LEAD SCORING CASE STUDY

Submitted by:
Arun, Prabanjan, Srushti

BUSINESS PROBLEM STATEMENT

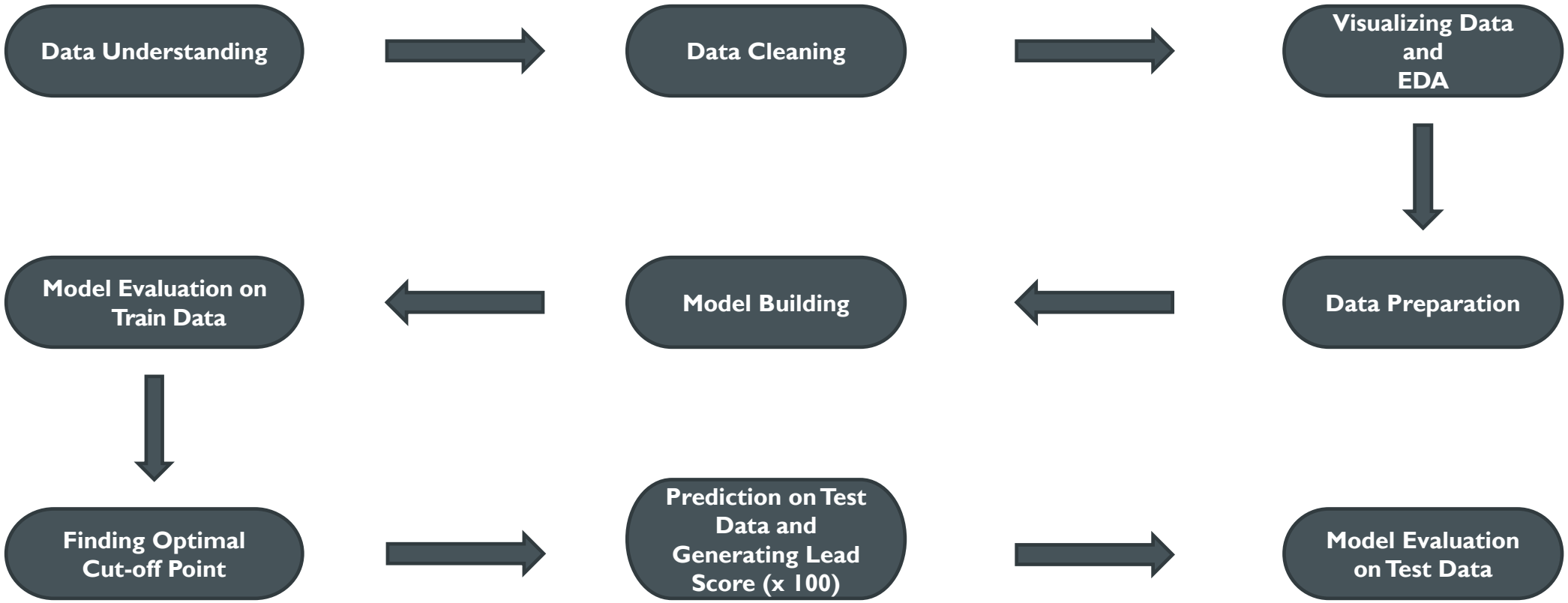
Business Problem Statement:

An education company named 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.

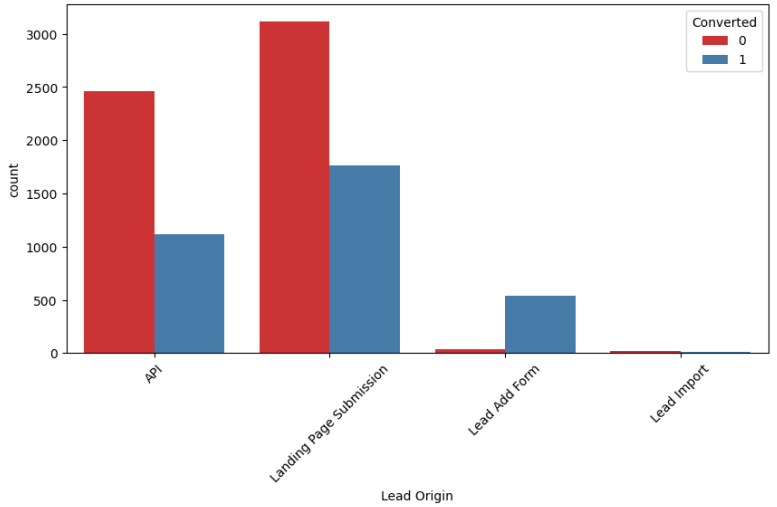
The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past 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 is around 30%.

