Logistic Regression Model-Lead Scoring Case Study

Logistic Regression Assignment

Introduction - Problem Statement

- X Education, a prominent online course provider, faces a challenge with its lead conversion rate.
- Despite attracting numerous leads, only a fraction convert into paying customers.
- The aim of this case study is to develop a logistic regression model for lead scoring, identifying potential leads more efficiently.
- Through this model, we intend to significantly improve the lead conversion rate from the current 30% to the targeted 80%.
- In this presentation, we will explore the data, delve into model building, and provide actionable recommendations for X Education's sales strategy.

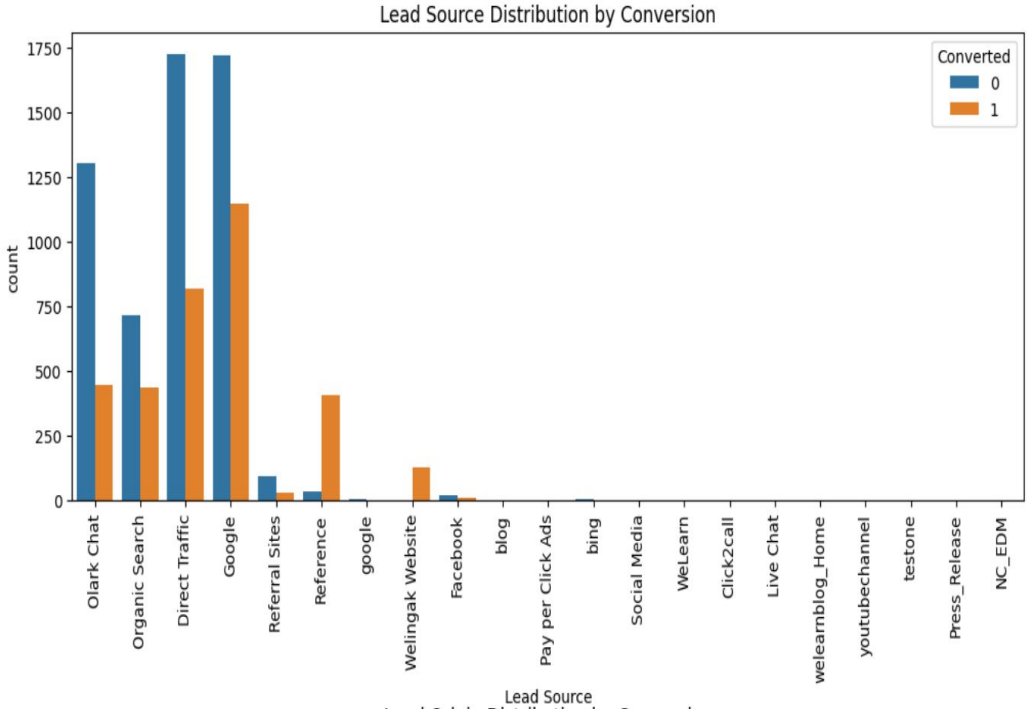
Key Objectives:

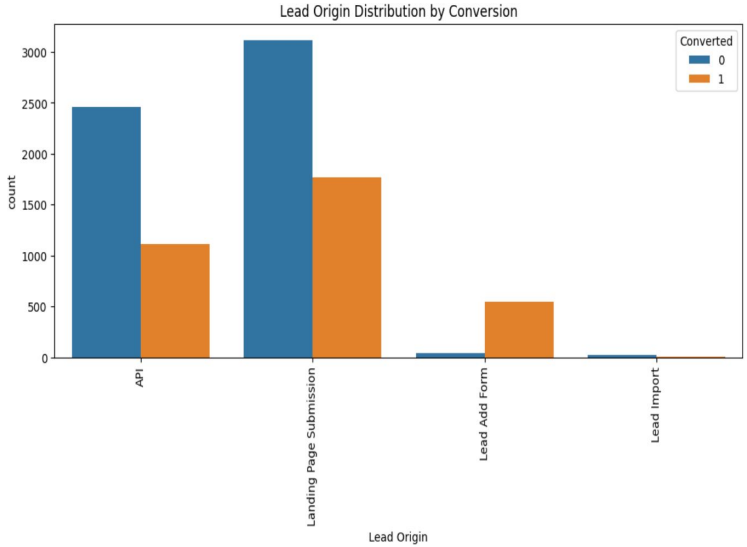
- Build a logistic regression model for lead scoring.
- Assign lead scores to identify potential hot leads.
- Achieve a target lead conversion rate of 80%.
- Provide solutions to additional problems posed by the company.
- Present a comprehensive analysis and recommendations for X Education.

