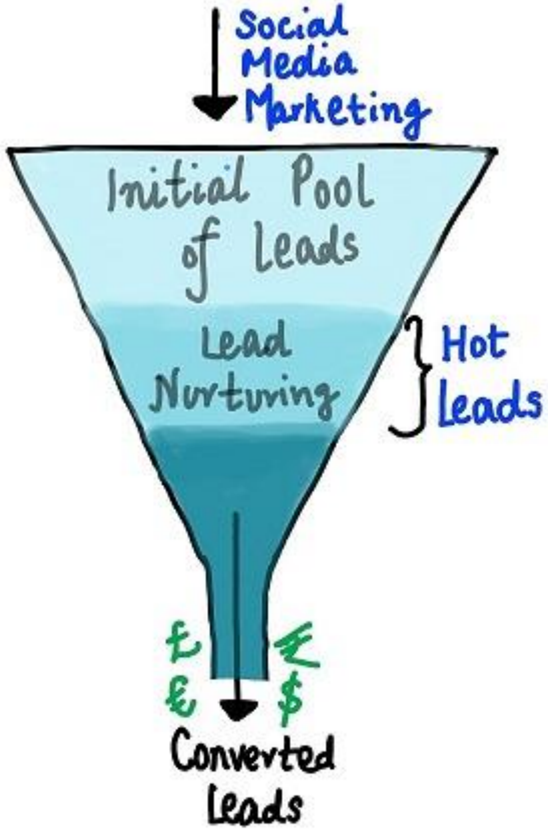
Improving Lead Conversion Rate Analysis



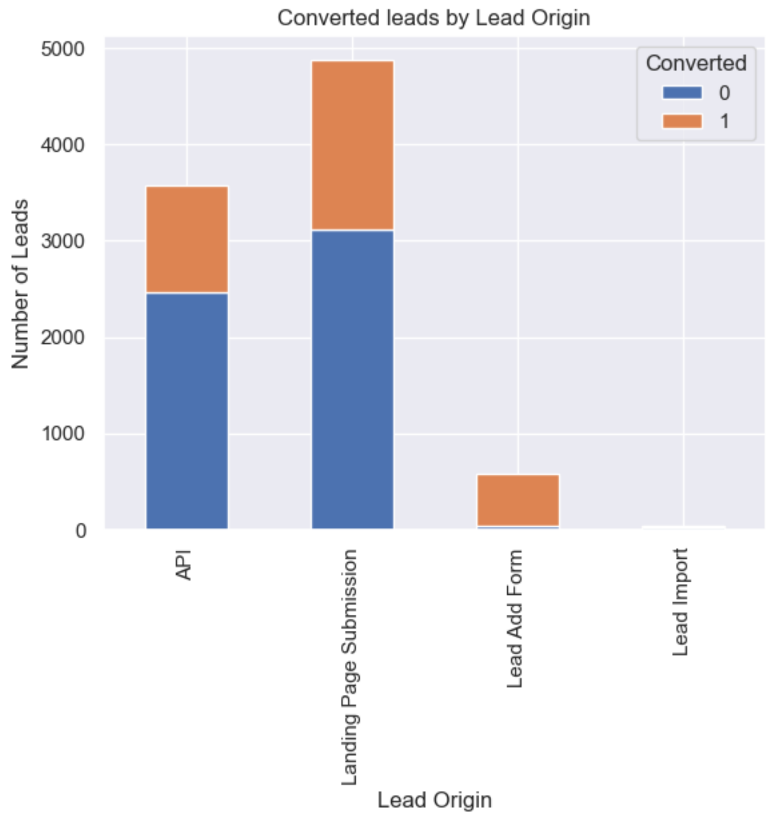
Introduction

- X Education is facing a challenge - while many people show interest in their online courses, only a small fraction end up becoming paying customers.
- The current conversion rate is about 37%. Our goal is to boost the conversion rate to around 80%, making their business more successful.

