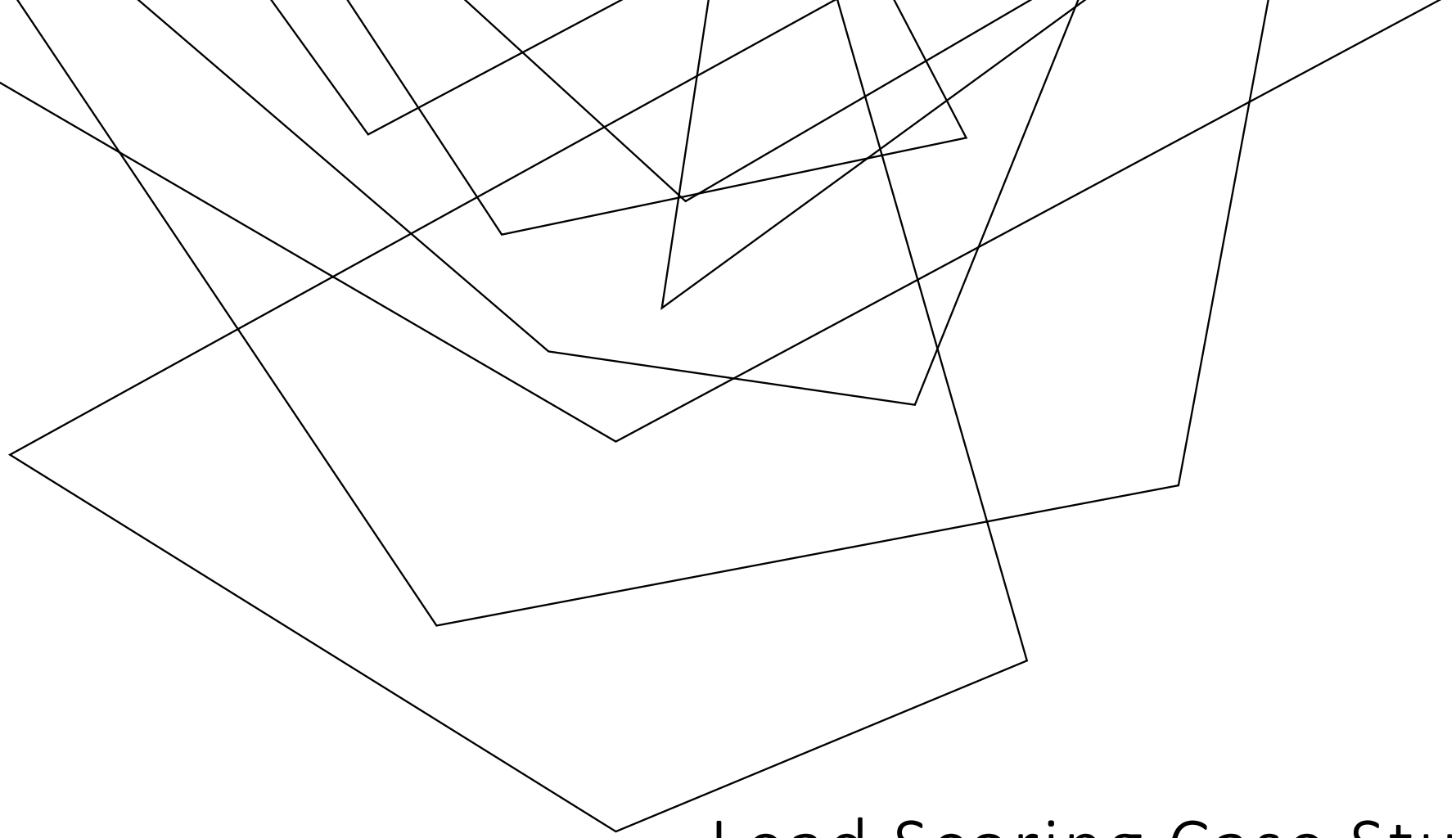


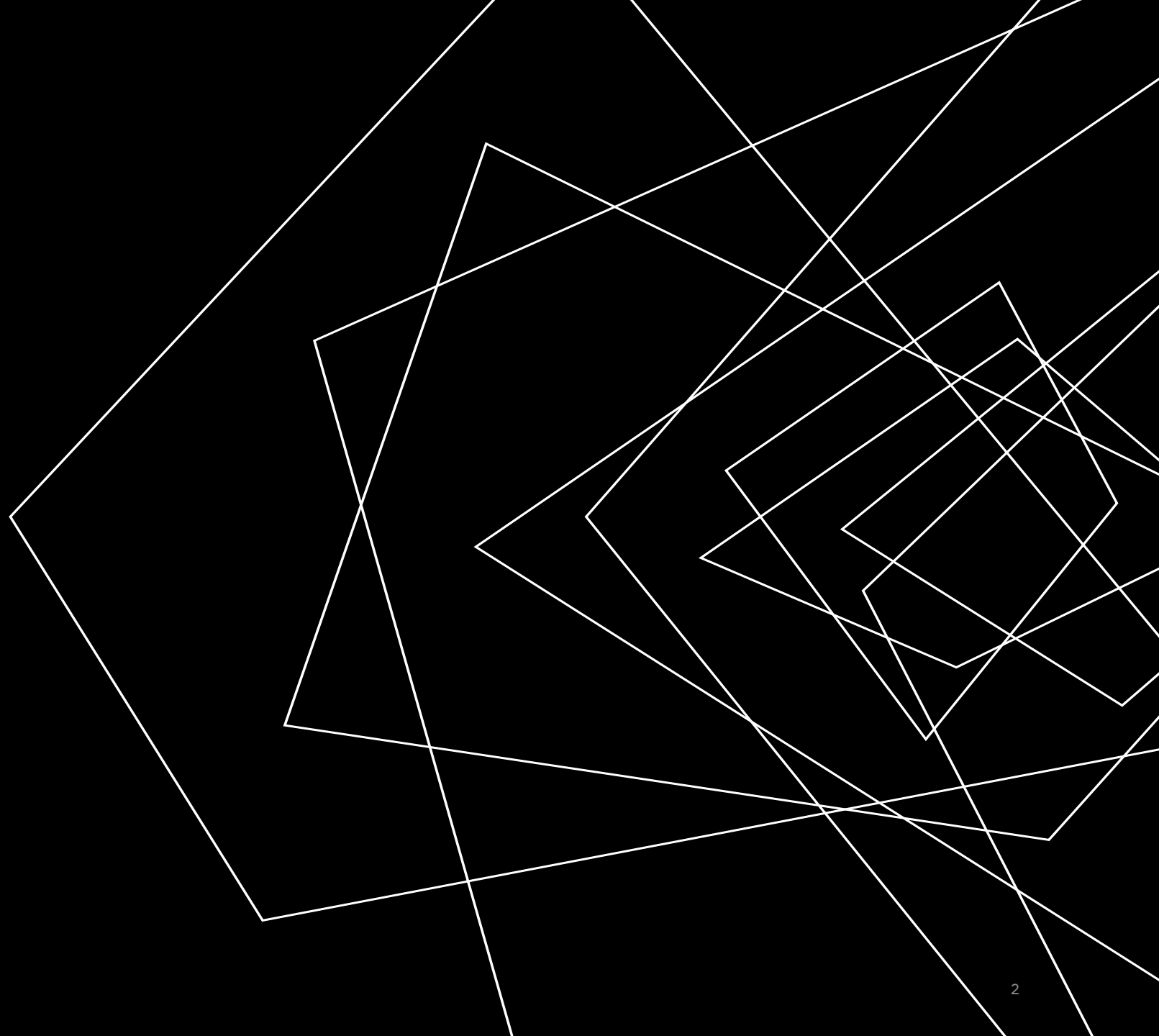
Lead Scoring Case Study

July 23, 2024

Vijay Viswanathan | Varsha K S

AGENDA

- Background and objectives
- Analysis Approach
- Executive Summary and Recommendations
- Technical results



Our understanding of the case study

Background

X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company generates leads via

- Filling form on their website
- Referrals

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education currently is around 30%. The CEO expects the target lead conversion rate to be around 80%.

To get a higher lead conversion, the company wishes to identify the most potential leads, also known as ‘Hot Leads’, so that the sales team can focus more on communicating with them rather than making calls to everyone.

Objectives

- To identify the most promising leads, i.e., the leads that are most likely to convert into paying customers

Additional Objectives

The company expects the solution model should also be able to adjust to certain changes as detailed below

- X Education has a period of 2 months every year during which the sales team get around 10 interns allotted to them. During this phase, they wish to make the lead conversion more aggressive. So, they want almost all the potential leads to be converted and hence, want to make phone calls to as much of such people as possible.

- Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So, during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e., they want to minimize the rate of useless phone calls

Analysis approach

Data understanding

- Import Necessary Libraries
- Load the data and Data dictionary

Data Preprocessing and Preparation

- Missing value treatment
- Outlier treatment
- Dummy variable creation for categorical variables

Exploratory data Analysis

- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Correlation analysis

Model building and Evaluation

- Test-train split of the data
- Feature Scaling
- Feature Selection
- Assessing the model

Results

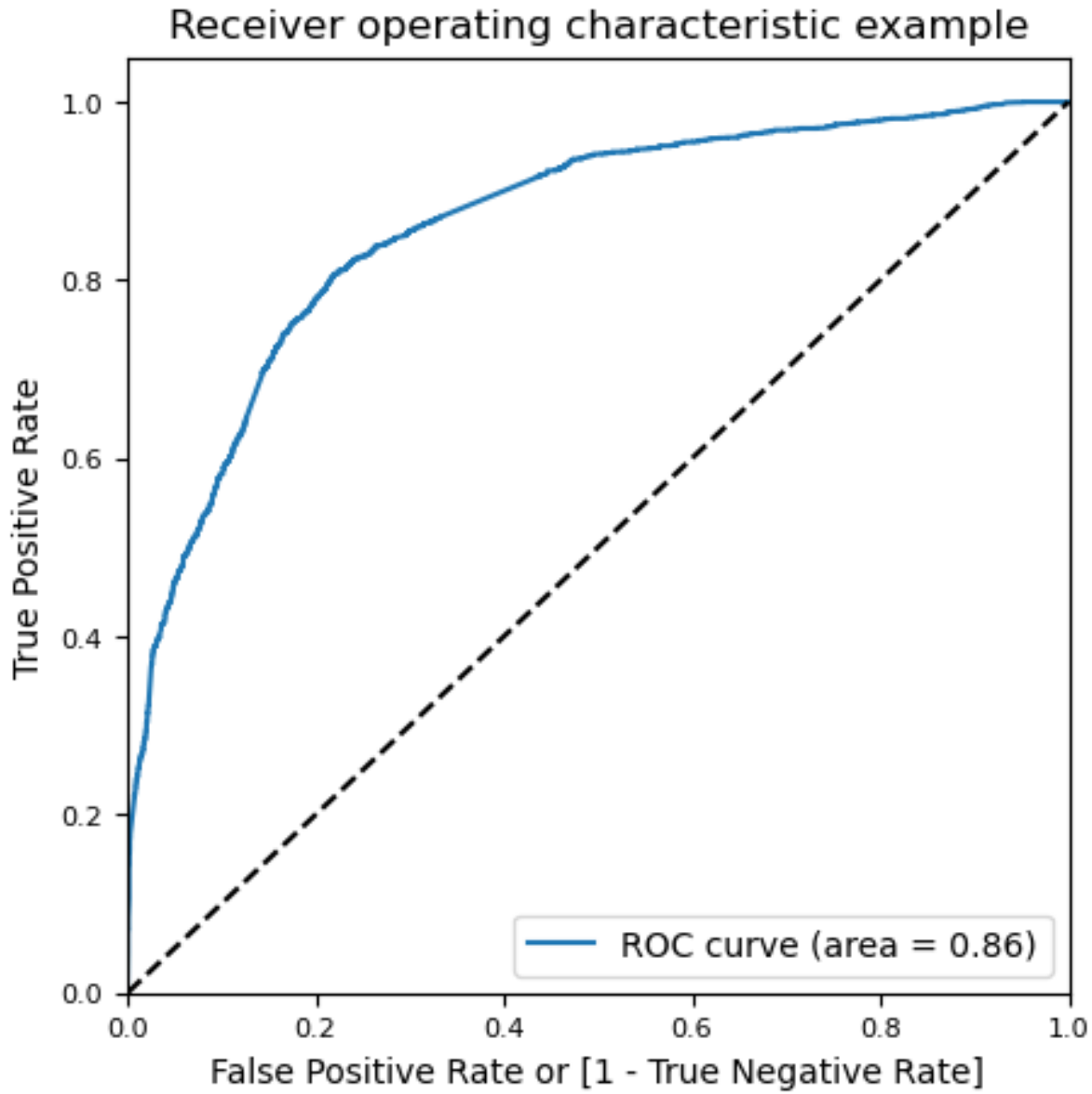
- Final model that will identify the most promising leads and

