LEAD SCORING CASE STUDY

USING LOGISTIC REGRESSION

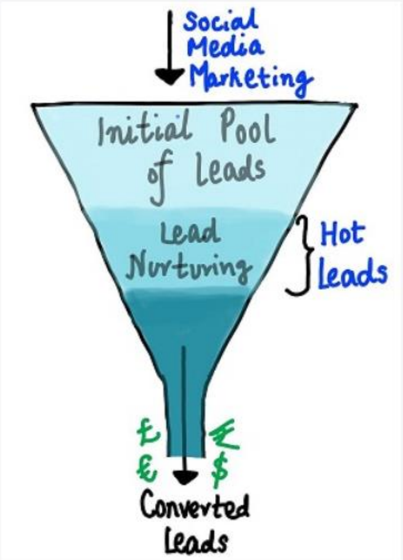
Submitted By-
Monica Tripathi
Muskan Dhamdere
Priya Mondal

CONTENTS

- Background & Problem Statement
- Objective
- Analysis Approach
- Data Visualisation and Insights
- Model Building
- Final Model Snippet
- Model Evaluation
- Inferences

BACKGROUND & PROBLEM STATEMENT

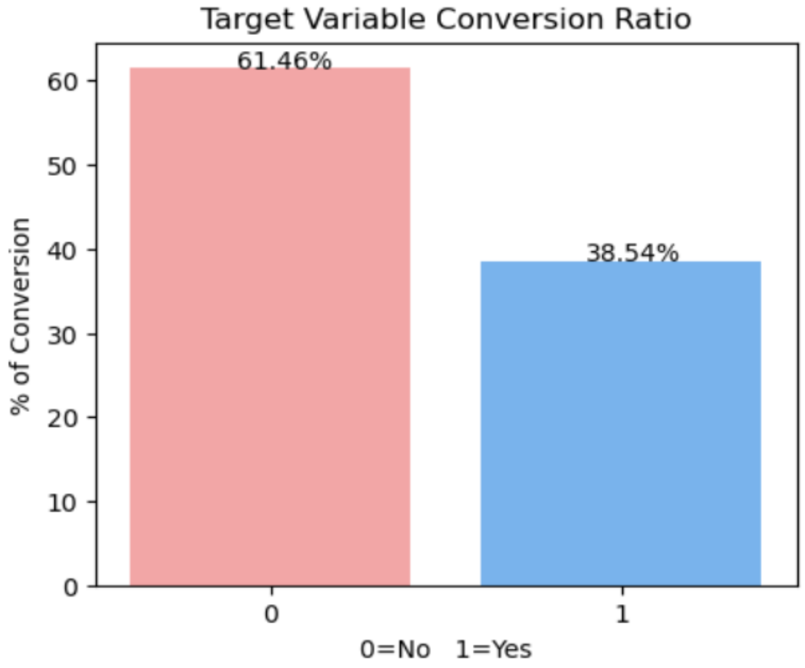
- X Education, an education company sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- Problem Statement : Company Uses various Marketing platforms, with which people land on their website, by filling up a form, these people gets converted to leads another source of leads is through past referrals. Major issue here is of Lead conversion i.e to predict the leads that are hot and most likely to convert into paying customers.



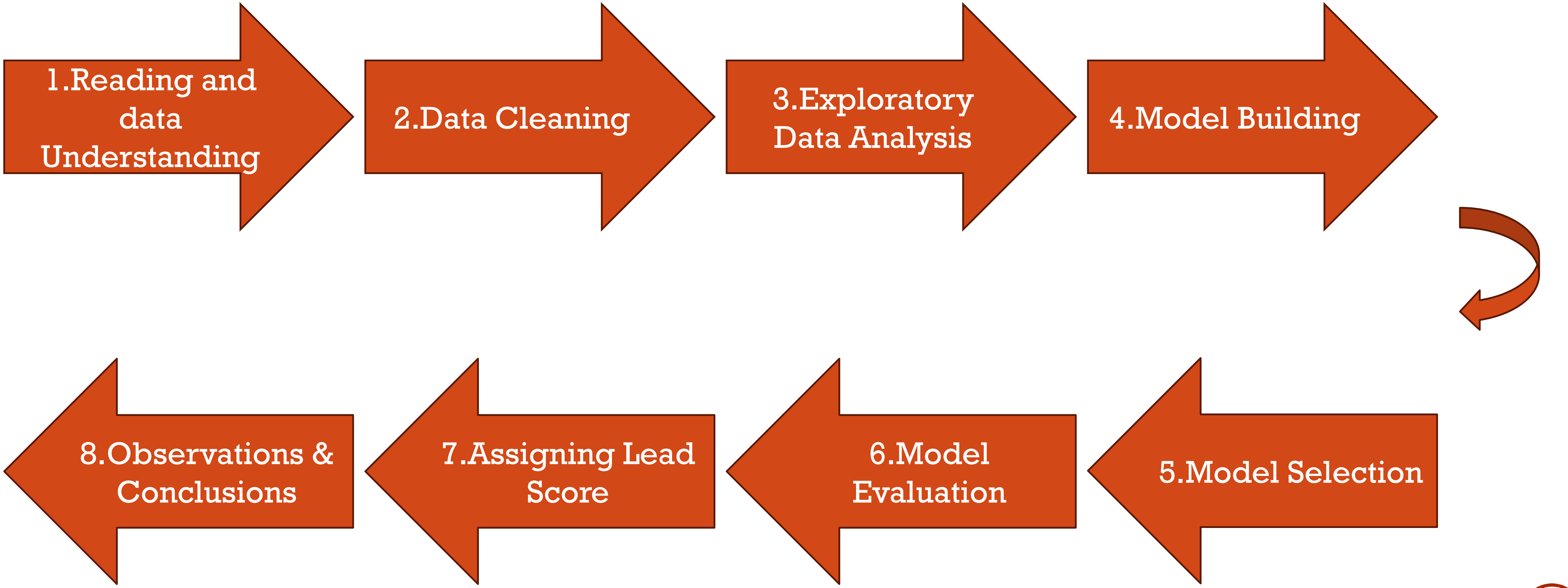
Lead Conversion Process - Demonstrated as a funnel

