LEAD SCORING CASE STUDY

PROBLEM STATEMENT

The problem

- X Education generates a high volume of leads but has a low conversion rate (~30%)
- The sales team needs a better system to focus on promising leads and improve efficiency.

Business goals

- Build a model to assign lead scores and identify "hot leads
- Increase the conversion rate close to 80% by prioritizing high-potential leads

DATA OVERVIEW

Data overview

- **Total Records:** 9,240 leads
- **Features:** 37 columns (e.g., Lead Source, Time on Website, etc.)
- **Target Variable:** Converted (1 = Converted, 0 = Not Converted)

Challenges

- Missing values in multiple columns
- Imbalanced dataset: ~36% leads converted
- Presence of redundant and uninformative features

```
Dataset Summary:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object
 11  Country                                        6779 non-null   object
 12  Specialization                                 7802 non-null   object
 13  How did you hear about X Education             7033 non-null   object
 14  What is your current occupation                6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                         9240 non-null   object
 17  Magazine                                       9240 non-null   object
 18  Newspaper Article                              9240 non-null   object
...
```

```
                                               Missing Values  Percentage
Lead Quality                                             4767   51.590909
Asymmetrique Activity Index                              4218   45.649351
Asymmetrique Profile Score                               4218   45.649351
Asymmetrique Activity Score                              4218   45.649351
Asymmetrique Profile Index                               4218   45.649351
Tags                                                     3353   36.287879
Lead Profile                                             2709   29.318182
What matters most to you in choosing a course            2709   29.318182
What is your current occupation                          2690   29.112554
Country                                                  2461   26.634199
How did you hear about X Education                       2207   23.885281
Specialization                                           1438   15.562771
City                                                     1420   15.367965
Page Views Per Visit                                      137    1.482684
TotalVisits                                               137    1.482684
Last Activity                                             103    1.114719
Lead Source                                                36    0.389610
```

DATA PREPARATION

Handling missing values

- Imputed categorical columns with the mode
- Imputed numerical columns with the median
- Dropped irrelevant features with high missing values

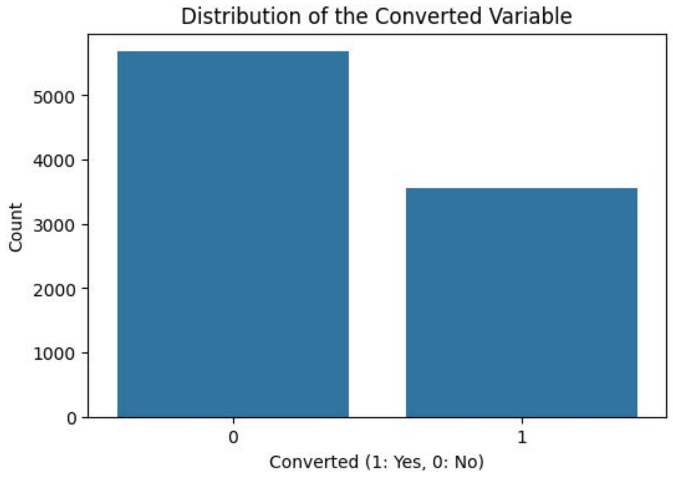
Feature engineering

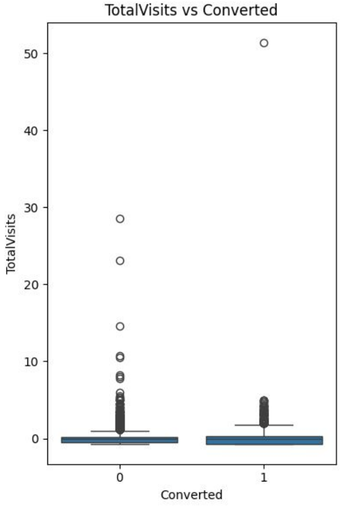
- Created dummy variables for categorical features
- Standardized numerical features (e.g., Total Visits, Page Views)