Executive Summary and Recommendations

- The model accuracy is ~80% i.e., the model is able to classify the potential lead conversion 80% of the time with an optimal cut off of 0.44%. Any lead with a probability of greater than 0.44% is deemed a hot lead.
- ‘Total Visits’, ‘Total Time spent on the website’, ‘Lead Origin Lead Add Form’ are the top three variables from the model that contribute most towards the probability of a lead getting converted. Hence the company can focus on leads who either spend more time on the website or visit it often for better conversion rates.
- Additionally leads originating from Lead Add Form or Welingak Website tend to convert at a higher rate, so company could focus on leads coming from these sources for better conversion rates
- When interns are available, threshold can be lowered for more leads. Train interns for effective communication.
- To minimize useless calls, threshold can be increased to reduce the rate of unnecessary phone calls unless it’s extremely necessary

TECHNICAL RESULTS

The area under the curve of the ROC is 0.86



Generalized Liner Model Regression Results

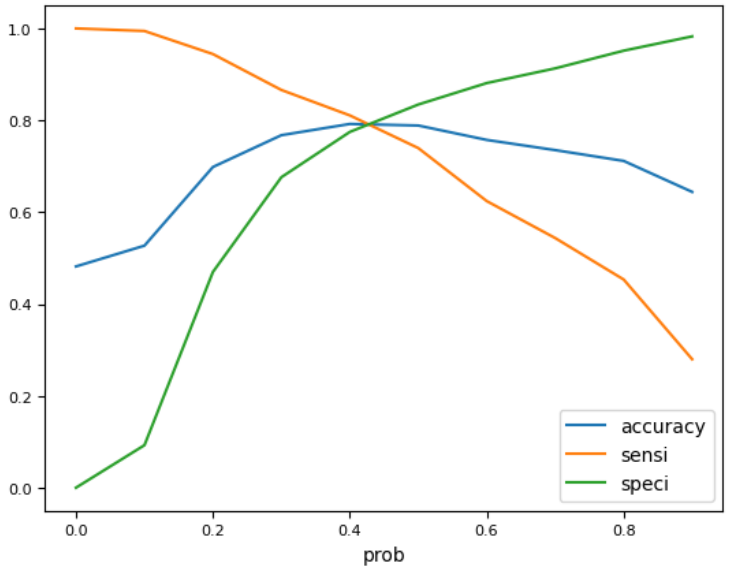
Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4449
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.1
Date:	Sun, 21 Jul 2024	**Deviance:**	4158.1
Time:	22:06:07	**Pearson chi2:**	4.80e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3642
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

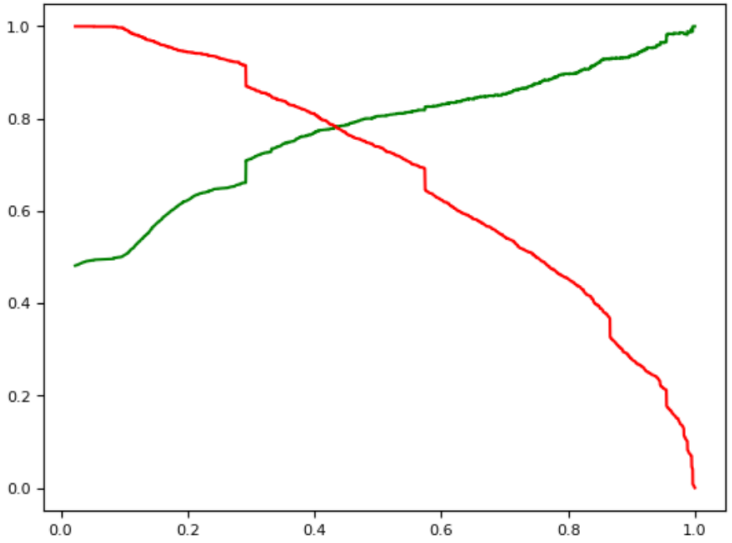
The optimal cutoff obtained was 0.44

Prob Vs accuracy |sensitivity | specificity

	prob	accuracy	sensi	speci
0.0	0.0	0.481731	1.000000	0.000000
0.1	0.1	0.527012	0.994416	0.092561
0.2	0.2	0.698274	0.944160	0.469723
0.3	0.3	0.767541	0.865984	0.676038
0.4	0.4	0.791975	0.810610	0.774654
0.5	0.5	0.788612	0.739414	0.834343
0.6	0.6	0.757229	0.624011	0.881055
0.7	0.7	0.735037	0.543509	0.913062
0.8	0.8	0.711500	0.453234	0.951557
0.9	0.9	0.644026	0.279665	0.982699