OBJECTIVE

- Build a Logistic Regression model and 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. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- Target lead conversion rate is around 80%.
- Current conversion percentage is 38.54%



ANALYSIS APPROACH



1.Reading and data Understanding

2.Data Cleaning

3.Exploratory Data Analysis

4.Model Building

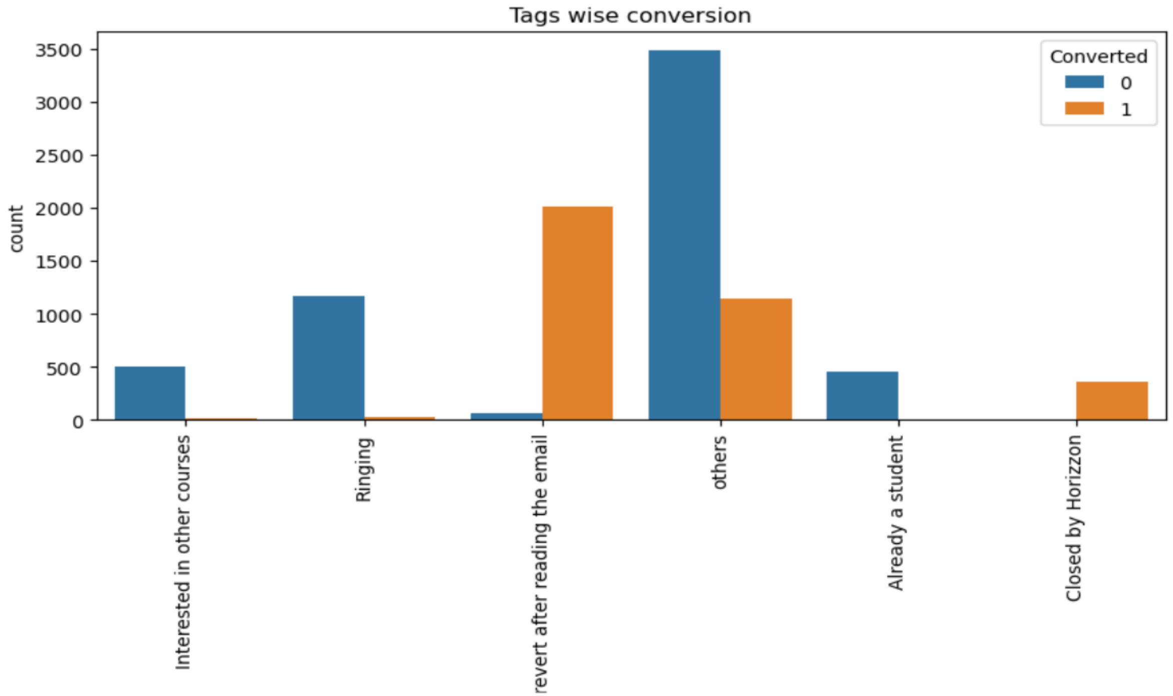
5.Model Selection

6.Model Evaluation

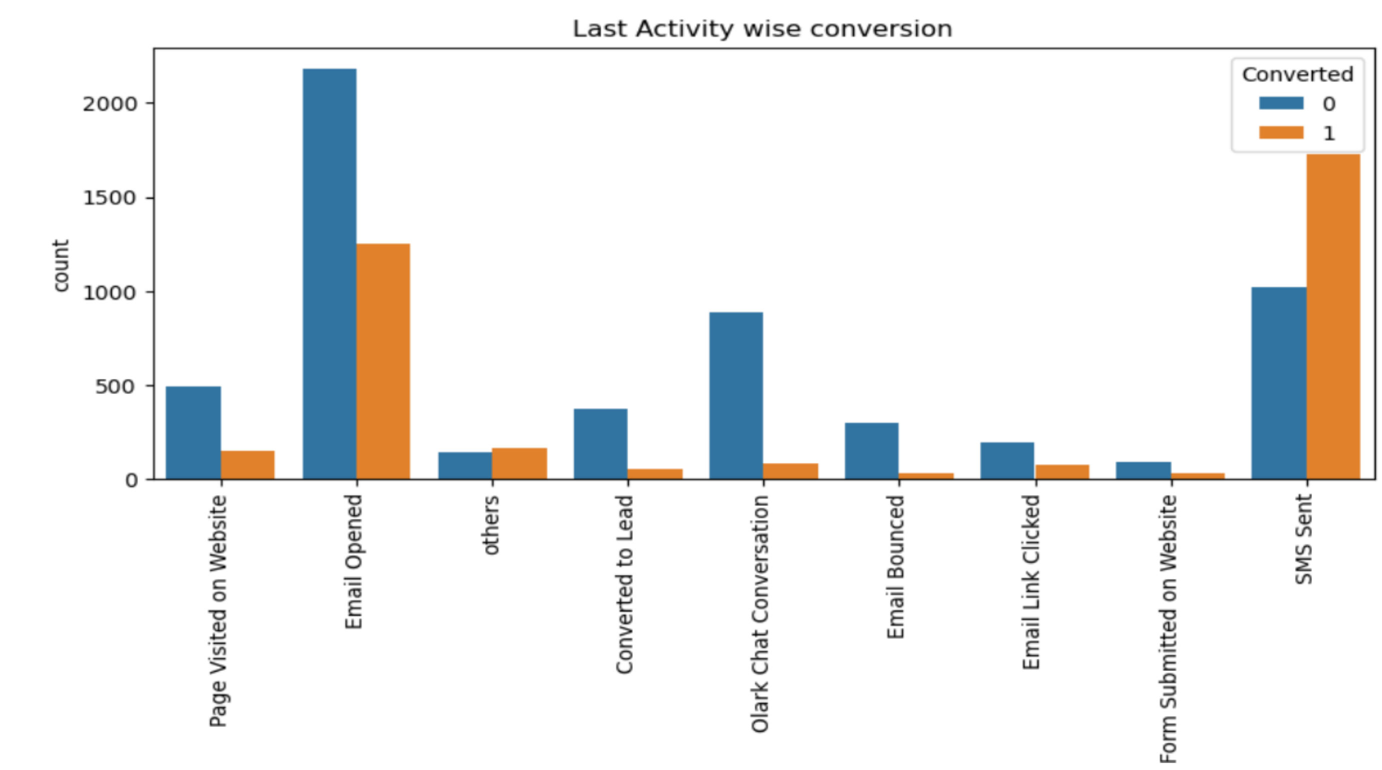
7.Assigning Lead Score

8.Observations & Conclusions

