Lead Score Case Study

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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Suggested Ideas for Lead Conversion

Leads Grouping

- Leads are grouped based on their propensity or likelihood to convert.
- This results in a focused group of hot leads.

Better Communication

- We could have a smaller pool of leads to communicate with, which would allow us to have a greater impact.

Boost Conversion

- We would have a greater conversion rate and be able to hit the 80% objective since we concentrated on hot leads that were more likely to convert.

Since we have a target of 80% conversion rate, we would want to obtain a high sensitivity in obtaining hot leads.

Analysis Approach-

Data Cleaning:
Loading Data Set, understanding & cleaning data

EDA:
Check imbalance, Univariate & Bivariate analysis

Data Preparation
Dummy variables, test-train split, feature scaling

Model Building:
RFE for top 15 feature, Manual Feature Reduction & finalizing model

Model Evaluation:
Confusion matrix, Cutoff Selection, assigning Lead Score

Predictions on Test Data:
Compare train vs test metrics, Assign Lead Score and get top features

Recommendation:
Suggest top 3 features to focus for higher conversion & areas for improvement

Data Cleaning

- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (tags, country)
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.

Data Cleaning

● Skewed category columns were checked and dropped to avoid bias in logistic regression models.

● Outliers in **TotalVisits** and **Page Views** Per Visit were treated and capped.

● Invalid values were fixed and data was standardized in some columns, such as lead source.

● Low frequency values were grouped together to “Others”.

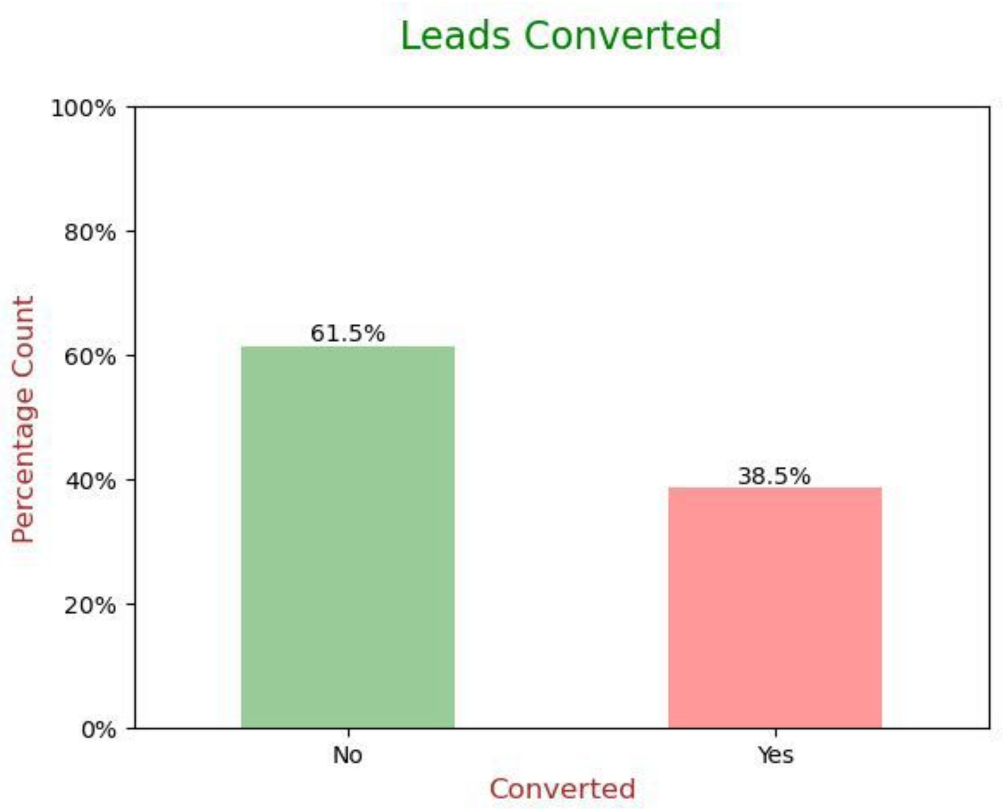
● Binary categorical variables were mapped.

