Predicting Lead Conversion: Logistic Regression for Lead Scoring

X Education
Lead Conversion
ML Assignment

Group Members:
Shayak Majumder
Shreya Dutta
Shweta

Problem Statement

Build a logistic regression model to predict the likelihood of lead conversion for a lead scoring system for X Education, enabling the company to prioritize leads based on the probability of conversion. This would help optimize resource allocation, enhancing sales efficiency.

Problem

- X Education struggles to efficiently prioritize leads, resulting in suboptimal resource allocation and slower conversions

Opportunity

- Optimizing lead conversion by predicting likelihood can increase sales efficiency and drive better targeting of high-potential leads

Solution

- Implement a logistic regression model to score leads, prioritizing high-conversion probability leads for targeted actions.

Analysis Approach

Dataset Overview

The dataset used for this project consisted of 9,240 leads, with 37 columns at first, which we whittled down to 14 features including Lead Number, demographics, and behavioural attributes. The target variable, 'Converted' (0 or 1), indicates whether a lead was successfully converted.

Step 1
Pre-processing data

Step 2
Data Visualization & EDA

Step 3
Building ML model

Step 4
Calculating performance metrics

Section 4

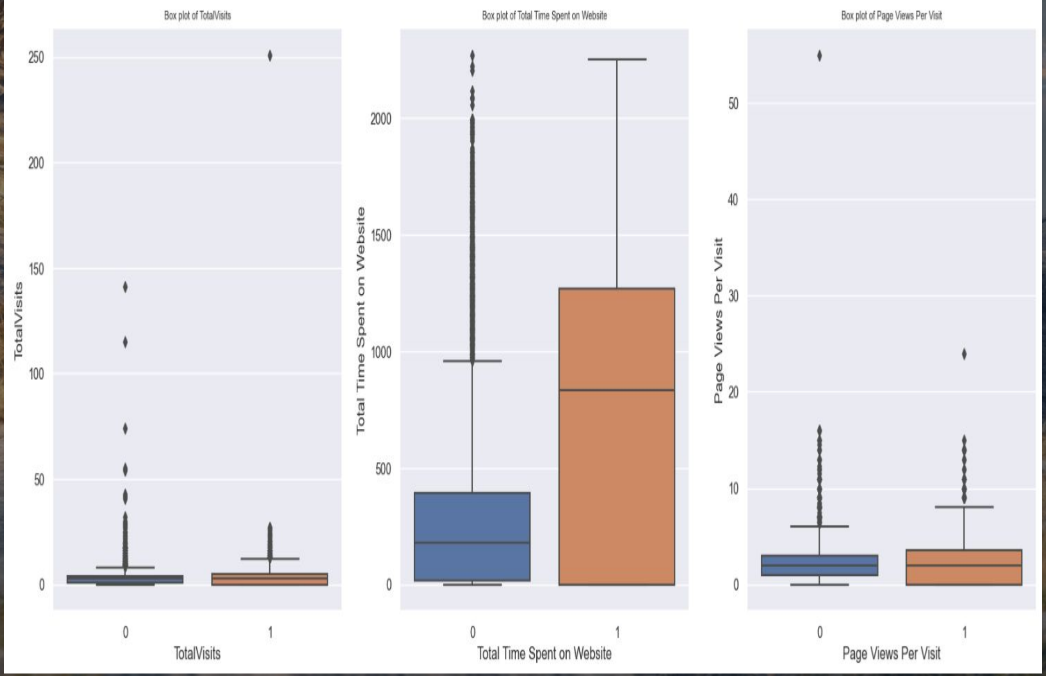
Pre-processing

Overview

Data preprocessing involved handling missing values by removing or imputing columns, excluding those with **over 40% missing data.**
Also, **columns with zero variance were dropped** as they weren't useful for ML model.
Lastly, **clubbed some data classes together** for easier readability for EDA.