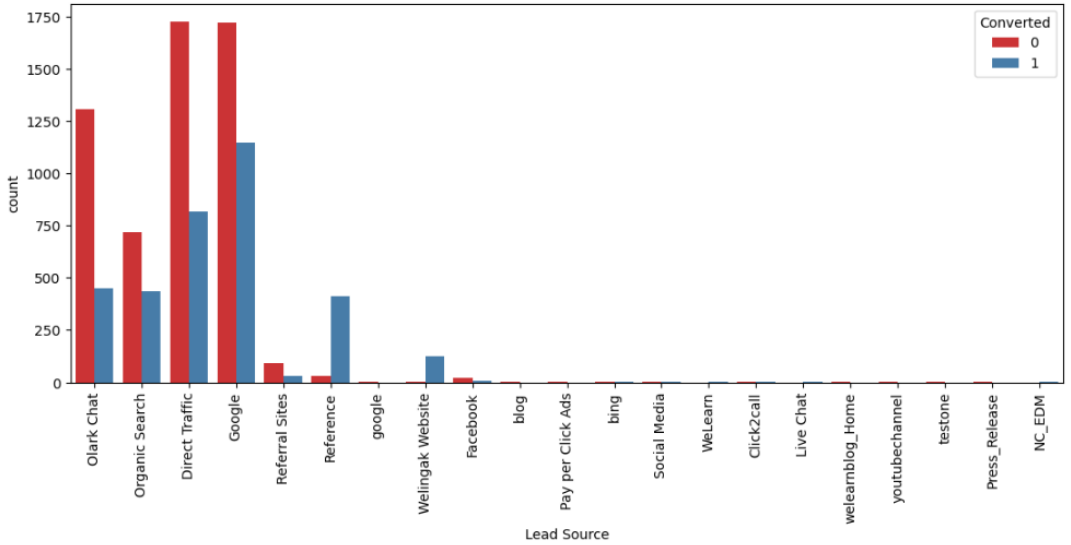
Goal: Building a logistic regression model to identify the Hot Leads also generating a column that contains Lead Score (Probability value 100). It'll help the business in achieving Higher Lead Conversion Rate.