Data Preparation

- Converted responses to binary by representing "YES" as 1 and "NO" as 0.
- Replaced instances of "Select" with NaN values.
- Removed columns with a significant number of missing values.
- Imputed missing values where necessary.
- Removed rows to mitigate the impact of missing values.

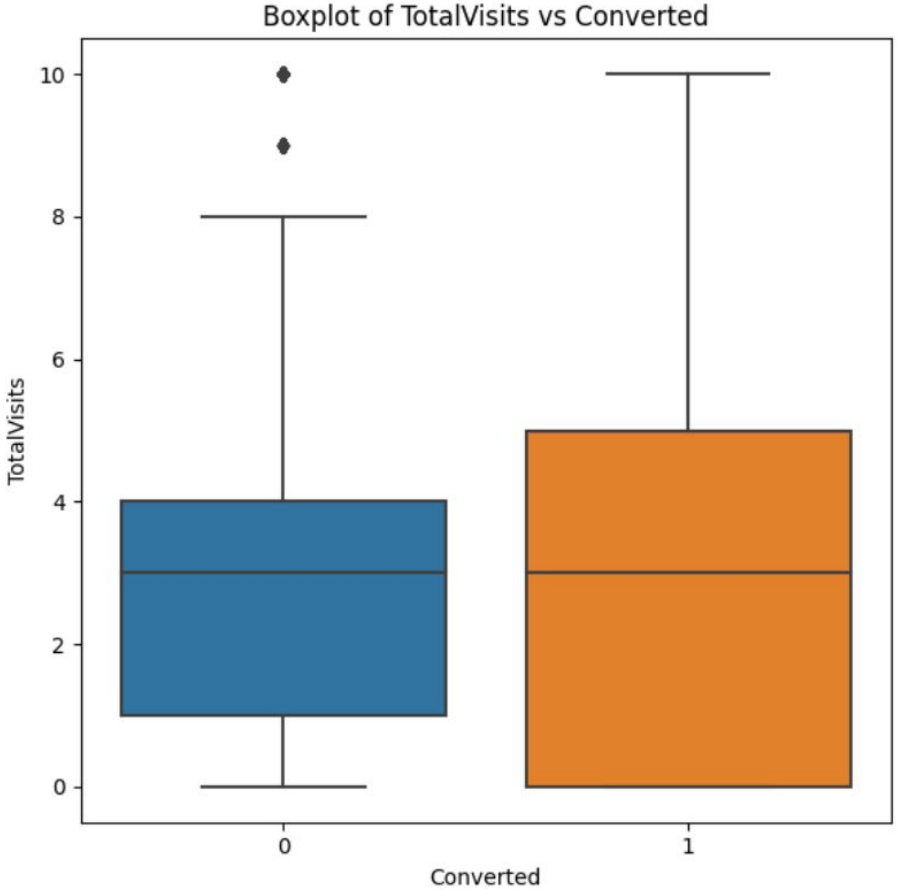
Exploratory Data Analysis:

