Lead Scoring Case Study by Tabassum Shaikh, Sushant Salunke, Huang-Yin Tso

Problem of Our Study

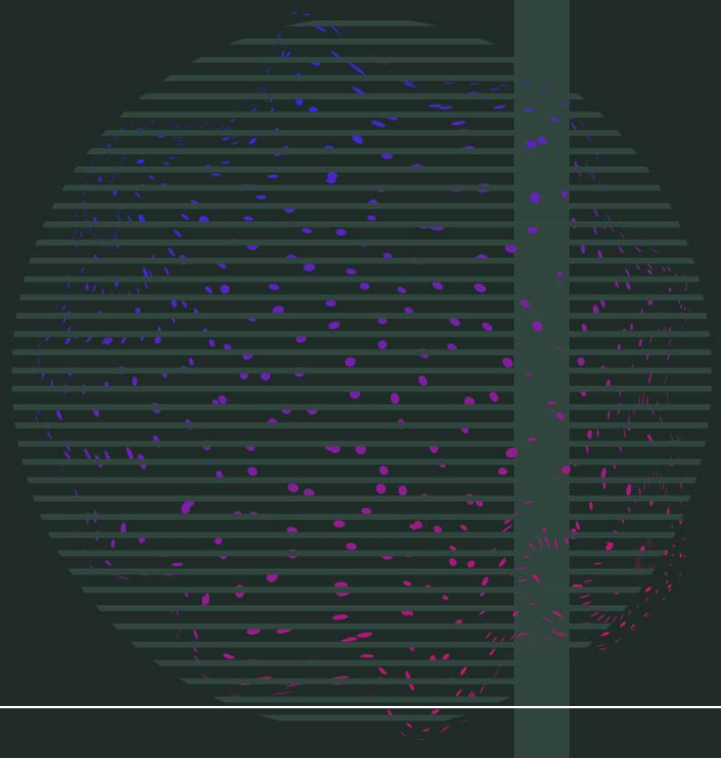
- X Education is an online course provider for industry professionals. The company gets leads from its website, online marketing, and referrals. Visitors on the website can browse courses, watch videos, or fill out a form with their contact details. When they provide contact information, they become leads. The sales team follows up with these leads through calls and emails, but only about 30% of them convert into paying customers.
- X Education has a low lead conversion rate. Out of 100 leads, only about 30 become paying customers. The sales team spends a lot of time contacting all leads, including those unlikely to convert. This makes the process inefficient. The company wants to find "Hot Leads"—leads that are more likely to convert—so the sales team can focus on these.

Business Objectives

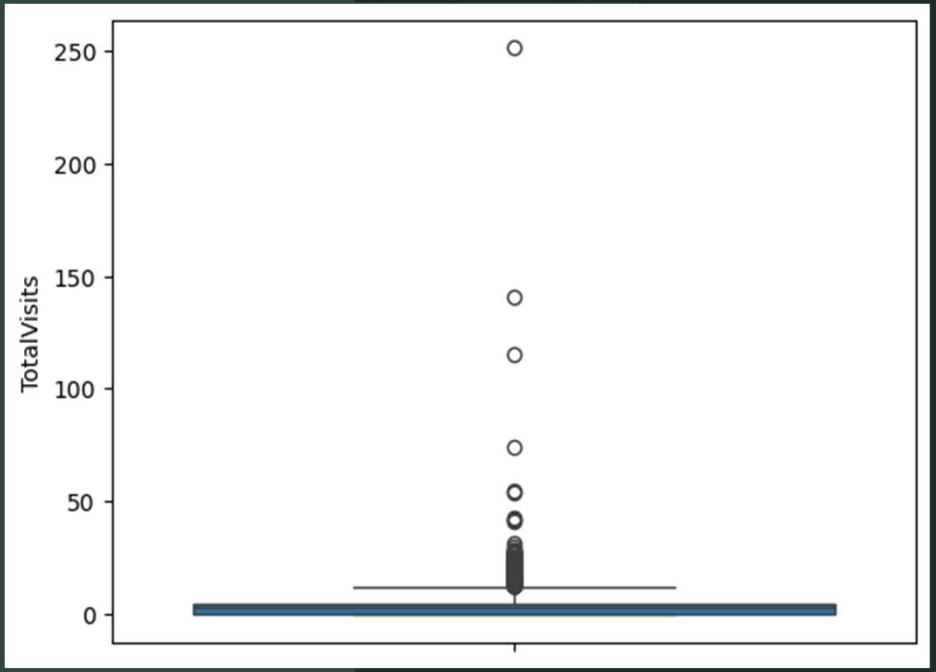
- X Education wants to focus its sales efforts on high-potential leads, known as "Hot Leads," to save time and resources. By identifying these promising leads, the company aims to improve the efficiency of its sales process and avoid spending unnecessary time on low-probability leads.
- The goal is to raise the lead conversion rate from the current 30% to 80%. This will result in more paying customers, higher revenue, and better overall performance of the sales and marketing strategies.