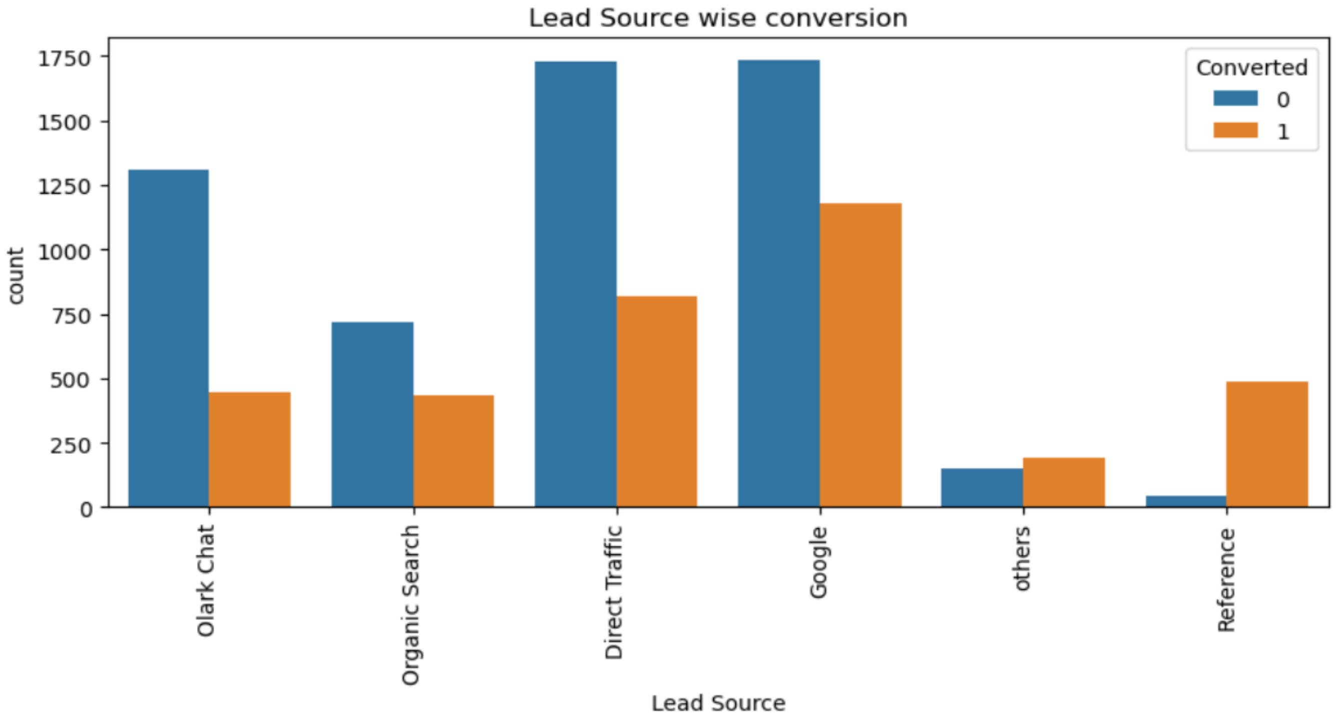
DATA VISUALISATION AND INSIGHTS



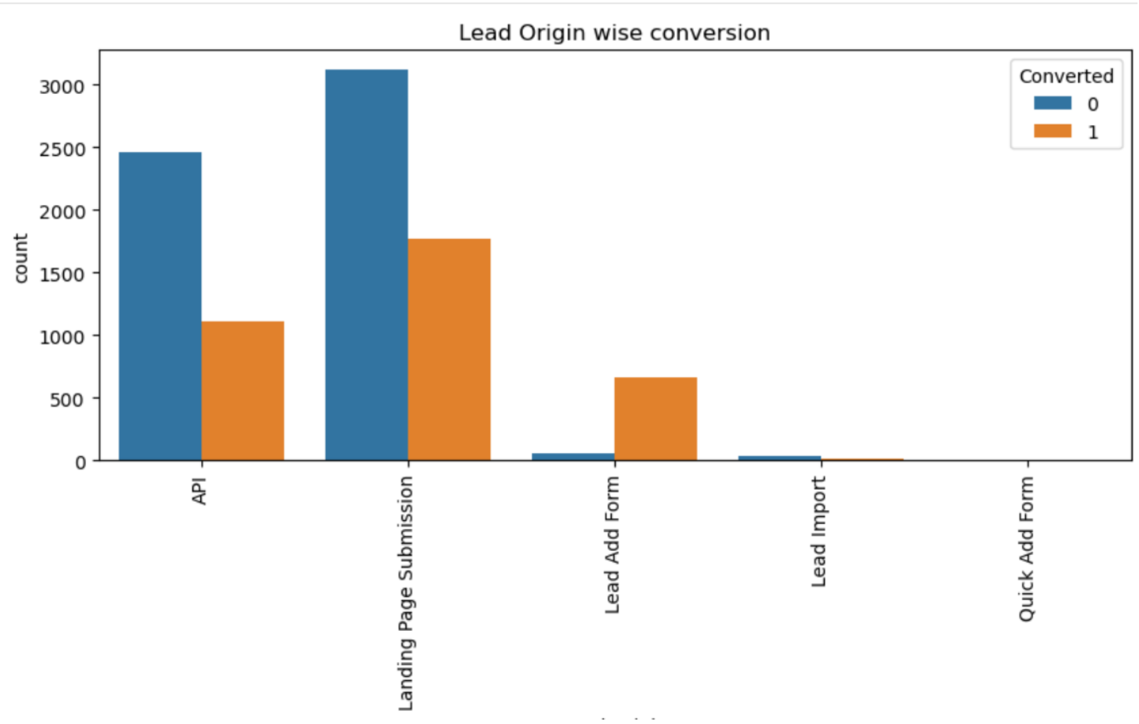
- **Tags** : Individuals tagged as 'will revert after reading email' have highest positive response rate, followed by others and closed by Horizzon.