● Other cleaning activities were performed to ensure data quality and accuracy.

○ Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead

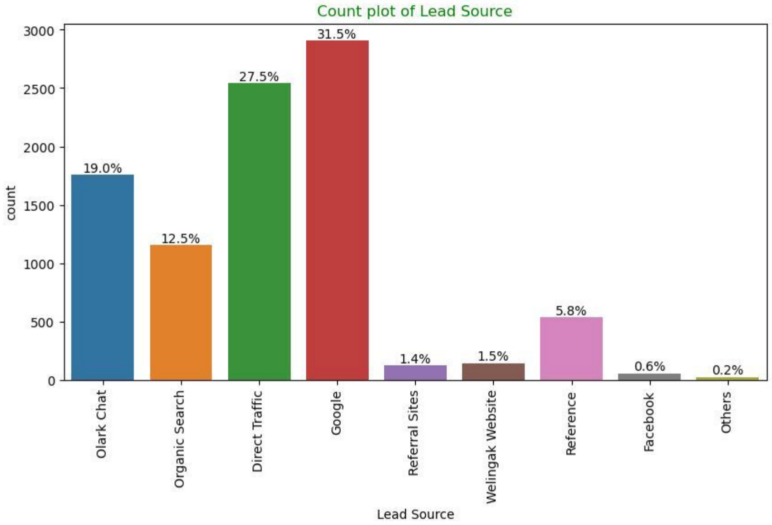
source has Google, google)

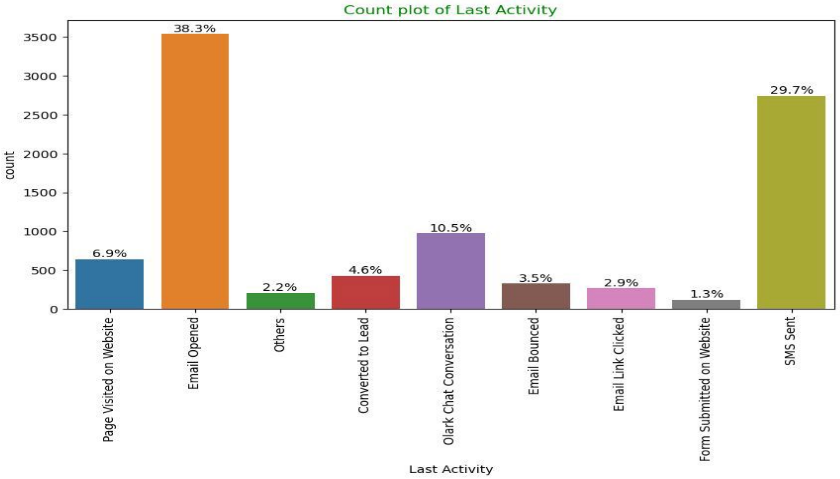
EDA



- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)

