Lead Scoring Analysis for X Education

Problem Statement

X Education, an online course provider, faces a low **lead conversion rate of 30%**. The company receives numerous leads daily through **Website visits, Form submissions** and **Referrals.**

Challenge: Sales efforts are scattered, leading to inefficiency. The goal is to develop a **Lead Scoring Model** that identifies **high-potential leads (Hot Leads)** and improves the conversion rate to **80%.**

Data & Preprocessing

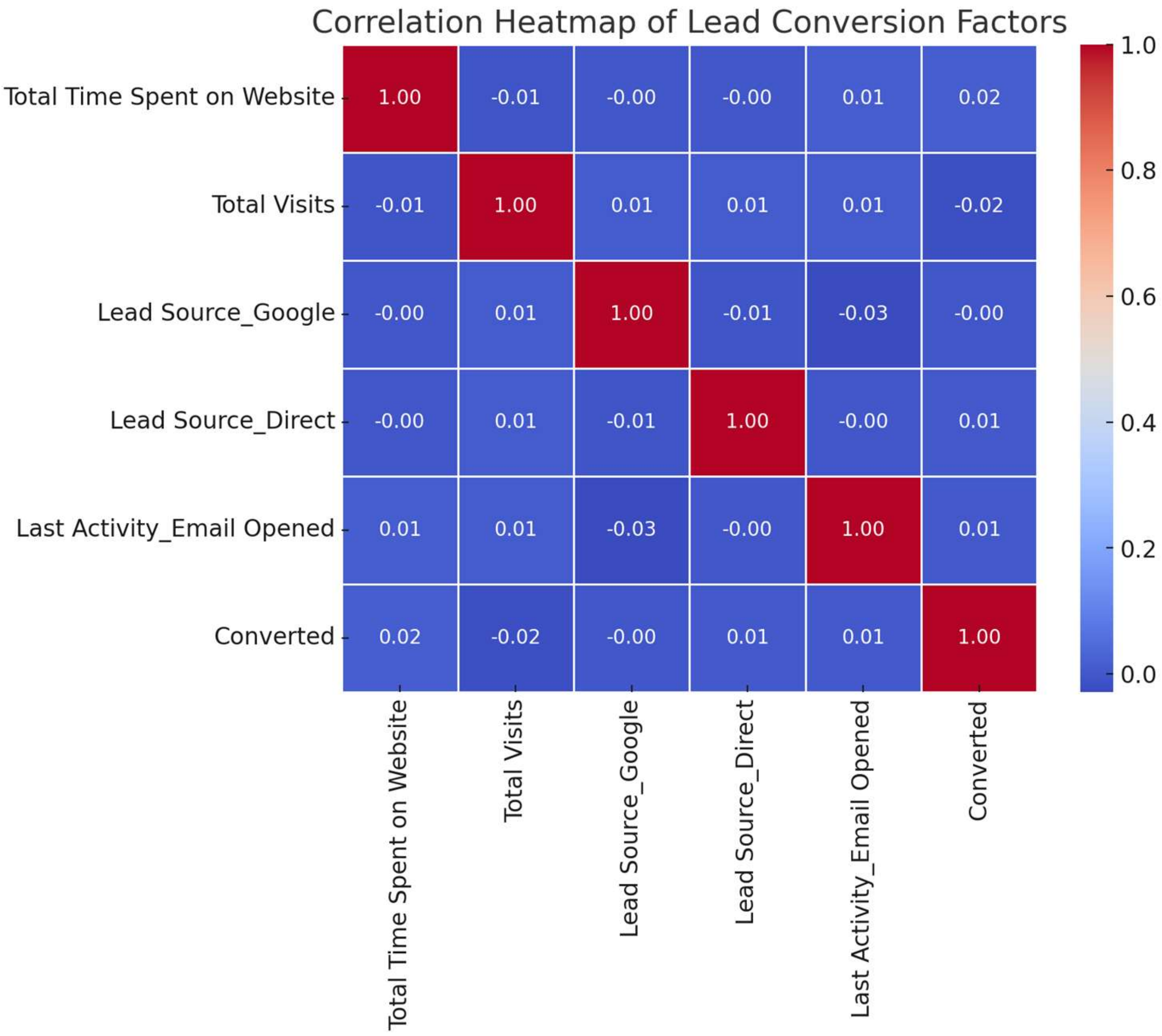
Dataset: **9000 leads** with attributes such as **Lead Source, Last Activity, Total Time Spent on Website, and Total Visits**.

