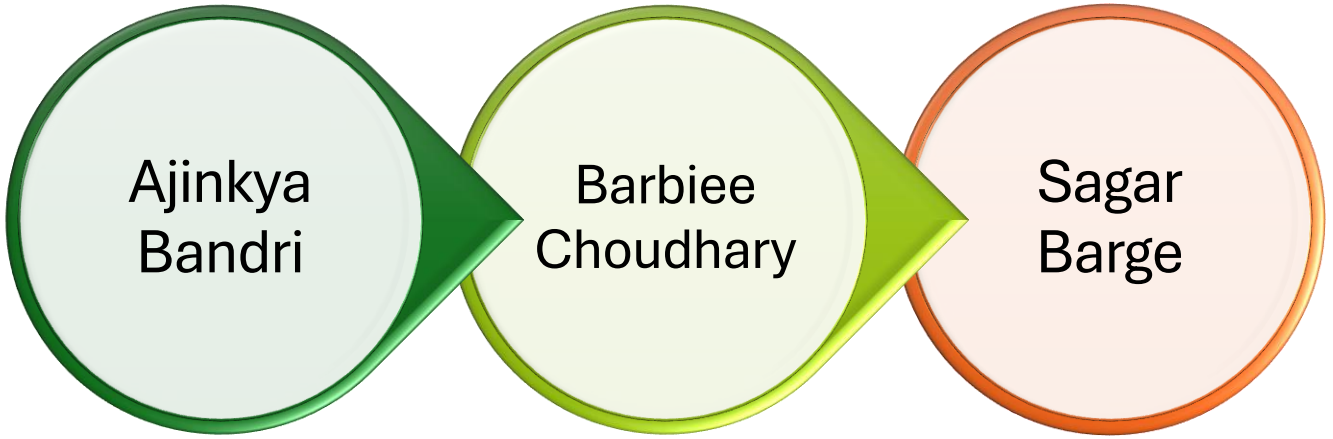
Case Study For Lead Scoring

Presented By



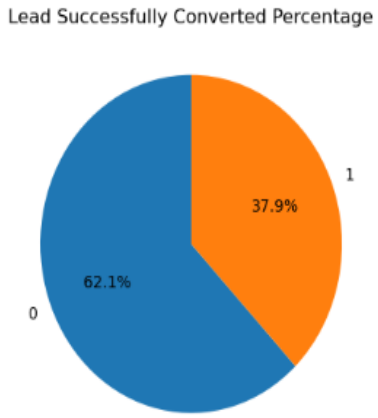
INTRODUCTION

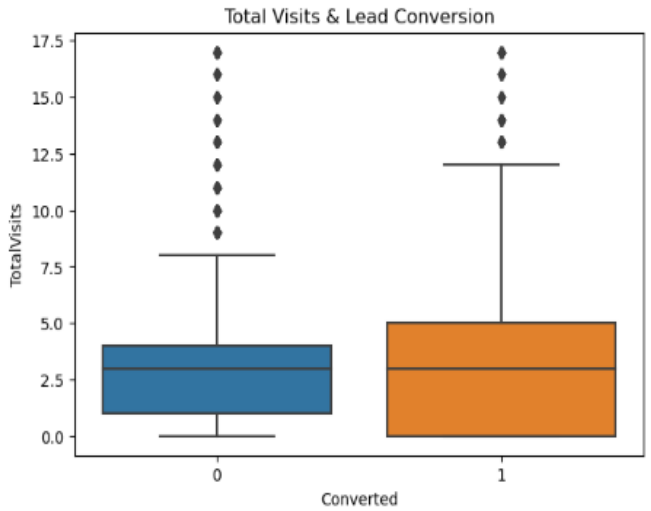
- This Case Study involves the in depth learning of behavior of the students who are tend to take admission for online courses.
- With the help of EDA & Model Building we have understood the outcomings of the case study and approach for growth in the business proposal.