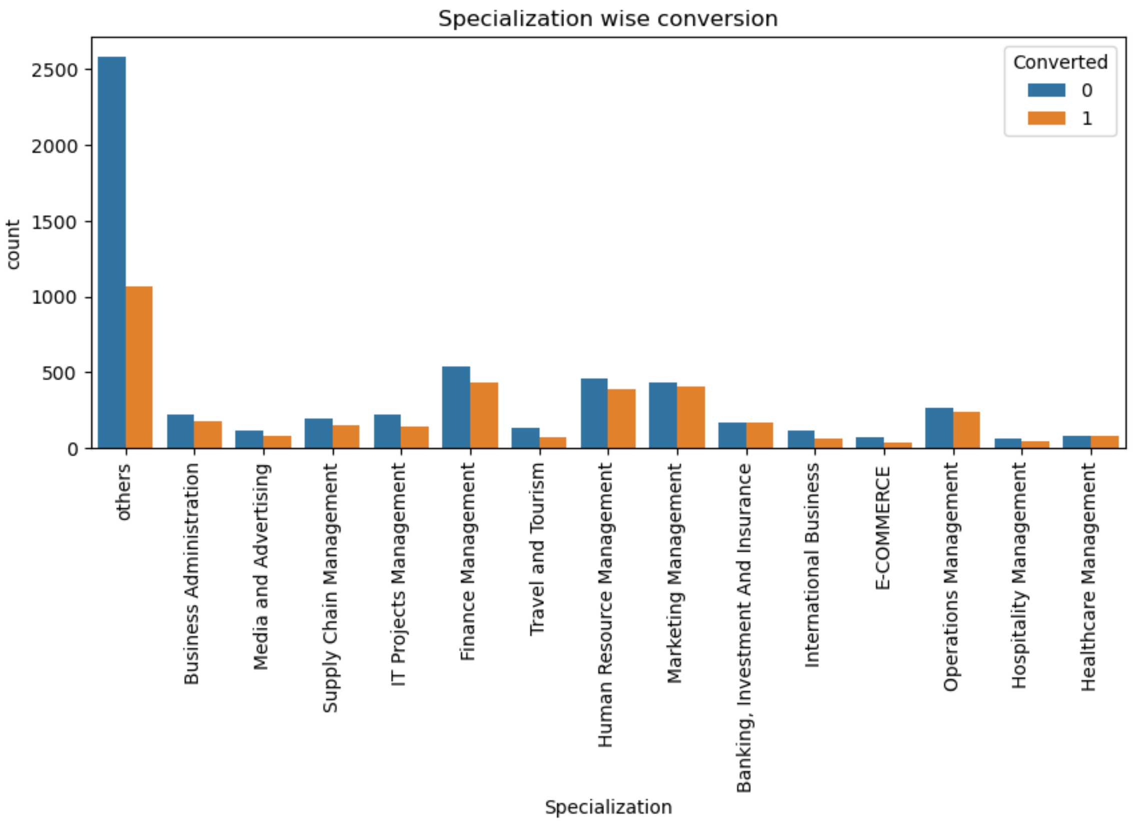
- **Last Activity** : SMS sent followed by Email opened has most conversions, but contrary to that email opened also has negative responses as well so, right now its hard to call it a reliable Last Activity for conversions. Olark chat is not giving any significant returns.



- **Lead source** : Lead sources through Google, followed by Direct Traffic and References leads to maximum conversions. Whereas Welingak and Reference are having more conversion rate and can be further optimized with more focus.

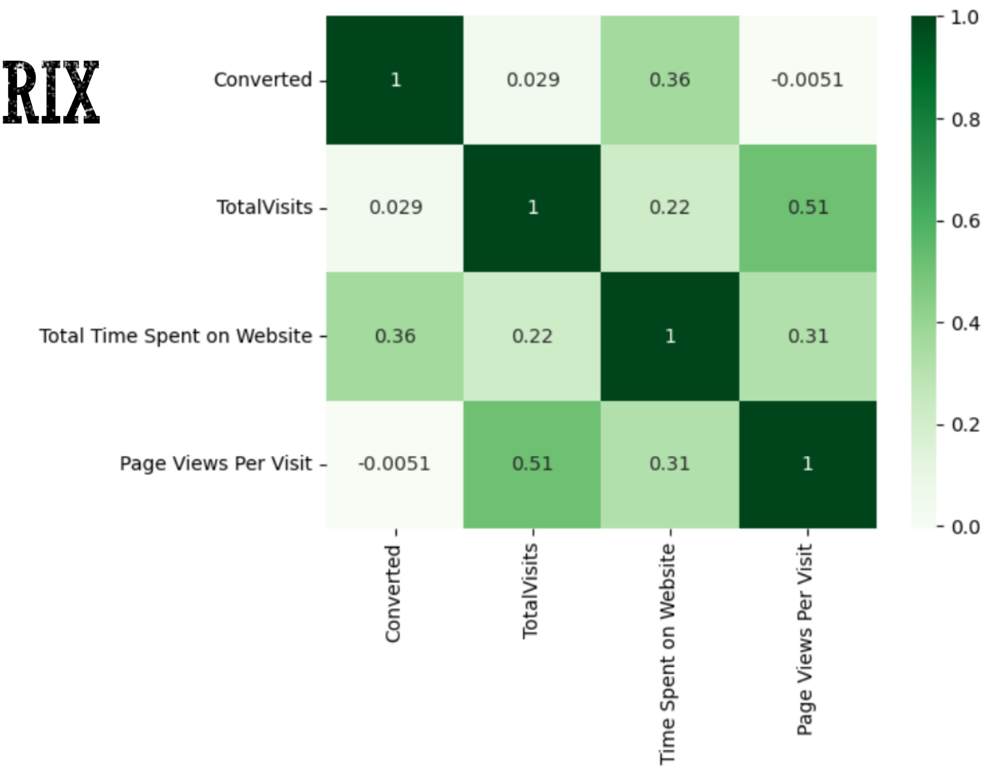


- **Lead Origin** : Most Potential conversions are through API, Landing Page Submission and Lead Add form, but that's also because these 3 have most responses also, so here we can say getting a huge responses is a natural way of more conversions. Lead Import and Quick Add From have less conversion.