Data Overview

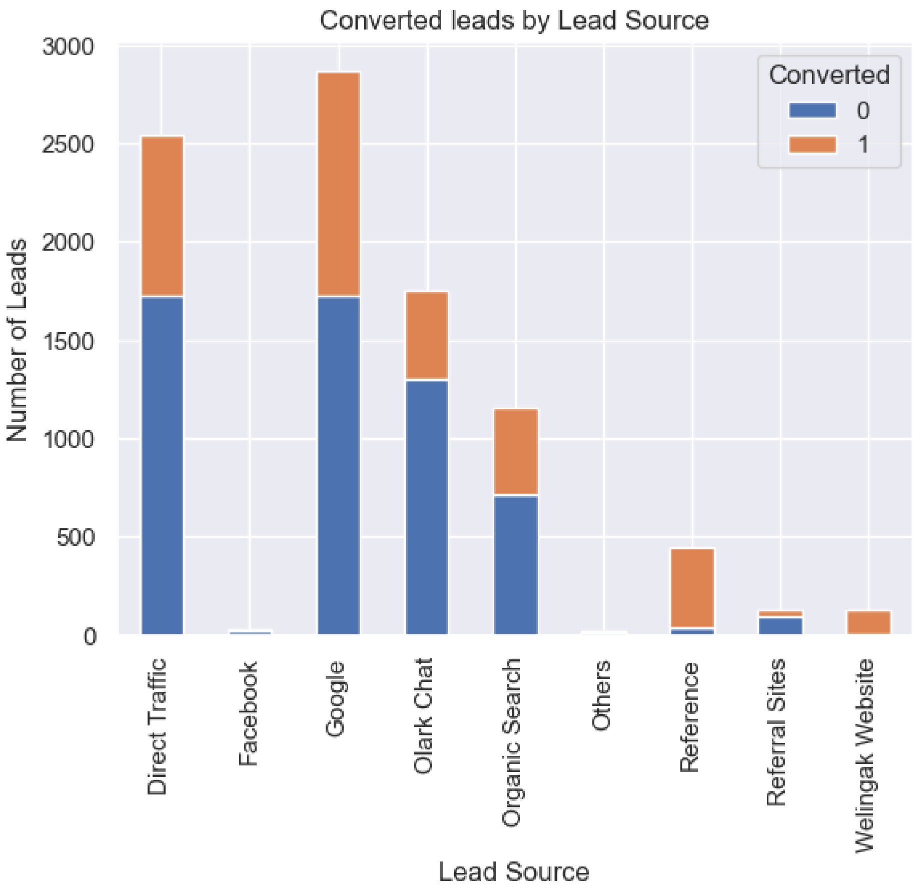
- Our dataset comprises 9,240 data points with 37 columns, each providing valuable insights into the leads and their behaviors.
- To ensure accurate and meaningful analysis, we performed thorough data preprocessing. This involved handling missing values, a common occurrence in real-world datasets. We also addressed the presence of 'Select' values, which often indicate unselected options and were treated as null values.
- Following preprocessing, our dataset was refined to include 9,074 rows and 26 columns. This process retained an impressive 98% of the original rows, ensuring that our analysis remains representative of the lead data.

Exploratory Data Analysis



- API and Landing Page Submission exhibit a conversion rate of 30-35%, with a substantial count of leads originating from these sources.
- Lead Add Form displays an impressive conversion rate of over 90%, although the total count of leads generated is not very high.
- Lead Import represents a category with a significantly lower count of leads.

