LEAD SCORE CASE STUDY

GROUP MEMBERS SEPT 30TH BATCH:

1. KAILA SAI PRANAVA KARTHIK
2. SHREYAS G
3. SOURAV KUMAR JHA

PROBLEM STATEMENT AND OBJECTIVE

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

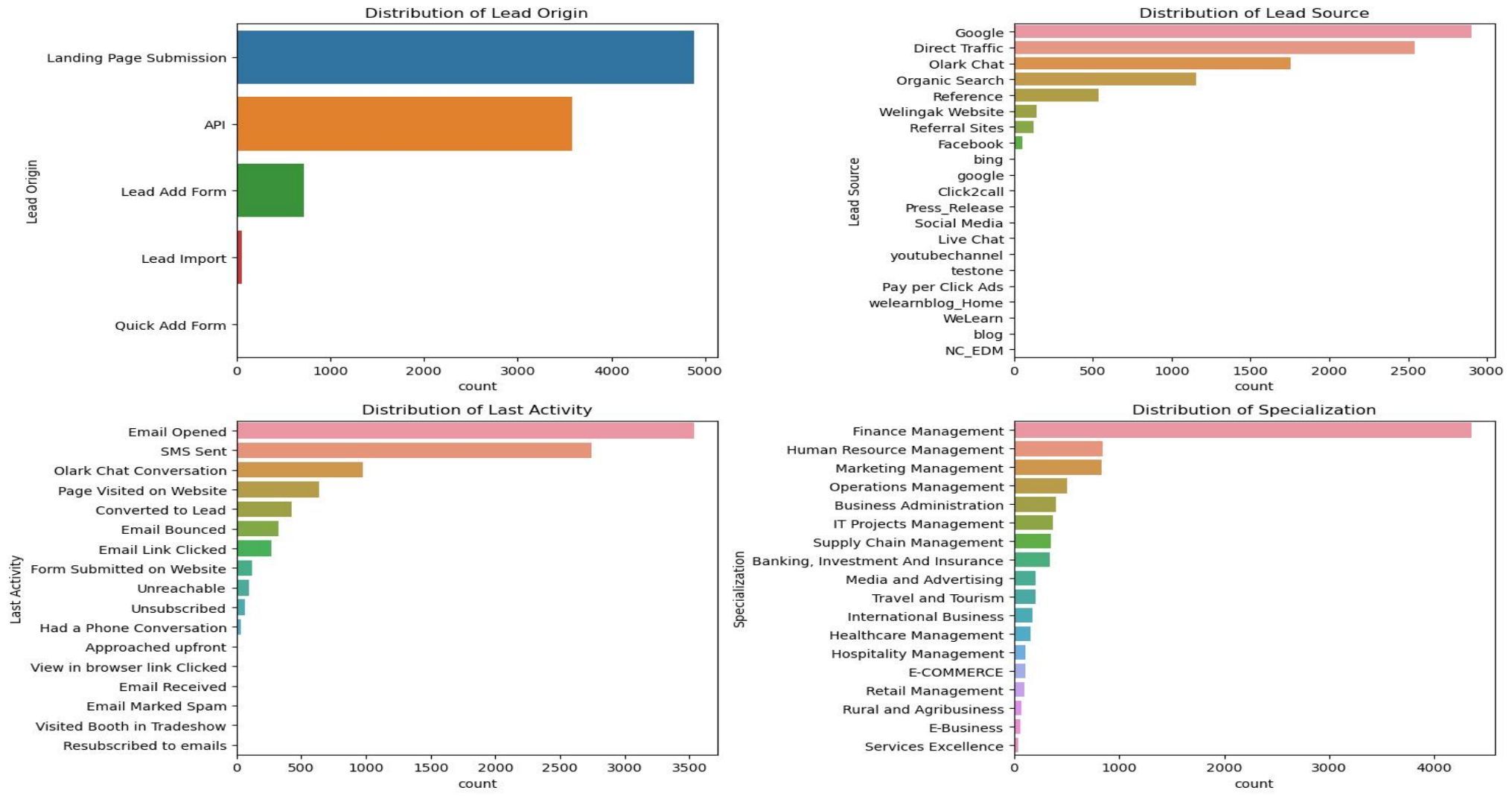
APPROACH

We started by loading and understanding the dataset, which contained 9,240 leads with 37 features. The dataset involved a combination of numerical and categorical fields like Lead Source, Last Activity, and Total Time Spent on Website.