- **Specialization** : Individuals from Finance, HR and Marketing backgrounds have most conversions, but people from various other backgrounds can also be observed.

CORRELATION MATRIX



	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Converted	1	0.029	0.36	-0.0051
TotalVisits	0.029	1	0.22	0.51
Total Time Spent on Website	0.36	0.22	1	0.31
Page Views Per Visit	-0.0051	0.51	0.31	1

- **INFERENCES**

1. Conversions has a negative correlation with Page Views per visit, conversion increases with less no of pages viewed per visit; it indicates that individuals more focused in a course might not explore many courses but stick to ones they are interested in.

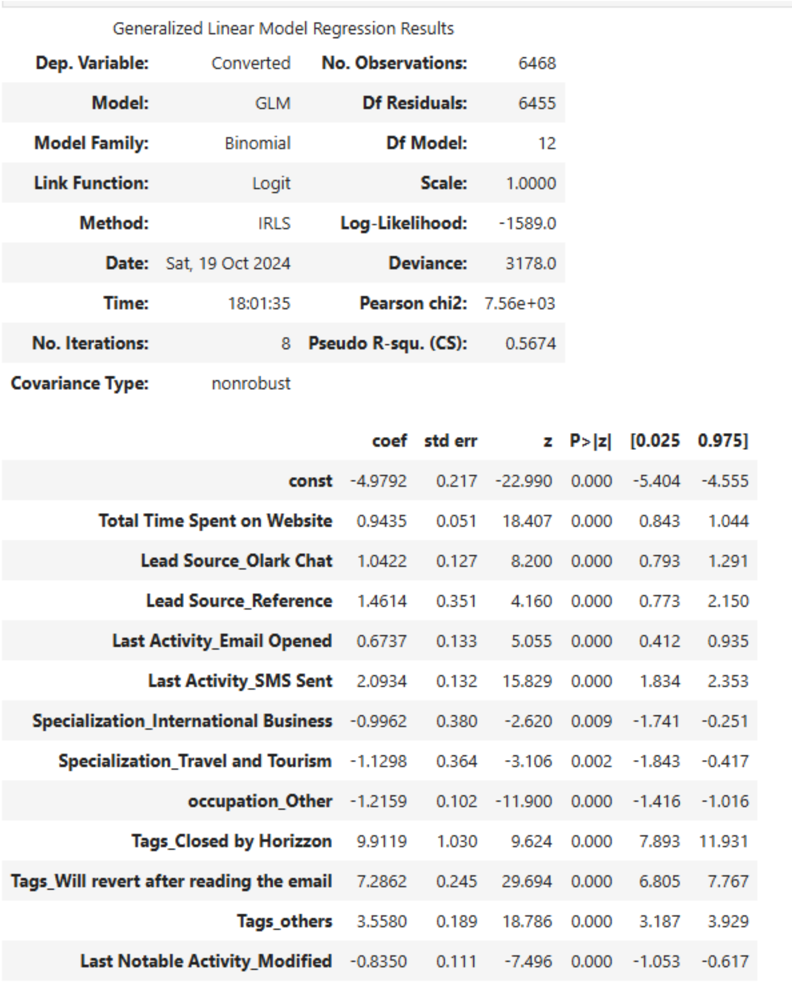
2. Total visit and Page views per visit shows a positive moderate correlation.

3. Converted has good correlation with Total time spent on Website.

MODEL BUILDING

- Splitting in train-test set
- Feature scaling in train set
- Feature selection using RFE
- Build 1st model
- Eliminate features based on high p-value and high VIF (only one feature at a time)
- Build final model
- Predict using train set
- Evaluate accuracy and other metrics
- Predict on test set
- Check area under ROC, evaluate accuracy and other metrics