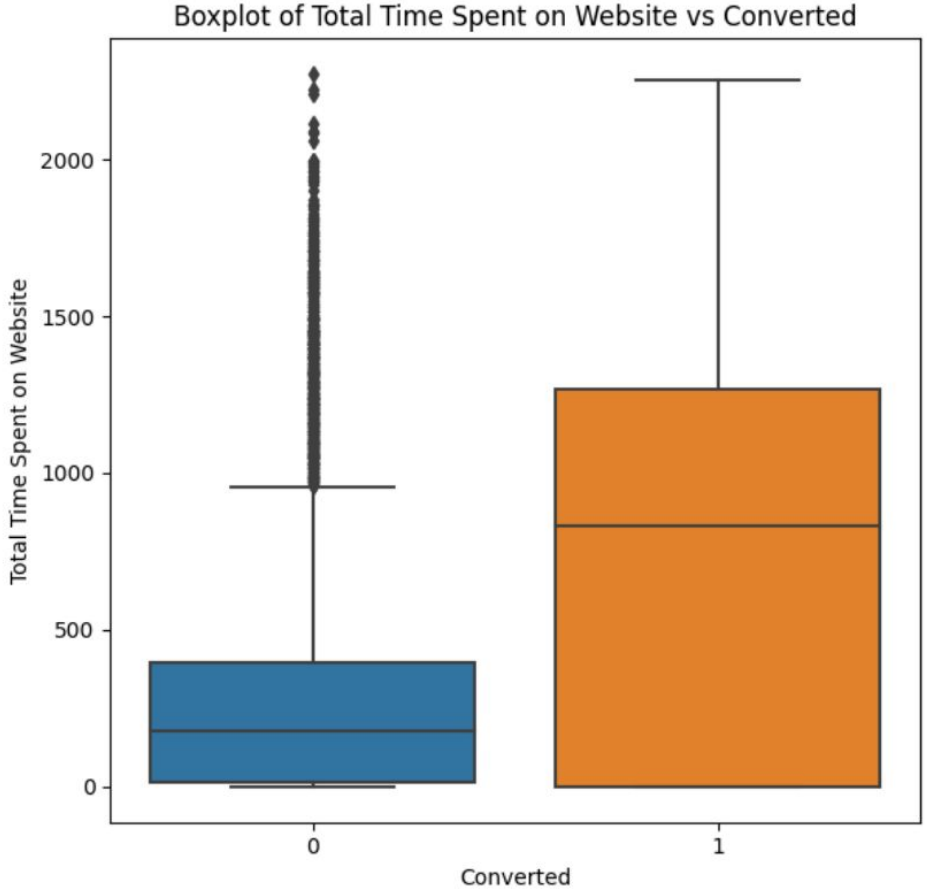




Observations:

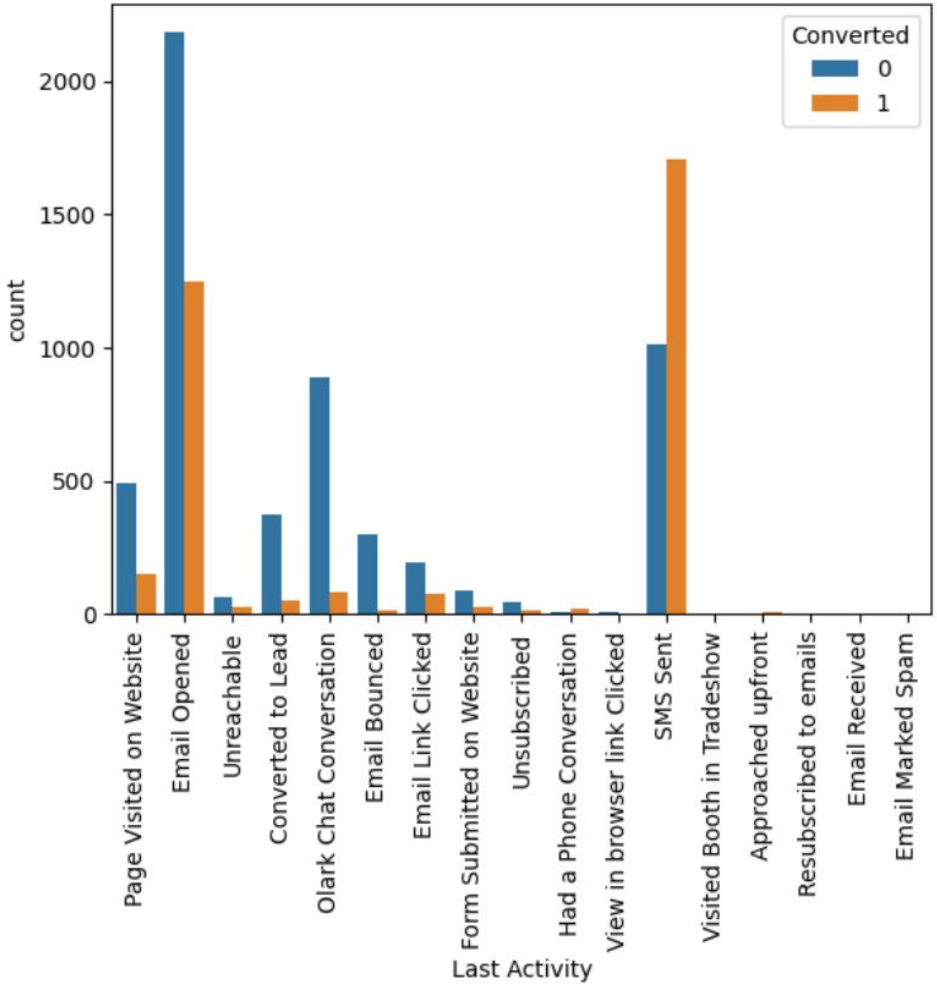
- API and Landing Page Submission exhibit a conversion rate of approximately 30%, with a notable number of leads generated.
- The Lead Add Form, despite having a low lead count, demonstrates a remarkably high conversion rate.
- Lead Import, with both low lead count and conversion rate, appears to have minimal impact in the analysis.

