Lead Scoring Case Study

Presented by:

Sudhamsha Sagar

Date Presented:

8th October 2024

Problem Statement

An education company called X Education sells online courses to professionals. Each day, many people visit their website to look at courses, fill out interest forms, or watch videos. These visitors come from different places, like ads on websites or search engines like Google, and some are referred by others. If someone fills out a form with their contact details, they are marked as a "lead." The company's sales team then contacts these leads through phone calls or emails, but only about 30% of the leads end up buying a course.

X Education wants to improve this lead conversion rate. They hope to find the most likely leads, called "Hot Leads," so the sales team can focus on them and improve the chances of converting more leads into customers.

Goal

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future

Problem Approach

Data Preparation	EDA	Dummy Variable Creation	Model Building
• Importing the data	• Cleaning the data by checking missing values and other info • Plotting graphs for better analysis	• Creating dummy variables as the data has categorical variables and we follow rfe method for model building	• Includes Test Train Split • Scaling if required but suggested • RFE • Sensitivity, Specificity • Model Evaluation for both test set

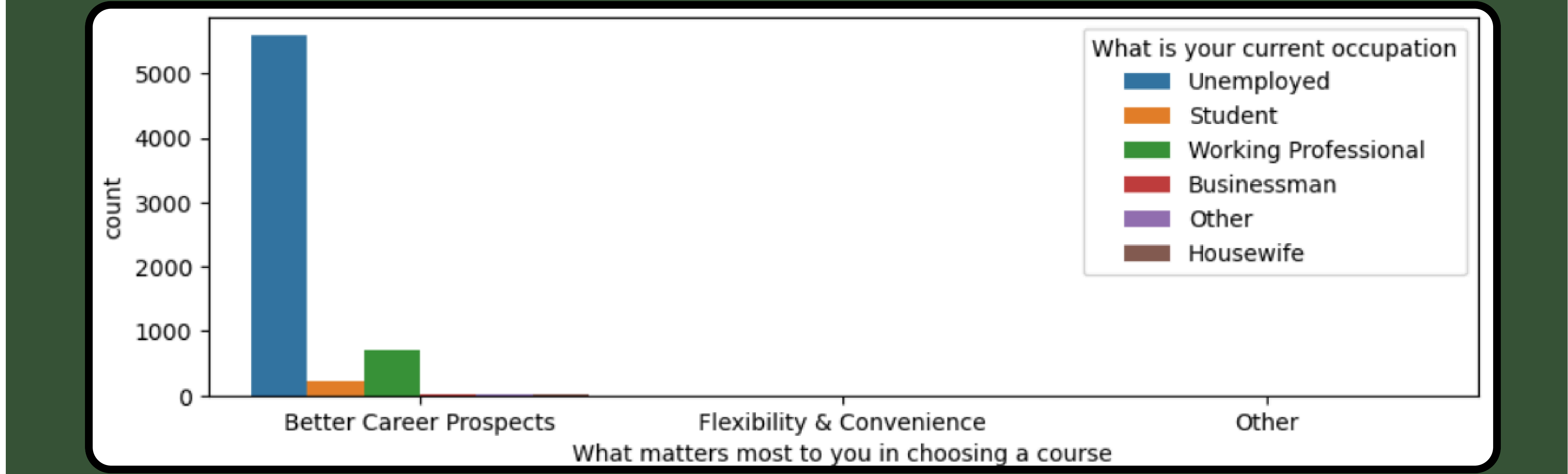
Data Understanding

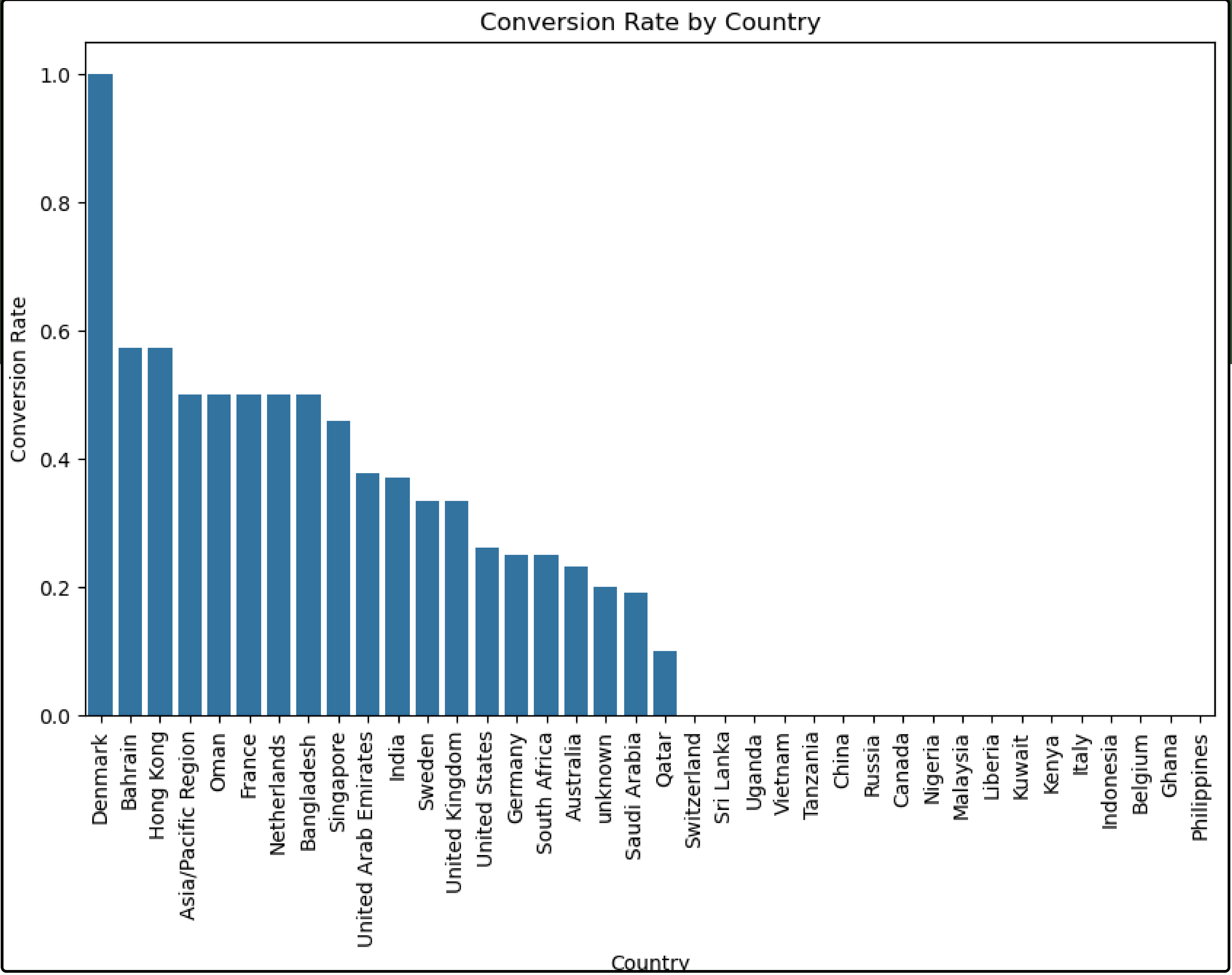
	Value
Shape of Data	9240 Rows, 37 Columns
Missing Values	17 Columns had missing values in which 9 columns had more than 30% of missing values which are dropped directly

Data Visualisation

- There is a relationship b/w these two variables.
- Umemployees are more but working professions joined the course in large number