Section 5

Data Visualization I



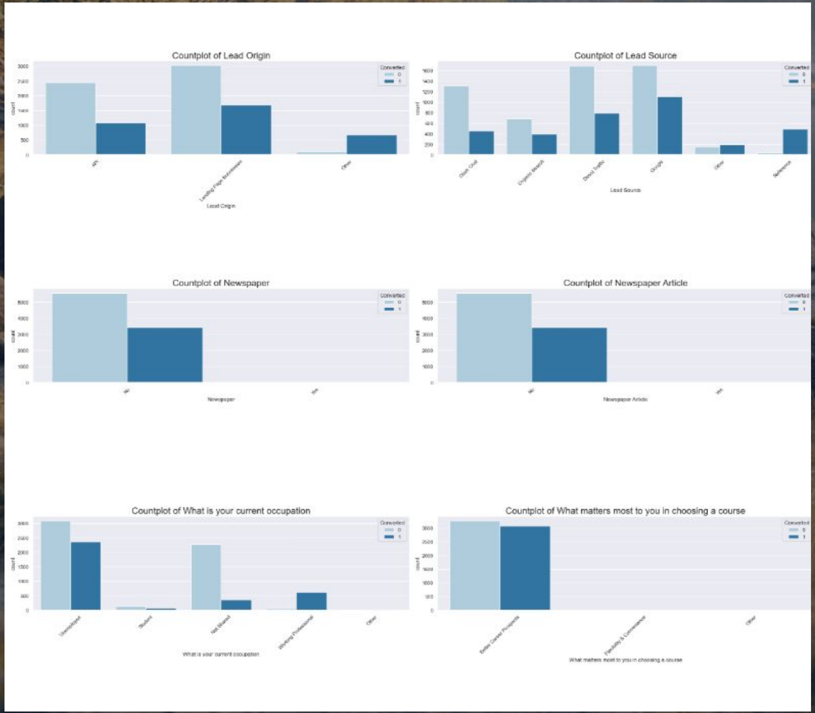
Overview

For Numerical Data, plotted pairplots and then box plots and inferred:

Median for **Total Time Spent On Website is visibly higher**, making these users the primary target for lead conversions.

Removed outliers after this step.

Data Visualization II

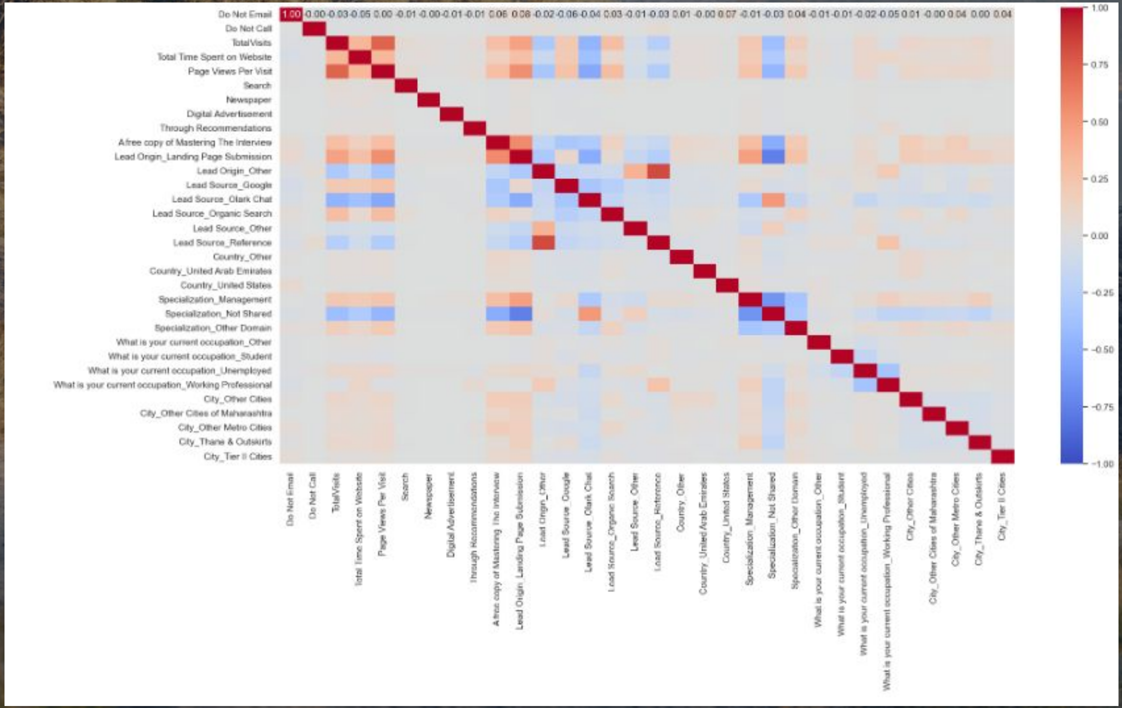


Overview

For Categorical Data, plotted countplots and inferred:

- Most of the customers are from **India** and from **Management specializations**.
- Users with +ve attributes in both **Last Activity** and **Last Notable Activity** have more conversions.
- The **Free Copy of Mastering The Interview has yielded more conversions** for those who have availed of the book.
- Users coming in through **Landing Page Submission** are showing more conversion success rates.
- **Google** and **Direct Traffic** are the biggest source of conversion.
- Odds of **Working Professional** and **Unemployed** to convert are high.

