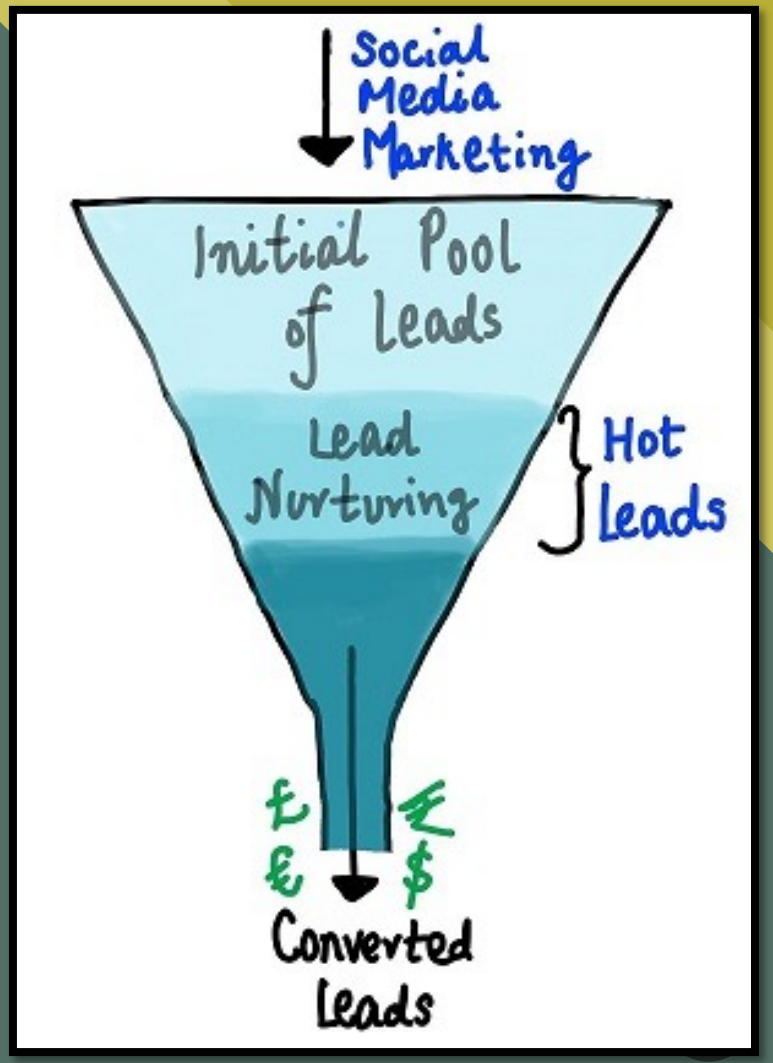
LEAD SCORING PROJECT

- Kamal Thampi

PROBLEM STATEMENT

- X Education, an online education company, generates many leads from its website and marketing efforts but has a low lead conversion rate (typically 30%).
- The company wants to identify "Hot Leads" that are most likely to convert into paying customers, so the sales team can focus on these high-potential leads.
- A typical lead conversion process is represented as a funnel, with many leads entering at the top and few converting to paying customers at the bottom.
- To improve the lead conversion rate, X Education needs you to build a model that assigns a lead score to each lead, so that those with higher scores are more likely to convert.
- The target lead conversion rate is set by the CEO to be around 80%.
- The goal is to select the most promising leads and increase the lead conversion rate by focusing on these high-scoring leads.



BUSINESS OBJECTIVE

- **Improve Lead Conversion Rate**
  Develop a logistic regression model to assign a lead score (0-100) that accurately identifies "Hot Leads" likely to convert into paying customers.
- **Target High-Potential Leads**
  Use the lead score model to prioritize and target leads with higher scores, increasing the chances of conversion.
- **Enhance Sales Efficiency**
  By focusing on high-scoring leads, the sales team can optimize their efforts, reducing unnecessary outreach and increasing conversions.
- **Future-Proof Model**
  Ensure that the logistic regression model is flexible enough to adapt to changing company requirements and external factors.