COMPLETE FRAMEWORK



Visualizing Data and EDA : Categorical Variables I

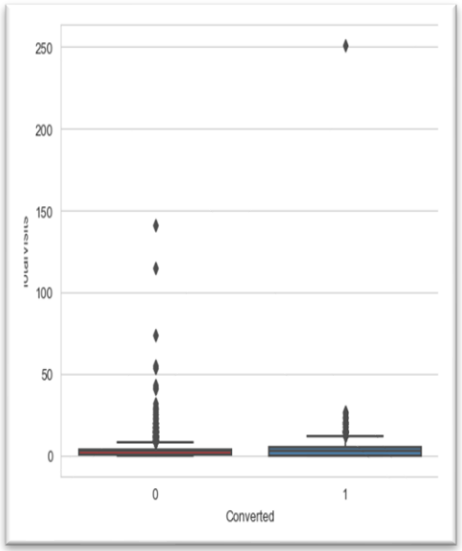


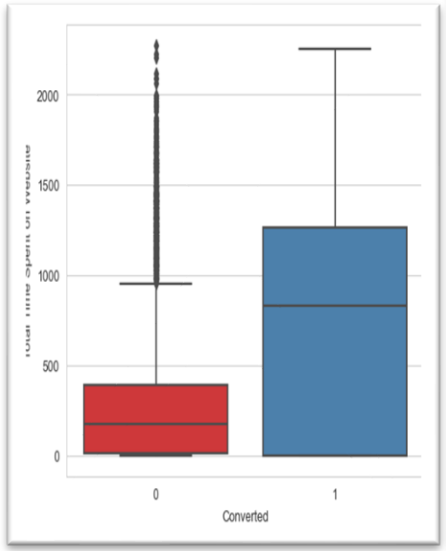


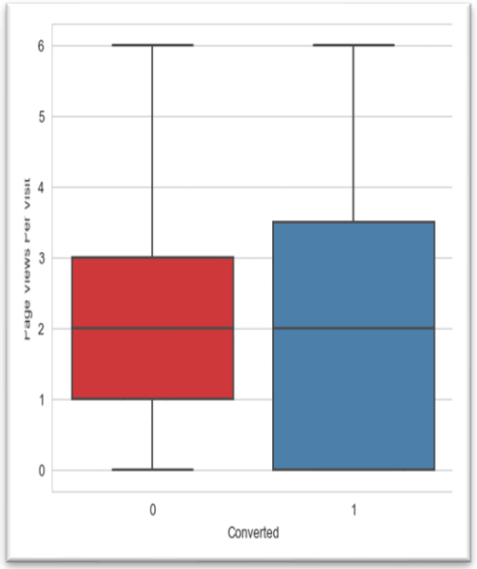
A. Lead Origin:
1. API and Landing Page Submission demonstrate a conversion rate of 30-35%, with a significant number of originated leads.
2. Lead Add Form shows a remarkably high conversion rate exceeding 90%, albeit with a relatively low lead count.
3. Lead Import contributes very few leads.
Enhancing the overall lead conversion rate entails prioritizing improvements in the conversion rates of API and Landing Page Submission, alongside increasing lead generation from the Lead Add Form.

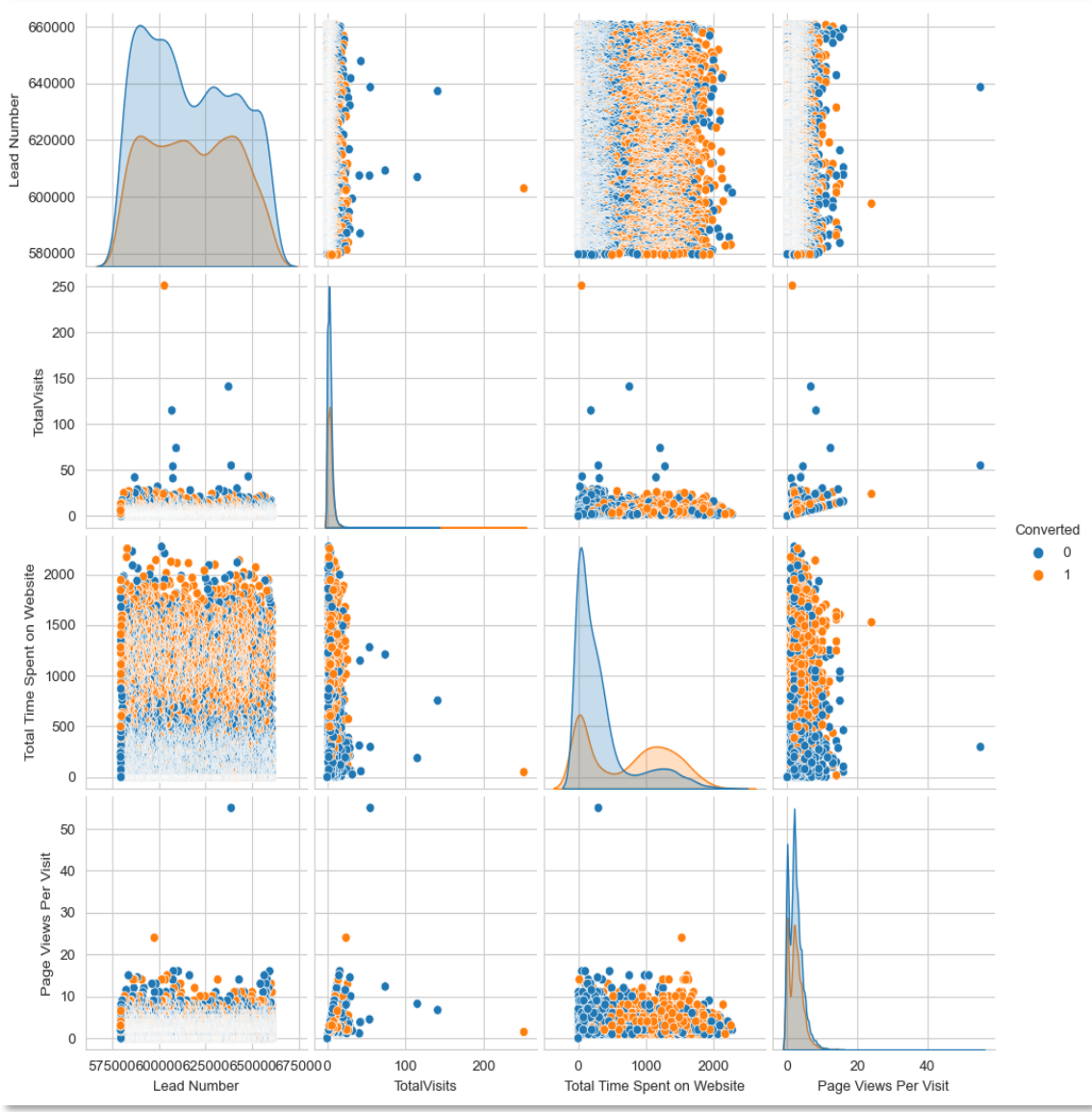
B. Lead Source:
1. Google and direct traffic are the primary sources generating the maximum number of leads.
2. Reference leads and leads through the Welingak website exhibit high conversion rates.
To improve the overall lead conversion rate, the focus should be on enhancing lead conversion from Olark Chat, organic search, direct traffic, and Google leads. Additionally, efforts should aim to generate more leads from reference sources and the Welingak website.