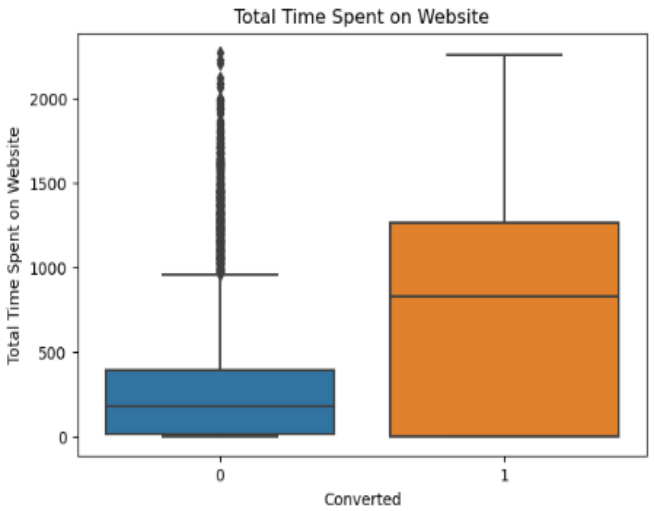
Steps of Analysis & Model Evaluation

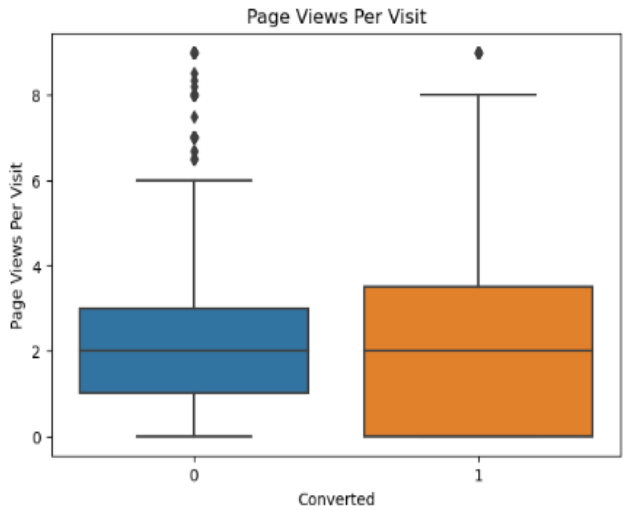
- Reading the data set
- Checking the Dataset
- Checking for missing values
- Cleaning the dataset
- Analysing the dataset with the help of EDA and different type of Charts
- Model Building
- Finding optimal cutoff
- Model evaluation
- Determining top features