FINAL MODEL SNIPPET



Generalized Linear Model Regression Results

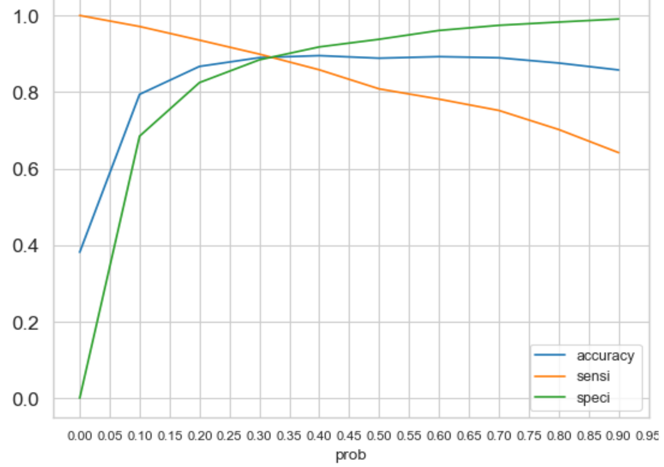
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6455
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1589.0
Date:	Sat, 19 Oct 2024	**Deviance:**	3178.0
Time:	18:01:35	**Pearson chi2:**	7.56e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.5674
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-4.9792	0.217	-22.990	0.000	-5.404	-4.555
Total Time Spent on Website	0.9435	0.051	18.407	0.000	0.843	1.044
Lead Source_Olark Chat	1.0422	0.127	8.200	0.000	0.793	1.291
Lead Source_Reference	1.4614	0.351	4.160	0.000	0.773	2.150
Last Activity_Email Opened	0.6737	0.133	5.055	0.000	0.412	0.935
Last Activity_SMS Sent	2.0934	0.132	15.829	0.000	1.834	2.353
Specialization_International Business	-0.9962	0.380	-2.620	0.009	-1.741	-0.251
Specialization_Travel and Tourism	-1.1298	0.364	-3.106	0.002	-1.843	-0.417
occupation_Other	-1.2159	0.102	-11.900	0.000	-1.416	-1.016
Tags_Closed by Horizzon	9.9119	1.030	9.624	0.000	7.893	11.931
Tags_Will revert after reading the email	7.2862	0.245	29.694	0.000	6.805	7.767
Tags_others	3.5580	0.189	18.786	0.000	3.187	3.929
Last Notable Activity_Modified	-0.8350	0.111	-7.496	0.000	-1.053	-0.617

- P-value for all features is 0 and highly significant.
- Features are also not Collinear as verified by VIF score.

MODEL EVALUATION- TRAIN SET

Sensitivity And Specificity For Various Probabilities



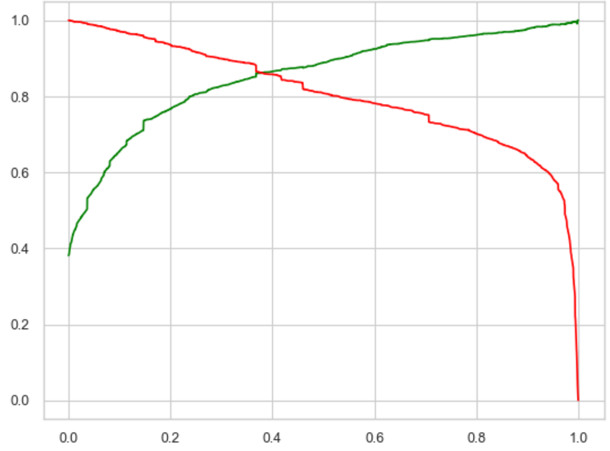
From the above curve, 0.33 is the optimum cutoff probability.

Confusion Matrix

3577 TN	425 FP
267 FN	2199 TP

Accuracy- 89.30%
Sensitivity- 89.17%
Specificity- 89.38%

Precision and Recall Trade - off

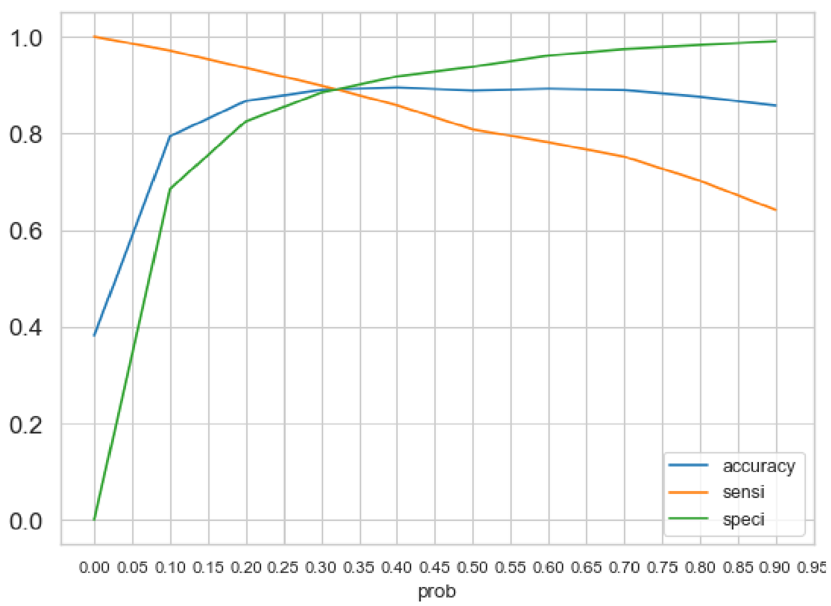


The above present Precision- Recall curve suggests an optimal cut off of 0.38

Accuracy- 89.47%
Precision- 86.23%
Recall 89.13%

MODEL EVALUATION- TEST SET

Sensitivity And Specificity For Various Probabilities



From the above curve, 0.33 is the optimum cutoff probability.