Visualizing Data and EDA : Numerical variables









Inferences:

1. The median value of Total Time Spent on Website is significantly higher for converted leads compared to the other group. The team should focus on targeting customers who spend more time on the website, as these leads have a higher likelihood of conversion
2. Total Time Spent on Website:

Leads spending more time on the website are more likely to be converted. To capitalize on this trend, efforts should focus on enhancing the website's engagement factors to encourage leads to spend more time.

3. Last Activity:

- 1. The majority of leads have their last activity recorded as Email opened.
- 2. Leads with their last activity recorded as SMS Sent exhibit an impressive conversion rate of nearly 60%.

4. City:

The majority of leads are from Mumbai, with approximately a 50% conversion rate.

Data Preparation

Data Preparation

03. Missing Value Imputation

Nominal categorical columns were imputed using the mode.

02. Train-test Split

The dataset has been divided into Train and Test sets in a 70:30 ratio. The Train dataset is utilized for training the model, while the Test dataset is used for evaluating the model's performance.

01. Outlier treatment

The team address outliers by replacing values below or equal to the 5th percentile with the 5th percentile value itself, and values above or equal to the 95th percentile with the 95th percentile value.

04. Categorical Variables Encoding

Categorical columns with 'Yes' and 'No' values have been encoded as 1 and 0, respectively. For columns with more than two categories, dummy variables were created, and the original column along with the first dummy variable for each column were dropped from the data-frame.

05. Scaling Features

Selected numerical variables were scaled using the Standard Scaler.

1. ** Prior to the Exploratory Data Analysis phase, we conducted basic data cleanup tasks, which included the deletion of columns with blank values exceeding 40%, imputation of missing values with mode values, and renaming of values where necessary.

Model Building : Approach

Model Building : Approach

1. Recursive Feature Elimination (RFE) has been used to get top 20 features

- Do Not Email
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Lead Origin_Lead Add Form
- Lead Origin_Lead Import
- Lead Source_Olark Chat
- Lead Source_Reference
- Lead Source_Welingak Website
- Last Activity_Olark Chat Conversation
- Last Activity_Other_Activity
- Last Activity_SMS Sent
- Last Activity_Unsubscribed
- Specialization_Others
- What is your current occupation_Housewife
- What is your current occupation_Student
- What is your current occupation_Unemployed
- What is your current occupation_Working Professional
- Last Notable Activity_Had a Phone Conversation
- Last Notable Activity_Modified
- Last Notable Activity_Unreachable

02

Team initially built a Logistic Regression model using sklearn's linear model with these 20 features.

Team then manually fine-tuned the model to ensure statistical significance by checking p-values (accepted if less than 0.05) and removed multicollinearity by examining Variance Inflation Factors (VIFs, accepted if less than 5).

03

04

A total of 9 models were built. After each iteration, we checked the p-values of all beta coefficients and VIFs, removing identified features in the subsequent model. Additionally, we evaluated the overall model accuracy and Confusion Matrix after each new model to compare performance with the previous version.

