Lead Scoring Project – X Education

BY- SHOBHIT GUPTA

Lead Scoring Project – X Education

- Business Context:
- X Education generates thousands of leads monthly, but most don’t convert.

- Challenge:
- Sales team bandwidth is limited and needs better prioritization.

- Objective:
- Use data-driven lead scoring to identify high-converting leads and optimize outreach.

Data Cleaning & Preparation

- - Dataset: ~9,240 leads with 37 features
- - Removed redundant columns and constants
- - 'Select' values treated as missing
- - Imputed missing countries with 'India'
- - Encoded categories using dummies and scaled numeric features

Data Cleaning & Preparation

Athe variable `City` won't be of any use in our analysis. So it's best that we drop it.

```
leads.drop(['City'], axis = 1, inplace = True)
```

[187] ✓ 0.0s

```
# Same goes for the variable 'Country'

leads.drop(['Country'], axis = 1, inplace = True)
```

[188] ✓ 0.0s

```
# Drop all the columns in which greater than 3000 missing va

for col in leads.columns:
    if leads[col].isnull().sum() > 3000:
        leads.drop(col, axis = 1, inplace=True)
```

[185] ✓ 0.0s

```
# Get the value counts of all the columns

for column in leads:
    print(leads[column].astype('category').value_counts())
    print('_____')
```

[203] ✓ 0.1s

```
Prospect ID
000104b9-23e4-4ddc-8caa-8629fe8ad7f4    1
a7a319ea-b6ae-4c6b-afc5-183b933d10b5    1
aa27a0af-eeab-4007-a770-fa8a93fa53c8    1
aa30ebb2-8476-41ce-9258-37cc025110d3    1
aa405742-17ac-4c65-b19e-ab91c241cc53    1
                                       ..
539eb309-df36-4a89-ac58-6d3651393910    1
539ffa32-1be7-4fe1-b04c-faf1bab763cf    1
53aabd84-5dcc-4299-bbe3-62f3764b07b1    1
53ac14bd-2bb2-4315-a21c-94562d1b6b2d    1
fffb0e5e-9f92-4017-9f42-781a69da4154    1
Name: count, Length: 9240, dtype: int64
_____
Lead Number
```

Data Cleaning & Preparation

```
leads.drop(['Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper',
            'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses',
            'Update me on Supply Chain Content', 'Get updates on DM Content',
            'I agree to pay the amount through cheque'], axis = 1, inplace = True)
```

[208] ✓ 0.0s

Column What matters most to you in choosing a course has one dominant value (Better Career Prospects: 6528 times) drop it.

```
leads['What matters most to you in choosing a course'].value_counts()
```

```
leads['How did you hear about X Education'].value_counts()
```

✓ 0.0s

```
How did you hear about X Education
Select                   5043
Online Search             808
Word Of Mouth             348
Student of SomeSchool     310
Other                     186
Multiple Sources          152
Advertisements             70
Social Media               67
Email                      26
SMS                        23
Name: count, dtype: int64
```

EDA - Conversion Overview

- - About 38.5% of leads converted
- - Converted leads spent ~12.3 minutes vs 5.5 mins by non-converted
- - Time spent is a strong indicator of interest

