

Lead Score Group Assignment

DS C72

Submitted by: Karthikein M R, Kruthika K S, and Pham Kieu Linh

Index

- **Problem Statement**
 - **Data preprocessing**
 - **Exploratory Data Analysis**
 - **Model Building**
 - **Model Evaluation**
 - **Conclusion**
-

Problem Statement

- **Company:** X Education, an online course provider.
 - **Challenge:** Low lead conversion rate (~30%) despite generating numerous leads daily.
 - **Objective:** Develop a **Lead Scoring Model** using **Logistic Regression**.
 - **Approach:** Assign a lead score between **0 and 100** to prioritize potential leads.
 - **Goal:** Improve sales efficiency and **increase the conversion rate to ~80%**.
-

Data Preprocessing

Summary

1. Handling Missing Data

- **Dropped columns with >45% missing values** if insignificant (e.g., Asymmetrique Index & Score). However Lead Quality was kept despite 51% missing as it might impact on conversion rate.
- Imputed missing values with **mode** or labeled as **'unknown'**
- Treated **'Select'** responses as missing and replaced them with **'unknown'**.

2. Feature Selection:

- Removed columns with **only one value** (e.g., Magazine, Supply Chain Content).
- Grouped **low-frequency categories** into **'Other'** to reduce dimensionality.

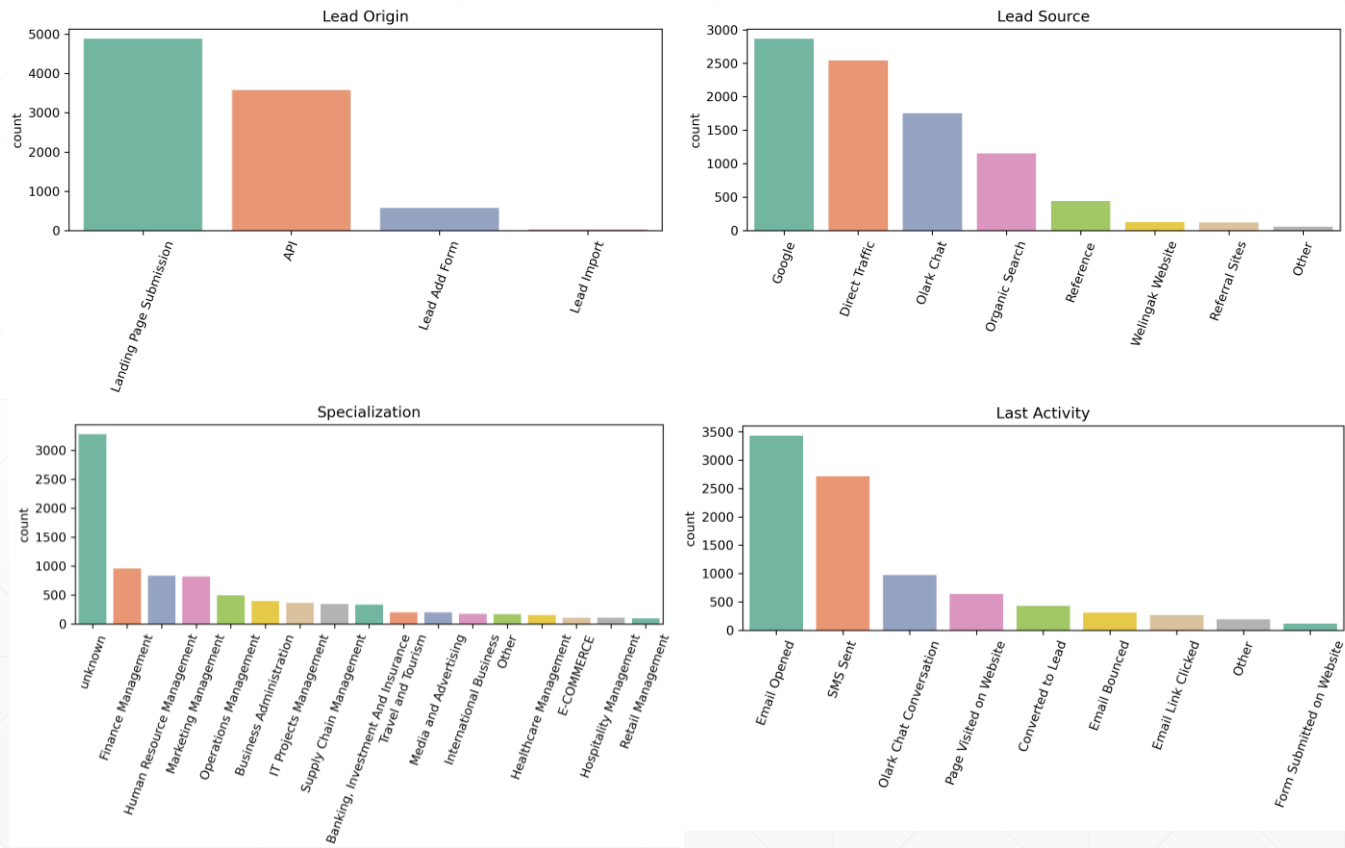
3. Outlier Treatment:

Outliers in Total Visits & Page Views Per Visit were identified but **retained** as they were **business-relevant**.

Exploratory Data Analysis

Univariate Analysis

Categorical columns



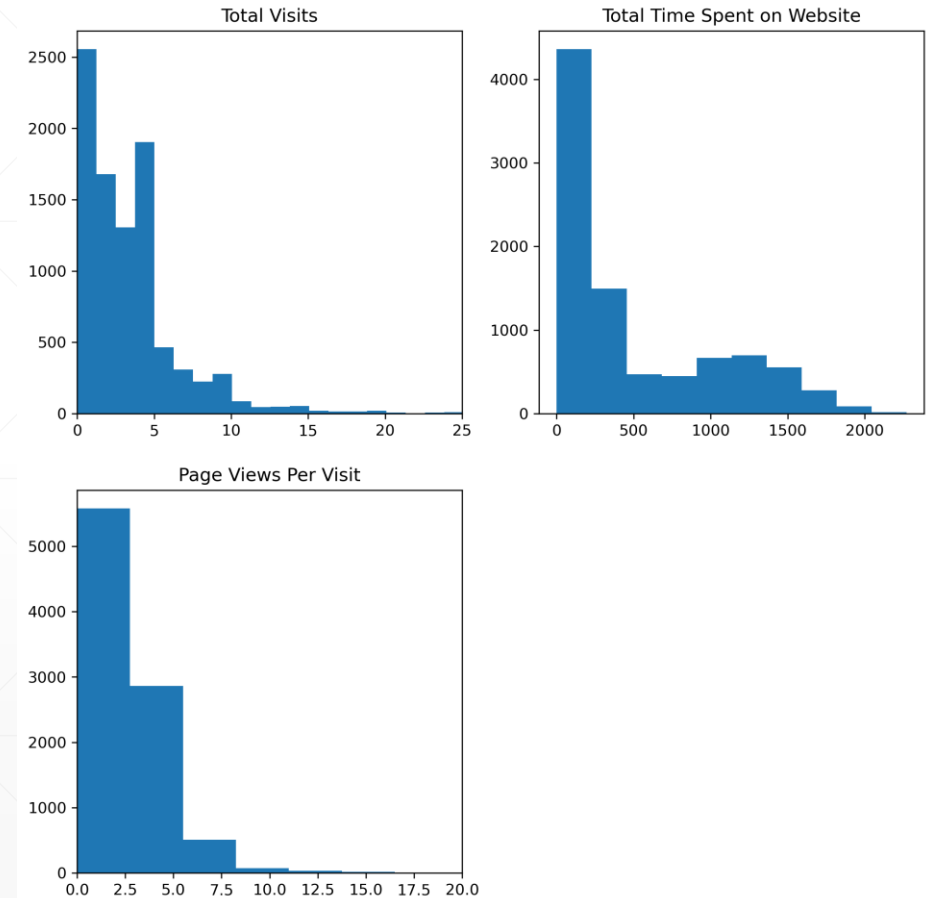
- **Lead Origin:** The majority of leads come from "Landing Page Submission," indicating that website visitors are a major source of potential customers.
- **Lead Source:** "Google" and "Direct Traffic" contribute significantly to lead generation, while other sources have lower representation.
- **Specialization:** The most common specializations among leads are "Marketing Management", "Human Resource Management" and "Finance Management," suggesting these fields have high interest in courses.
- **Last Activity:** "Email Opened" and "SMS Sent" are the most frequent activities, implying that email and SMS marketing play a crucial role in engagement.