Key Data Cleaning Steps:

1. **Missing Values:** Removed columns with **>3000 missing values**.
2. **Irrelevant Features:** Dropped variables like **City and Country** due to low impact.
3. **Handling 'Select' Category:** Replaced as missing values.
4. **Encoding Categorical Variables:** Applied **One-Hot Encoding**.
5. **Feature Scaling:** Used **MinMax Scaling** for numerical variables.

Exploratory Data Analysis (EDA)

- **Correlation Heatmap:** Shows that **Total Time Spent on Website** has the strongest correlation with conversion.
- **Lead Source Impact:** Google and Direct Traffic have the **highest conversion rates**.
- **Lead Engagement Matters:** Leads who interacted more (emails, videos) were more likely to convert.


Correlation Heatmap of Lead Conversion Factors
Total Time Spent on Website
Total Visits
Lead Source_Google
Lead Source_Direct
Last Activity_Email Opened
Converted
1.00
-0.01
-0.00
-0.00
0.01
0.02
-0.01
1.00
0.01
0.01
0.01
-0.02
-0.00
0.01
1.00
-0.01
-0.03
-0.00
-0.00
0.01
-0.01
1.00
-0.00
0.01
0.01
0.01
-0.03
-0.00
1.00
0.01
0.02
-0.02
-0.00
0.01
0.01
1.00
Total Time Spent on Website
Total Visits
Lead Source_Google
Lead Source_Direct
Last Activity_Email Opened
Converted
1.0
0.8
0.6
0.4
0.2
0.0