EDA
● Univariate Analysis – Categorical Variables



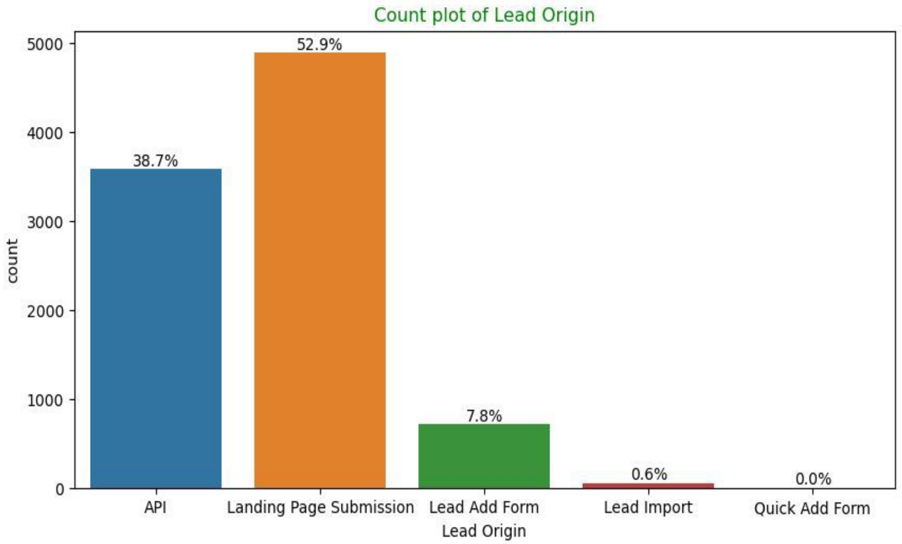


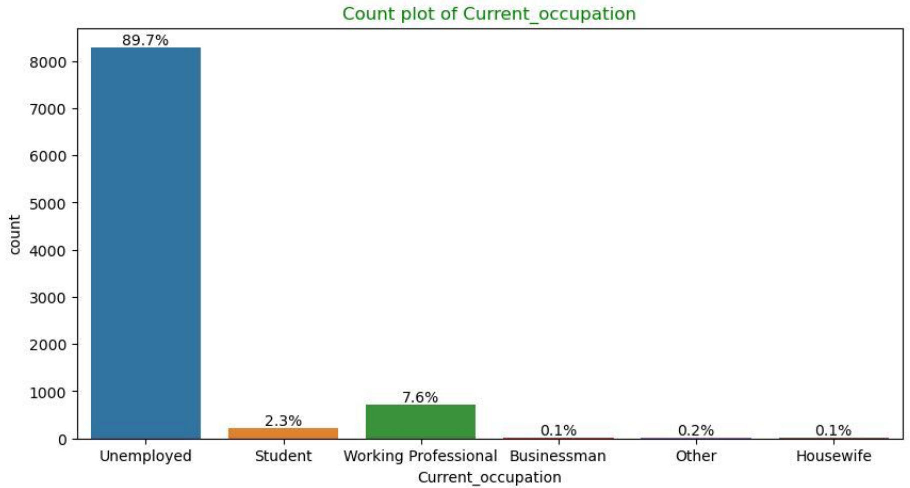
Lead Source: 58% Lead source is from Google & Direct Traffic combined.

Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities.

EDA

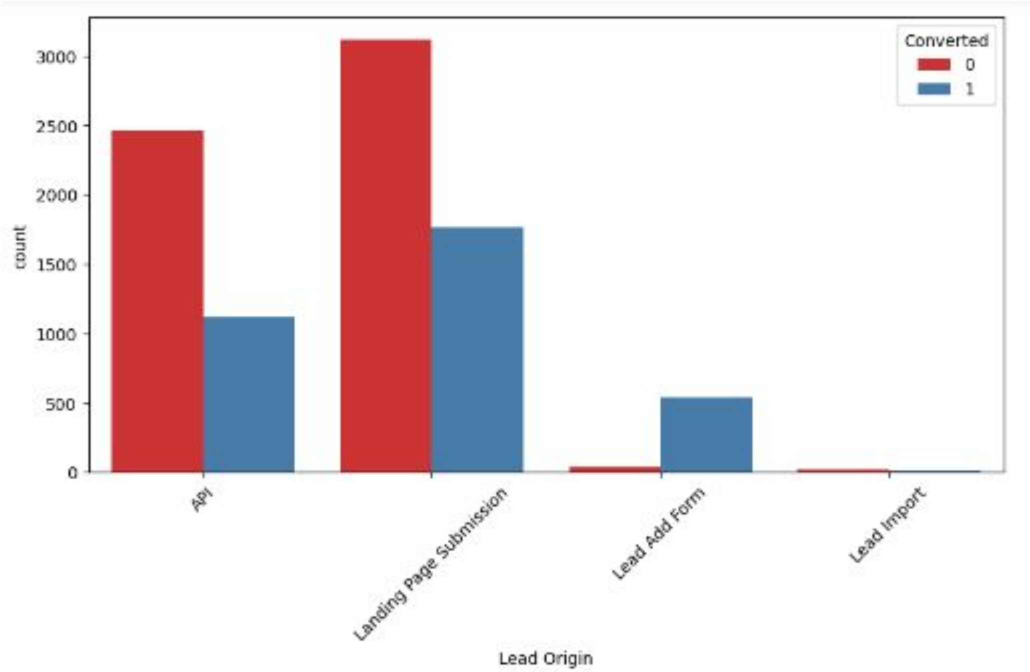
● Univariate Analysis – Categorical Variables





Lead Origin: "Landing Page Submission" identified 53% of customers, "API" identified 39%.