Exploratory Data Analysis

Numerical columns

- **Skewed Distribution:** All three metrics—Total Visits, Total Time Spent on Website, and Page Views Per Visit—are highly right-skewed, indicating that most users have low engagement, while a few spend significantly more time or visit more pages.
- **Majority of Users Have Low Interaction:** Most users have fewer than 5 total visits, less than 500 seconds spent on the website, and under 3 page views per visit, suggesting that engagement strategies may be needed to retain visitors and increase conversions.

Univariate Analysis

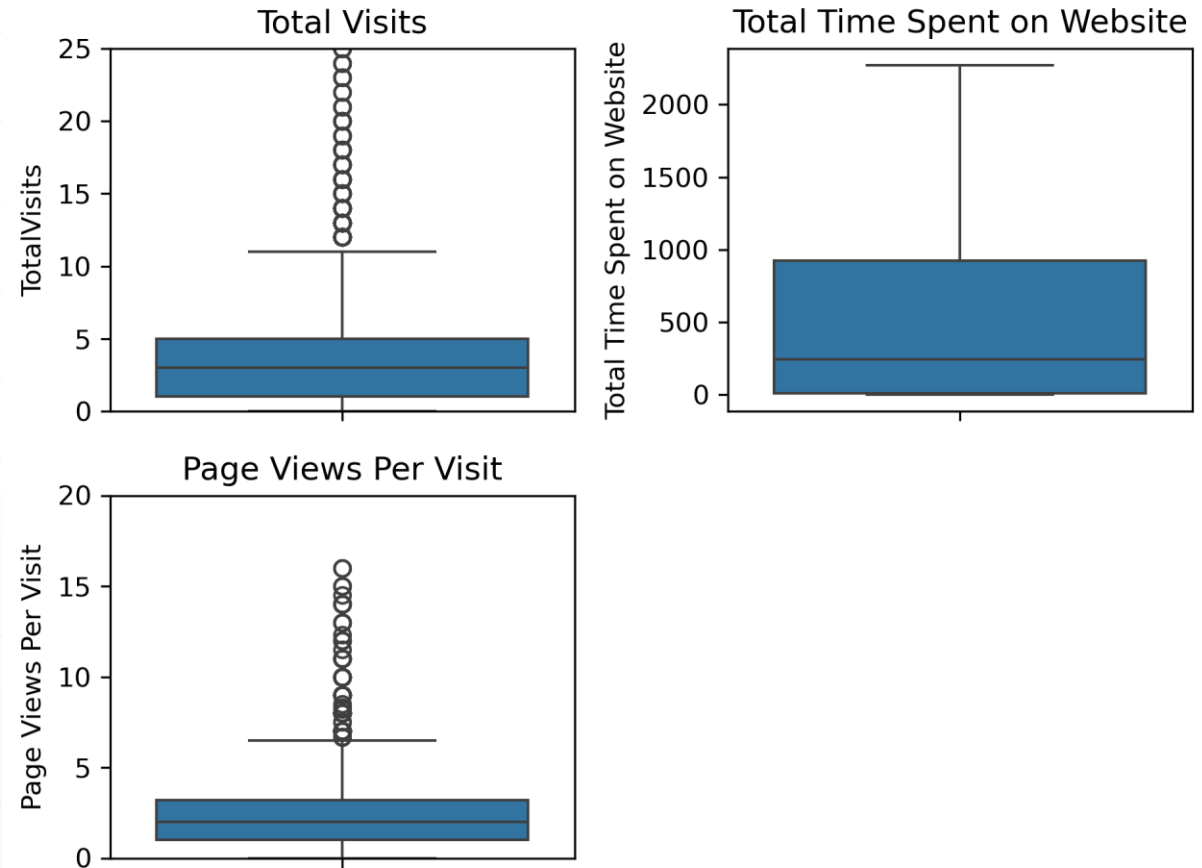


Exploratory Data Analysis

Numerical columns for checking outliers

- **Presence of Outliers:** The Total Visits and Page Views Per Visit metrics have a significant number of outliers, indicating that a small group of users engage with the website much more frequently than the majority. These high-activity users may represent highly interested leads.
- In this case, the outliers are not due to incorrect data but are logically valid from a business perspective. Therefore, they are retained in the analysis