EDA Analysis





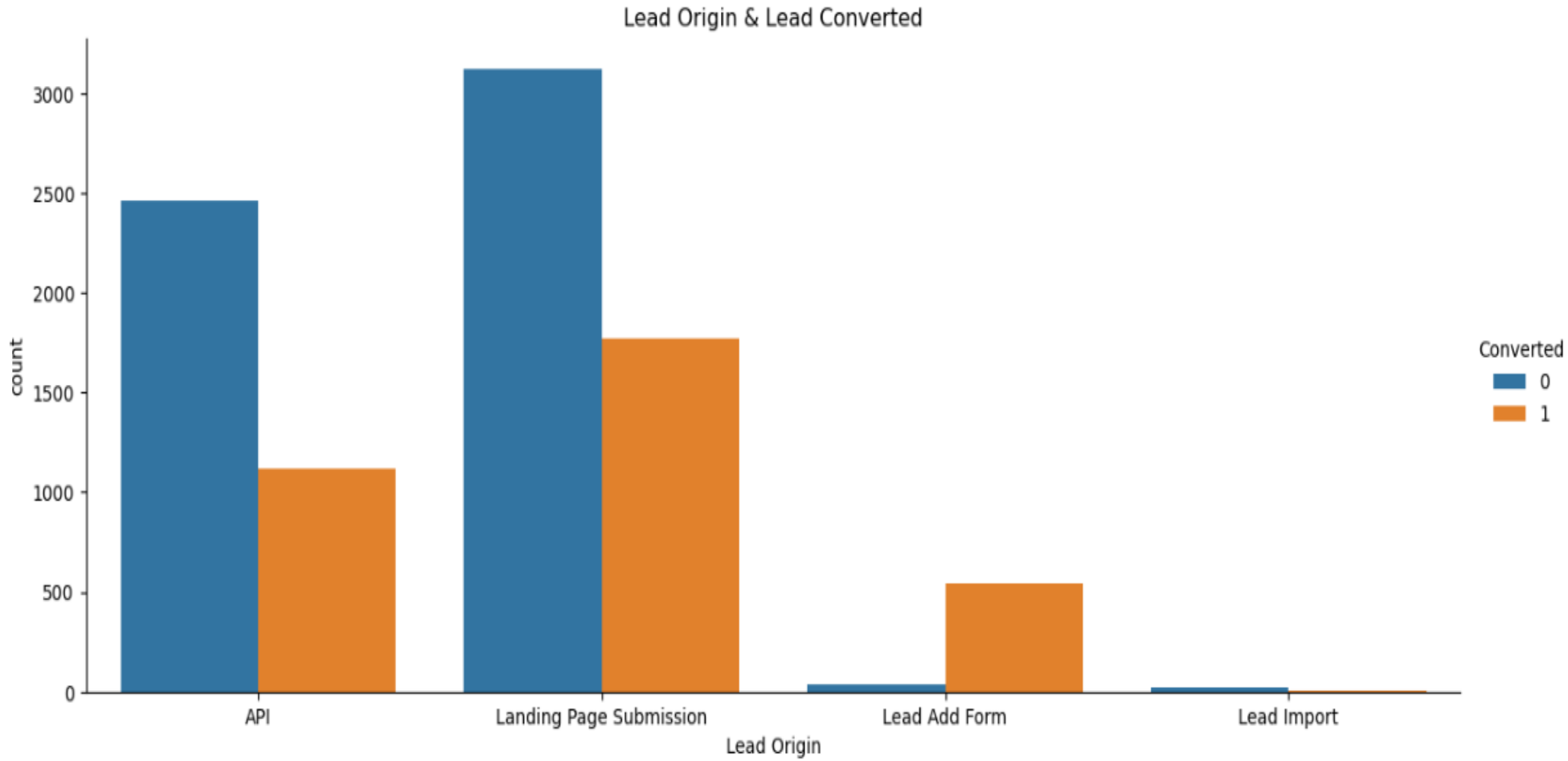




- After cleaning and visualizing the dataset, we found that there is a positive correlation between the amount of time spent visiting the website and the lead conversion rate.
- Also Lead conversion rate (denoted by 1) is about 37.9%
- The more time a person spends on the website, the higher the likelihood of conversion.

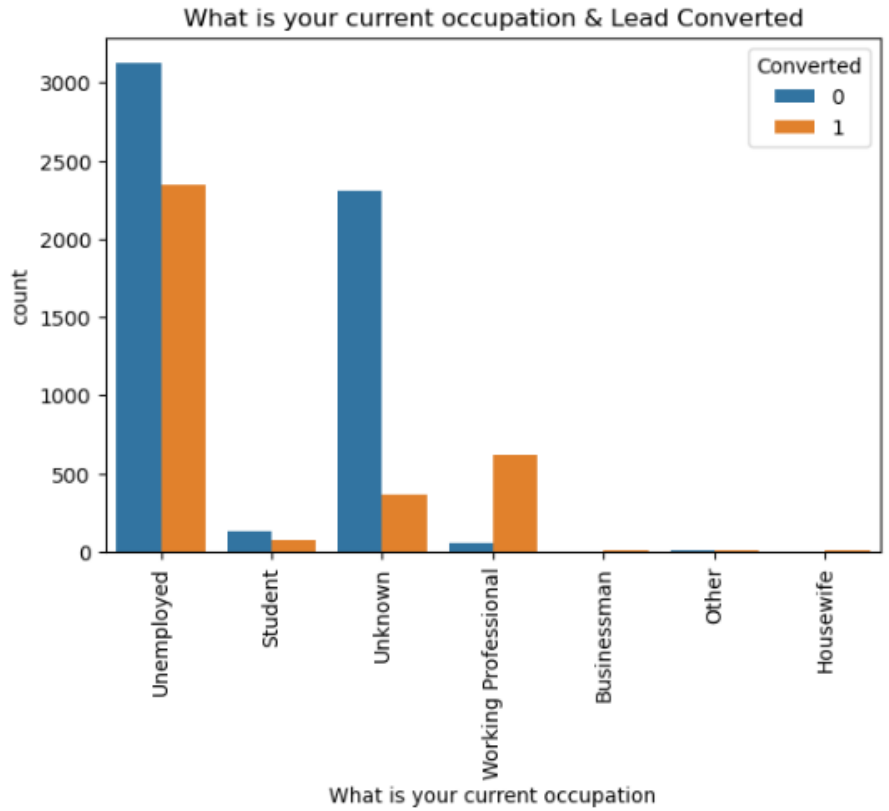
EDA Analysis

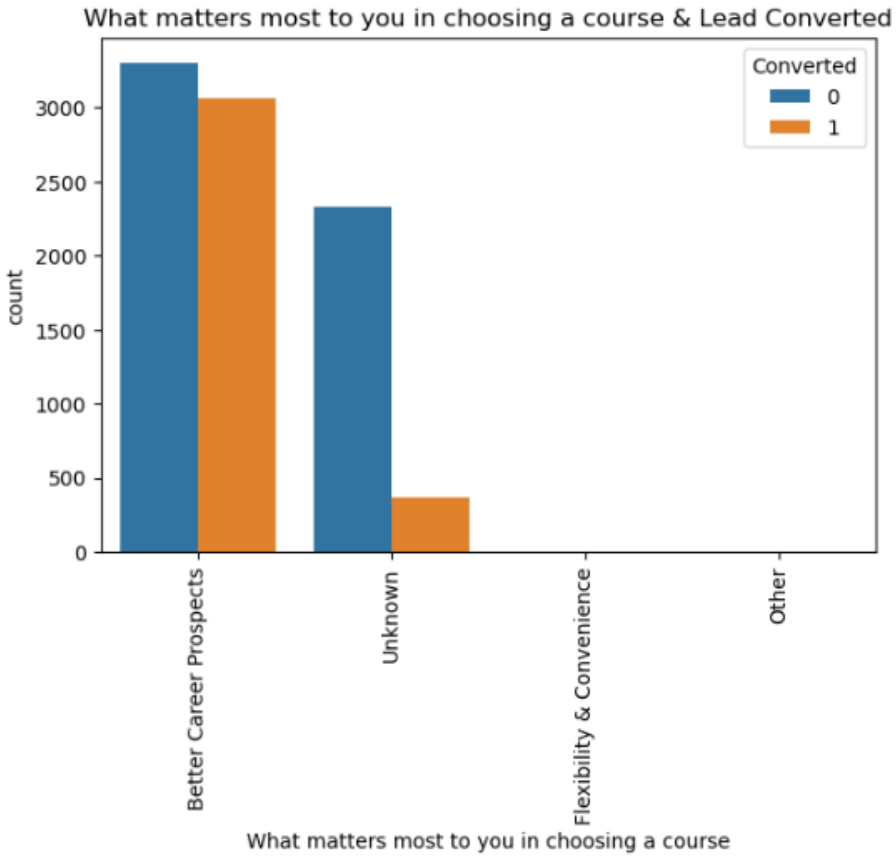
- API & Landing Page Submission having higher non-lead conversion count.
- But Lead Add Form has high conversion count.