Current_occupation: It has 90% of the customers as Unemployed.

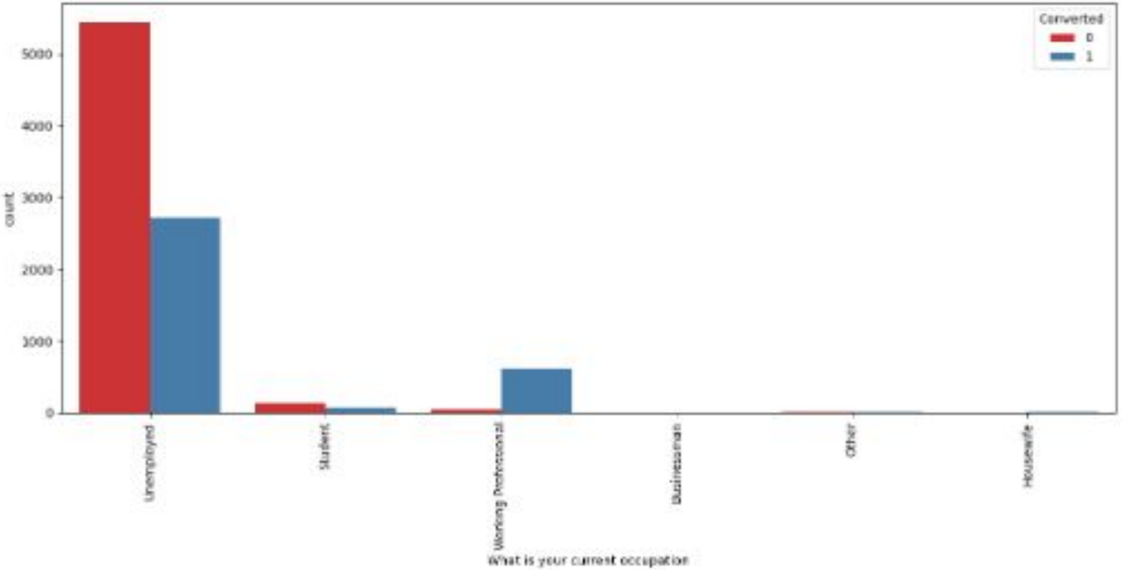


Lead Origin:

● Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%.

● The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.

EDA – Bivariate Analysis for Categorical Variables

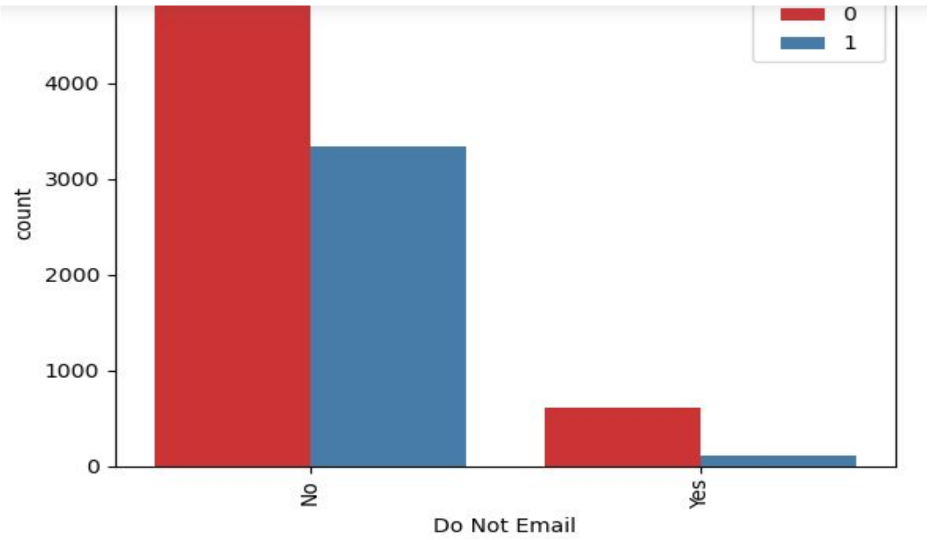


What is your current occupation:

● Around 90% of the customers are Unemployed, with lead conversion rate (LCR) of 34%.