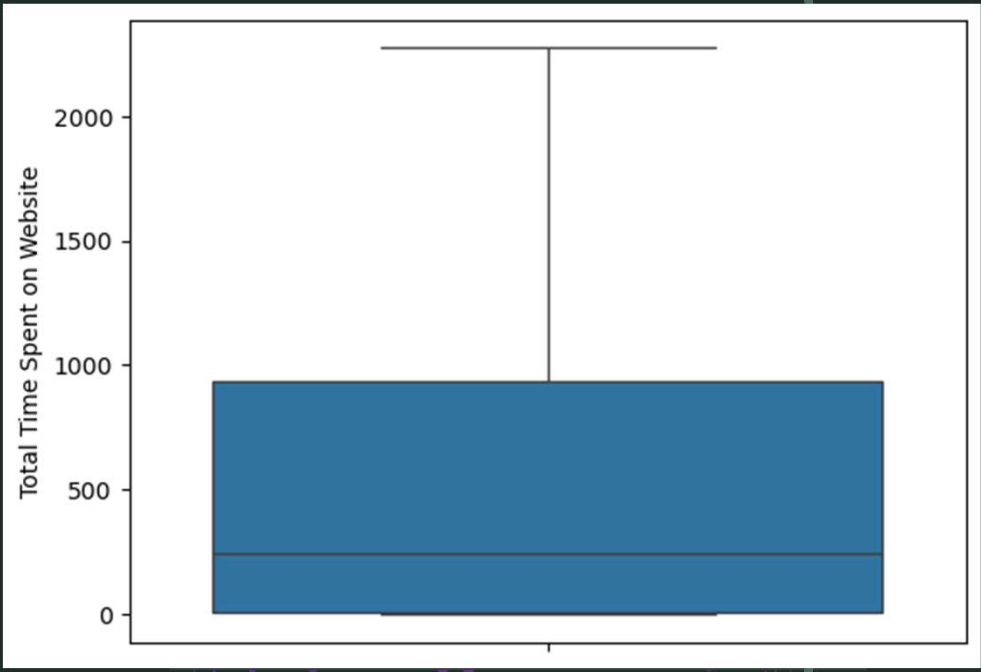
Problem Approach

- Importing and Merging Data
- Inspecting the Dataframe
- Data Cleaning and Preparation
- Data Visualization
- Bivariate and Multivariate Analysis
- Creating Dummy Variables
- Model Building
- Evaluating Our Model
- Making Predictions on Test Set

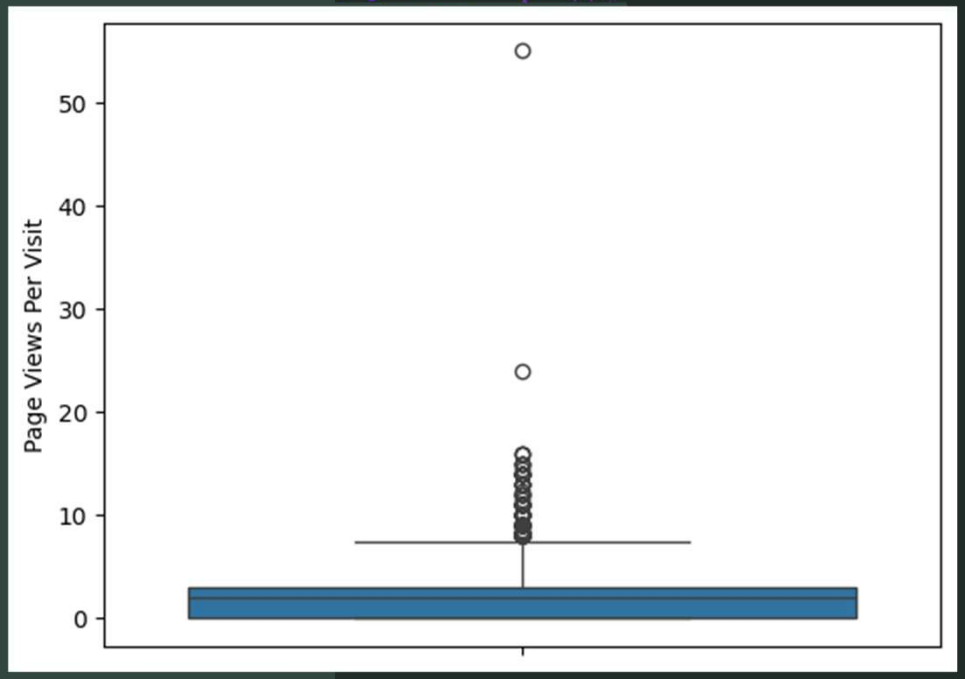


Data Cleaning and Preparation





Data Cleaning and Preparation



The three boxplots represent the following:

1.Page Views Per Visit:

This plot shows significant outliers, with most data points clustered below 10. A few extreme outliers (above 50) indicate visitors with unusually high page views.