EDA Analysis

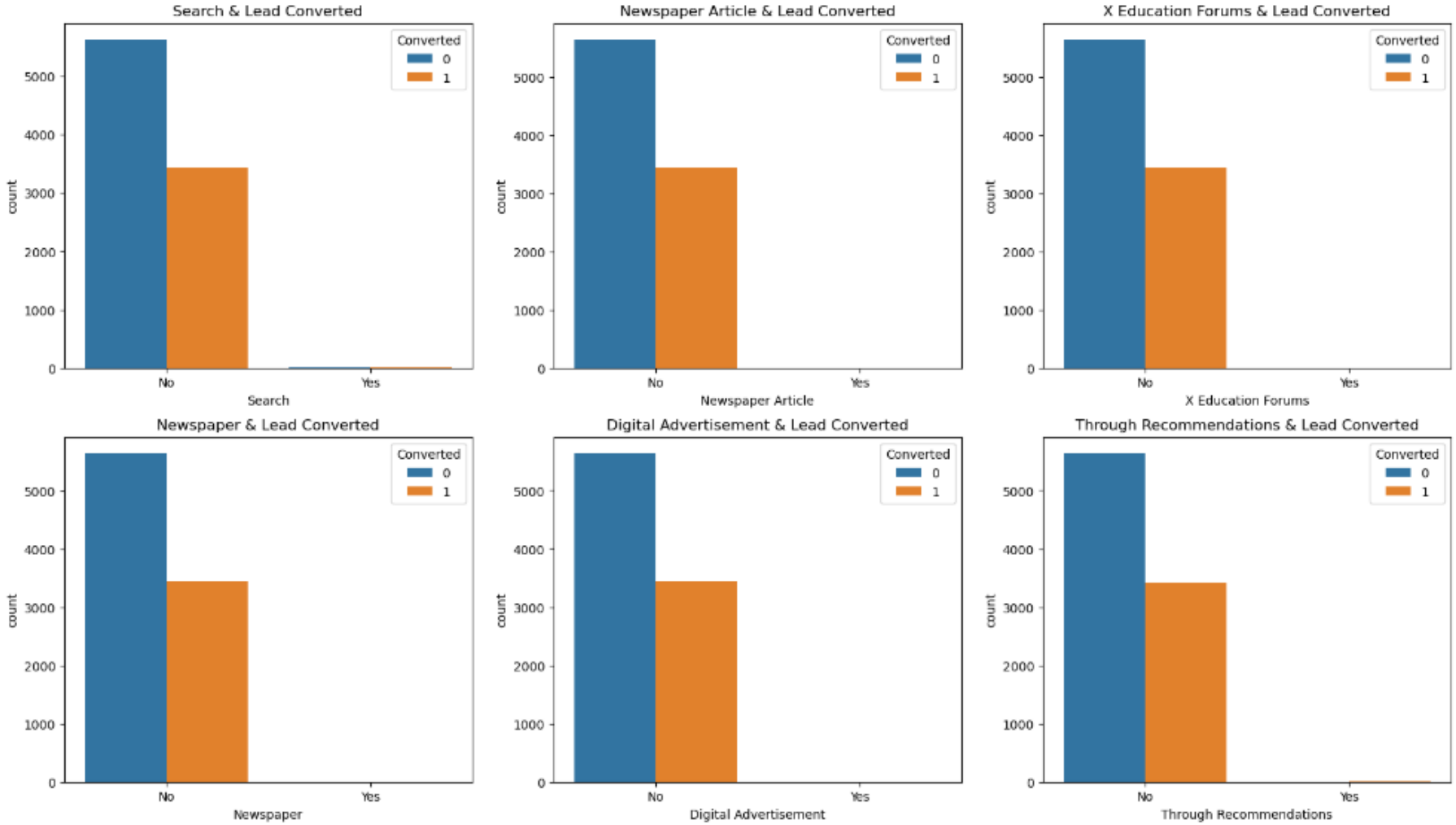
- The majority of individuals in the dataset are unemployed students.
- Despite the high count of unemployed students, working professionals show a higher propensity to enroll in courses.
- Working professionals are primarily motivated by the prospect of better career opportunities.