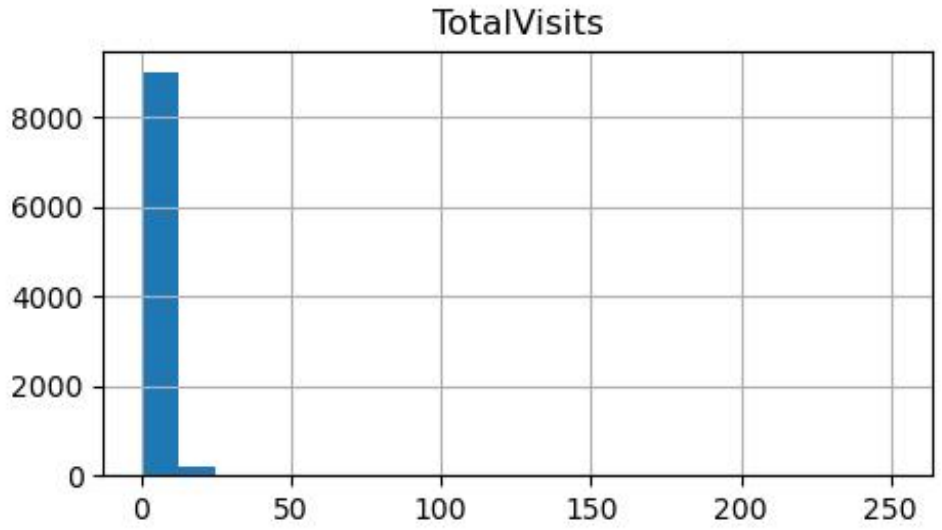
Key steps in preprocessing were:

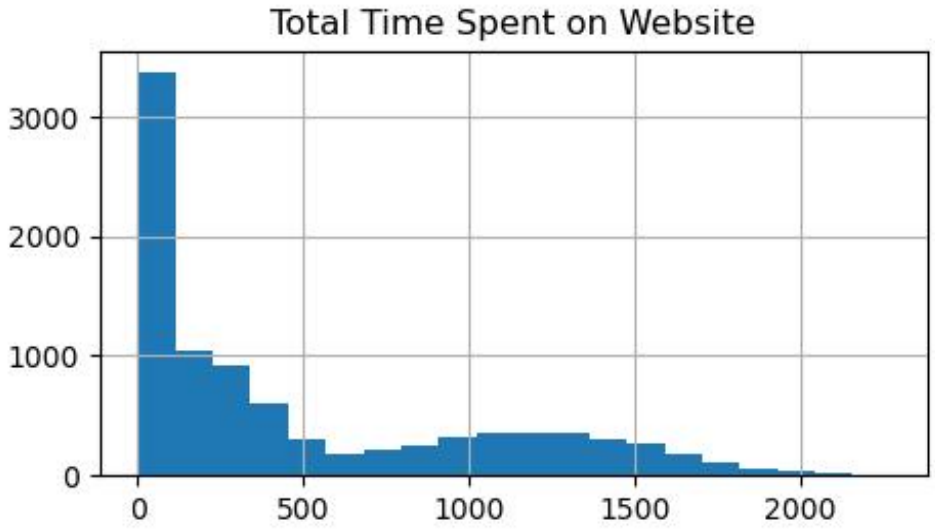
- **Handling Missing Values**: Columns with excessive missing values were dropped, while others were imputed using the mode.
- **Encoding Categorical Variables**: Converted categorical features into numerical representations.
- **Feature Scaling**: Standardized numerical variables to bring them onto the same scale.
- **Splitting Data**: Divided the dataset into training (70%) and testing (30%) sets.