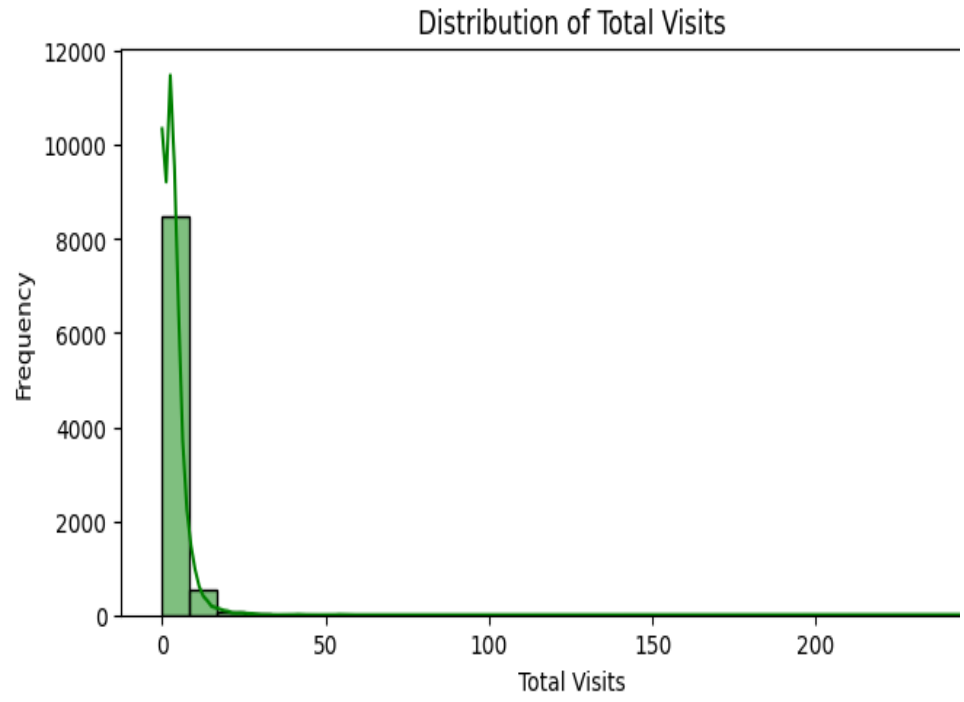
EDA - Lead Sources & Origin

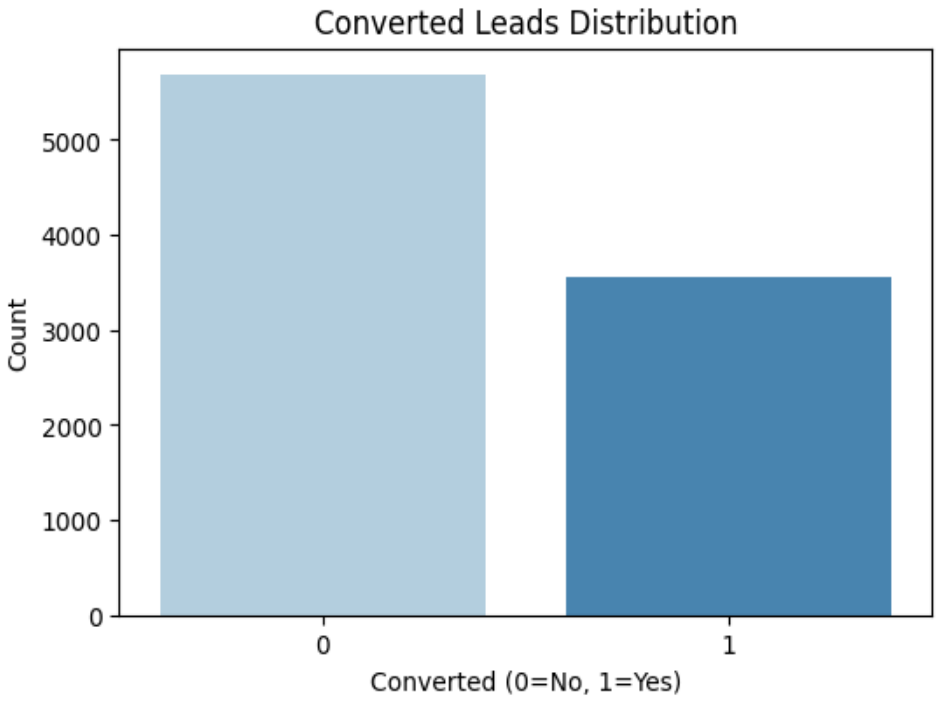
- - Referrals converted at 91–92%
- - Organic/Google leads at 38–40%
- - Chat/ads at ~25%, lower conversion
- - ‘Lead Add Form’ leads had highest conversion at ~92%