Confusion Matrix

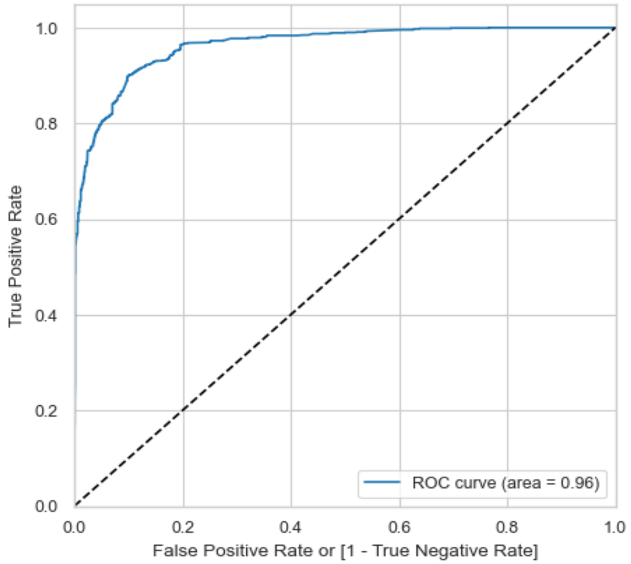
1501 **TN**	**176** **FP**
107 FN	988 TP

Test threshold is 0.33

Accuracy- 89.79%
Sensitivity- 90.22%
Specificity- 89.50%

Precision- 84.87%
Recall- 90.22%

MODEL EVALUATION- SUMMARISED



ROC Curve

METRICS	TRAIN	TEST
ACCURACY	89.30%	89.79%
SENSITIVITY/RECALL:	89.17%	90.23%
SPECIFICITY	89.38%	89.51%
PRECISION	83.80%	84.88%
FALSE POSITIVE RATE (FPR)	10.62%	10.49%

- At a cut off of 0.33; Area under curve is 0.96 which is very close to ideal AUC of 1.

INFERENCES

- Twelve features were selected as the most significant in predicting the conversion.
- ‘Tags will revert after reading the email’ and ‘Tags closed by Horizzon’ have the most positive impact on conversion and further engagement will improve lead conversion.
- Occupation Others, Specialization Travel and Tourism and Specialization International Business have negative impact on conversions, maybe looking into these profile in the end or putting in least effort is recommended here.
- More Focus on Last Activity Email opened, Total time spent on website, Lead source Olark chat have great potential and with increased and interactive engagement we can have more conversions.
- As cutoff 0.33 is likely better aligned with our goal of identifying the maximum number of leads that can convert. It gives a slightly higher recall (89.17%), meaning we’ll identify more potential customers. The trade-off is a marginally higher false positive rate and slightly lower precision, but since our focus is on maximizing lead capture, this is acceptable.
- Overall Model Looks good and aligns with our business goal to achieve a minimum of 80% conversion.

SCENARIO 1

X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

- Given the performance of the logistic regression model and the goal of aggressively converting leads during a 2-month intern phase, the company can take following measures:
- Leverage High Sensitivity: Since sensitivity (recall) on the test set is 90.23%, it indicates that the model is effective at identifying most of the actual conversions (true positives). This means that the model can successfully predict leads who are likely to convert. So, use the model’s predictions to prioritize phone calls to leads predicted as "1" (i.e., potential conversions). With interns available, focus on making personal phone calls to as many of these leads as possible.
- Reduce decision Threshold: Experiment with different thresholds (e.g., 0.3, 0.25) and analyse the precision-recall trade-off. A lower threshold will result in more leads being predicted as conversions, providing the sales team with more contacts to follow up on.
- Prioritize leads with the highest predicted conversion probabilities.

- Assign specific segments to each intern (based on lead characteristics) and track their success rate. Identify high-performers and optimize resource allocation accordingly.
- This strategy will allow X Education to take full advantage of the interns and ensure they maximize the number of leads converted during this aggressive sales push.

SCENARIO 2

- 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. Suggest a strategy they should employ at this stage.
- Strategy to Minimize Useless Phone Calls: 1. Increase the Decision Threshold this will make the model more conservative, only marking leads with very high probabilities as likely to convert. This threshold will decrease false positives, 2. Prioritize high-value leads based on strong positive indicators (like high-impact tags or lead sources), 3. Focus on Precision, meaning that when the model predicts a lead as a conversion, it's highly likely to be correct.

THANKYOU