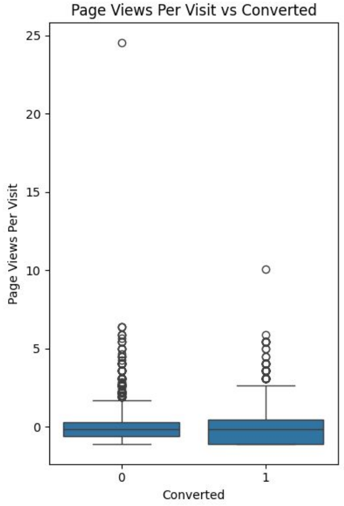
Final cleaned dataset

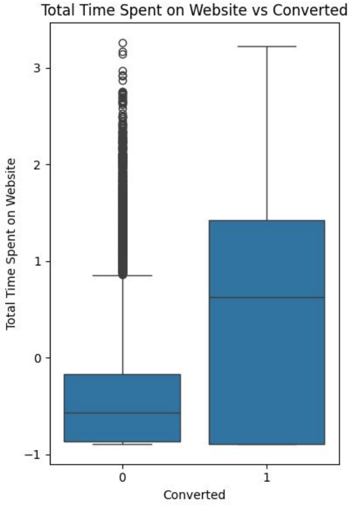
- **Rows:** 9,240
- **Features:** 9,384 (after encoding).

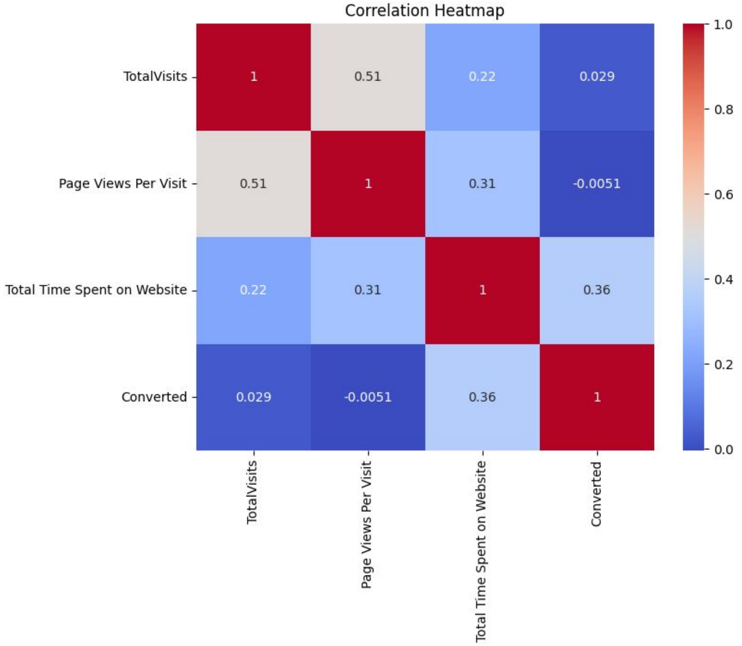
EXPLORATORY DATA ANALYSIS

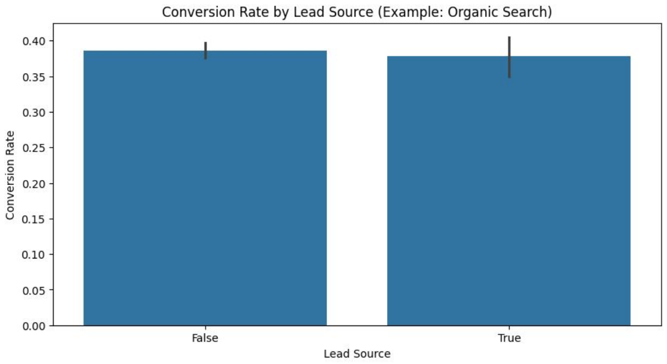












The EDA revealed key insights into lead conversion patterns:

- The overall conversion rate is approximately **36%,** with a moderate imbalance between converted and non-converted leads
- Among the numerical features, Total Time Spent on Website shows a strong positive association with conversion, indicating it is a critical predictor. In contrast, TotalVisits and Page Views Per Visit display weak relationships and potential outliers
- Analysis of categorical features (e.g., Lead Source) suggests that some categories may drive higher conversion rates, requiring deeper investigation
- The correlation heatmap highlights that Total Time Spent on Website is the strongest numerical driver of conversion, while multicollinearity between TotalVisits and Page Views Per Visit should be addressed

MODEL DEVELOPMENT

Chosen model

Logistic Regression

Simple, interpretable, and effective for binary classification.

Performance metrics

- **Accuracy:** 80.5%
- **Precision:** 72.6%
- **Recall:** 79.4%
- **AUC-ROC:** 87% (excellent model discrimination)

Model Coefficient and Intercept

- **Intercept:** -0.19
- Since the intercept is negative, it indicates that **without any feature influence, the base probability of conversion is low**
- This aligns with business reality, as leads typically need engagement before conversion

Odds = e^(-0.19) = 0.83 (approx)

Probability = 0.83 / (1+0.83) = 45.4% (approx)

This means that, on average, a lead without key influencing features has a **45.4% probability** of conversion.