EDA - Lead Engagement & Activity

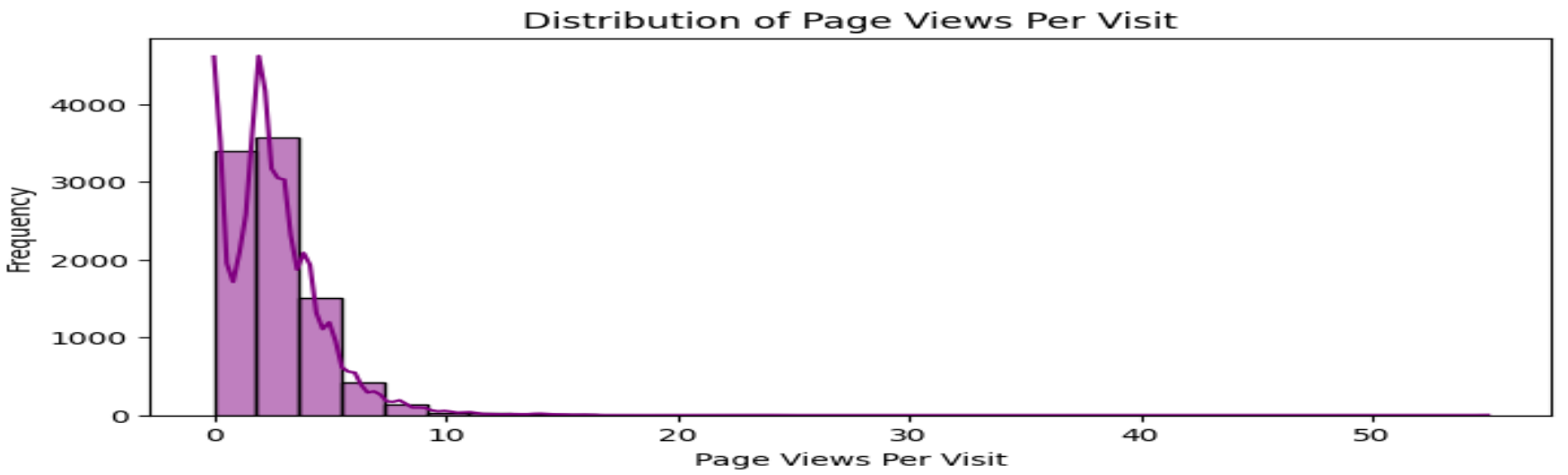
- - SMS/Call last activities → High conversion (63–73%)
- - Email opened ~36%, Olark chat ~8.6%
- - Opt-outs converted at only 16%
- - Personal engagement matters most

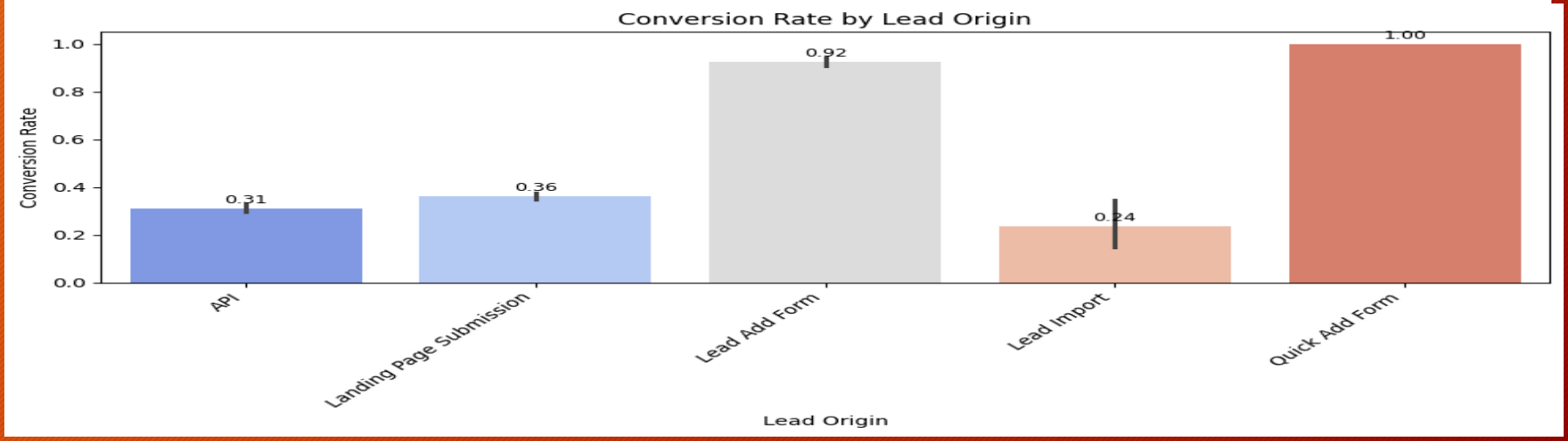
EDA -


Distribution of Total Visits
Frequency
Total Visits
0
2000
4000
6000
8000
10000
12000
0
50
100
150
200