METHODOLOGY

Data Cleaning

- Check for duplicate data
- Check and handle NA and missing values
- Handle missing values either by dropping columns or imputation
- Check and handle outliers (if any)

Exploratory Data Analysis (EDA)

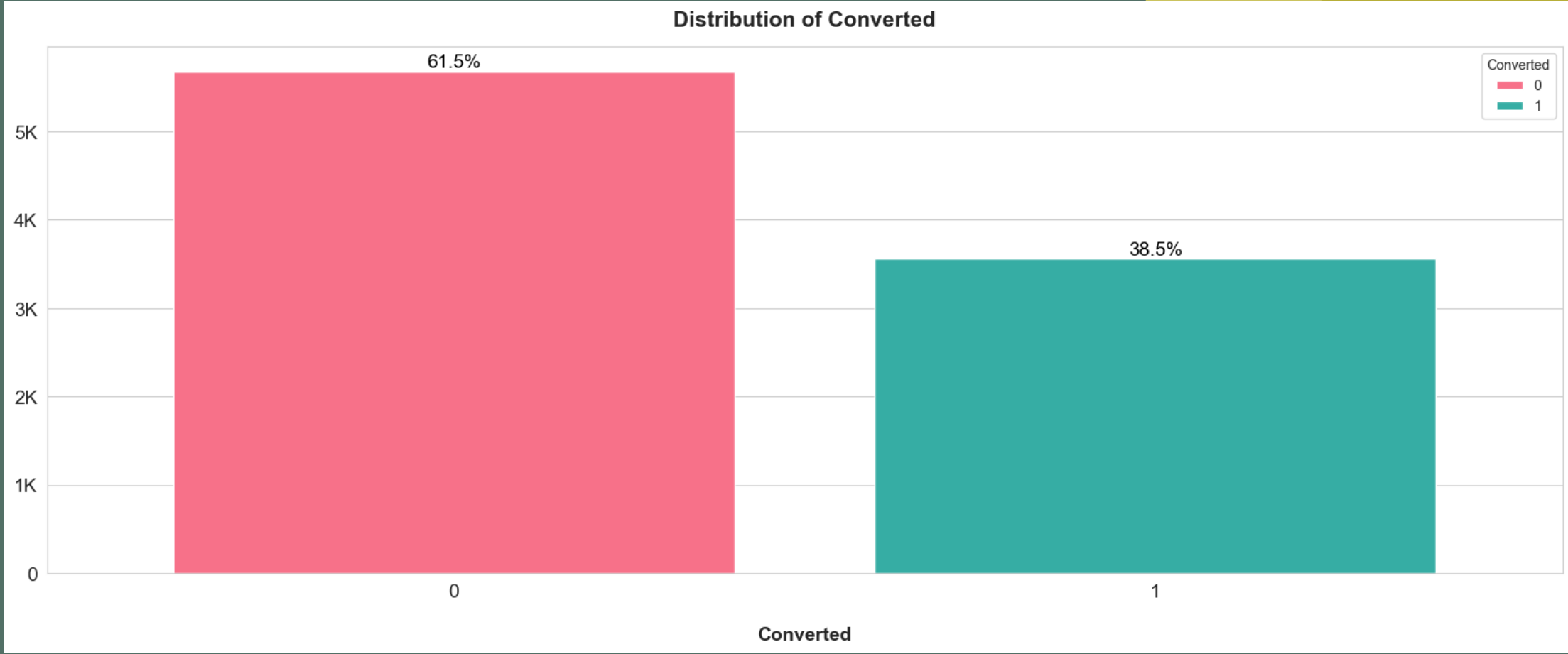
- Check for imbalances in a dataset, value counts, distribution of variables, etc.
- Handle imbalances in the target variable using SMOTE
- Feature scaling using StandardScaler
- Dummy variable encoding of the data
- Create a vanilla logistic regression model and another one with Hyperparameters
- Model Validation with Test data
- Get probability scores
- Change threshold to improve recall

DATA CLEANING

- Leads dataset with 37 Rows and 9,240 columns
- Dropped ‘Prospect ID’ features due it non-relevance
- Dropped singleton features or columns like 'Magazine', '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’
- Handle imbalanced features by either grouping them into ‘Others’ or other relevant categories.
- Handle missing values by either dropping or imputation. For e.g., impute the missing country as ‘India’ when the city is Mumbai.