2.Total Time Spent on Website:

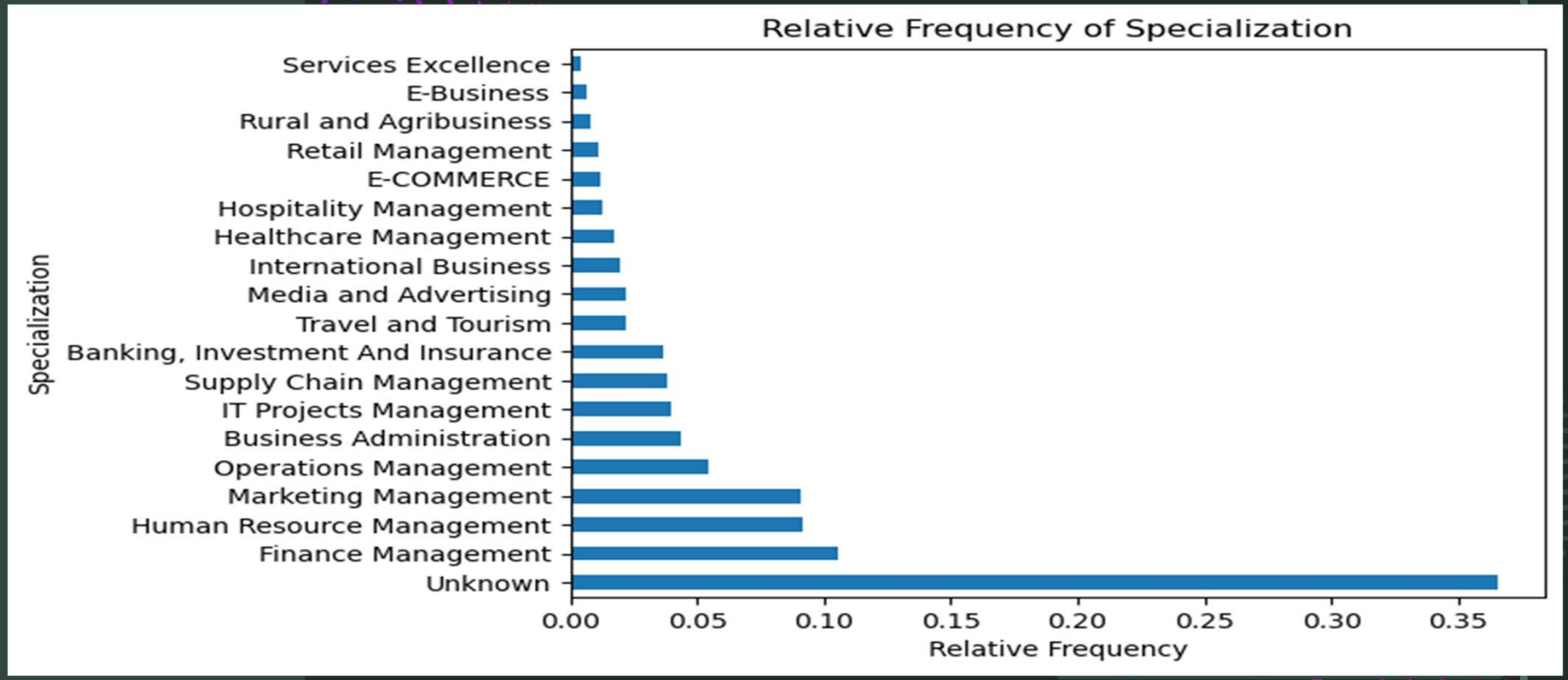
Data is more spread out, ranging from near 0 to over 2,200 seconds. There are no major outliers, but the median time spent appears relatively low, suggesting most users spend minimal time.

3.Total Visits:

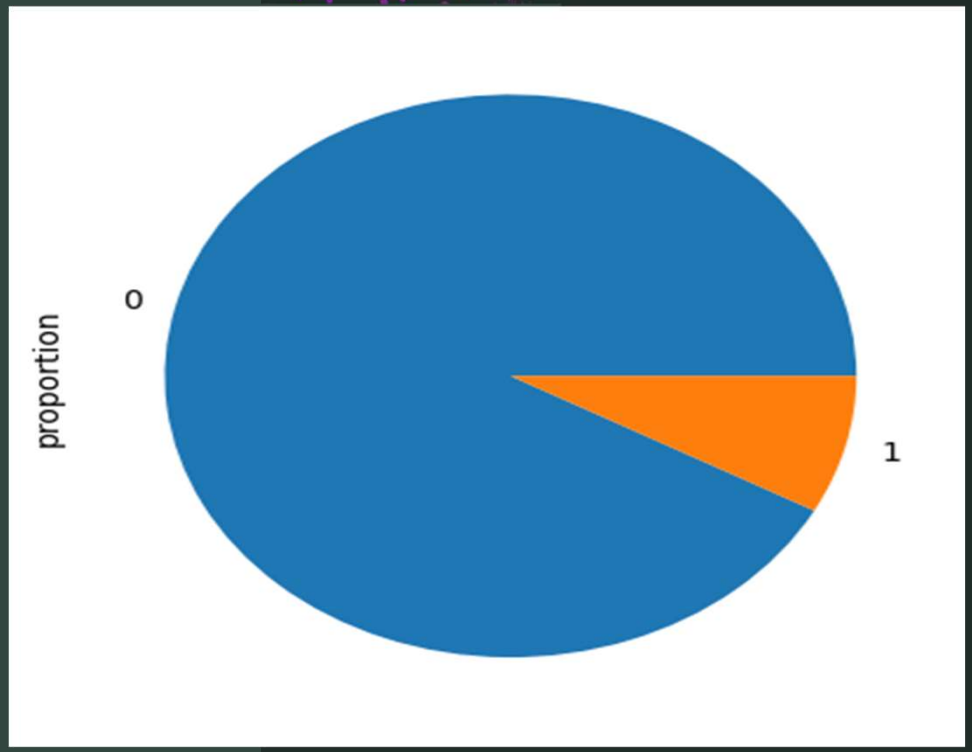
This plot reveals heavy outliers, with most users having fewer than 10 visits. Outliers beyond 50 suggest some visitors repeatedly accessed the website.