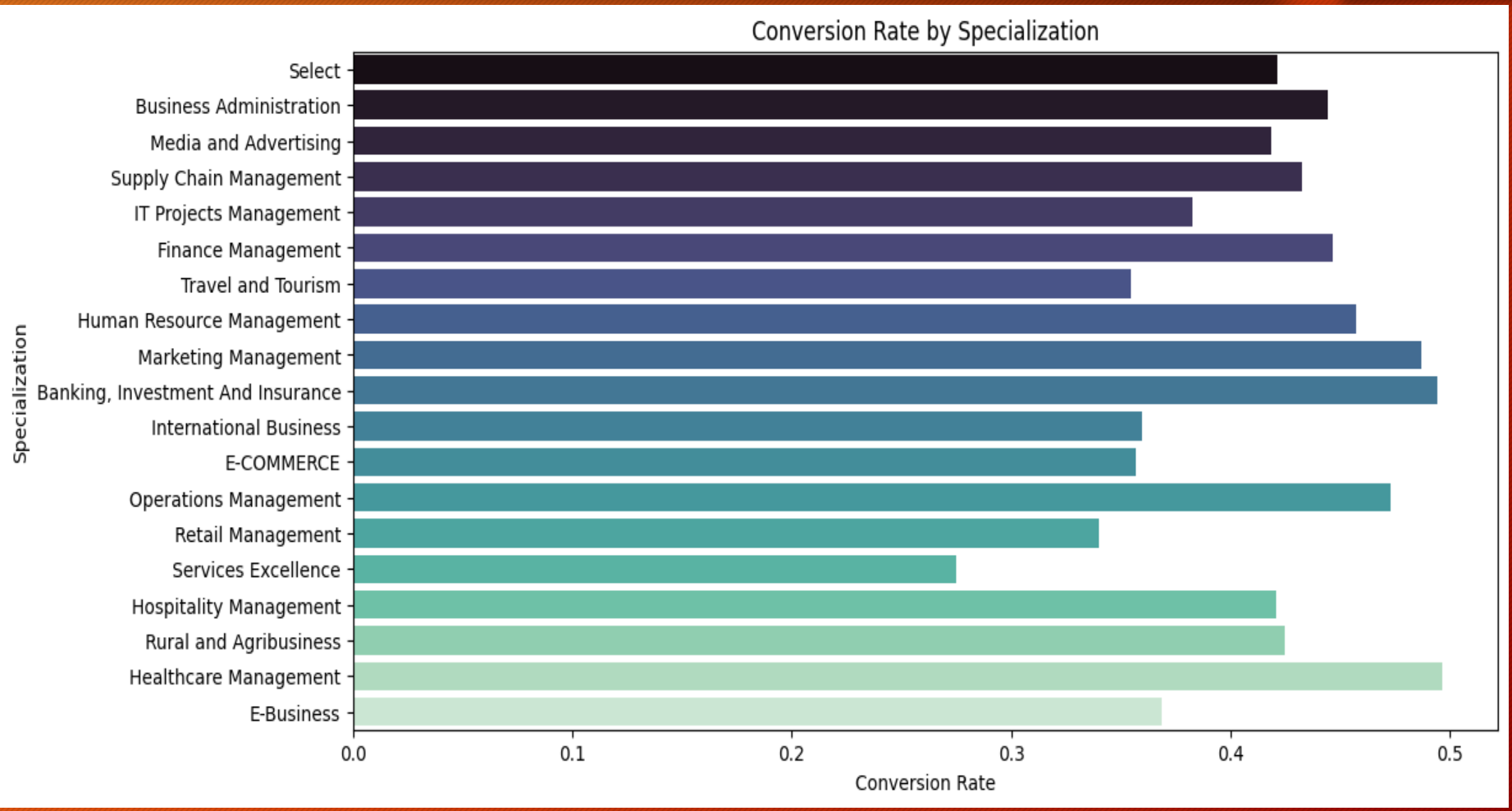

Converted Leads Distribution
Count
Converted (0=No, 1=Yes)
0
1000
2000
3000
4000
5000
0
1