Exploratory Data Analysis



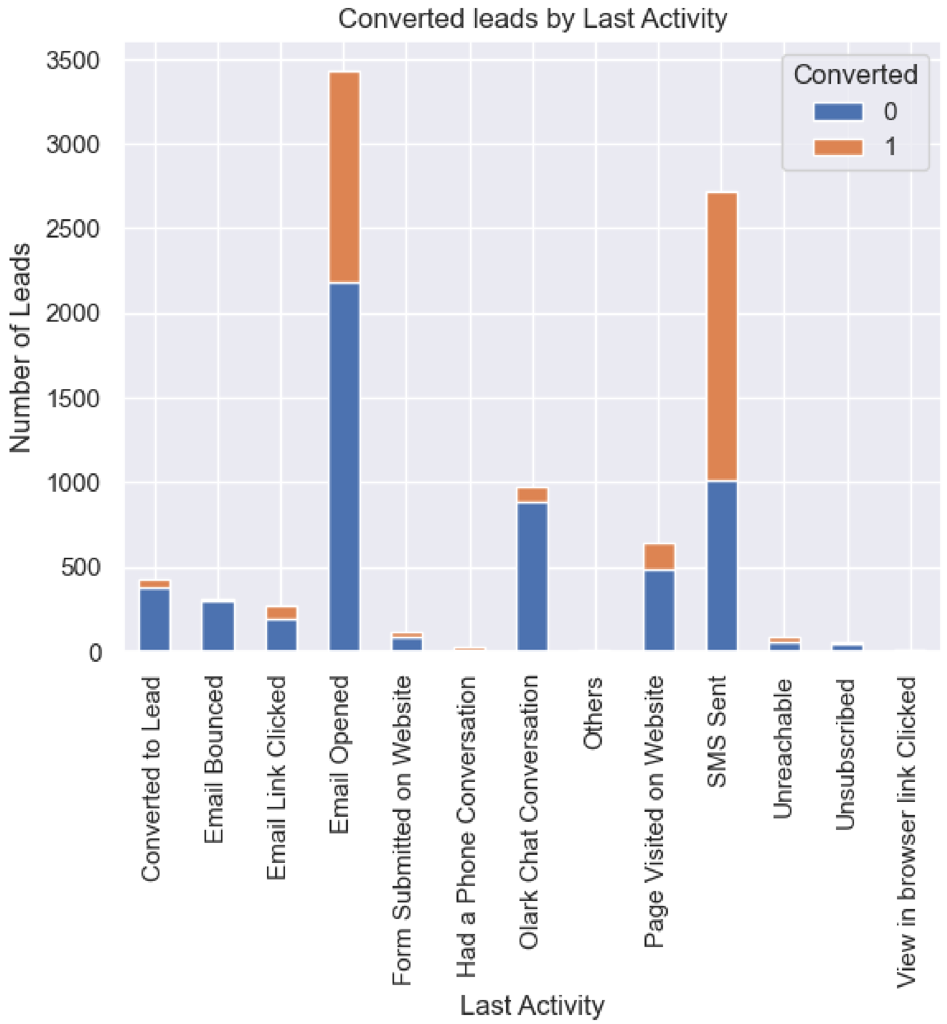
The "Lead Source" that contributes the most to converted leads is "Google“.

"Direct Traffic" and "Olark Chat" are also good sources of converted leads.

"Reference" and "Welingak Website" have a high number of conversions relative to their count of leads, making them potential strong lead sources.

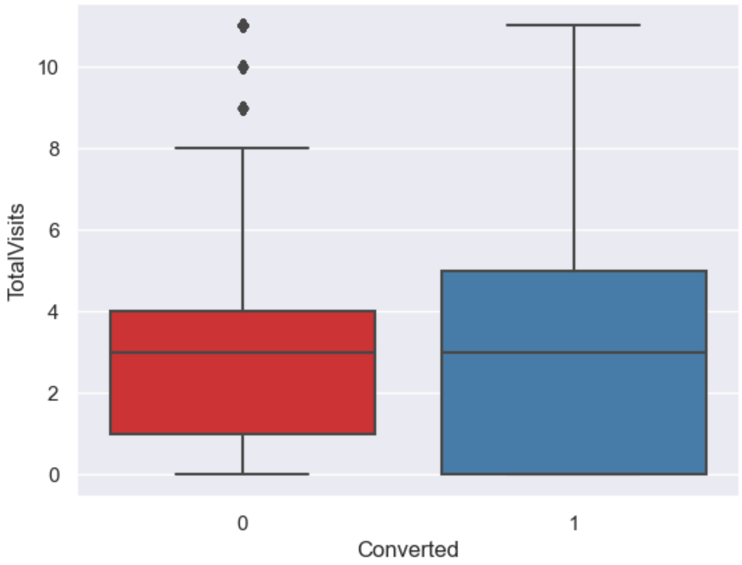
"Facebook" and "Others" have a relatively lower number of conversions compared to the count of leads, indicating room for improvement in converting leads from these sources.