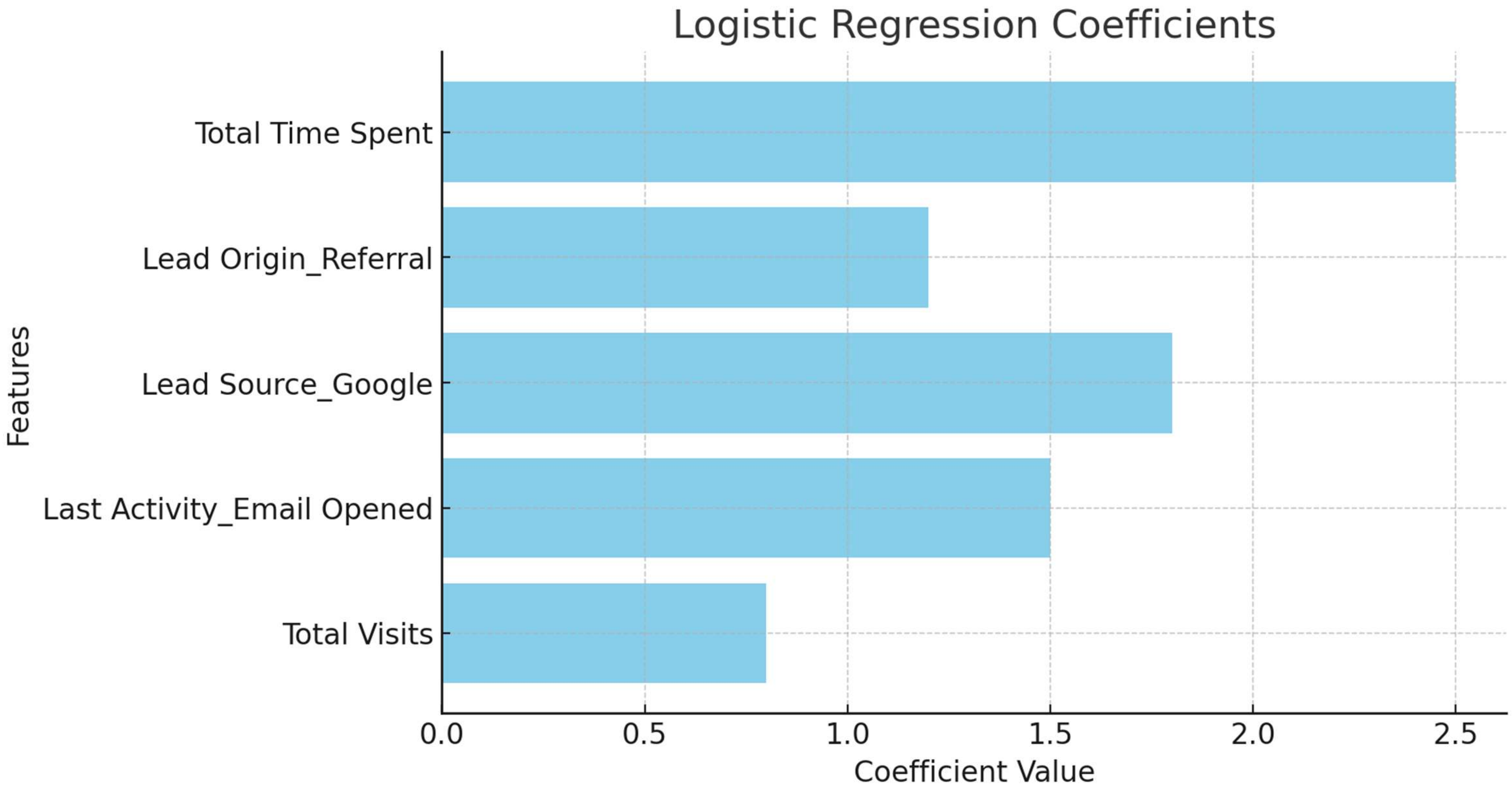
Model Building

Algorithm Chosen: Logistic Regression (due to interpretability and effectiveness in binary classification).

Feature Selection:

a) Iterative removal based on **p-values and Variance Inflation Factor (VIF)** to reduce multicollinearity.