Model 1

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.8338	0.637	1.309	0.190	-0.414	2.082
Do Not Email	-1.6759	0.191	-8.796	0.000	-2.049	-1.302
Total Time Spent on Website	1.1081	0.041	27.194	0.000	1.028	1.188
Lead Origin_Landing Page Submission	-1.1219	0.130	-8.663	0.000	-1.376	-0.868
Lead Origin_Lead Add Form	1.6019	0.915	1.751	0.080	-0.191	3.395
Lead Origin_Lead Import	0.9059	0.480	1.888	0.059	-0.035	1.846
Lead Source_Olark Chat	1.1250	0.124	9.082	0.000	0.882	1.368
Lead Source_Reference	1.7697	0.938	1.887	0.059	-0.069	3.608
Lead Source_Welingak Website	4.2961	1.165	3.687	0.000	2.012	6.580
Last Activity_Olark Chat Conversation	-0.9504	0.172	-5.531	0.000	-1.287	-0.614
Last Activity_Other_Activity	1.8717	0.537	3.483	0.000	0.818	2.925
Last Activity_SMS Sent	1.3454	0.076	17.766	0.000	1.197	1.494
Last Activity_Unsubscribed	1.4083	0.483	2.917	0.004	0.462	2.355
Specialization_Others	-1.1410	0.126	-9.052	0.000	-1.388	-0.894
What is your current occupation_Housewife	21.7588	1.53e+04	0.001	0.999	-2.99e+04	2.99e+04
What is your current occupation_Student	-0.5518	0.673	-0.820	0.412	-1.871	0.767
What is your current occupation_Unemployed	-1.0059	0.634	-1.587	0.113	-2.248	0.236
What is your current occupation_Working Professional	1.6281	0.660	2.466	0.014	0.334	2.922
Last Notable Activity_Had a Phone Conversation	1.4204	1.223	1.161	0.246	-0.978	3.818
Last Notable Activity_Modified	-0.8675	0.082	-10.620	0.000	-1.028	-0.707
Last Notable Activity_Unreachable	1.5785	0.476	3.316	0.001	0.645	2.512

- Since Pvalue of 'What is your current occupation_Housewife' is very high, we can drop this column.

Model-4

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4409	0.240	1.836	0.066	-0.030	0.912
Do Not Email	-1.6789	0.191	-8.807	0.000	-2.053	-1.305
Total Time Spent on Website	1.1067	0.041	27.196	0.000	1.027	1.186
Lead Origin_Landing Page Submission	-1.1290	0.129	-8.745	0.000	-1.382	-0.876
Lead Origin_Lead Add Form	1.5974	0.914	1.747	0.081	-0.195	3.390
Lead Origin_Lead Import	0.8993	0.480	1.874	0.061	-0.041	1.840
Lead Source_Olark Chat	1.1178	0.124	9.029	0.000	0.875	1.360
Lead Source_Reference	1.7790	0.938	1.897	0.058	-0.059	3.617
Lead Source_Welingak Website	4.3023	1.165	3.693	0.000	2.019	6.586
Last Activity_Olark Chat Conversation	-0.9478	0.172	-5.518	0.000	-1.284	-0.611
Last Activity_Other_Activity	2.2295	0.463	4.816	0.000	1.322	3.137
Last Activity_SMS Sent	1.3427	0.076	17.728	0.000	1.194	1.491
Last Activity_Unsubscribed	1.4093	0.483	2.919	0.004	0.463	2.356
Specialization_Others	-1.1534	0.126	-9.171	0.000	-1.400	-0.907
What is your current occupation_Unemployed	-0.6003	0.213	-2.818	0.005	-1.018	-0.183
What is your current occupation_Working Professional	2.0282	0.283	7.161	0.000	1.473	2.583
Last Notable Activity_Modified	-0.8740	0.081	-10.725	0.000	-1.034	-0.714
Last Notable Activity_Unreachable	1.5774	0.475	3.318	0.001	0.646	2.509

- After removing the columns "Last Notable Activity_Had a Phone Conversation" and "What is your current occupation_Student" due to their high p-values, we proceed to model number 4.
- In model 4, we decided to remove the feature "What is your current occupation_Student" due to its high p-value.