● While Working Professional contribute only 7.6% of total customers with almost 92% Lead conversion rate (LCR).

EDA – Bivariate Analysis for Categorical Variables



Do Not Email:

- 92% of the people has opted that they don't want to be emailed about the course & 40% of them are converted to leads.

EDA – Bivariate Analysis for Categorical Variables

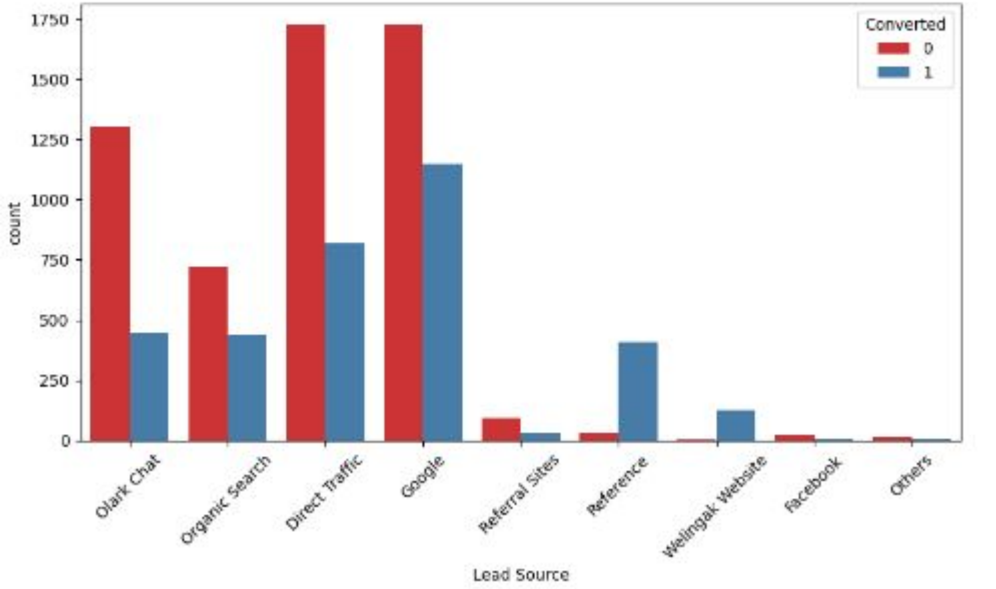
Lead Source:

● Google has LCR of 40% out of 31% customers,

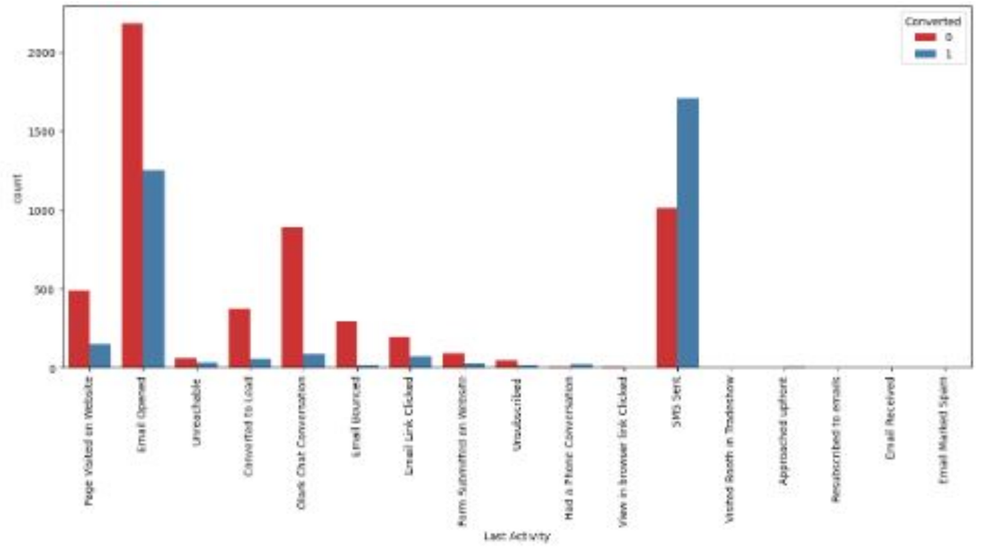
● Direct Traffic contributes 32% LCR with 27% customers, which is lower than Google,