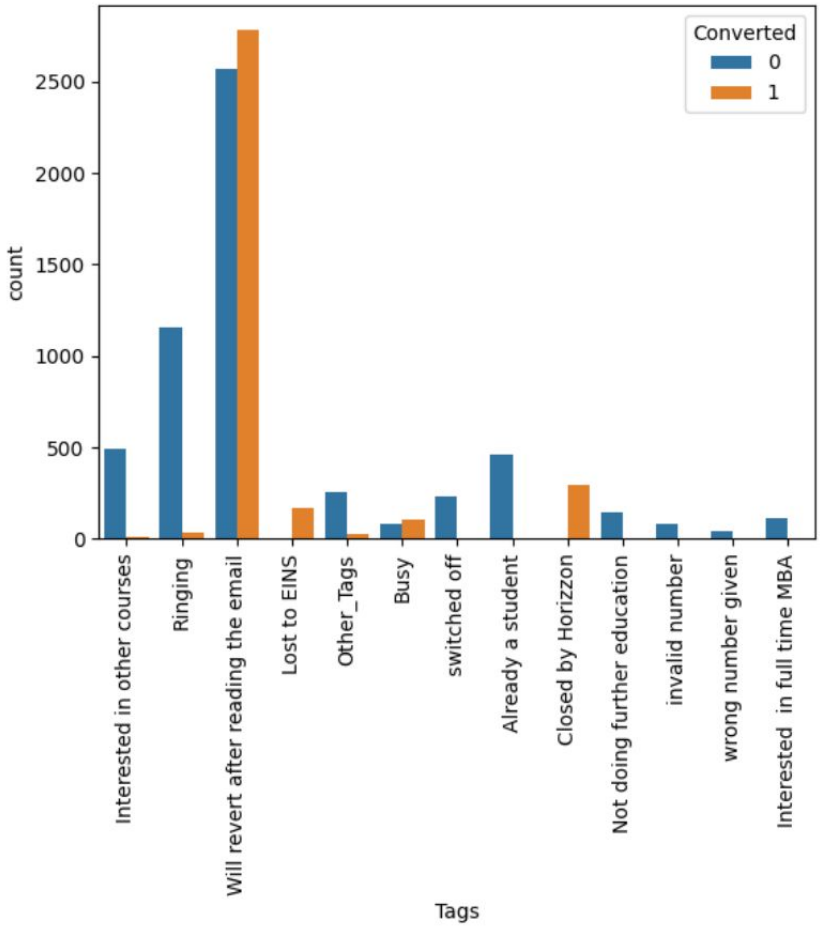

Boxplot of TotalVisits vs Converted
TotalVisits
Converted
0
1
0
2
4
6
8
10


Boxplot of Total Time Spent on Website vs Converted
Total Time Spent on Website
Converted
0
1
0
500
1000
1500
2000

Observations:

- The median values for both conversion and non-conversion cases are identical, providing inconclusive insights based on this information.
- Users who spend more time on the website tend to have a higher likelihood of conversion.


Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity


Converted
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
Other_Tags
Busy
switched off
Already a student
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Tags

Observations:

Last Activity:

- The highest count of last activities is recorded for "Email Opened."
- The maximum conversion rate is observed for leads with the last activity being "SMS Sent."

Tags:

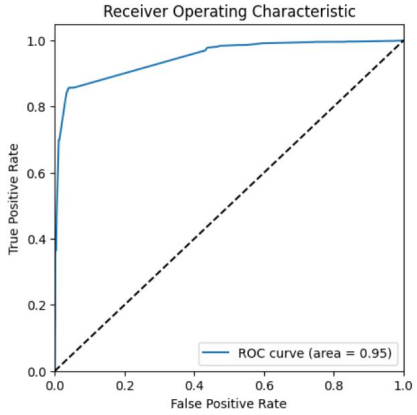
- The lead tags 'Will revert after reading the email' and 'Closed by Horizzon' exhibit a high conversion rate.