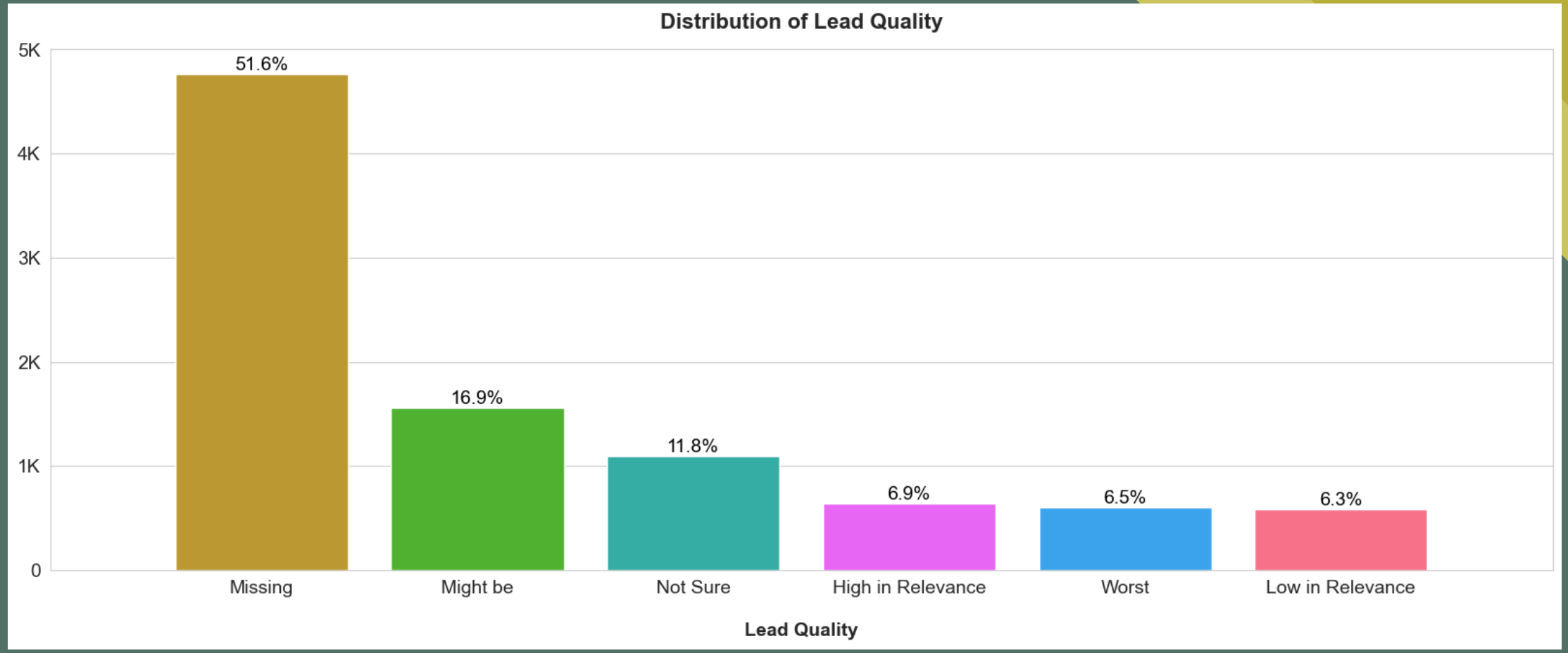
EXPLORATORY DATA ANALYSIS (EDA) - I

Checking for data imbalance in the Target variable



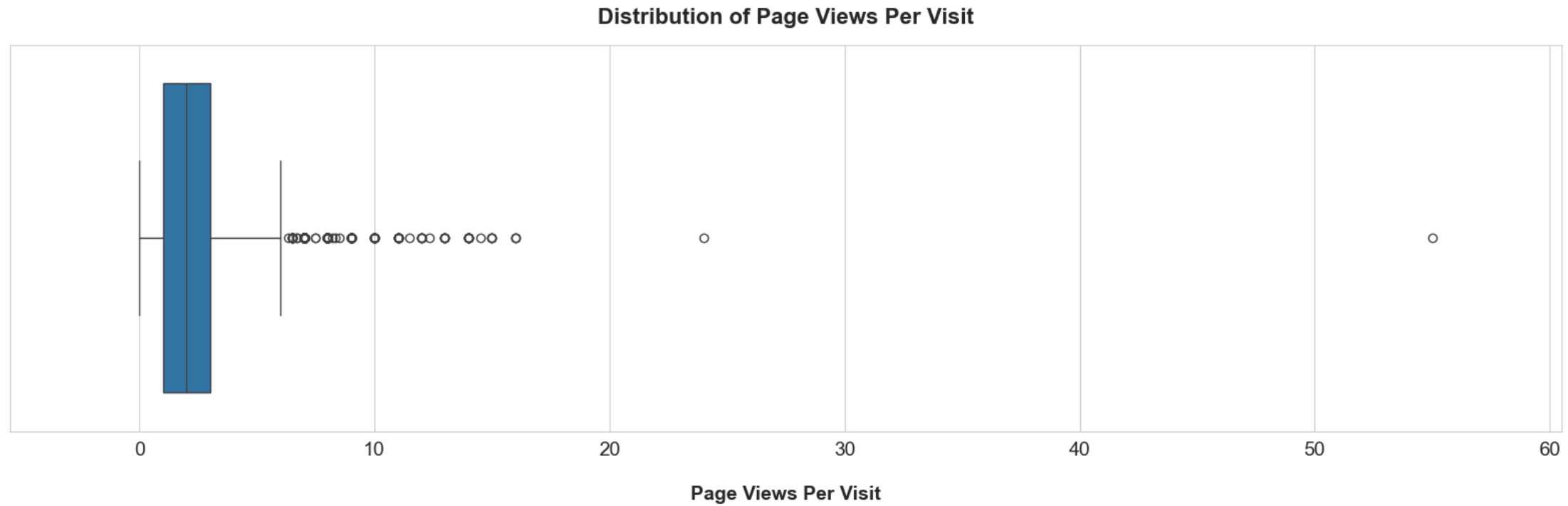
EXPLORATORY DATA ANALYSIS (EDA) - II

Dropping features with missing values



EXPLORATORY DATA ANALYSIS (EDA) - III

Checking for Outliers



DATA CONVERSION

- The dataset was split into test and train data with a 30:70 ratio
- Train data was oversampled using SMOTE to balance the Target variables
- Dummy variables were created for object-type variables
- Numerical variables were then normalized