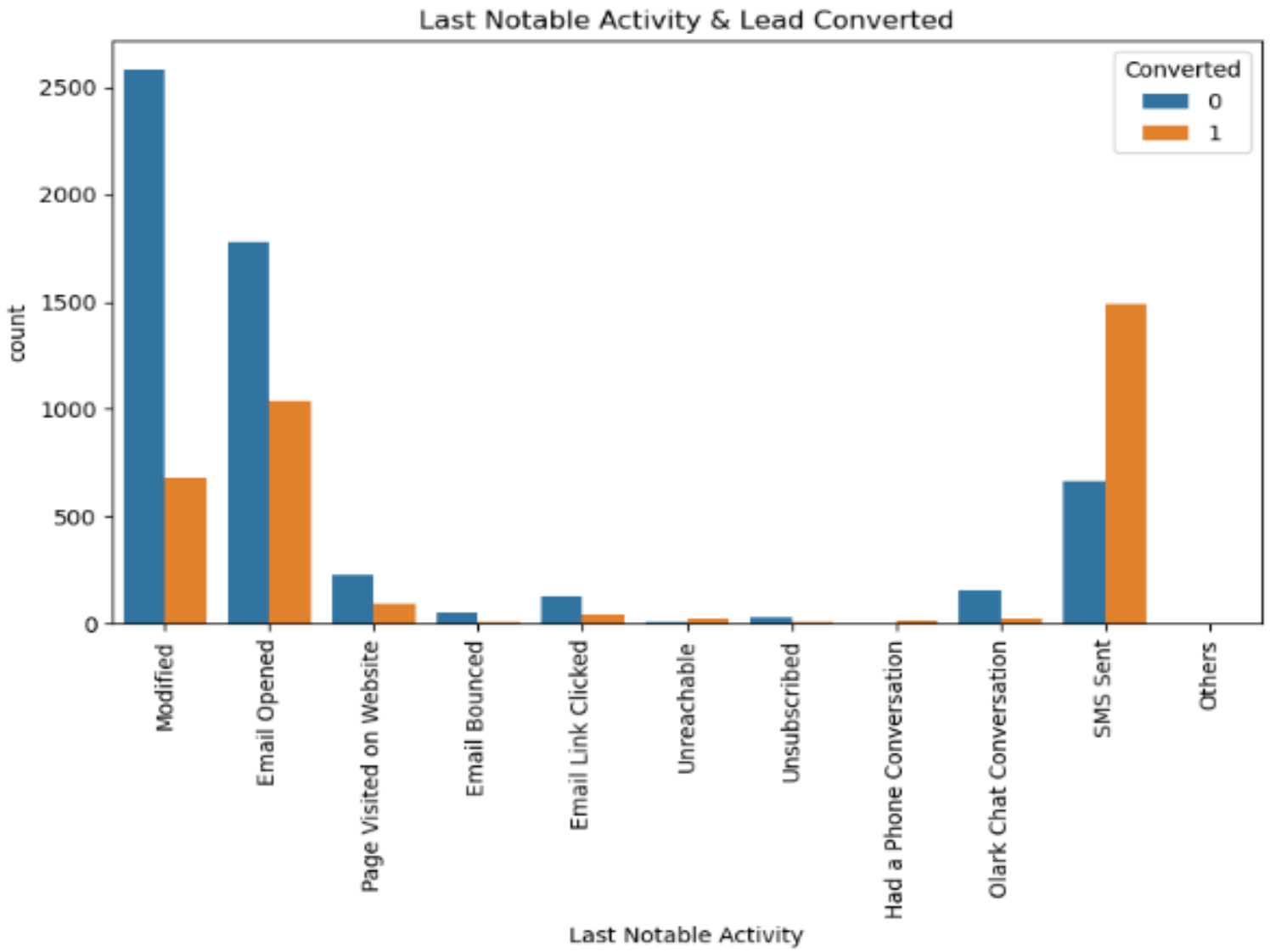


EDA Analysis

- Further analyzing the dataset, we came to know that various mediums through which Leads are converted.
- The below graphs shows us various sources the lead have searched for the program.



EDA Analysis



- After performing analysis on the Last notable activity, we found out some exciting insights.
- Higher Lead conversion rate is through SMS sent on the mobile devices.
- Below that e-mail communication has high conversion rate after SMS activity.