Summary:

- To enhance the overall lead conversion rate, focus on improving the conversion rates of leads originating from 'API' and 'Landing Page Submission,' as well as increasing the number of leads from 'Lead Add Form.'
- Additionally, target efforts to enhance the conversion rates of leads from sources like 'Google,' 'Olark Chat,' 'Organic Search,' and 'Direct Traffic.' Increase the lead count from sources like 'Reference' and 'Welingak Website' to positively impact overall conversion.
- Optimize website design to make it more engaging, increasing the time users spend on the site.
- For leads with the last activity as 'Email Opened,' consider making direct calls to improve conversion. Additionally, focus on increasing the count of leads with the last activity as 'SMS Sent.'

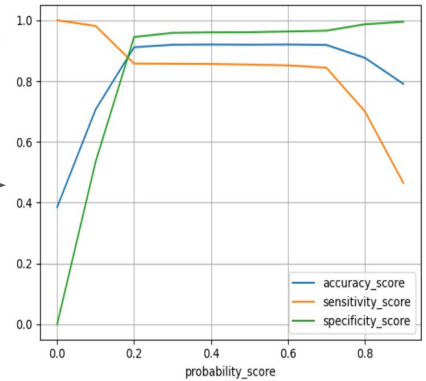
Logistic Regression Model Building

- Splitting the data into test and training sets
- The dataset has been divided into training and test sets with a ratio of 70:30.
- Utilizing Recursive Feature Elimination (RFE) to select the top 15 variables.
- Building the model by excluding variables with a p-value greater than 0.05 and Variance Inflation Factor (VIF) greater than 5.
- Generating predictions on the test dataset.
- Achieving an overall accuracy of 91.0%.



Roc Curve

Optimum Cut-off



Evaluation:

* Computed accuracy, sensitivity, and specificity across different probability cutoffs ranging from 0.1 to 0.9.

* Analyzing the graph and considering other metrics, the optimal cutoff point is identified to be 0.19.

probability_score	accuracy_score	sensitivity_score	specificity_score	precision_score
0.0	0.385136	1.000000	0.000000	0.385136
0.1	0.706503	0.981194	0.534443	0.568990
0.2	0.911195	0.857318	0.944942	0.907007
0.3	0.919383	0.856909	0.958515	0.928255
0.4	0.920170	0.856092	0.960307	0.931080
0.5	0.919540	0.854047	0.960563	0.931342
0.6	0.920170	0.851594	0.963124	0.935339
0.7	0.918753	0.843827	0.965685	0.939035
0.8	0.876397	0.700327	0.986684	0.970538
0.9	0.790742	0.464841	0.994878	0.982714

TRAIN DATA - CONFUSION MATRIX

PREDICTED ACTUAL	**NOT CONVERTED**	**CONVERTED**
NOT CONVERTED	3690	215
CONVERTED	349	2097

ACCURACY	0.911
PRECISION	0.907
SENSITIVITY	0.857
SPECIFICITY	0.944

Prediction:

TESTDATA - CONFUSION MATRIX

PREDICTED ACTUAL	**NOT CONVERTED**	**CONVERTED**
NOT CONVERTED	1640	94
CONVERTED	157	832

ACCURACY	0.907
PRECISION	0.898
SENSITIVITY	0.857
SPECIFICITY	0.945

Conclusion:

1. **Effective Lead Scoring:** The Logistic Regression Model, with a cutoff probability of 0.19, provides an efficient lead scoring system for X Education.

2. **Top Influential Features:** 'Tags_Lost to EINS,' 'Tags_Closed by Horizzon,' and 'Lead Quality_Worst' emerged as the most influential features, impacting conversion positively or negatively.

3. **High Predictive Accuracy:** Achieving a Sensitivity of 0.841 and Precision of 0.89, the model accurately identifies and prioritizes potential leads for conversion.

4. **Strategic Focus Areas:** Recommendations include optimizing conversion rates from specific lead sources, improving user engagement on the website, and targeting leads with specific last activities.

5. **Actionable Insights:** The Logistic Regression Model equips X Education with actionable insights to enhance lead conversion strategies and achieve the target conversion rate of 80%.