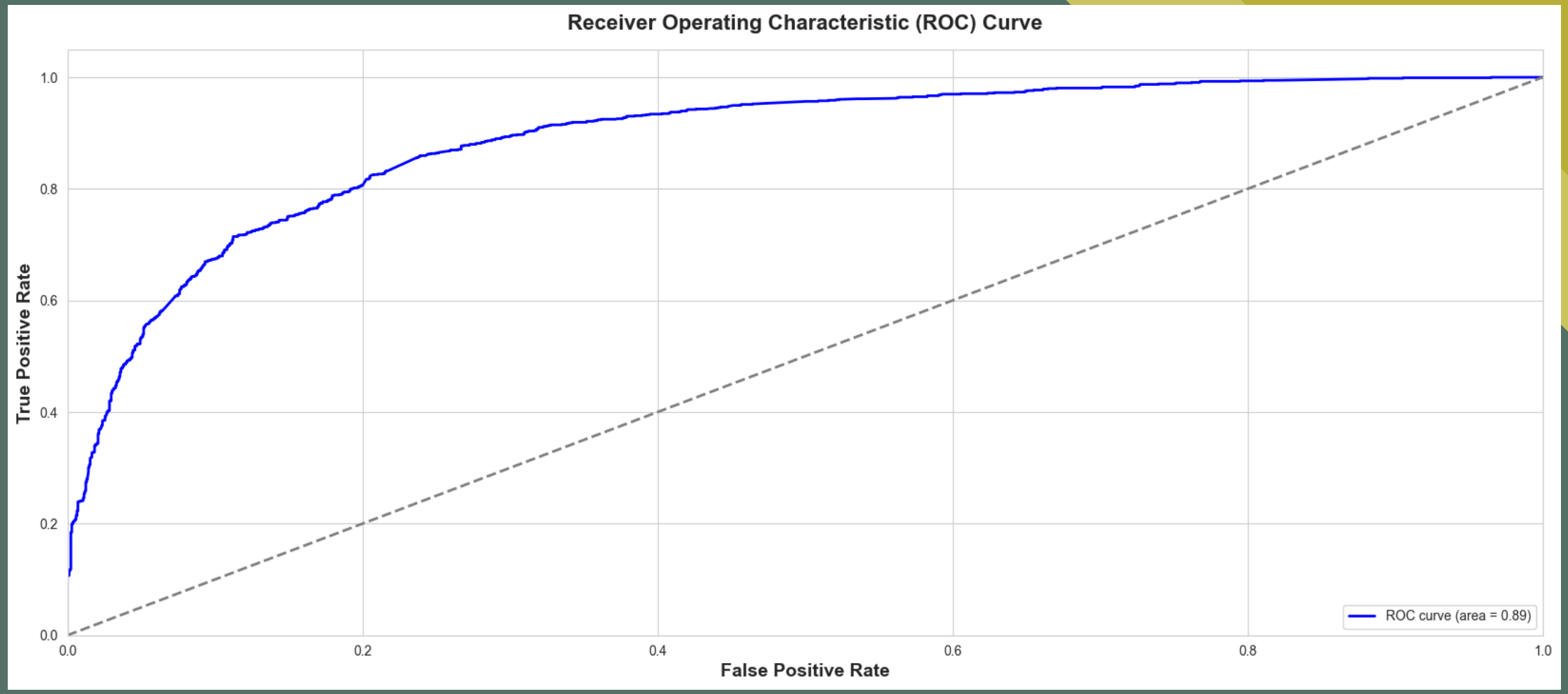
MODEL BUILDING

- Create a vanilla Logistic Regression Model
- Normalize the test dataset to validate the model with roc_auc score of 0.8882
- Created another model with Hyperparameters using GridSearchCV