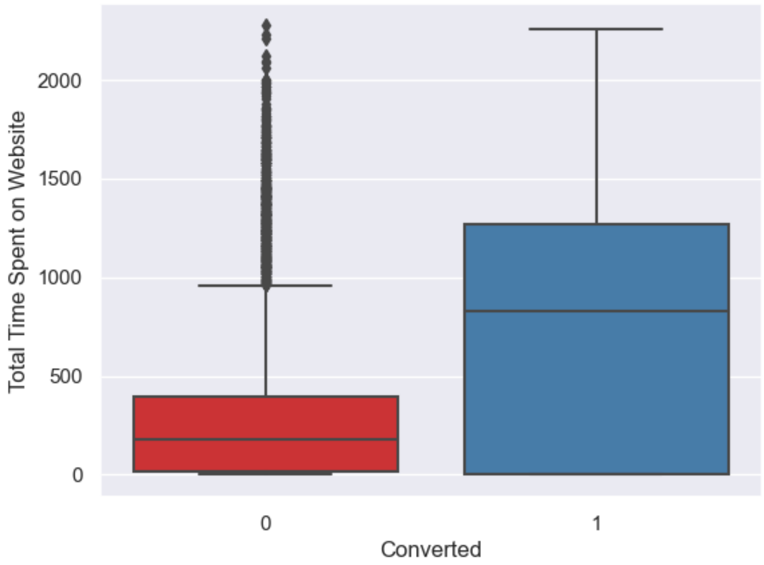
Exploratory Data Analysis

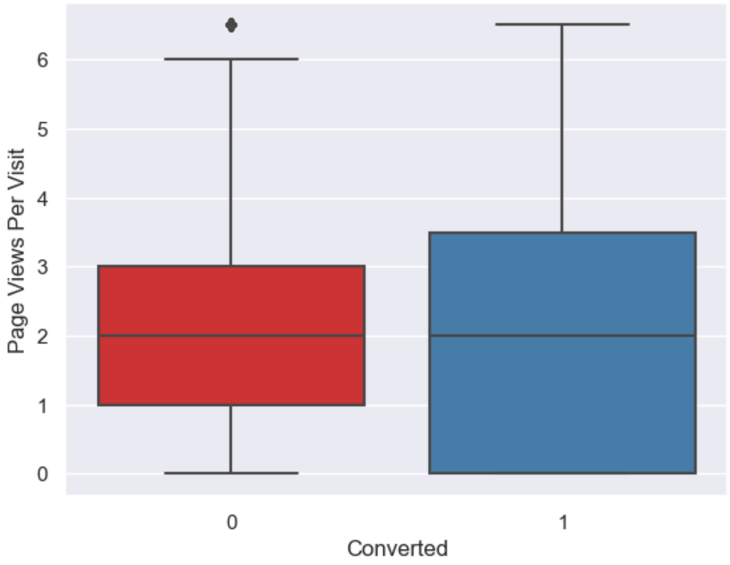


- Leads who have the last activity of "Email Opened" and "SMS Sent" show a higher conversion rate, while leads who have the last activity of "Email Bounced," "Form Submitted on Website," and "Unsubscribed" have a relatively lower conversion rate.

Exploratory Data Analysis



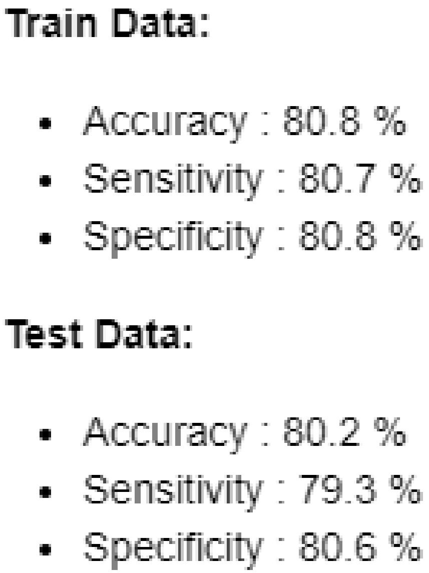




Regarding Numeric Variables, I observed the presence of outliers, and I effectively managed them using the Interquartile Range (IQR) method.

Additionally, I discovered a compelling trend: leads who invest more time on the website show a higher likelihood of conversion. To capitalize on this insight, it's advisable to enhance website engagement strategies, as they can potentially result in increased lead interaction and conversion rates.

