend to be converted.
- Leads spending more time in website are likely to be converted.

EDA


Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted


Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Lead Add Form
Lead Import
Quick Add Form


Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source


Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

Model Performance

The final model was fit after taking in only relevant features based on p-value and VIF.

	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

Model Performance

For Train Data set

Model Accuracy value is 81.91 %

Model Sensitivity value is 70.73 %

Model Specificity value is 88.93 %

Model Precision value is 80.04 %

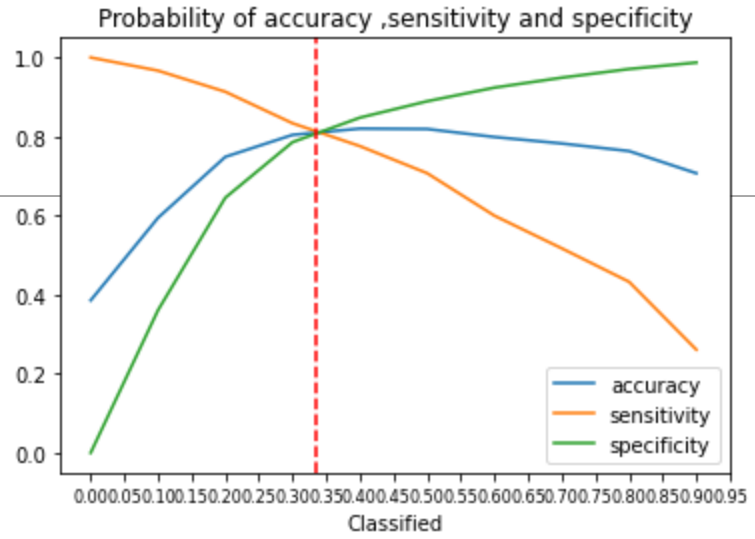
Model Recall value is 70.73 %

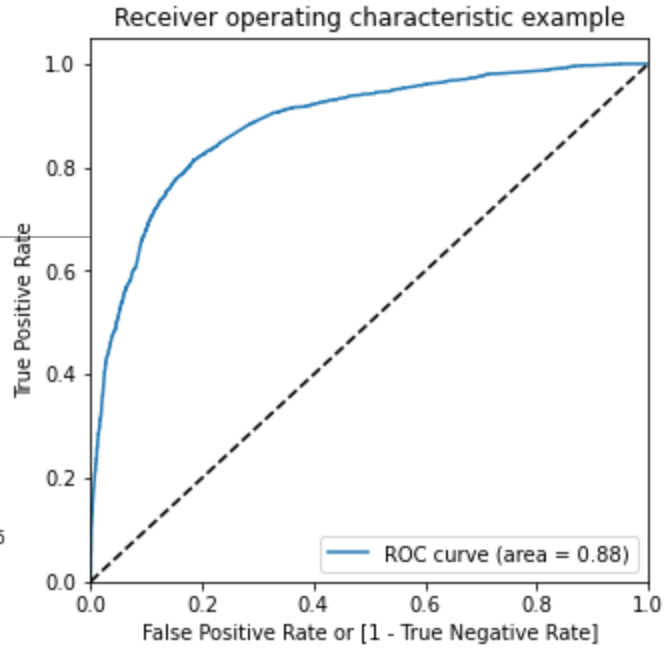
Model True Positive Rate (TPR) 70.73 %

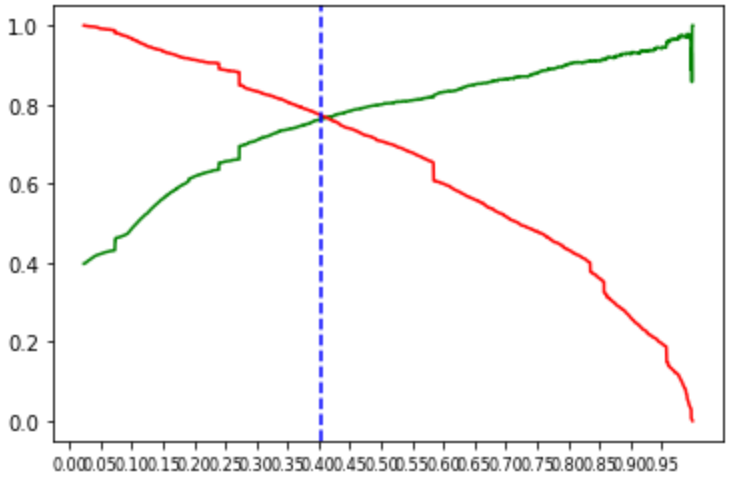
Model False Positive Rate (FPR) is 11.07 %

Model Positive Prediction Value is 80.04 %

Model Negative Prediction value is 82.88 %







Precision recall trade-off

74.93

Model Performance

For Test Data set

Model Accuracy value is 81.1 %

Model Sensitivity value is 76.48 %

Model Specificity value is 83.99 %

Model Precision value is 74.93 %

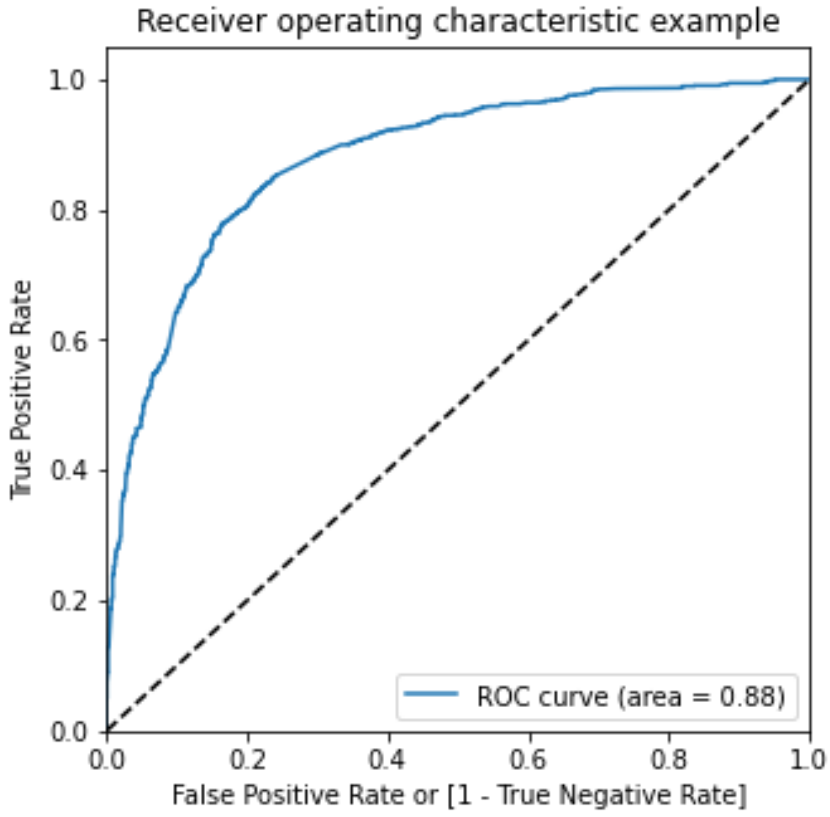
Model Recall value is 76.48 %

Model True Positive Rate (TPR) 76.48 %

Model False Positive Rate (FPR) is 16.01 %

Model Positive Prediction Value is 74.93 %

Model Negative Prediction value is 85.09 %



Conclusion

Working professionals tends to be converted most whereas unemployed leads are less likely to be converted despite of greater lead numbers.

Leads spending more time on the website gets converted more and should be given high focus.

Lead Add Form seems to be the best Lead Source.