- The roc_auc score for the new model against test data was 0.8887
- The second model stats are as shown below,

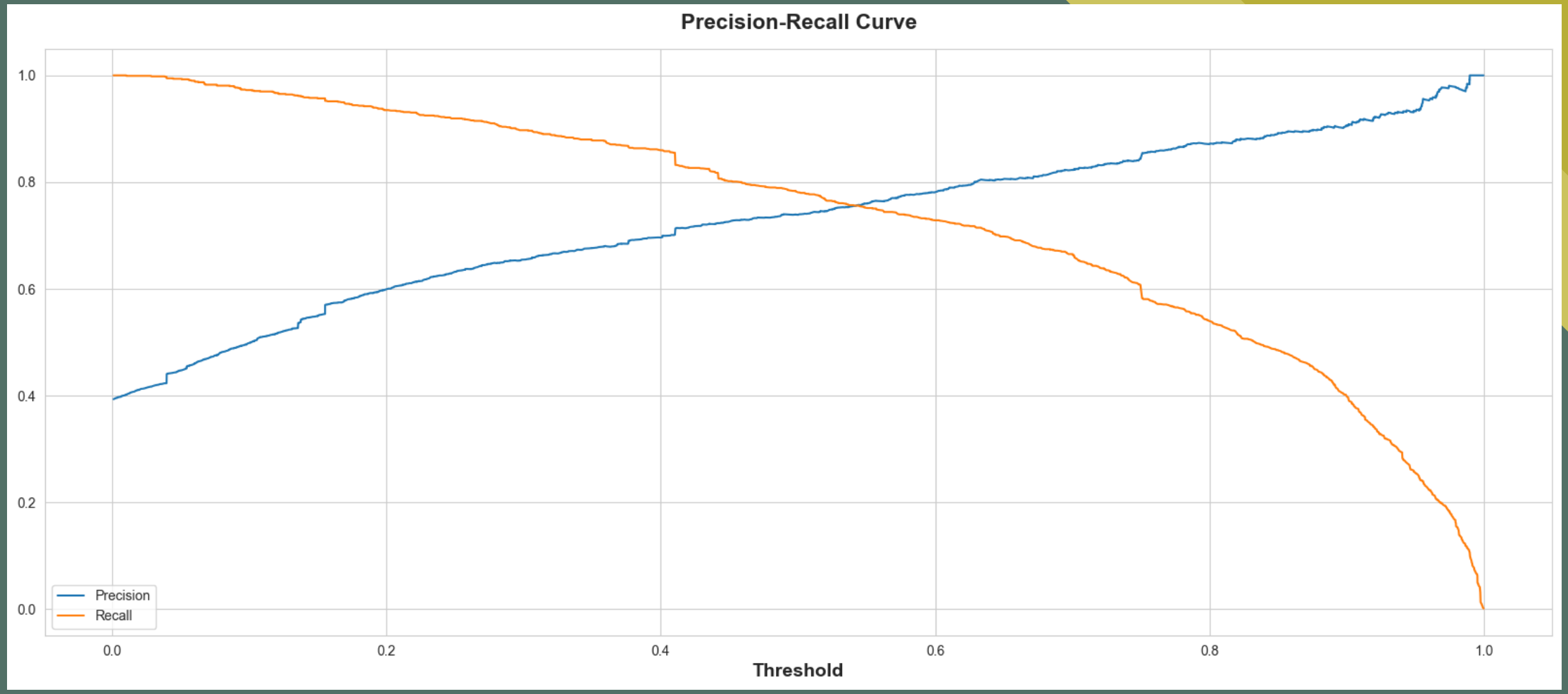
Accuracy	Precision	Recall	F1-Score	AUC-ROC
0.805195	0.738715	0.780734	0.759144	0.888733