Model Building

- Running Logistic regression on all dataset and deleting columns one by one it will be time consuming.
- So, we using here RFE (Recursive Feature Elimination) model removes the unwanted features/columns.

	Feature	Support	Rank
0	Do Not Email	True	1
73	Last Notable Activity_SMS Sent	True	1
68	Last Notable Activity_Had a Phone Conversation	True	1
58	Lead Profile_Student of SomeSchool	True	1
54	Lead Quality_Worst	True	1
...	...	...	...
62	City_Other Metro Cities	False	58
23	Lead Source_testone	False	59
41	Specialization_Supply Chain Management	False	60
29	Specialization_Finance Management	False	61
18	Lead Source_WeLearn	False	62

Model Building

- After building Model 1 and performing VIF, we found out that VIF value for the above model is below 5.
- But p-value for the Model 1 was higher than 0.05 so we must build another model to meet our desired results.

	Feature	VIF
9	Lead Quality_Worst	1.59
12	Last Notable Activity_SMS Sent	1.59
2	Lead Origin_Lead Add Form	1.57
6	Lead Quality_Might be	1.57
10	Lead Profile_Student of SomeSchool	1.55
3	Lead Source_Olark Chat	1.49
8	Lead Quality_Not known	1.45
4	Lead Source_Welingak Website	1.36
1	Total Time Spent on Website	1.29
5	What is your current occupation_Working Profes...	1.29
0	Do Not Email	1.17
7	Lead Quality_Not Sure	1.13
14	Last Notable Activity_Unsubscribed	1.06
11	Last Notable Activity_Had a Phone Conversation	1.01
13	Last Notable Activity_Unreachable	1.00

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:	-2189.6
Date:	Tue, 23 Jul 2024	Deviance:	4379.2
Time:	11:08:27	Pearson chi2:	6.46e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4746
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.0487	0.126	8.313	0.000	0.801	1.296
Do Not Email	-1.5424	0.203	-7.609	0.000	-1.940	-1.145
Total Time Spent on Website	1.1038	0.045	24.468	0.000	1.015	1.192
Lead Origin_Lead Add Form	2.8038	0.248	11.321	0.000	2.318	3.289
Lead Source_Olark Chat	1.3014	0.113	11.498	0.000	1.080	1.523
Lead Source_Welingak Website	3.7389	0.766	4.884	0.000	2.238	5.239
What is your current occupation_Working Professional	1.7725	0.220	8.065	0.000	1.342	2.203
Lead Quality_Might be	-1.4814	0.154	-9.620	0.000	-1.783	-1.180
Lead Quality_Not Sure	-3.3797	0.168	-20.058	0.000	-3.710	-3.049
Lead Quality_Not known	-3.2085	0.138	-23.323	0.000	-3.478	-2.939
Lead Quality_Worst	-5.0934	0.409	-12.443	0.000	-5.896	-4.291
Lead Profile_Student of SomeSchool	-0.9144	0.649	-1.408	0.159	-2.187	0.358
Last Notable Activity_Had a Phone Conversation	2.6638	1.195	2.230	0.026	0.322	5.005
Last Notable Activity_SMS Sent	1.7377	0.090	19.336	0.000	1.562	1.914
Last Notable Activity_Unreachable	1.8472	0.545	3.391	0.001	0.780	2.915
Last Notable Activity_Unsubscribed	1.3586	0.640	2.124	0.034	0.105	2.612

Model Building

- After building Model 2 and performing VIF we found out that VIF value for the above model is below 5.
- Also, p-value for the Model 2 is below 0.05 so we will not build another model as we have achieved our desired results.