Building ML Models I



Overview

First, after train-test split, dealt with **remaining null values** for train data, **imputing** with mode or median.

Then **plotted heatmap** to check and **drop features with high correlations.**

Building ML Models II

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	6298
Model:	GLM	Df Residuals:	6284
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2815.3
Date:	Sat, 18 Jan 2025	Deviance:	5630.5
Time:	09:34:38	Pearson chi2:	6.73e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3536
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.4097	0.135	-25.166	0.000	-3.675	-3.144
Do Not Email	-1.1678	0.154	-7.565	0.000	-1.470	-0.865
TotalVisits	0.8645	0.241	3.581	0.000	0.391	1.338
Total Time Spent on Website	3.8051	0.137	27.863	0.000	3.537	4.073
Lead Origin_Landing Page Submission	-0.3381	0.109	-3.089	0.002	-0.553	-0.124
Lead Origin_Other	3.5027	0.180	19.428	0.000	3.149	3.856
Lead Source_Google	0.2416	0.079	3.067	0.002	0.087	0.396
Lead Source_Olark Chat	1.2019	0.135	8.896	0.000	0.937	1.467
Specialization_Management	0.3793	0.099	3.846	0.000	0.186	0.573
Specialization_Other Domain	0.4641	0.121	3.847	0.000	0.228	0.701
What is your current occupation_Other	1.7291	0.483	3.580	0.000	0.783	2.676
What is your current occupation_Student	1.0709	0.231	4.645	0.000	0.619	1.523
What is your current occupation_Unemployed	1.3479	0.086	15.709	0.000	1.180	1.516
What is your current occupation_Working Professional	3.6556	0.187	19.499	0.000	3.288	4.023

	Features	VIF
3	Lead Origin_Landing Page Submission	4.73
7	Specialization_Management	3.63
1	TotalVisits	2.95
11	What is your current occupation_Unemployed	2.85
2	Total Time Spent on Website	2.18
8	Specialization_Other Domain	1.96
5	Lead Source_Google	1.54
4	Lead Origin_Other	1.45
12	What is your current occupation_Working Profes...	1.42
6	Lead Source_Olark Chat	1.21
0	Do Not Email	1.09
10	What is your current occupation_Student	1.06
9	What is your current occupation_Other	1.02

Overview

Then, used **RFE** to prune unnecessary features. Then **manually dropped one feature at a time** based on **p-values**, until there were **no values over 0.05**. Lastly, checked to make sure **no VIFs are above 5**.
Given on left is our final model, **Model 8/logmod8**. *Given above* is our final VIF counts.

Performance Metrics

```
array([[3100,  790],
       [ 539, 1869]], dtype=int64)
```

As we can see,

True Positives (TP): 1869 - Correctly predicted positive cases.

True Negatives (TN): 3100 - Correctly predicted negative cases.

False Positives (FP): 790 - Negative cases incorrectly predicted as positive.

False Negatives (FN): 539 - Positive cases incorrectly predicted as negative.

Overview

The following metrics were used to evaluate the model:

- **Accuracy**: Proportion of correctly classified instances | **Sensitivity**: Ability of the model to correctly identify converted leads | **Specificity**: Ability of the model to identify non-converted leads | **Confusion matrix:** This was done to identify positive and negative cases and the nuances in between

The above was done by randomly selecting a cutoff of 0.5 first and then using **ROC** to decide a **cutoff of 0.3**.

Key Results:

- The final model achieved an **accuracy of 78.89% (~79%)**, sensitivity of **77.6% (~78%)**, and specificity of **79.7% (~80%)** on the training set. The f**inal confusion matrix is *shared above***.
- The lead scoring system assigned scores from 0 to 100 based on conversion probabilities, helping prioritize high-potential leads.

Suggestions For X Education

Deploy & Monitor

Integrate the lead scoring system into CRM platforms to guide sales teams. Regularly update and retrain the model with new data to maintain performance.

Prediction will show lead scores from 0 to 100, the higher the better

	Lead Number	Converted	Predicted	Conversion_Prob	Lead Score
2810	632785	1.0	1	0.999542	99.954230
97	659545	1.0	1	0.999509	99.950896
3182	629524	1.0	1	0.999259	99.925891
1931	640972	1.0	1	0.999241	99.924143
852	651501	1.0	1	0.999226	99.922558

Top Recommendations for Lead Conversion (based on final ML Model features):

- Focus on Leads with High Total Website Engagement
- Target "Working Professionals" and "Unemployed" Leads
- Leverage Google and Olark Chat as Lead Sources
- Maximize Outreach for Management Specialization Leads
- Capitalize on "Landing Page Submission" Leads


Thank
You