Model-5

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4578	0.240	1.907	0.056	-0.013	0.928
Do Not Email	-1.6806	0.191	-8.816	0.000	-2.054	-1.307
Total Time Spent on Website	1.1047	0.041	27.190	0.000	1.025	1.184
Lead Origin_Landing Page Submission	-1.1473	0.129	-8.907	0.000	-1.400	-0.895
Lead Origin_Lead Import	0.8826	0.480	1.838	0.066	-0.059	1.824
Lead Source_Olark Chat	1.1108	0.124	8.993	0.000	0.869	1.353
Lead Source_Reference	3.3614	0.243	13.840	0.000	2.885	3.837
Lead Source_Welingak Website	5.8902	0.730	8.073	0.000	4.460	7.320
Last Activity_Olark Chat Conversation	-0.9522	0.172	-5.544	0.000	-1.289	-0.616
Last Activity_Other_Activity	2.2254	0.463	4.808	0.000	1.318	3.133
Last Activity_SMS Sent	1.3427	0.076	17.732	0.000	1.194	1.491
Last Activity_Unsubscribed	1.4077	0.483	2.916	0.004	0.462	2.354
Specialization_Others	-1.1652	0.126	-9.273	0.000	-1.411	-0.919
What is your current occupation_Unemployed	-0.5974	0.213	-2.804	0.005	-1.015	-0.180
What is your current occupation_Working Professional	2.0280	0.283	7.158	0.000	1.473	2.583
Last Notable Activity_Modified	-0.8745	0.081	-10.736	0.000	-1.034	-0.715
Last Notable Activity_Unreachable	1.5728	0.475	3.308	0.001	0.641	2.505

	Features	VIF
12	What is your current occupation_Unemployed	9.72
2	Lead Origin_Landing Page Submission	5.74
11	Specialization_Others	3.99
4	Lead Source_Olark Chat	2.24
14	Last Notable Activity_Modified	1.86
13	What is your current occupation_Working Profes...	1.66
9	Last Activity_SMS Sent	1.63
7	Last Activity_Olark Chat Conversation	1.59
5	Lead Source_Reference	1.46
1	Total Time Spent on Website	1.32
0	Do Not Email	1.21
6	Lead Source_Welingak Website	1.11
10	Last Activity_Unsubscribed	1.08

- Dropping the column 'What is your current occupation_Unemployed' because it has high VIF

Model-6

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1106	0.127	-0.868	0.385	-0.361	0.139
Do Not Email	-1.6767	0.191	-8.786	0.000	-2.051	-1.303
Total Time Spent on Website	1.1047	0.041	27.207	0.000	1.025	1.184
Lead Origin_Landing Page Submission	-1.1519	0.129	-8.935	0.000	-1.405	-0.899
Lead Origin_Lead Import	0.8640	0.480	1.799	0.072	-0.077	1.805
Lead Source_Olark Chat	1.1164	0.124	9.037	0.000	0.874	1.359
Lead Source_Reference	3.3731	0.243	13.906	0.000	2.898	3.848
Lead Source_Welingak Website	5.8819	0.730	8.063	0.000	4.452	7.312
Last Activity_Olark Chat Conversation	-0.9437	0.172	-5.502	0.000	-1.280	-0.608
Last Activity_Other_Activity	2.2075	0.463	4.767	0.000	1.300	3.115
Last Activity_SMS Sent	1.3276	0.075	17.609	0.000	1.180	1.475
Last Activity_Unsubscribed	1.3822	0.483	2.863	0.004	0.436	2.328
Specialization_Others	-1.1774	0.126	-9.356	0.000	-1.424	-0.931
What is your current occupation_Working Professional	2.6063	0.195	13.382	0.000	2.225	2.988
Last Notable Activity_Modified	-0.8814	0.081	-10.826	0.000	-1.041	-0.722
Last Notable Activity_Unreachable	1.5571	0.474	3.284	0.001	0.628	2.486