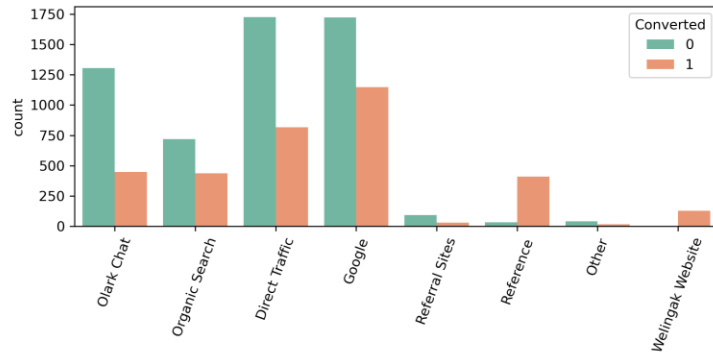
Univariate Analysis



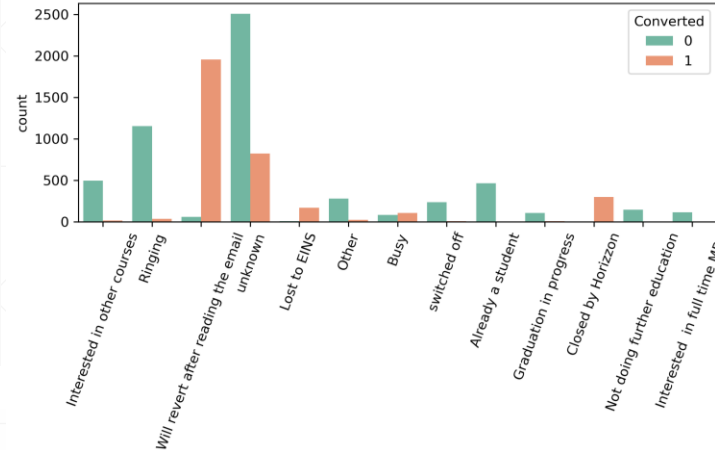
Exploratory Data Analysis

Bivariate Analysis

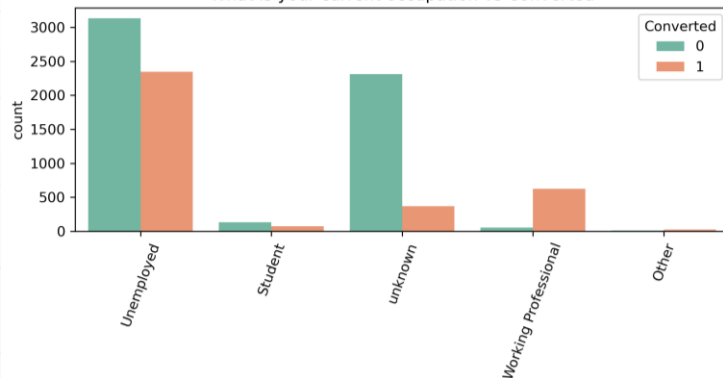
Lead Source vs Converted



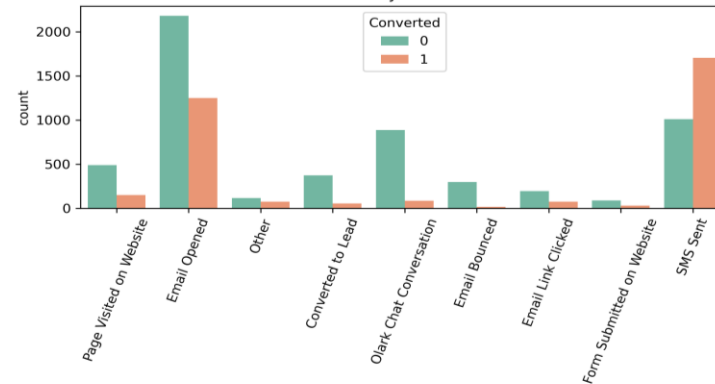
Tags vs Converted



What is your current occupation vs Converted



Last Activity vs Converted



- **Lead Source Impact:** Leads coming from the "Welingak Website" and "Reference" have a significantly higher conversion rate compared to other sources, indicating a strong trust factor.
- **Tags Influence:** Leads tagged as "Will revert after reading the email" shows a higher likelihood of conversion, suggesting that engagement signals are valuable in predicting lead quality.
- **Occupation Matters:** Working professionals have a noticeably higher conversion rate than students or unemployed individuals, likely due to their financial capability to invest in education.
- **Last Activity Effect:** Leads whose last activity involved "SMS Sent" have a higher conversion rate, emphasizing the effectiveness of direct and personalized communication.

Model Building

Approach

A Logistic Regression Model was chosen for its interpretability and effectiveness in binary classification. The dataset was split into train (70%) and test (30%). Key steps included:

- **Data Preprocessing:** Applied one-hot encoding and Min-Max scaling.
 - **Feature Selection:** Used Recursive Feature Elimination (RFE) and Variance Inflation Factor (VIF) to retain relevant predictors and eliminate multicollinearity.