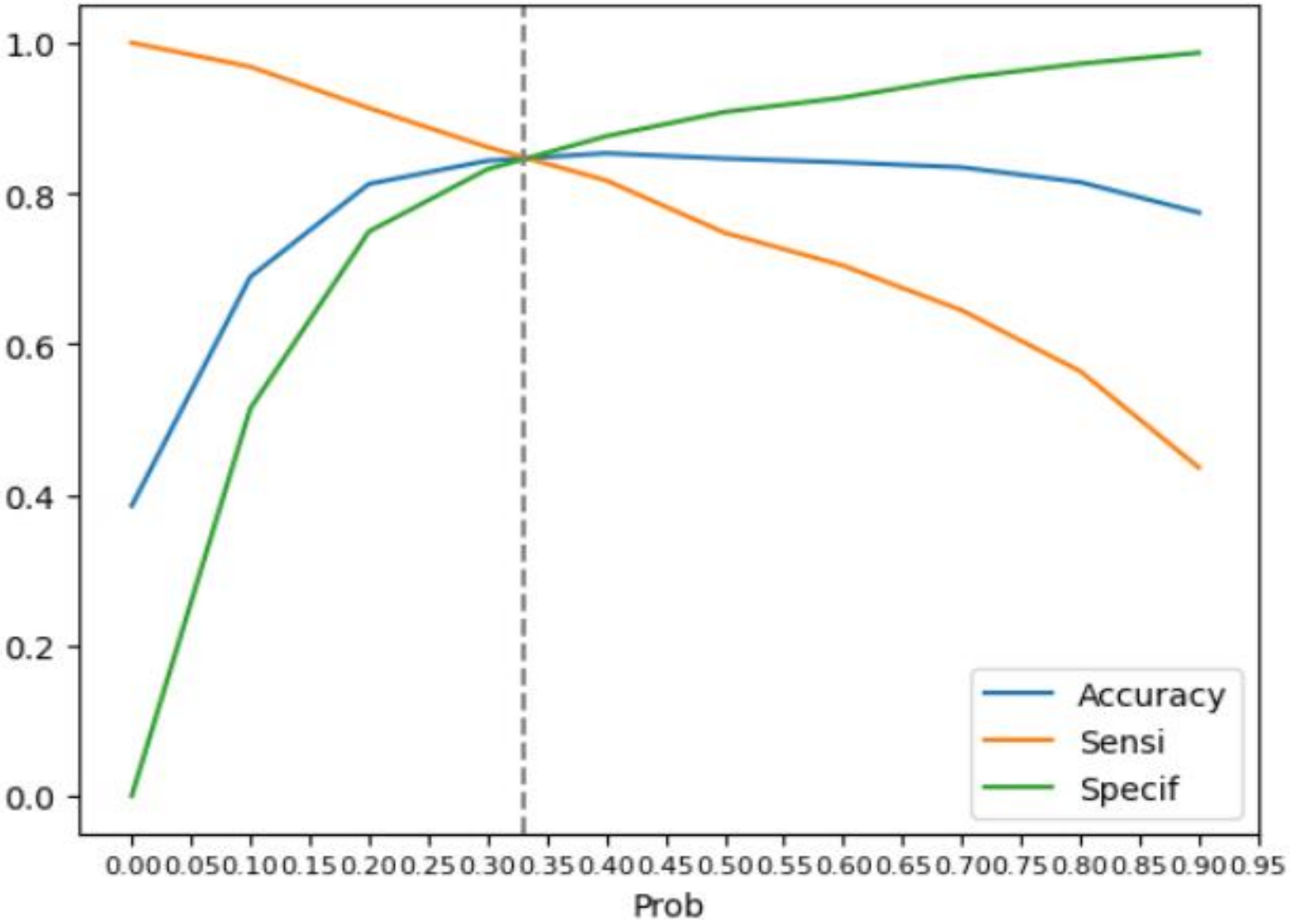
	Feature	VIF
11	Last Notable Activity_SMS Sent	1.59
2	Lead Origin_Lead Add Form	1.57
6	Lead Quality_Might be	1.57
3	Lead Source_Olark Chat	1.49
8	Lead Quality_Not known	1.45
4	Lead Source_Welingak Website	1.36
1	Total Time Spent on Website	1.29
5	What is your current occupation_Working Profes...	1.29
0	Do Not Email	1.17
7	Lead Quality_Not Sure	1.13
13	Last Notable Activity_Unsubscribed	1.06
9	Lead Quality_Worst	1.05
10	Last Notable Activity_Had a Phone Conversation	1.01
12	Last Notable Activity_Unreachable	1.00

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6336
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2190.7
Date:	Tue, 23 Jul 2024	Deviance:	4381.4
Time:	11:08:27	Pearson chi2:	6.41e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4744
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.0404	0.126	8.264	0.000	0.794	1.287
Do Not Email	-1.5401	0.203	-7.598	0.000	-1.937	-1.143
Total Time Spent on Website	1.1031	0.045	24.474	0.000	1.015	1.191
Lead Origin_Lead Add Form	2.8059	0.248	11.325	0.000	2.320	3.292
Lead Source_Olark Chat	1.2983	0.113	11.480	0.000	1.077	1.520
Lead Source_Welingak Website	3.7355	0.766	4.879	0.000	2.235	5.236
What is your current occupation_Working Professional	1.7690	0.220	8.056	0.000	1.339	2.199
Lead Quality_Might be	-1.4745	0.154	-9.590	0.000	-1.776	-1.173
Lead Quality_Not Sure	-3.3733	0.168	-20.047	0.000	-3.703	-3.043
Lead Quality_Not known	-3.2004	0.137	-23.310	0.000	-3.470	-2.931
Lead Quality_Worst	-5.3445	0.389	-13.735	0.000	-6.107	-4.582
Last Notable Activity_Had a Phone Conversation	2.6658	1.194	2.232	0.026	0.325	5.006
Last Notable Activity_SMS Sent	1.7396	0.090	19.361	0.000	1.564	1.916
Last Notable Activity_Unreachable	1.8489	0.544	3.396	0.001	0.782	2.916
Last Notable Activity_Unsubscribed	1.3594	0.639	2.127	0.033	0.107	2.612

Finding Out Optimal Cutoff



- After performing various analysis on Train dataset including parameter such as Accuracy, Sensitivity & Specific, we found out the optimal cutoff value.
- Optimal cutoff value for the above parameters is 0.33.