● Organic Search also gives 37.8% of LCR, but the contribution is by only 12.5% of customers,

● Reference has LCR of 91%, but there are only around 6% of customers through this Lead Source.



EDA – Bivariate Analysis for Categorical Variables

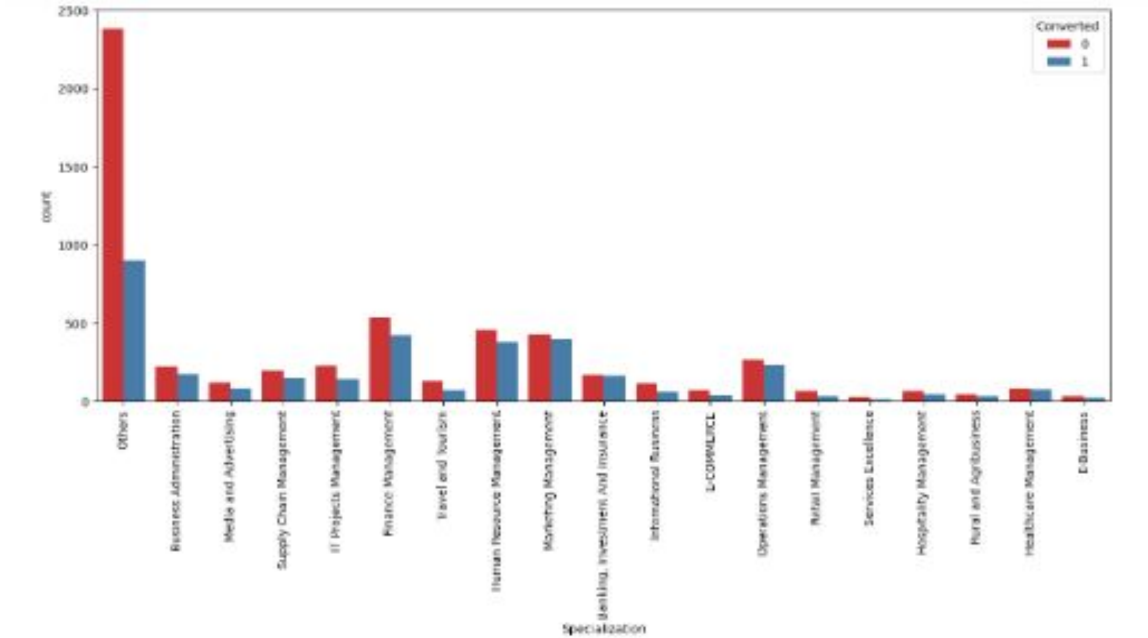


Last Activity:

● 'SMS Sent' has high lead conversion rate of 63% with 30% contribution from last activities,

● 'Email Opened' activity contributed 38% of last activities performed by the customers, with 37% lead conversion rate.

EDA – Bivariate Analysis for Categorical Variables



Specialization:

● Marketing Management, HR Management, Finance Management shows good contribution in Leads conversion than other specialization.

Data Preparation before Model building

● Binary level categorical columns were already mapped to 1 / 0 in previous steps

● Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation

● Splitting Train & Test Sets

○ 70:30 % ratio was chosen for the split

● Feature scaling

○ Standardization method was used to scale the features

● Checking the correlations

○ Predictor variables which were highly correlated with each other were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add Form).

Model Building

Feature Selection

● The data set has lots of dimension and large number of features.