 - The **cut-off point was chosen based on the Sensitivity-Specificity Intersection**. Since this problem requires balancing between converted and non-converted customers while the dataset is not highly imbalanced, selecting the intersection of Sensitivity and Specificity as the optimal cut-off ensures a well-generalized model that minimizes both false positives and false negatives
-

Model Building

Result

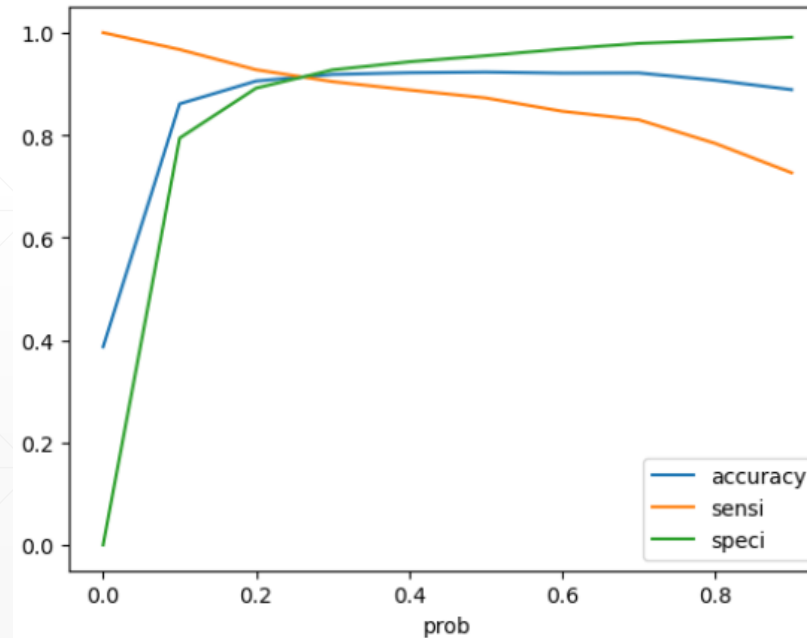
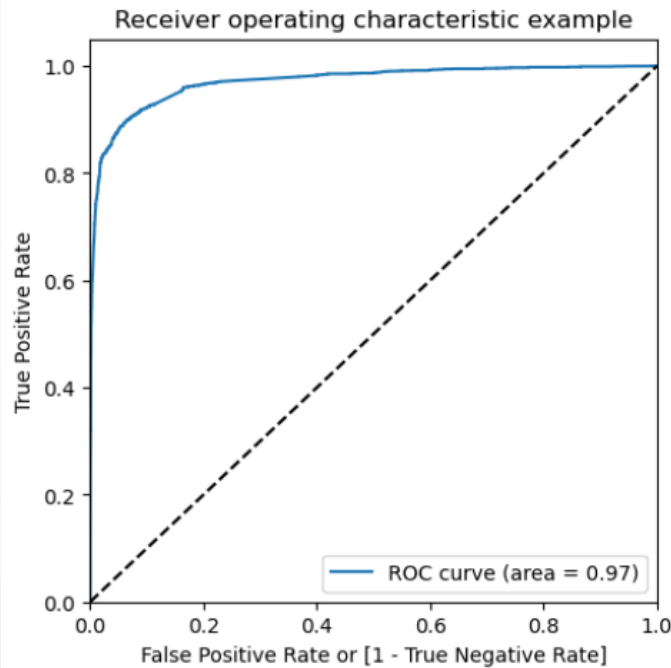
	coef	std err	z	P> z	[0.025	0.975]
const	-3.0170	0.134	-22.516	0.000	-3.280	-2.754
Total Time Spent on Website	4.7992	0.255	18.791	0.000	4.299	5.300
Lead Source_Welingak Website	4.9626	1.022	4.858	0.000	2.960	6.965
Last Activity_Email Bounced	-1.9022	0.487	-3.906	0.000	-2.857	-0.948
Country_unknown	1.4938	0.145	10.284	0.000	1.209	1.778
What is your current occupation_unknown	-0.9436	0.118	-7.976	0.000	-1.175	-0.712
Tags_Busy	0.7279	0.244	2.978	0.003	0.249	1.207
Tags_Closed by Horizzon	6.2138	0.736	8.441	0.000	4.771	7.657
Tags_Lost to EINS	5.8633	0.601	9.762	0.000	4.686	7.040
Tags_Ringing	-3.6896	0.247	-14.961	0.000	-4.173	-3.206
Tags_Will revert after reading the email	4.5790	0.199	23.006	0.000	4.189	4.969
Tags_switched off	-4.6898	0.745	-6.295	0.000	-6.150	-3.230
Lead Quality_Worst	-3.8314	0.565	-6.782	0.000	-4.939	-2.724
Last Notable Activity_Other	1.3356	0.418	3.198	0.001	0.517	2.154
Last Notable Activity_SMS Sent	2.6957	0.128	21.052	0.000	2.445	2.947