- This shows that our test prediction has accuracy, precision, and recall scores in an acceptable range.
- Lead score is created on the test dataset to identify hot leads – the higher the lead score higher the chance of conversion, the lower the lead score lower the chance of getting converted.

ROC CURVE


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.89)

PRECISION-RECALL CURVE



KEY INSIGHTS - I

- **Total Time Spent on Website** is the most influential feature with a high positive coefficient of 1.1755, indicating that the more time a user spends on the website, the more likely they are to convert.
- **Occupation - Working Professional** has a significant positive impact (coefficient: 0.6201), suggesting that working professionals are more likely to convert compared to other occupations.
- **Last Activity - SMS Sent** and **Last Activity - Email Opened** are important indicators of lead conversion with coefficients of 0.5771 and 0.4956, respectively. Engaging leads through SMS and email is crucial.
- **Missing Occupation Information** (coefficient: 0.5637) and **Missing Specialization Information** (coefficient: 0.4798) surprisingly contribute positively to lead conversion, which might indicate that the absence of this data correlates with higher engagement or priority.
- **Lead Origin - Lead Add Form** (coefficient: 0.4875) and **Lead Origin - Landing Page Submission** (coefficient: 0.4152) are significant, highlighting the importance of where leads originate from in predicting their likelihood to convert.
- **Lead Source - Olark Chat** (coefficient: 0.4715) and **Lead Source - Welingak Website** (coefficient: 0.4709) are strong predictors, indicating that these sources provide high-quality leads.
- **Last Notable Activity - SMS Sent** (coefficient: 0.3976) and **Last Notable Activity - Others** (coefficient: 0.2480) are also significant, reinforcing the importance of recent interactions in lead conversion.

KEY INSIGHTS - II

- **Page Views Per Visit** (coefficient: 0.2179) and **Total Visits** (coefficient: 0.1336) both contribute positively, indicating that higher engagement on the website is a good predictor of conversion.
- **Specific Specializations** such as Retail Management (coefficient: 0.1131) and Hospitality Management (coefficient: 0.1114) also play a role, highlighting that certain specializations have a higher likelihood of converting.
- **Geographical Factors**: Leads from certain countries like Saudi Arabia (coefficient: 0.0970) and Bahrain (coefficient: 0.0956) are slightly more likely to convert compared to others.
- **Other Engagement Channels** such as Newspaper (coefficient: 0.0832) and Organic Search (coefficient: 0.0813) have smaller but positive impacts on lead conversion.
- **Least Impactful Features**: Some features such as Digital Advertisement (coefficient: 0.0079), and recommendations (coefficient: 0.0119), have very minimal impact on conversion rates.

RECOMMENDATIONS

- Focus on High-Engagement Channels: Enhance and optimize channels like SMS, email, and live chat (Olark Chat) which have shown a significant positive impact on lead conversion.
- Leverage Time Spent on Website: Develop strategies to increase the time users spend on the website, such as engaging content and interactive elements, as it is the strongest predictor of conversion.
- Occupation-Based Targeting: Target working professionals more aggressively since their likelihood to convert is higher.
- Form Origin Optimization: Prioritize leads generated from specific forms like Lead Add Form and Landing Page Submissions.
- Handle Missing Data Strategically: Consider that missing information in certain fields (Occupation and Specialization) might still lead to high conversion rates. This might suggest other underlying factors at play.
- Utilize Geographical Insights: Customize marketing strategies for leads from regions like Saudi Arabia and Bahrain to maximize conversion rates.
- By focusing on these insights, X Education can more effectively allocate resources and tailor their marketing and engagement strategies to improve lead conversion rates significantly.

THANK YOU