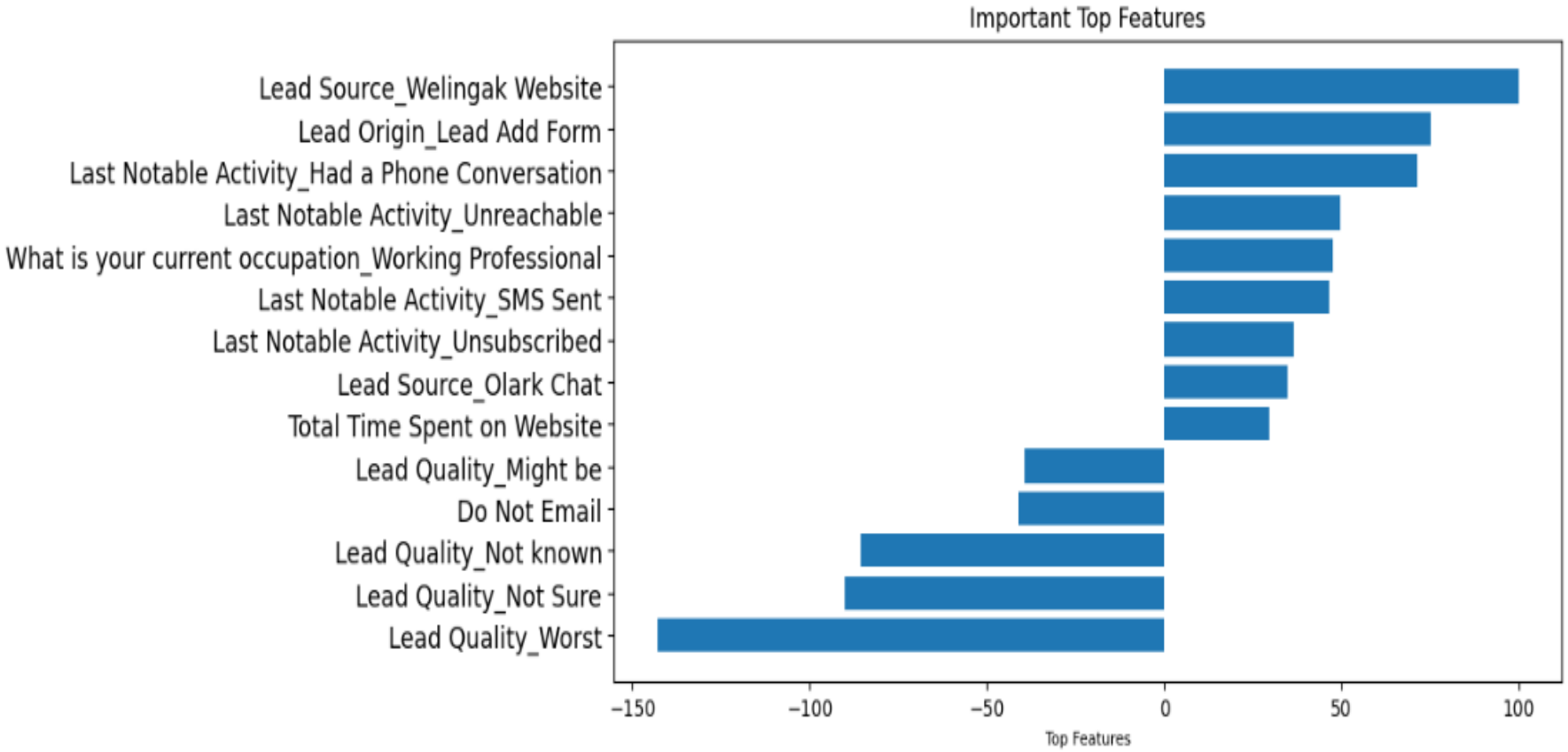
Model Evaluation & Conclusions

- After performing model evaluation on Train Dataset & Test Dataset, we found out some interesting insights.
- Sensitivity value for train data is 80%. And for train-test dataset it is 79% .
- Accuracy values are also ~80 % which shows that the model perform well on test dataset also.

Recommendations

Important features to improve lead conversion rate in X-Education company also find the graph for better understanding

- **Most recent action was via phone conversation could lead to business.**
- **Obtaining more leads from leads who have engaged with the "Lead Add Form," since they have a better likelihood of converting.**
- **Company can focus more on Lead Source Welingak website to get more leads.**
- **Company can focus on working professionals to get more numbers of leads.**
- **Last Notable Activity is also important lead.**



Thank You!!!!