From a business standpoint, the following variables have a significant influence on lead conversion:

- **Time Spent on Website** (coefficient = 4.7992): The longer a person stays on the website, the more likely they are to convert.
- **Lead Source - Welingak Website** (coefficient = 4.9626): Leads coming from this website have a much higher chance of conversion.
- **Tags - Lost to EINS** (coefficient = 5.8633): Even though these leads were marked as "lost," they still have a strong chance of converting.

Model Evaluation

- The default threshold of **0.5** was initially used but was found to be suboptimal for balancing model performance.
- The **ROC Curve** shows that the model performs well with an **AUC of 0.97**, indicating a strong ability to distinguish between converted and non-converted leads.
- The **Sensitivity-Specificity Trade-off Curve** was used to determine the best cut-off



Final Decision:

- **0.25 was selected as the optimal cut-off** as it offers the best balance between correctly identifying converted leads (high sensitivity) and avoiding false positives (maintaining specificity).
- This cut-off ensures that the sales team can prioritize high-potential leads effectively while minimizing wasted outreach on unlikely conversions.

Model Building

Result

Evaluation Metric	Train Set	Test Set
Accuracy	91.30%	90.40%
Sensitivity (Recall)	91.80%	91.70%
Specificity	91%	89.70%
False Positive Rate	9%	10%
Positive Predictive Value	86.50%	83%
Negative Predictive Value	94.60%	95.10%

The model demonstrates strong generalization with minimal overfitting, as the accuracy remains high and consistent between the train set (91.3%) and test set (90.4%). Additionally, sensitivity (91.8% train, 91.7% test) and specificity (91% train, 89.7% test) indicate a well-balanced performance in identifying both converted and non-converted leads.

Suggestion

- 1. Prioritize High-Scoring Leads:** Focus outreach (calls, emails) on leads with scores 80+ to maximize conversion efficiency.
 - 2. Optimize Marketing Efforts:** Invest in Welingak Website, website engagement, SMS campaign and follow-ups for leads tagged as "Will Revert After Reading Email" to improve conversion rates.
 - 3. Adjust Lead Handling by Business Phases:** Lower cut-off (e.g. 0.25) during peak sales for wider outreach; raise it (0.35 - 0.5) during slow periods to focus on high-quality leads.
-