Outlier treatment may help improve model accuracy.

Specialization – Leads from unknown, Finance, HR and Marketing management are high frequency to convert

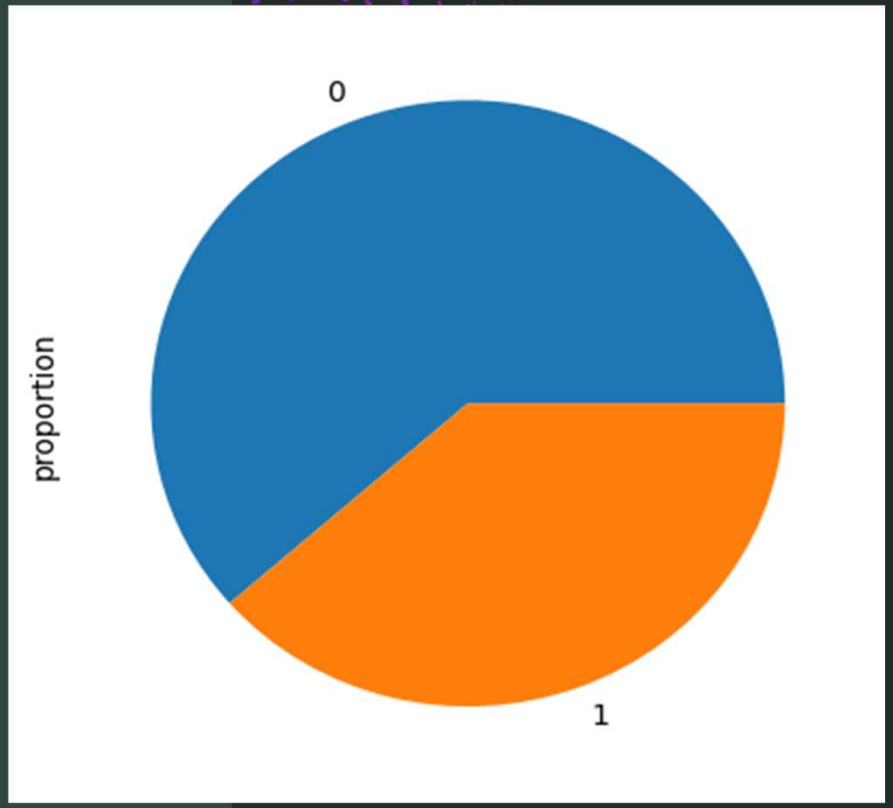


Categorical Ordered Univariate Analysis



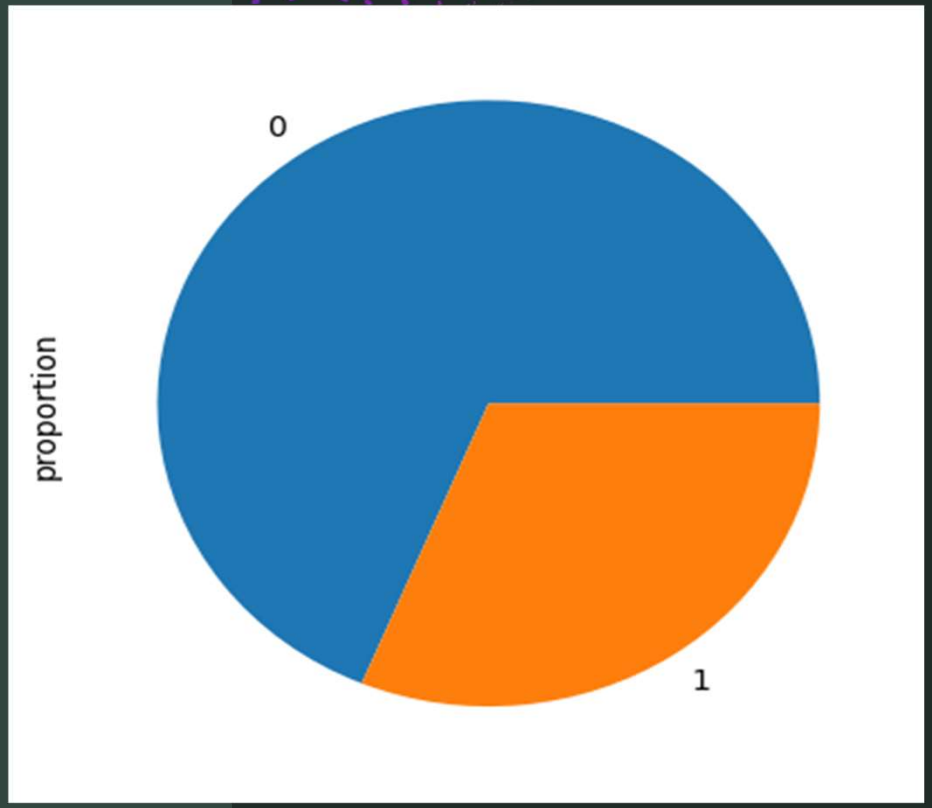
An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not (No = 0, Yes =1).