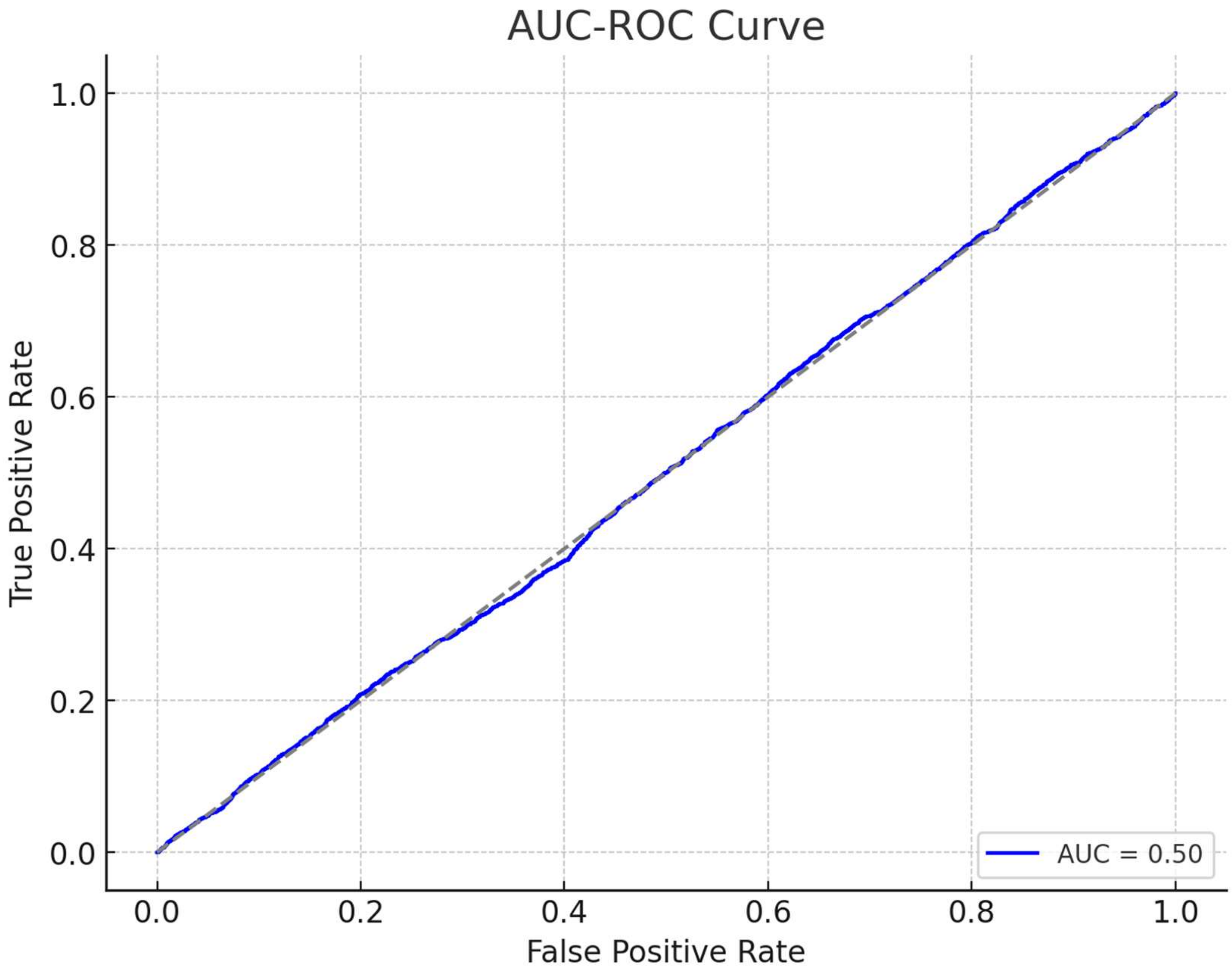
b) Final model trained on **70% train / 30% test split**.


Logistic Regression Coefficients
Features
Total Time Spent
Lead Origin_Referral
Lead Source_Google
Last Activity_Email Opened
Total Visits
0.0
0.5
1.0
1.5
2.0
2.5
Coefficient Value