Top Coefficients

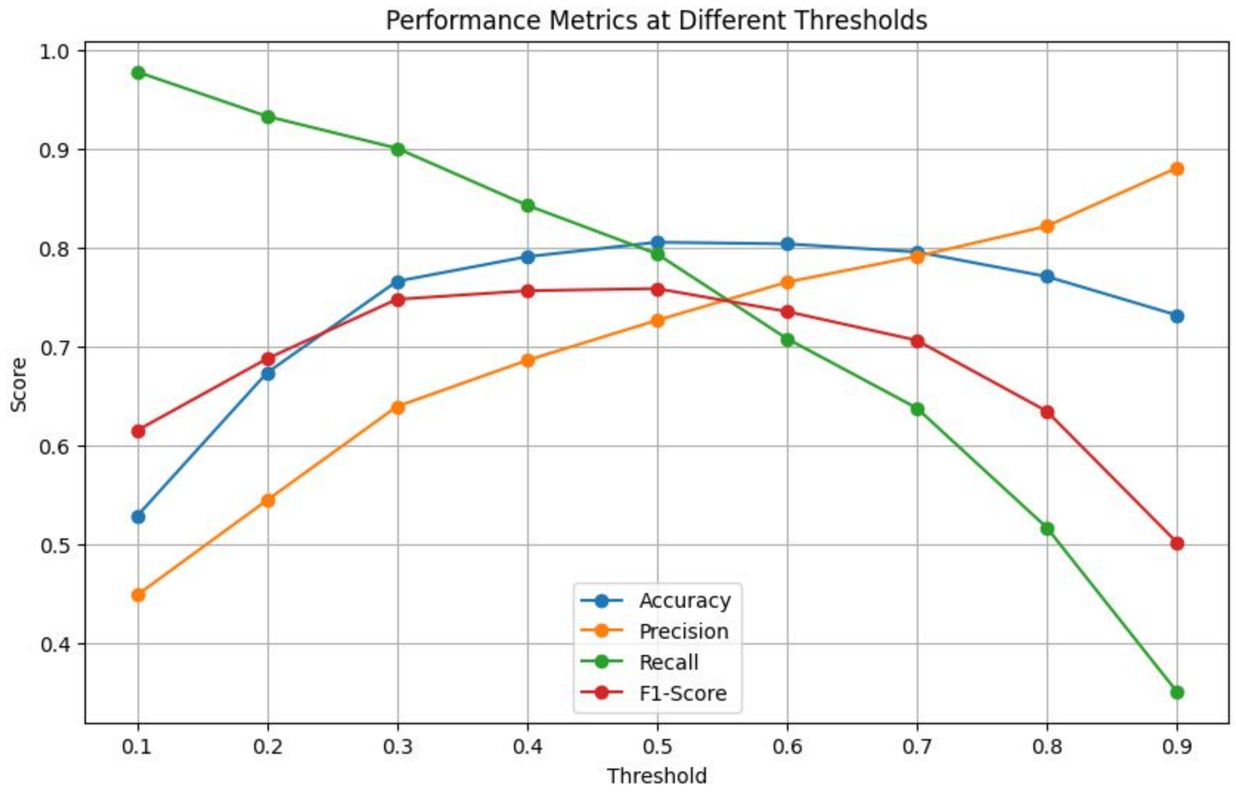
Feature	Coefficient	Interpretation
Lead Origin_Lead Add Form	2.14	Strongly increases conversion probability
Total Time Spent on Website	1.13	The longer a user spends on the website, the more likely they are to convert
Do Not Email_Yes	-1.32	Leads marked as "Do Not Email" are significantly less likely to convert
Specialization_Select	-1.31	Leads who did not specify specialization are less likely to convert

THRESHOLD OPTIMIZATION

Objective

Adjust the classification threshold to balance precision and recall

Threshold	Accuracy	Precision	Recall	F1-Score
0.1	0.528139	0.448454	0.977528	0.614841
0.2	0.67316	0.544262	0.932584	0.687371
0.3	0.765693	0.639083	0.900281	0.747522
0.4	0.790584	0.685714	0.842697	0.756144
0.5	0.805195	0.726221	0.793539	0.758389
0.6	0.803571	0.764795	0.707865	0.73523
0.7	0.795455	0.790941	0.63764	0.706065
0.8	0.770563	0.821429	0.516854	0.634483
0.9	0.731602	0.880282	0.351124	0.502008