Categorical Ordered Univariate Analysis



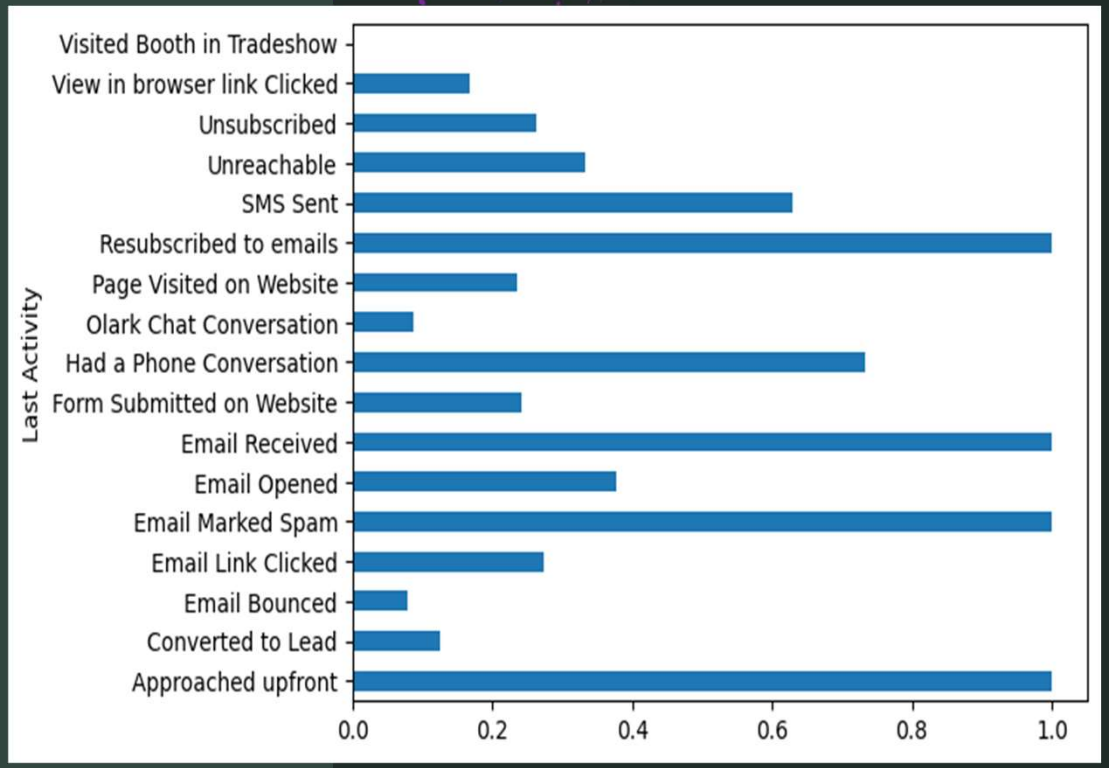
Variable indicating whether a lead has been successfully converted or not (No = 0, Yes =1).

Categorical Ordered Univariate Analysis



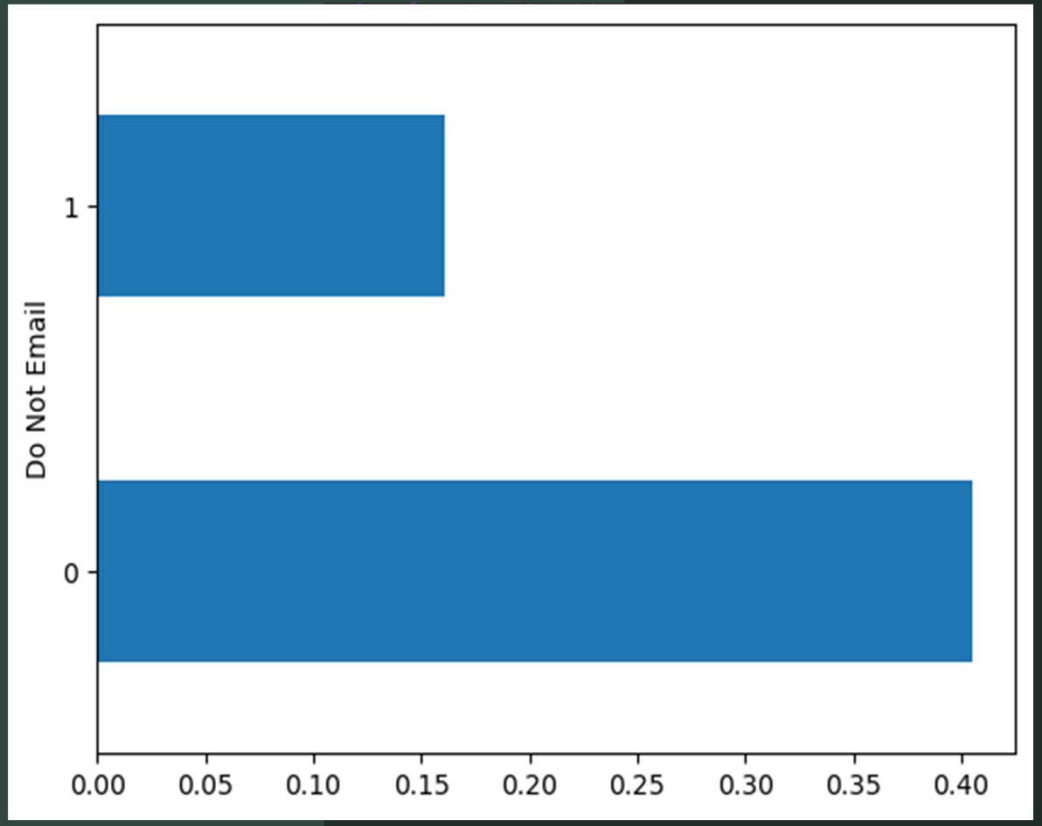
Variable indicating whether the customer wants a free copy of 'Mastering the Interview' or not. (No = 0, Yes =1).