Occupation v/s Reason for Joining

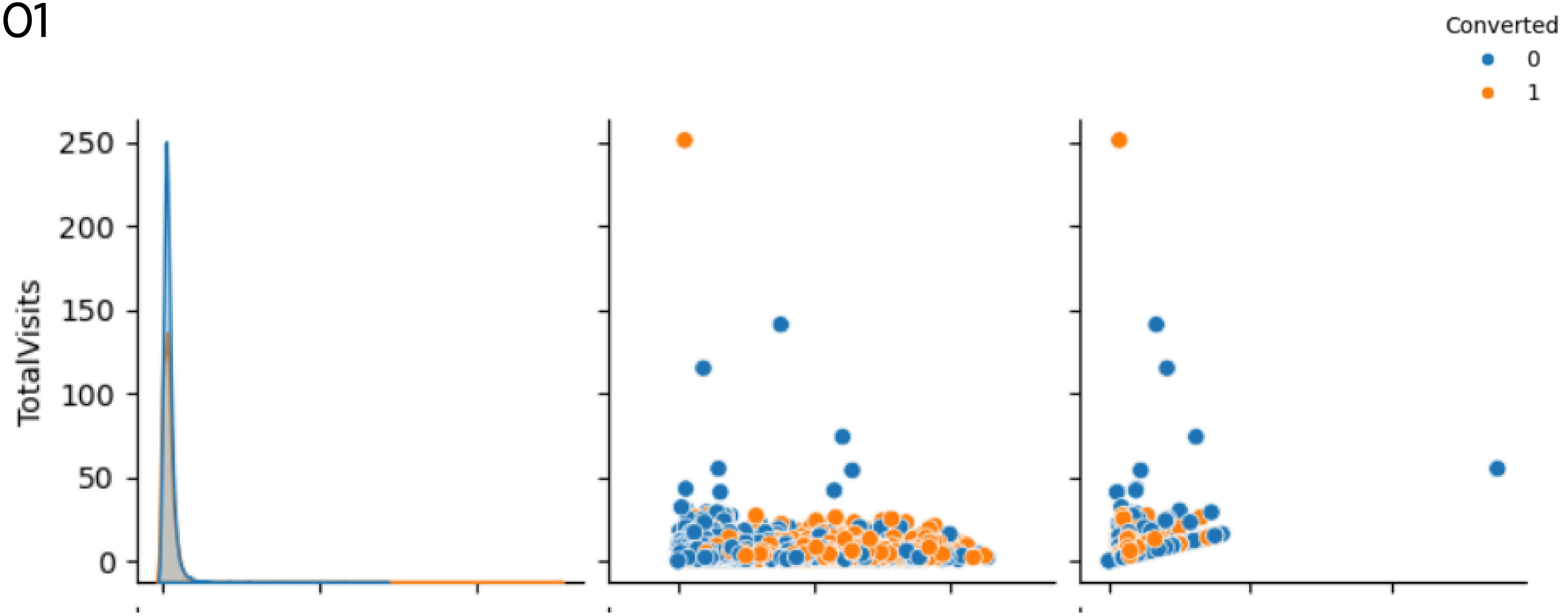




There is no significant insight can be drawn on Conversion Rate using Country. So have dropped this along with the City Column

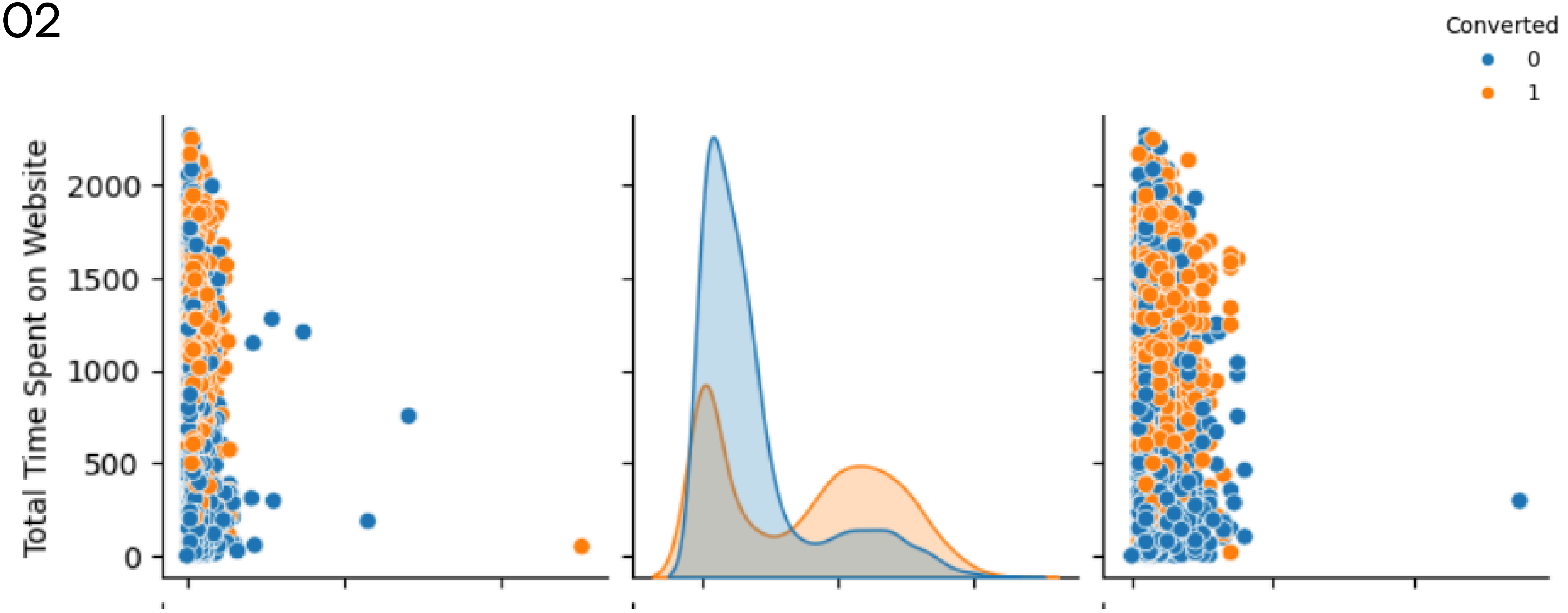
Relationship between Total Visits, Total Time Spent on Website, and Page Views Per Visit in relation to the conversion status of leads.