- Dropping the column 'Lead Origin_Lead Import' because it has high Pvalue

Model-7

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0717	0.126	-0.570	0.569	-0.318	0.175
Do Not Email	-1.6783	0.191	-8.798	0.000	-2.052	-1.304
Total Time Spent on Website	1.0976	0.040	27.211	0.000	1.019	1.177
Lead Origin_Landing Page Submission	-1.1863	0.128	-9.291	0.000	-1.437	-0.936
Lead Source_Olark Chat	1.0915	0.123	8.905	0.000	0.851	1.332
Lead Source_Reference	3.3401	0.242	13.812	0.000	2.866	3.814
Lead Source_Welingak Website	5.8588	0.729	8.033	0.000	4.429	7.288
Last Activity_Olark Chat Conversation	-0.9485	0.171	-5.531	0.000	-1.285	-0.612
Last Activity_Other_Activity	2.1988	0.463	4.752	0.000	1.292	3.106
Last Activity_SMS Sent	1.3250	0.075	17.587	0.000	1.177	1.473
Last Activity_Unsubscribed	1.3784	0.482	2.858	0.004	0.433	2.324
Specialization_Others	-1.1983	0.126	-9.536	0.000	-1.445	-0.952
What is your current occupation_Working Professional	2.6064	0.195	13.389	0.000	2.225	2.988
Last Notable Activity_Modified	-0.8816	0.081	-10.833	0.000	-1.041	-0.722
Last Notable Activity_Unreachable	1.5470	0.474	3.264	0.001	0.618	2.476

	Features	VIF
10	Specialization_Others	2.17
3	Lead Source_Olark Chat	2.03
12	Last Notable Activity_Modified	1.79
2	Lead Origin_Landing Page Submission	1.70
6	Last Activity_Olark Chat Conversation	1.59
8	Last Activity_SMS Sent	1.57
1	Total Time Spent on Website	1.29
4	Lead Source_Reference	1.24
0	Do Not Email	1.21
11	What is your current occupation_Working Profes...	1.19
5	Lead Source_Welingak Website	1.09

- In models 7 and 8, we are eliminating the variables 'Last Activity_Unsubscribed' and 'Last Notable Activity_Unreachable' to streamline and reduce unnecessary variables.