EDA -

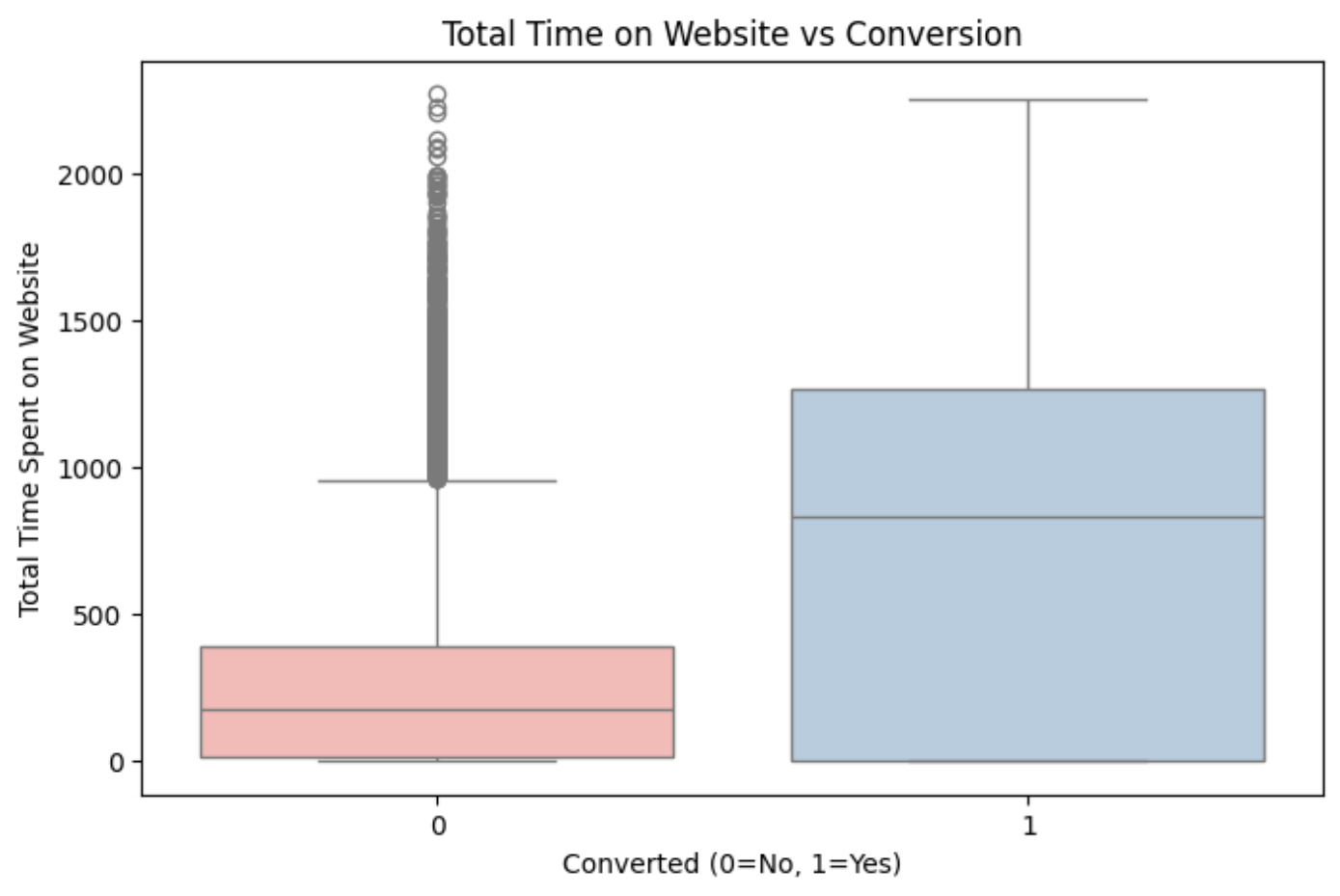

Distribution of Page Views Per Visit
Frequency
4000
3000
2000
1000
0
0
10
20
30
40
50
Page Views Per Visit


Conversion Rate by Lead Origin
Conversion Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.31
0.36
0.92
0.24
1.00
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin

EDA -


Conversion Rate by Specialization
Specialization
Select
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
0.0
0.1
0.2
0.3
0.4
0.5
Conversion Rate

EDA-


Total Time on Website vs Conversion
Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted (0=No, 1=Yes)

Model Building Approach

- - Logistic Regression with RFE + VIF
- - Final features: Time on Website, Total Visits, etc.
- - Dummy encoding for categoricals
- - Scaled numerical features

Model Performance (Final)

- - Accuracy: 65.59%
- - Precision: 65.96%
- - Recall: 59.05%
- - ROC AUC: 86%
- - Optimal cutoff: 0.44

Strategic Recommendations

- - Score every incoming lead
- - Prioritize leads above cutoff
- - Use email/SMS for low scorers (nurture flow)
- - Invest more in referral/organic channels
- - Monitor and retrain model regularly

- Business Recommendations:

- - 📌 Focus on leads spending more than 500 seconds on the website, as they show strong conversion signals.
- - 🚫 Avoid spending too much effort on leads with >30 visits and low engagement time.
- - 🎯 Adjust probability threshold dynamically:
- 0.30–0.40 → Aggressive outreach (intern support, growth months)
- 0.60–0.70 → Conservative mode (after targets are met)

Conclusion

- - X Education can now target hot leads effectively
- - Improved conversion rates with less manual effort
- - The lead scoring model enables smart allocation of sales effort.
- By targeting high-conversion leads, X Education can improve conversion
- rates efficiently and dynamically adjust strategies based on resources.