Last Lead Activity



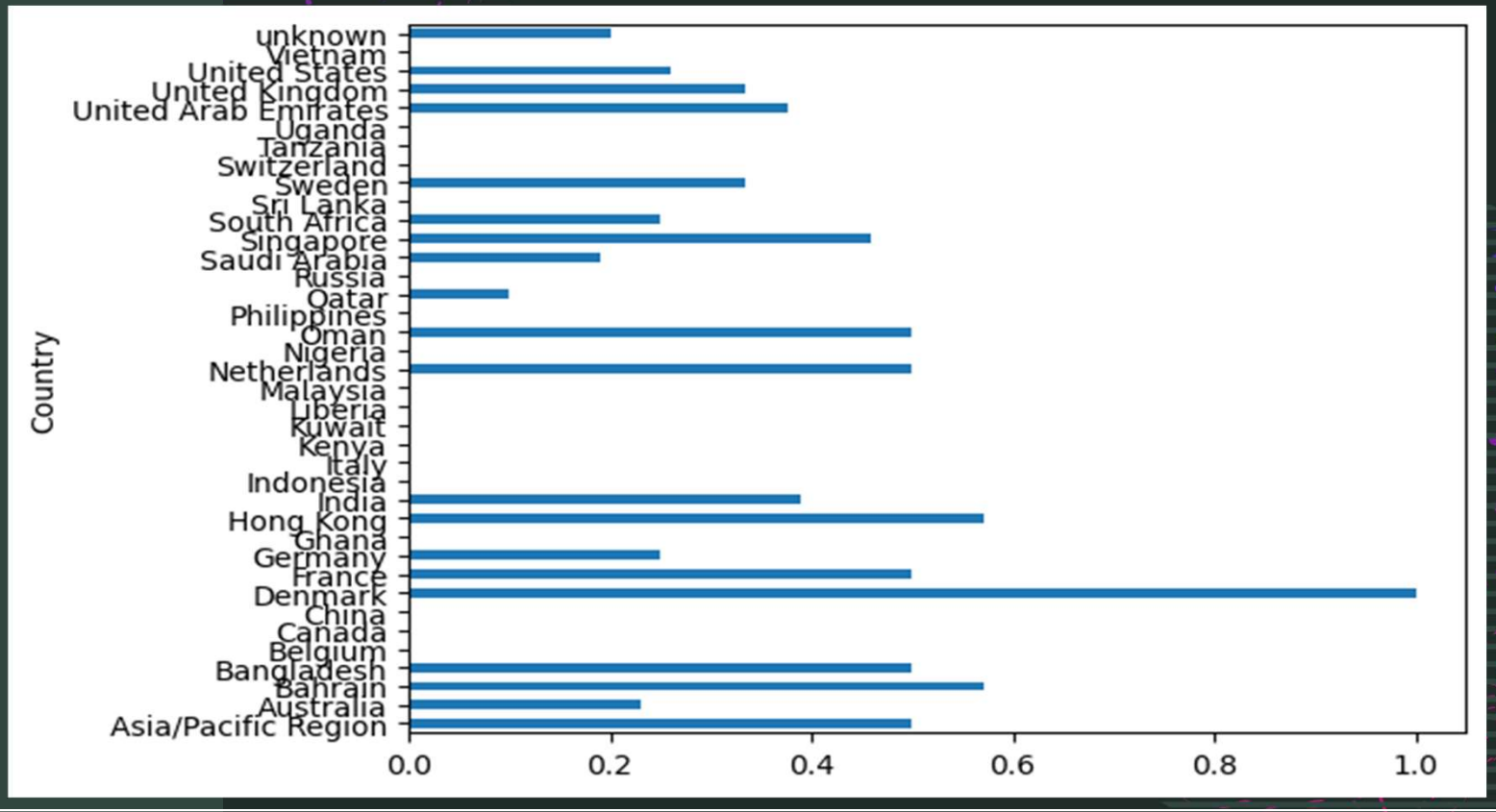
From this graph we can see Emails, SMS and phone conversation have high convert rate.