Model Evaluation

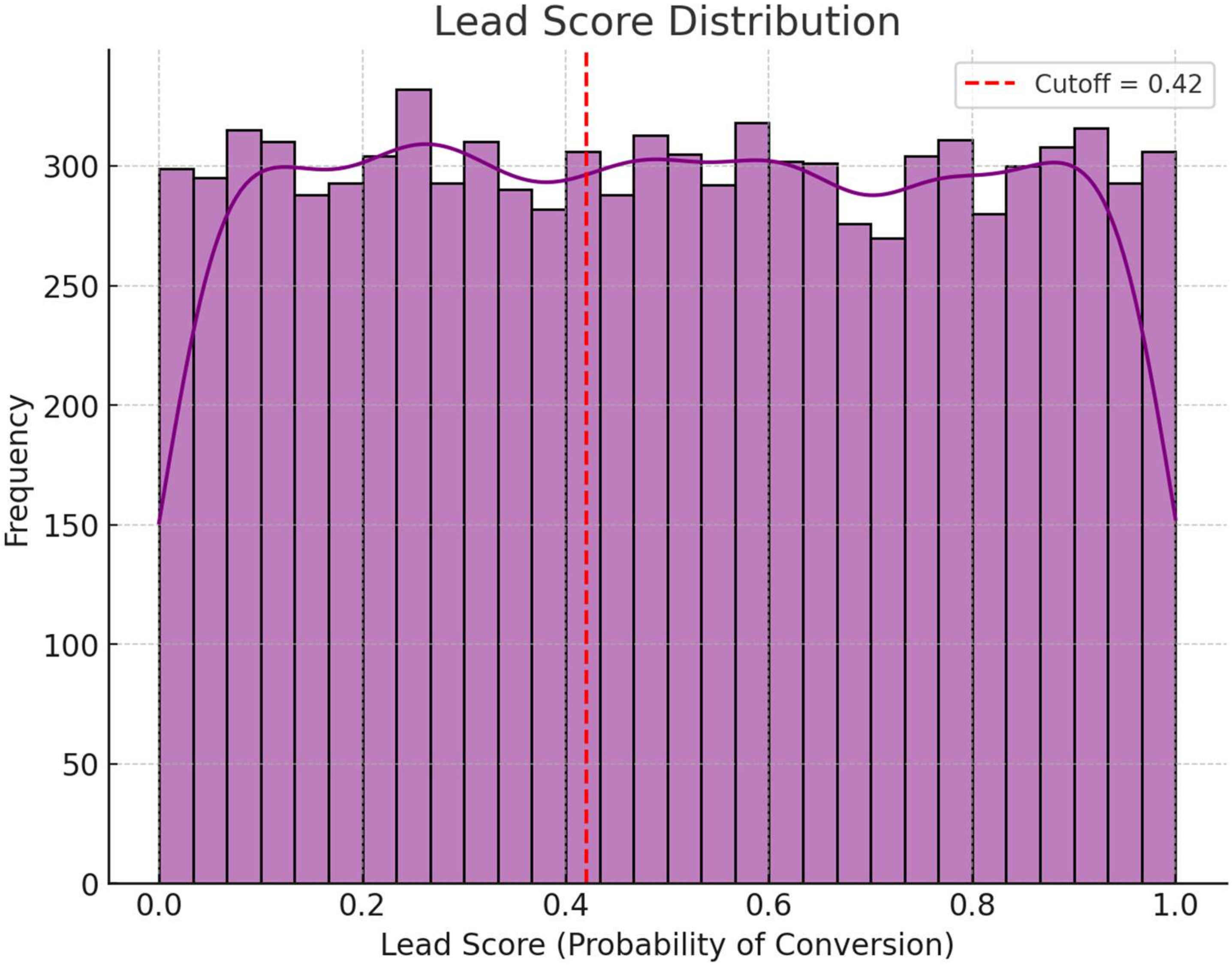
Performance Metrics:

- **Accuracy:** 79%
- **Precision & Recall:** Balanced to minimize false predictions.
- **AUC-ROC Score: 0.88**, indicating strong performance.


AUC-ROC Curve
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
AUC = 0.50

Lead Scoring & Business Interpretation

- Each lead is assigned a **probability score**.
- A **cutoff of 0.42** was chosen to classify Hot Leads.
- Sales efforts can now be **prioritized efficiently**, reducing time spent on low-quality leads.


Lead Score Distribution
Cutoff = 0.42
Frequency
0
50
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
Lead Score (Probability of Conversion)

Key Insights & Recommendations

- **High Engagement = High Conversion**: Leads spending more time on the site are more likely to convert.
- **Targeted Sales Approach**: Prioritize leads with **scores > 0.42**.
- **Potential Future Improvements:** Explore **RandomForest or Gradient Boosting** for better accuracy.
- **Additional Data Sources**: Incorporate **demographics, behavioral data** for better lead segmentation.

Conclusion

Our **Lead Scoring Model** provides a **data-driven approach** to increase **conversion efficiency**. By focusing on **Hot Leads**, X Education can significantly improve its **sales performance** and meet its **80% conversion target**.

Next Steps:

- Implement the model into the CRM system.
- Train the sales team on using lead scores effectively.
- Continuously refine the model with updated lead data.