● This will reduce model performance and might take high computation time.

● Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.

● Then we can manually fine tune the model.

● RFE outcome

Pre RFE – 48 columns & Post RFE – 20 columns

Model Building

● Manual Feature Reduction process was used to build models by dropping variables with p – value

greater than 0.05.

●Model 7 looks stable after seven iteration with:

○ significant p-values within the threshold (p-values < 0.05) and

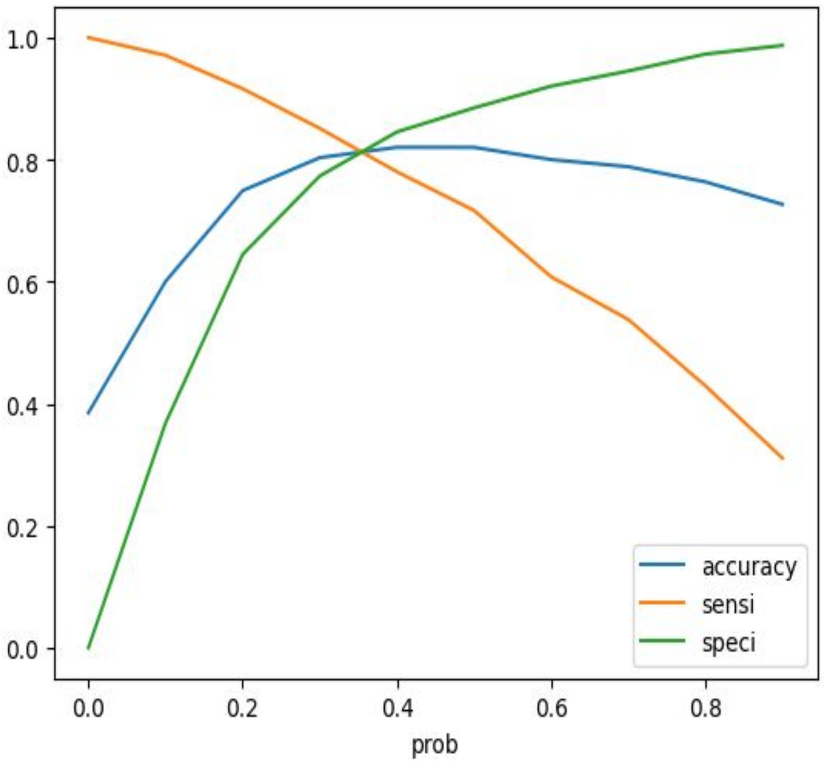
○ No sign of multicollinearity with VIFs less than 5

● Hence logm7 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

Model Evaluation

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:**	-2580.7
Date:	Tue, 23 May 2023	**Deviance:**	5161.3
Time:	19:06:36	**Pearson chi2:**	6.36e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4057
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1406	0.127	-1.108	0.268	-0.389	0.108
Do Not Email	-1.6984	0.191	-8.887	0.000	-2.073	-1.324
Total Time Spent on Website	1.1171	0.040	27.686	0.000	1.038	1.196
Lead Origin_Landing Page Submission	-1.1961	0.128	-9.339	0.000	-1.447	-0.945
Lead Source_Olark Chat	1.1430	0.124	9.242	0.000	0.901	1.385
Lead Source_Reference	3.4019	0.243	14.026	0.000	2.927	3.877
Lead Source_Welingak Website	5.9684	0.732	8.158	0.000	4.535	7.402
Last Activity_Olark Chat Conversation	-1.0216	0.173	-5.914	0.000	-1.360	-0.683
Last Activity_Other_Activity	2.1646	0.461	4.691	0.000	1.260	3.069
Last Activity_SMS Sent	0.7940	0.157	5.047	0.000	0.486	1.102
Last Activity_Unreachable	0.7494	0.310	2.415	0.016	0.141	1.358
Last Activity_Unsubscribed	1.4180	0.480	2.952	0.003	0.476	2.360
Specialization_Others	-1.1989	0.126	-9.514	0.000	-1.446	-0.952
What is your current occupation_Working Professional	2.6042	0.195	13.337	0.000	2.221	2.987
Last Notable Activity_Modified	-0.6922	0.097	-7.138	0.000	-0.882	-0.502
Last Notable Activity_SMS Sent	0.6910	0.177	3.894	0.000	0.343	1.039



It was decided to go ahead with 0.35 as cutoff after checking evaluation metrics coming from both plots

Model Evaluation

.