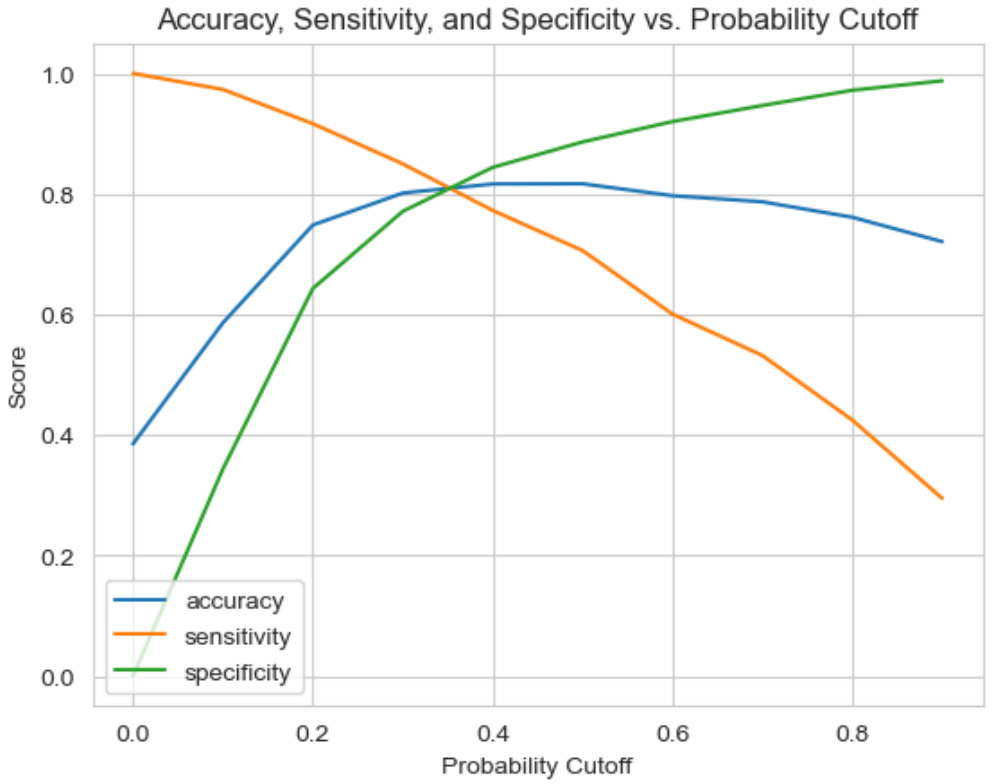
Model-9

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0376	0.125	-0.300	0.764	-0.283	0.208
Do Not Email	-1.5218	0.177	-8.611	0.000	-1.868	-1.175
Total Time Spent on Website	1.0954	0.040	27.225	0.000	1.017	1.174
Lead Origin_Landing Page Submission	-1.1940	0.128	-9.360	0.000	-1.444	-0.944
Lead Source_Olark Chat	1.0819	0.122	8.847	0.000	0.842	1.322
Lead Source_Reference	3.3166	0.241	13.747	0.000	2.844	3.789
Lead Source_Welingak Website	5.8115	0.728	7.981	0.000	4.384	7.239
Last Activity_Olark Chat Conversation	-0.9613	0.171	-5.610	0.000	-1.297	-0.625
Last Activity_Other_Activity	2.1751	0.463	4.699	0.000	1.268	3.082
Last Activity_SMS Sent	1.2942	0.075	17.308	0.000	1.148	1.441
Specialization_Others	-1.2025	0.125	-9.582	0.000	-1.448	-0.957
What is your current occupation_Working Professional	2.6083	0.194	13.454	0.000	2.228	2.988
Last Notable Activity_Modified	-0.9004	0.081	-11.097	0.000	-1.059	-0.741

	Features	VIF
9	Specialization_Others	2.16
3	Lead Source_Olark Chat	2.03
11	Last Notable Activity_Modified	1.78
2	Lead Origin_Landing Page Submission	1.69
6	Last Activity_Olark Chat Conversation	1.59
8	Last Activity_SMS Sent	1.56
1	Total Time Spent on Website	1.29
4	Lead Source_Reference	1.24
10	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.13
5	Lead Source_Welingak Website	1.09
7	Last Activity_Other_Activity	1.01

- Model-9 is finalized as our ultimate model since all variable p-values are 0 and VIF values are low for each variable. Our final model comprises 12 variables.

Determining the Optimal Probability Threshold and Assessing Performance on Training Data



Model Evaluation on Train Dataset

```
Confusion Matrix :
 [[3183  722]
 [ 462 1984]]
Accuracy : 0.8135726657219335
Sensitivity : 0.8111201962387572
Specificity : 0.8151088348271447
Precision : 0.7331855136733185
```

- From the depicted plot, it's evident that 0.35 serves as the optimal cutoff probability for our model.

- In this context, sensitivity holds significant importance because, for the education company to maximize revenue, it's preferable to capture as many potential conversions as possible. While there may be instances of falsely predicting some leads, prioritizing sensitivity ensures a higher likelihood of converting leads overall.

Forecasting & Creating Lead Scoring (Business Requirement)

	Prospect ID	Converted	Converted_prob	final_predicted	Lead_Score
1	1490	1	0.969057	1	97
8	4223	1	0.916621	1	92
16	1946	1	0.924467	1	92
21	2461	1	0.992551	1	99
23	5822	1	0.997991	1	100
...	...	...	...	...	...
2694	1566	1	0.947723	1	95
2699	6461	1	0.961562	1	96
2703	5741	1	0.908283	1	91
2715	6299	1	0.871977	1	87
2720	6501	1	0.854745	1	85

- Utilizing Model 9, we computed probabilities on the Test dataset and applied a cutoff of 0.32 to predict the pred_Converted values (0 or 1).
- Following the business specifications, we've introduced a Lead Score column for the leads, ranging from 0 to 100. A higher score indicates a hotter lead, implying a higher likelihood of conversion, while a lower score suggests a colder lead with lower conversion potential. The Lead Score is generated by multiplying the pred_Converted probability by 100.

Model Evaluation on Test data

Model Evaluation on Test Dataset

```
Confusion Matrix :
 [[1409  325]
 [ 197  792]]
Accuracy : 0.8082996694821888
Sensitivity : 0.8008088978766431
Specificity : 0.8125720876585929
Precision : 0.7090420769919427
```

Feature Importance

Lead Source_Welingak Website	5.811465
Lead Source_Reference	3.316598
What is your current occupation_Working Professional	2.608292
Last Activity_Other_Activity	2.175096
Last Activity_SMS Sent	1.294180
Total Time Spent on Website	1.095412
Lead Source_Olark Chat	1.081908
const	-0.037565
Last Notable Activity_Modified	-0.900449
Last Activity_Olark Chat Conversation	-0.961276
Lead Origin_Landing Page Submission	-1.193957
Specialization_Others	-1.202474
Do Not Email	-1.521825

dtype: float64

- The accuracy and sensitivity are both high, and based on the insights from the logistic regression analysis, the above categories should be given appropriate importance.

Observations & Reccomendations

Reccomendations

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.

- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.