Model Building



Train Data:

- Accuracy : 80.8 %
- Sensitivity : 80.7 %
- Specificity : 80.8 %

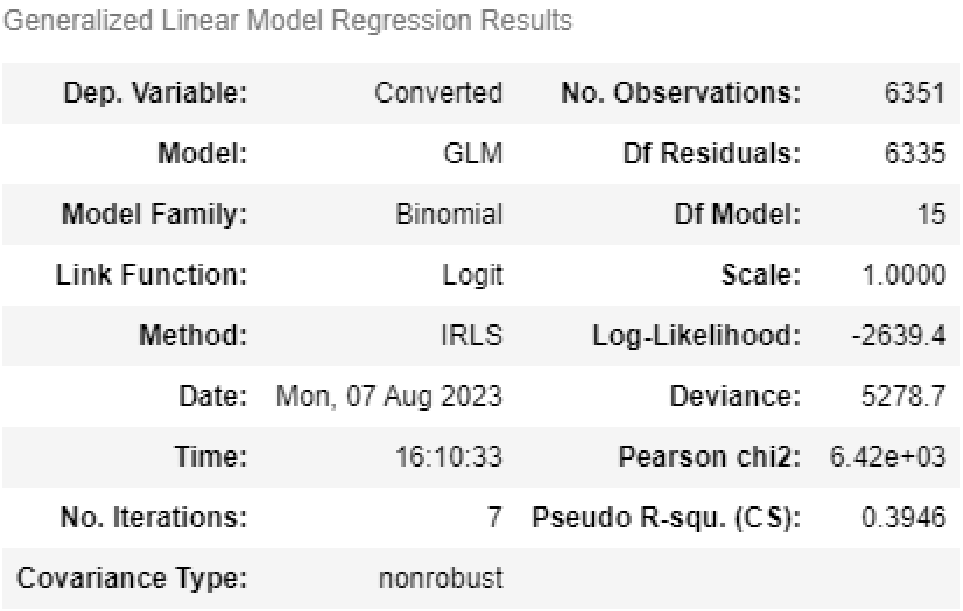
Test Data:

- Accuracy : 80.2 %
- Sensitivity : 79.3 %
- Specificity : 80.6 %

- Utilizing logistic regression, we constructed a predictive model that assigns a lead score ranging from 0 to 100. This score acts as an indicator of conversion probability, empowering the sales team to focus their efforts where they matter most.
- Selection of significant variables using p-values and VIF
- The model demonstrates solid performance with accuracy around 80% for both train and test data. While sensitivity is slightly lower in the test set, the model maintains consistent specificity levels.