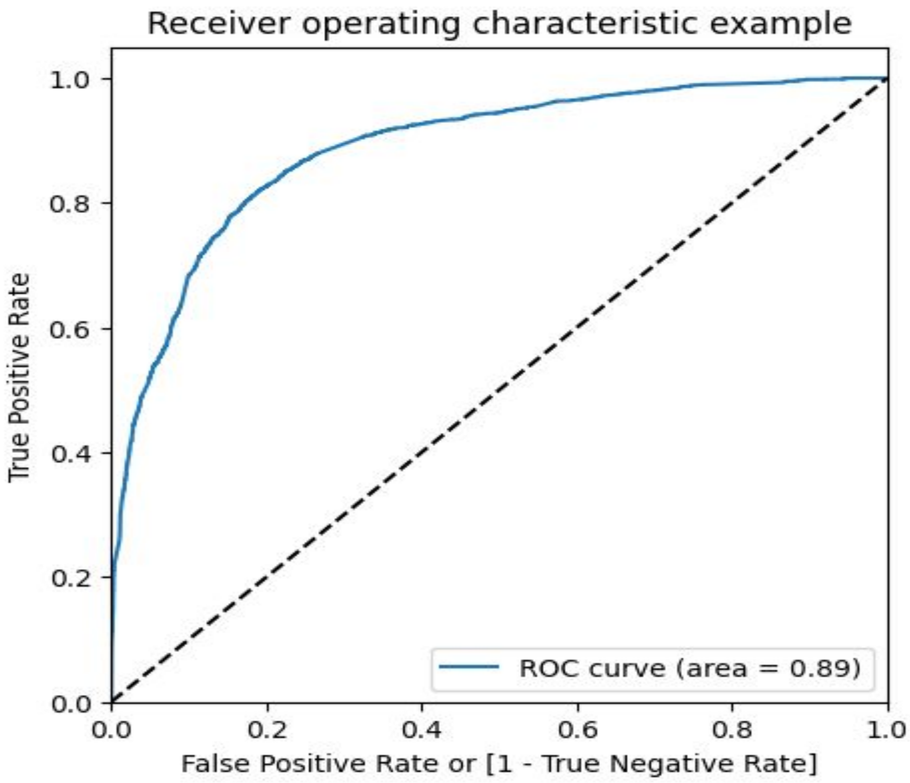
●ROC Curve –t

● Area under ROC curve is 0.87 out of 1 which indicates a good predictive model.

● The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.

Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.

The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low



Model Evaluation

Conclusion

• As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.

• We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.

Lead Source_Welingak Website	5.914695
Lead Source_Reference	3.392774
What is your current occupation_Working Professional	2.618774
Last Activity_Other_Activity	2.226927
Last Activity_Unsubscribed	1.380067
Last Activity_SMS Sent	1.328999
Lead Source_Olark Chat	1.141863
Total Time Spent on Website	1.118245
Last Activity_Unreachable	0.811978

Conclusion-

We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:

Last Activity_Olark Chat Conversation	-0.922916
Lead Origin_Landing Page Submission	-1.190922
Specialization_Others	-1.197650
Do Not Email	-1.676398

Conclusion-

To increase our Lead Conversion Rates

● Focus on features with positive coefficients for targeted marketing strategies.

● Develop strategies to attract high-quality leads from top-performing lead sources.

● Optimize communication channels based on lead engagement impact.

● Engage working professionals with tailored messaging.

● More budget/spend can be done on Welingak Website in terms of advertising, etc.

● Incentives/discounts for providing reference that convert to lead, encourage providing more references.

● Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

To identify areas of improvement

● Analyze negative coefficients in specialization offerings.

● Review landing page submission process for areas of improvement.