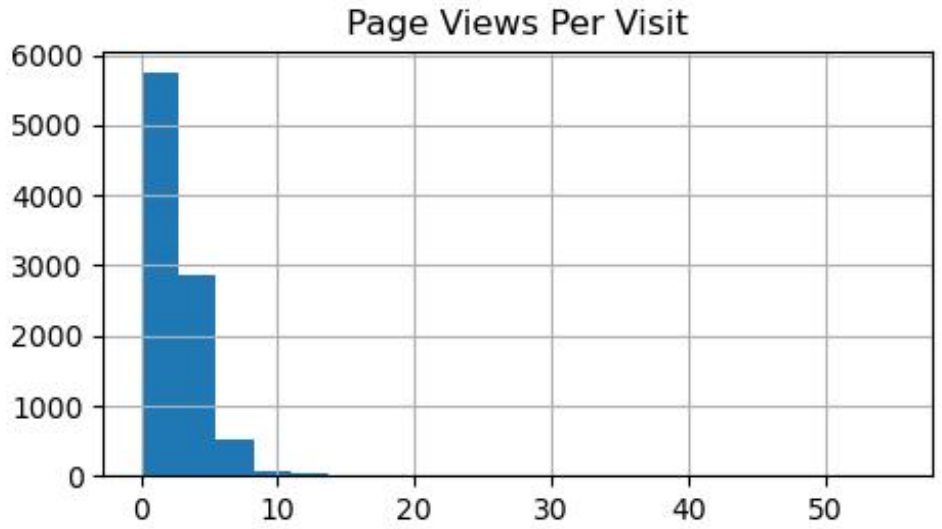
EDA


Distribution of Lead Origin
Lead Origin
Landing Page Submission
API
Lead Add Form
Lead Import
Quick Add Form
0
1000
2000
3000
4000
5000
count
Distribution of Lead Source
Lead Source
Google
Direct Traffic
Olark Chat
Organic Search
Reference
Welingak Website
Referral Sites
Facebook
bing
google
Click2call
Press_Release
Social Media
Live Chat
youtubechannel
testone
Pay per Click Ads
welearnblog_Home
WeLearn
blog
NC_EDM
0
500
1000
1500
2000
2500
3000
count
Distribution of Last Activity
Last Activity
Email Opened
SMS Sent
Olark Chat Conversation
Page Visited on Website
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unreachable
Unsubscribed
Had a Phone Conversation
Approached upfront
View in browser link Clicked
Email Received
Email Marked Spam
Visited Booth in Tradeshow
Resubscribed to emails
0
500
1000
1500
2000
2500
3000
3500
count
Distribution of Specialization
Specialization
Finance Management
Human Resource Management
Marketing Management
Operations Management
Business Administration
IT Projects Management
Supply Chain Management
Banking, Investment And Insurance
Media and Advertising
Travel and Tourism
International Business
Healthcare Management
Hospitality Management
E-COMMERCE
Retail Management
Rural and Agribusiness
E-Business
Services Excellence
0
1000
2000
3000
4000
count