Precision-Recall Curve

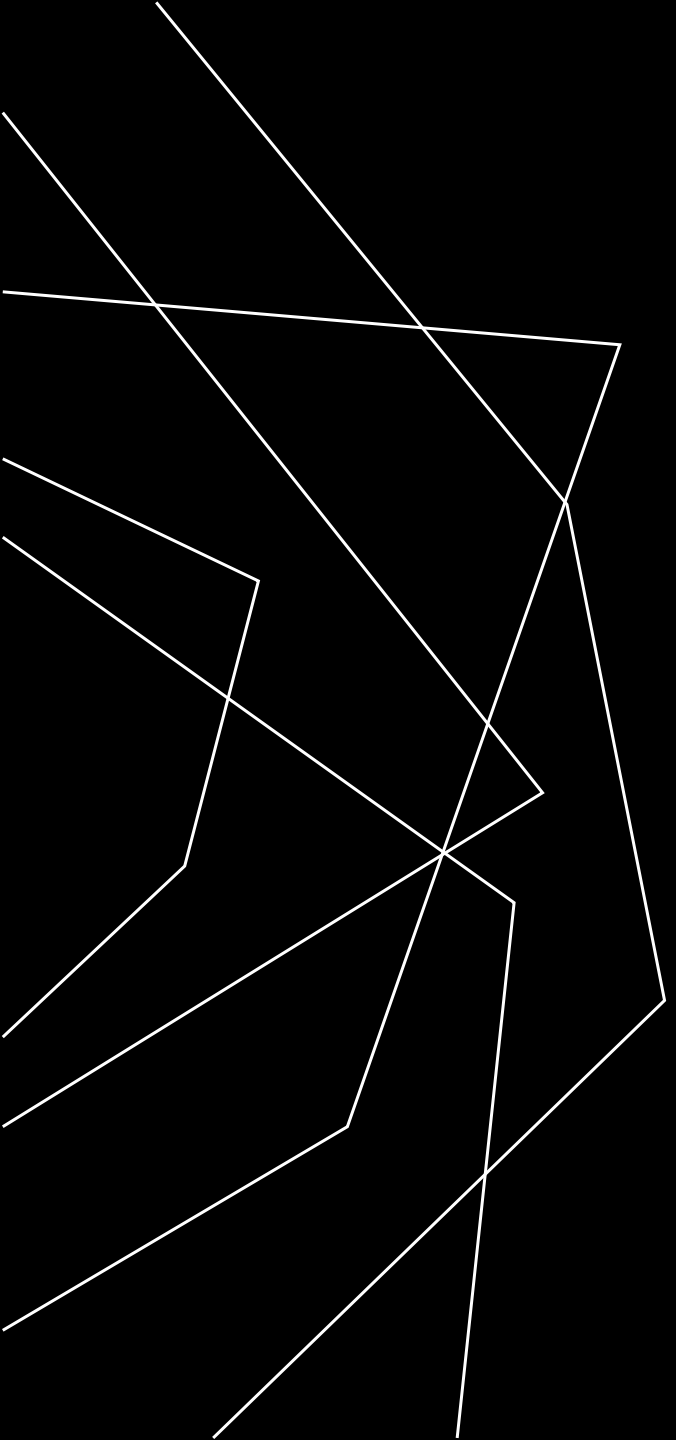


The following features appeared significant in predicting the lead conversion as per the model

Features
• What is your current occupation_Unemployed
• Total Time Spent on Website
• TotalVisits
• Last Activity_SMS Sent
• Lead Origin_Lead Add Form
• Lead Source_Olark Chat
• Lead Source_Welingak Website
• Do Not Email_Yes
• What is your current occupation_Student
• Last Activity_Had a Phone Conversation
• Last Notable Activity_Unreachable

Model Evaluation scores

Data	Overall Accuracy	Sensitivity	Specificity	Precision	Recall
Train data	79.08%	79.3%	78.8%	78.4%	77.7%
Test data	78.45%	77.9%	78.9%	78.2%	76.7%



THANK YOU