Do not Email

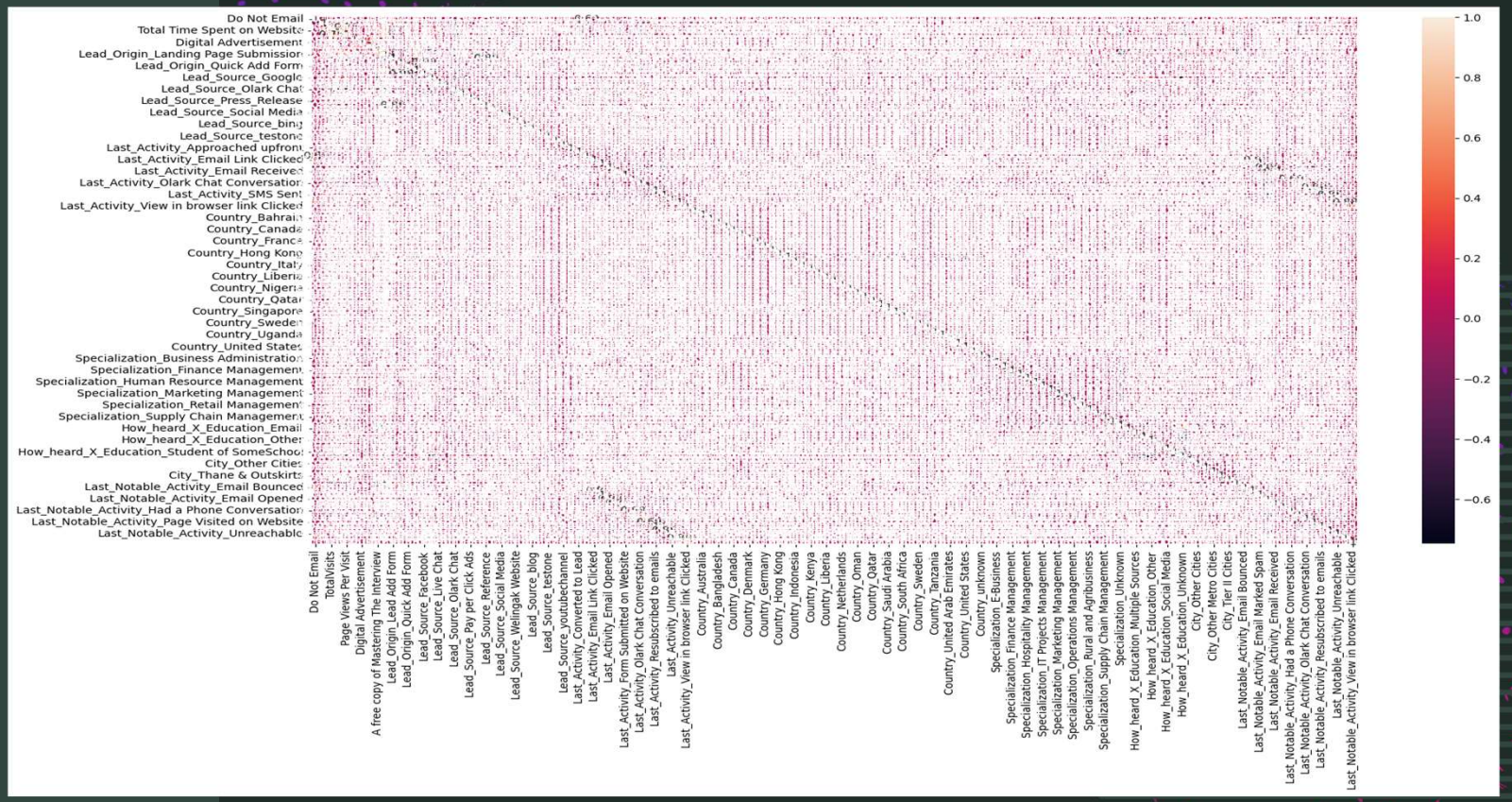


Most of peoples don’t want to receive email about courses.