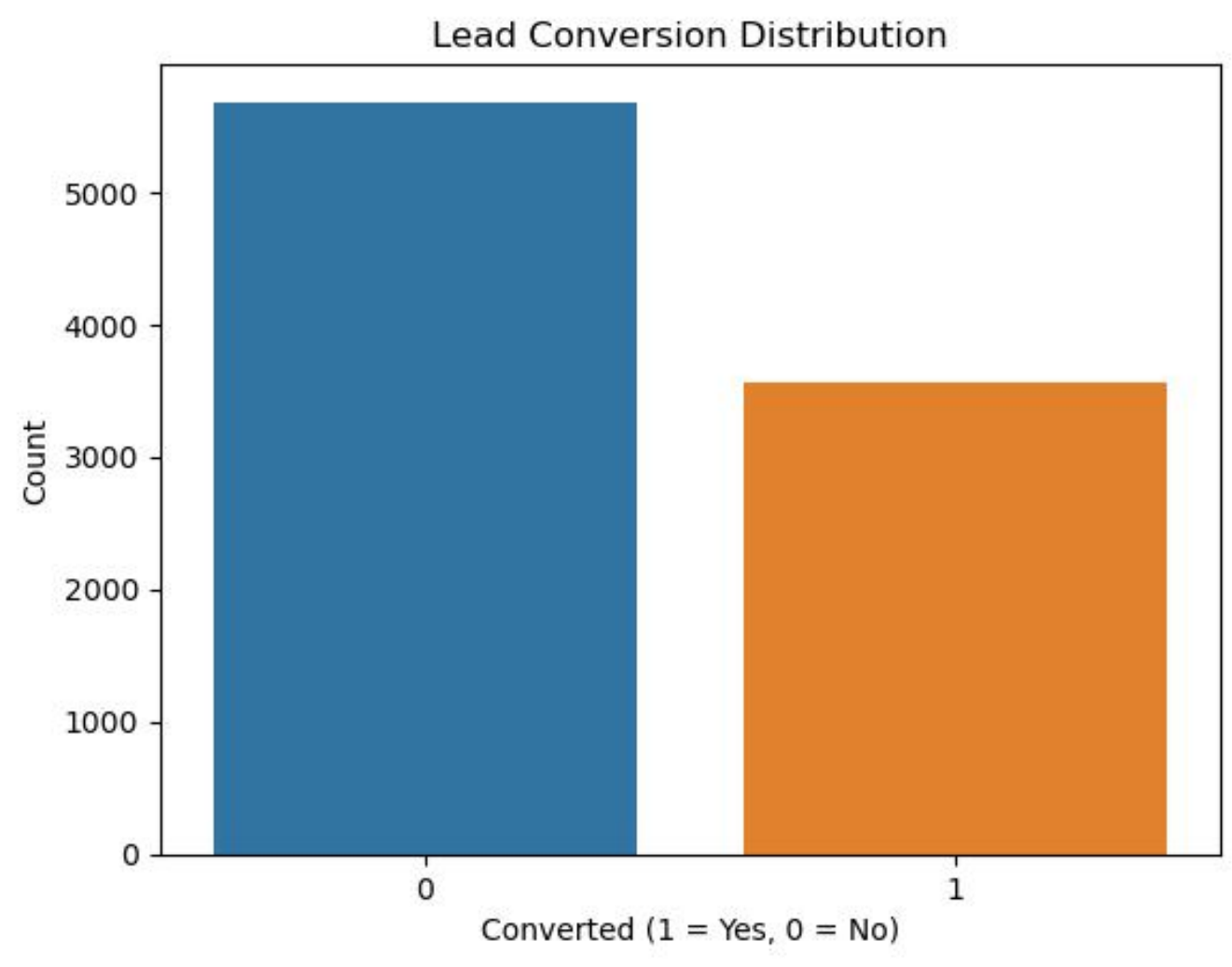
NUMERICAL VARIABLE RELATION







LEAD CONVERSION DISTRIBUTION



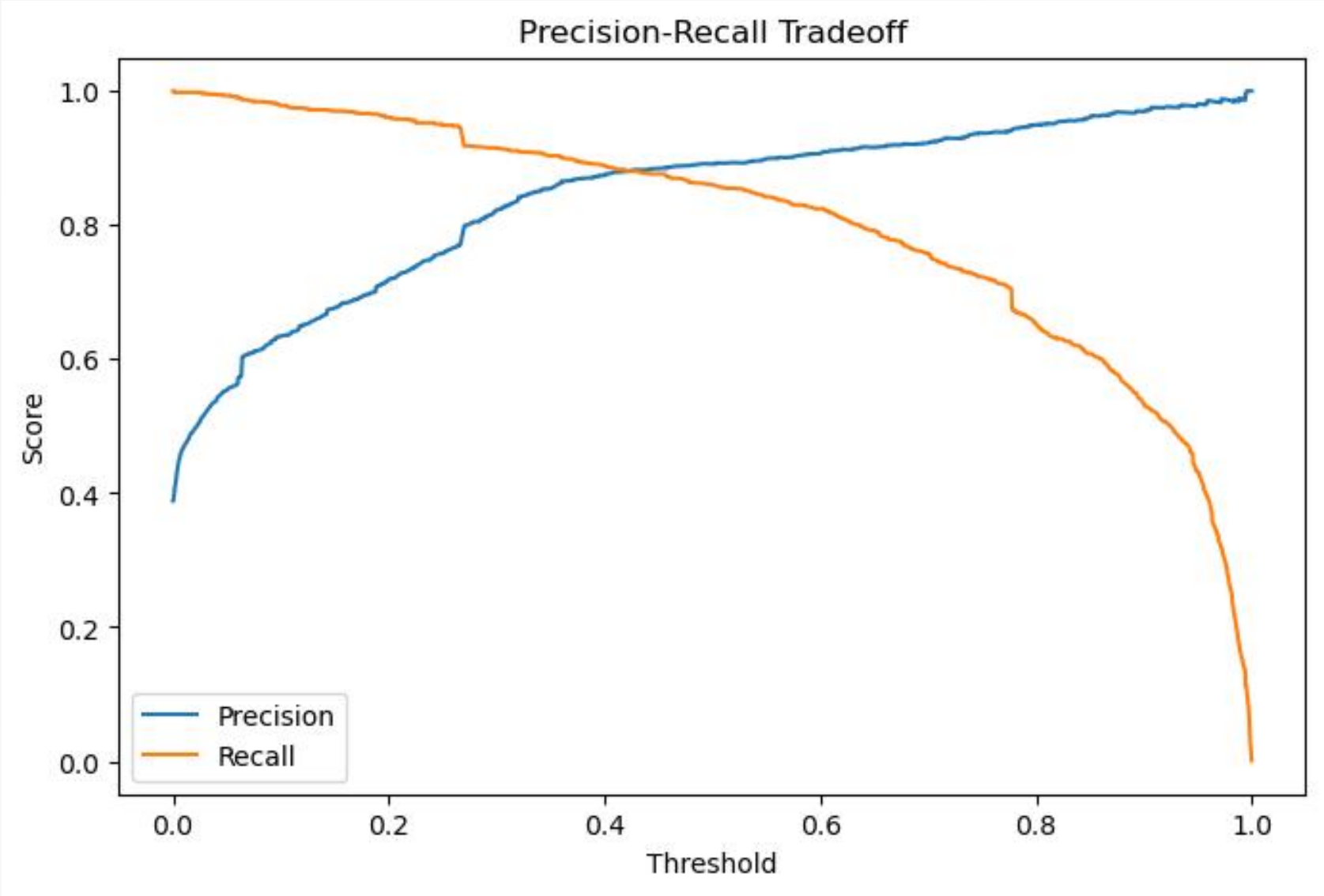
MODEL BUILDING

We selected Logistic Regression, which is a widely used classification model for binary outcomes, because it is easy to interpret and it is also quite efficient. The model was trained to predict whether a lead would convert into a paying customer (1) or not (0). Decision Trees model was considered but it was overfitted.

Key insights from model performance:

- **Accuracy**: ~91% on test data
- **ROC-AUC Score**: 0.96 (which means it's a strong predictive model)
- **Optimal Threshold**: 0.42 (this was so that we can balance precision and recall)

PRECISION-RECALL TRADEOFF



KEY FINDINGS

The top three numerical features contributing to lead conversion are the following:

- **Total Time Spent on Website** – More time spent means higher the interest.
- **Lead Origin** – Leads from "API" and "Landing Page Submission" converted more.
- **Last Activity** – Engagement through emails/SMS increased chances of conversion.

The top categorical features that should be prioritized for higher conversion rates are:

- **Lead Source** – Organic Search, Direct Traffic, and Google performed well.
- **Lead Quality** – "High" and "Might Be" had better conversion chances.
- **Specialization** – Some fields, like Finance and Banking, had higher conversion.

STRATEGIC RECOMMENDATIONS

For Aggressive Lead Conversion During Intern Hiring Period

- Lowering the Lead Score Threshold (like from 50 to 30) would increase the pool of leads.
- Prioritizing High-Engagement Leads (those who interacted with emails, watched videos, or revisited the site).
- Applying a Multi-Touch Approach – Combining emails, SMS, and calls to improve the response rates.

For Reducing Unnecessary Calls After Achieving the Targets

- Increasing the Lead Score Threshold (like focusing only on leads scoring 80+).
- Prioritizing Referral Leads, as they have a higher chances of conversion.
- Automating Follow-ups via emails and SMS before assigning a sales call.

CONCLUSION

- **Data Quality Matters:** Missing values and inconsistent categories require careful preprocessing.
- **Feature Importance Insights:** User behavior (time spent, last activity) they have a significant impact on lead conversion.
- **Threshold Selection is Crucial:** Adjusting the probability threshold will help in balancing the business needs (more calls vs fewer, high-quality leads).
- **Model Simplicity vs Interpretability:** Logistic regression worked well due to its clarity and strong results, making it preferable over other complex models.
- By implementing this model, X Education can improve their efficiency by focusing on high-scoring leads, potentially increasing conversions by **10-20%**.

This project gave useful insights into lead prioritization, helping businesses increase their efficiency and focus efforts on high-converting leads.

THANK YOU