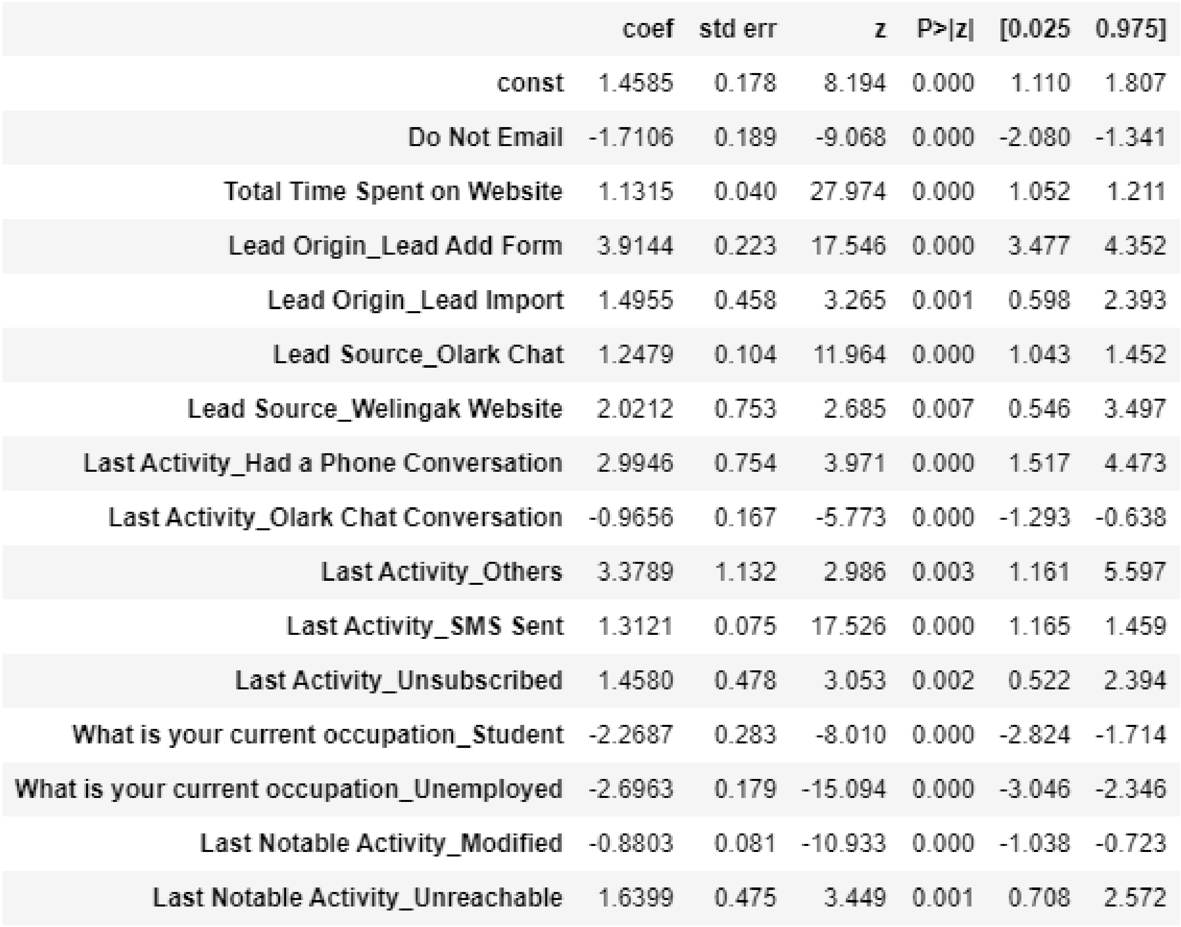
Model Building

Final model summary :



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:	-2639.4
Date:	Mon, 07 Aug 2023	Deviance:	5278.7
Time:	16:10:33	Pearson chi2:	6.42e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3946
Covariance Type:	nonrobust		



	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4585	0.178	8.194	0.000	1.110	1.807
Do Not Email	-1.7106	0.189	-9.068	0.000	-2.080	-1.341
Total Time Spent on Website	1.1315	0.040	27.974	0.000	1.052	1.211
Lead Origin_Lead Add Form	3.9144	0.223	17.546	0.000	3.477	4.352
Lead Origin_Lead Import	1.4955	0.458	3.265	0.001	0.598	2.393
Lead Source_Olark Chat	1.2479	0.104	11.964	0.000	1.043	1.452
Lead Source_Welingak Website	2.0212	0.753	2.685	0.007	0.546	3.497
Last Activity_Had a Phone Conversation	2.9946	0.754	3.971	0.000	1.517	4.473
Last Activity_Olark Chat Conversation	-0.9656	0.167	-5.773	0.000	-1.293	-0.638
Last Activity_Others	3.3789	1.132	2.986	0.003	1.161	5.597
Last Activity_SMS Sent	1.3121	0.075	17.526	0.000	1.165	1.459
Last Activity_Unsubscribed	1.4580	0.478	3.053	0.002	0.522	2.394
What is your current occupation_Student	-2.2687	0.283	-8.010	0.000	-2.824	-1.714
What is your current occupation_Unemployed	-2.6963	0.179	-15.094	0.000	-3.046	-2.346
Last Notable Activity_Modified	-0.8803	0.081	-10.933	0.000	-1.038	-0.723
Last Notable Activity_Unreachable	1.6399	0.475	3.449	0.001	0.708	2.572

Key Insight

The company should prioritize making calls to leads with the following characteristics based on the coefficients from the final model:

Higher probabilities of conversion (Positive coefficients):

- Lead Origin: Lead Add Form, Lead Import
- Lead Source: Olark Chat, Welingak Website
- Last Activity: Had a Phone Conversation, SMS Sent

Conclusion: We opted for logistic regression in lead scoring due to its strong predictive ability in determining lead conversions. It leverages the probability of conversion to assign lead scores effectively. Logistic regression provides clear insights into the relationship between variables and the likelihood of conversion. This transparency empowers us to understand and communicate the factors influencing lead conversion, crucial for informed decision-making.