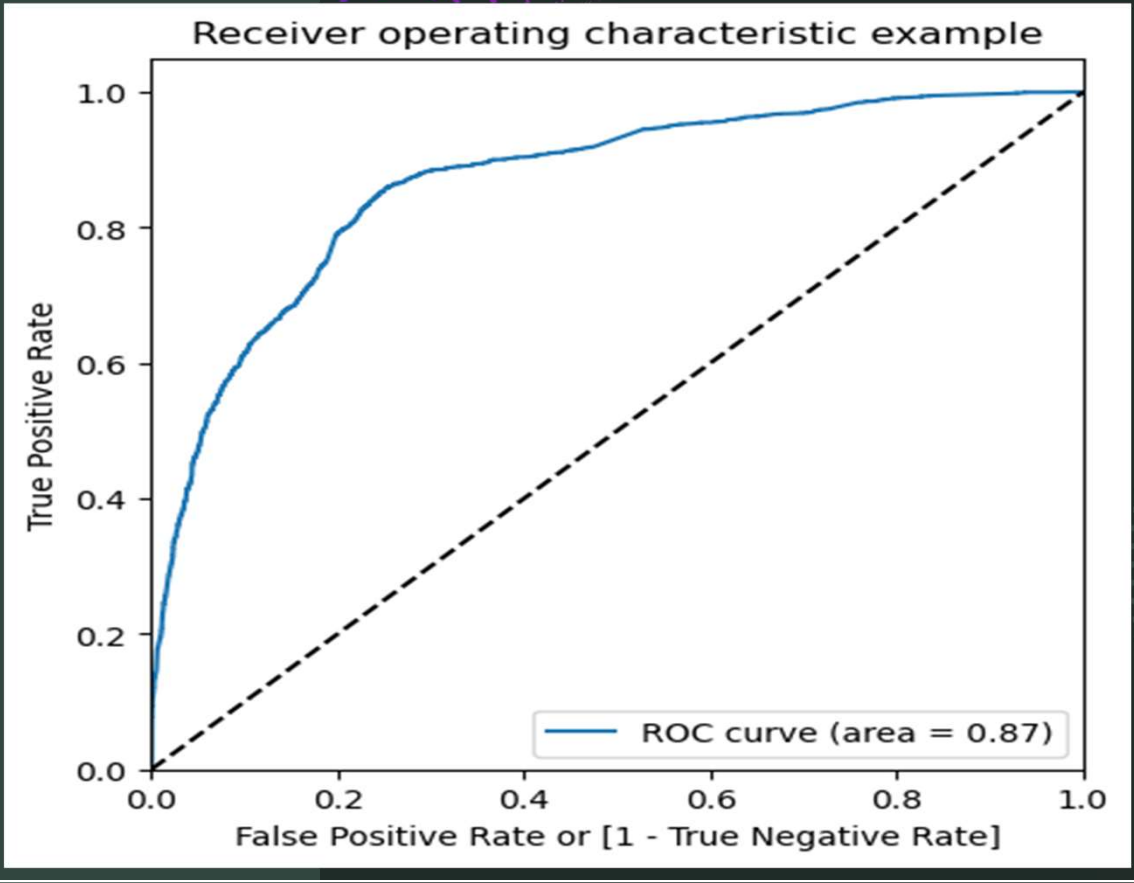
Where are Customers Coming from? Denmark, Bahrain and Hong Kong



Correlation – there is no correlation between the variables



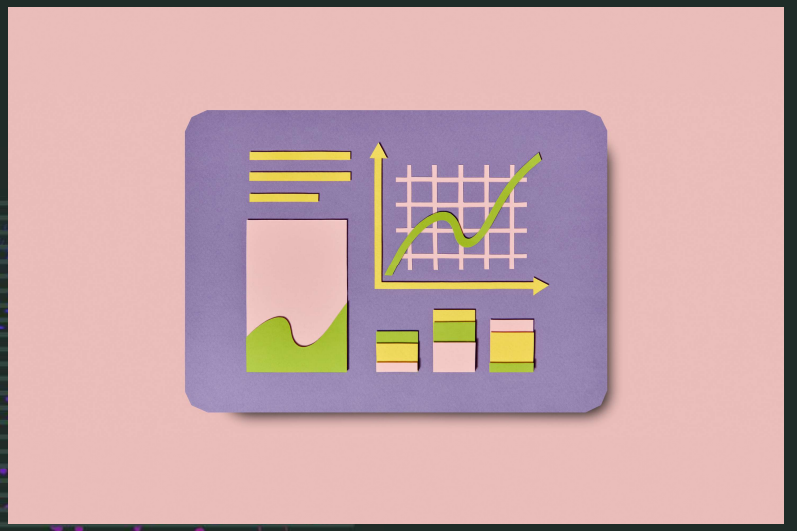
Model Evaluation – ROC Curve



The **ROC curve** measures how well a model can separate two classes. An **AUC value of 0.87** means the model is very good at predicting which leads will convert. The AUC ranges from **0 to 1**, where **0.5** is random guessing and **1** is perfect. A score of 0.87 shows the model has an **87% chance** of correctly identifying converted and non-converted leads. This means the model balances **True Positive Rate** (sensitivity) and **False Positive Rate** well. Overall, the model is reliable and can help prioritize high-potential leads, making the sales process more efficient.

Our Observations

- We used logistic regression to train the model with 15 selected features. After training, the model's performance was evaluated on the training dataset, achieving an accuracy of **78.99%**. This result indicates that the model is learning meaningful patterns from the data and provides a good starting point for further improvement.



- On the test data, we experimented with different decision thresholds to optimize the model's performance. Initially, a threshold of **0.42** resulted in an accuracy of **79.65%**. We tested other thresholds, including **0.3, 0.4, and 0.5**, and found that a threshold of **0.4** achieved the best accuracy of **80.45%**. This improvement demonstrated that the model generalizes well to unseen data. The results suggest the model is not overfitting and can make reliable and accurate predictions on new leads.

Conclusion

Our approach focused on cleaning and analyzing lead data to predict conversions effectively. We used **logistic regression** to train the model with selected features. Key insights emerged from the graphs:

1.Country Analysis: Certain countries have higher lead conversion rates, indicating they should be prioritized. Especially customers from Denmark, Bahrain and Hong Kong.

2.Do Not Email: Most of clients prefer not to receive email, X education would like to explore different marketing mean.

3.Last Lead Activity: Features like “SMS Sent” and “Email Received” show strong associations with conversion.

4.Page Views and Visits: Outliers indicate some visitors are highly engaged but rare.

The model achieved a solid performance, identifying critical factors to prioritize high-potential leads efficiently.

The End