Default threshold of 0.5 was chosen because:

- Offers a strong balance between precision and recall

- Allows the sales team to prioritize promising leads while minimizing wasted effort

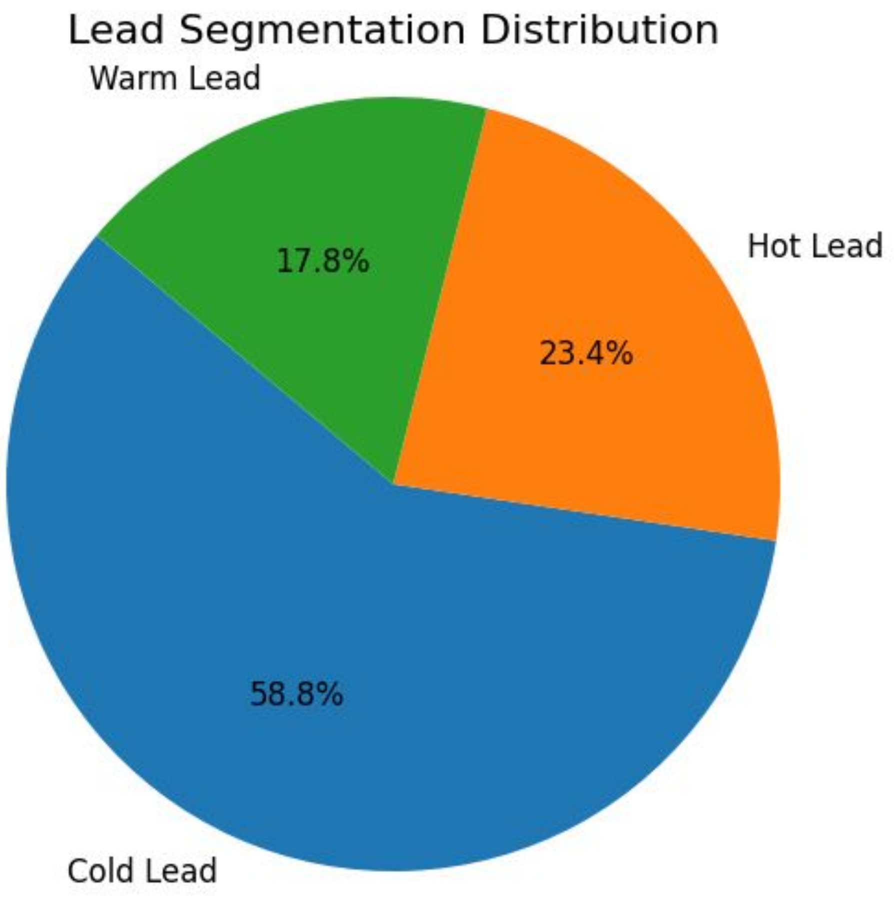
- Already aligns well with the business's target conversion rate (~79.4% recall)

LEAD SCORING

Each lead is assigned a score (0–100) based on its likelihood to convert

Lead segmentation

- **Hot Leads (Score ≥ 80):** 2,165 leads – immediate focus
- **Warm Leads (50 ≤ Score < 80):** 1,644 leads – nurture with targeted content
- **Cold Leads (Score < 50):** 5,431 leads – low priority



CONCLUSIONS

Key Achievements

- Developed a robust model with:
  - AUC-ROC: 87% (strong discriminatory power).
  - Balanced precision and recall to meet business needs.
- Improved lead prioritization and sales efficiency.
- Expected to significantly boost conversion rates and optimize resource allocation.

Business recommendations

Prioritize Leads

1. **Hot Leads:**
   - Immediate manual follow-up by the sales team
   - Allocate more resources for personalized outreach
2. **Warm Leads:**
   - Engage with nurturing campaigns (e.g., emails, webinars)
   - Move them to the "Hot" category over time
3. **Cold Leads:**
   - Use automated re-engagement campaigns
   - Minimize manual effort to save resources

Next steps

- **CRM Integration:**
  - Export lead scores and categories to streamline sales processes
- **Continuous Improvement:**
  - Monitor lead conversion trends and update the model periodically
- **Advanced Models (Optional):**
  - Explore more advanced models for improved accuracy and insights

THANK YOU