01



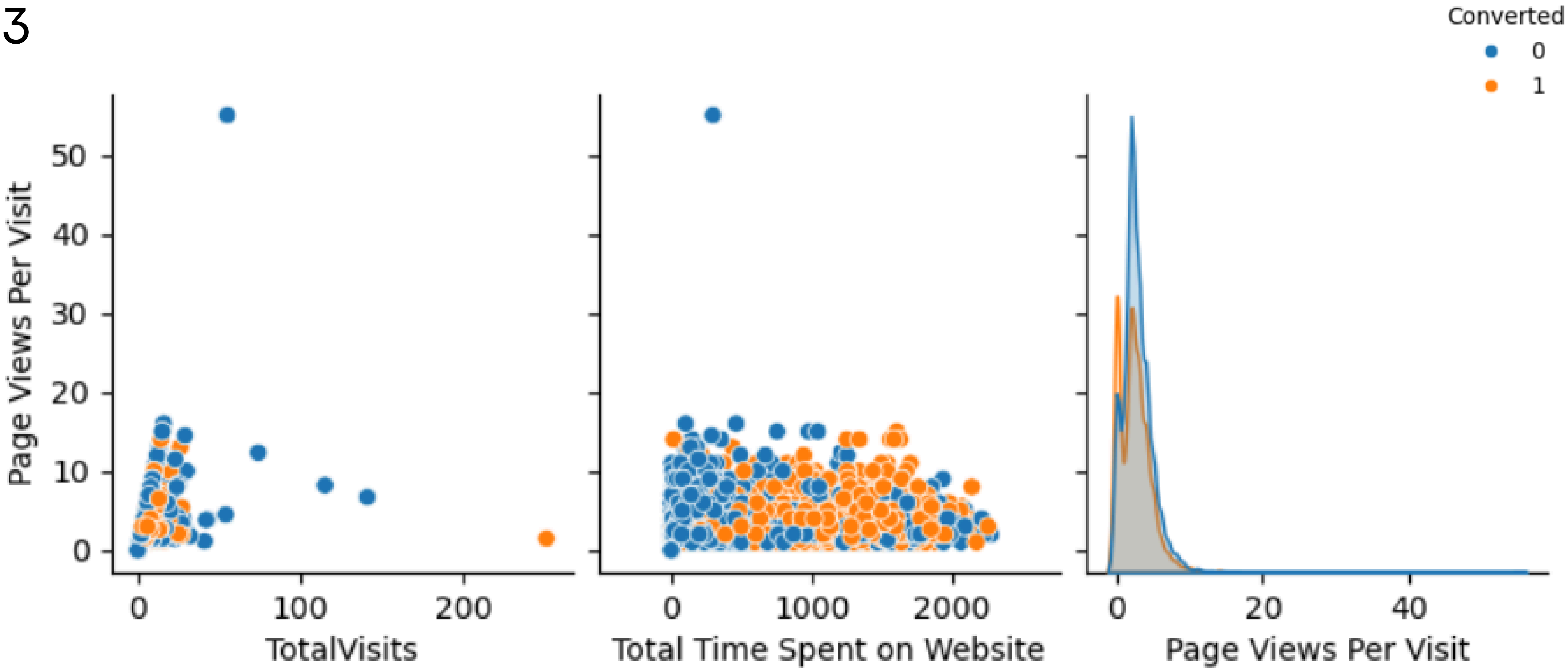
Relationship between Total Visits, Total Time Spent on Website, and Page Views Per Visit in relation to the conversion status of leads.

02



Relationship between Total Visits, Total Time Spent on Website, and Page Views Per Visit in relation to the conversion status of leads.

03



Insights on Pairplots

1.Distribution of Variables:

Each variable shows a right-skewed distribution, indicating that a small number of users have very high values for total visits, time spent, and page views, while most users have lower values. This suggests the presence of outliers or a long tail in the dataset.

2.Conversion Status:

In many of the scatter plots, it appears that converted leads (orange dots) tend to have higher values for Total Visits, Total Time Spent on Website, and Page Views Per Visit compared to non-converted leads (blue dots).

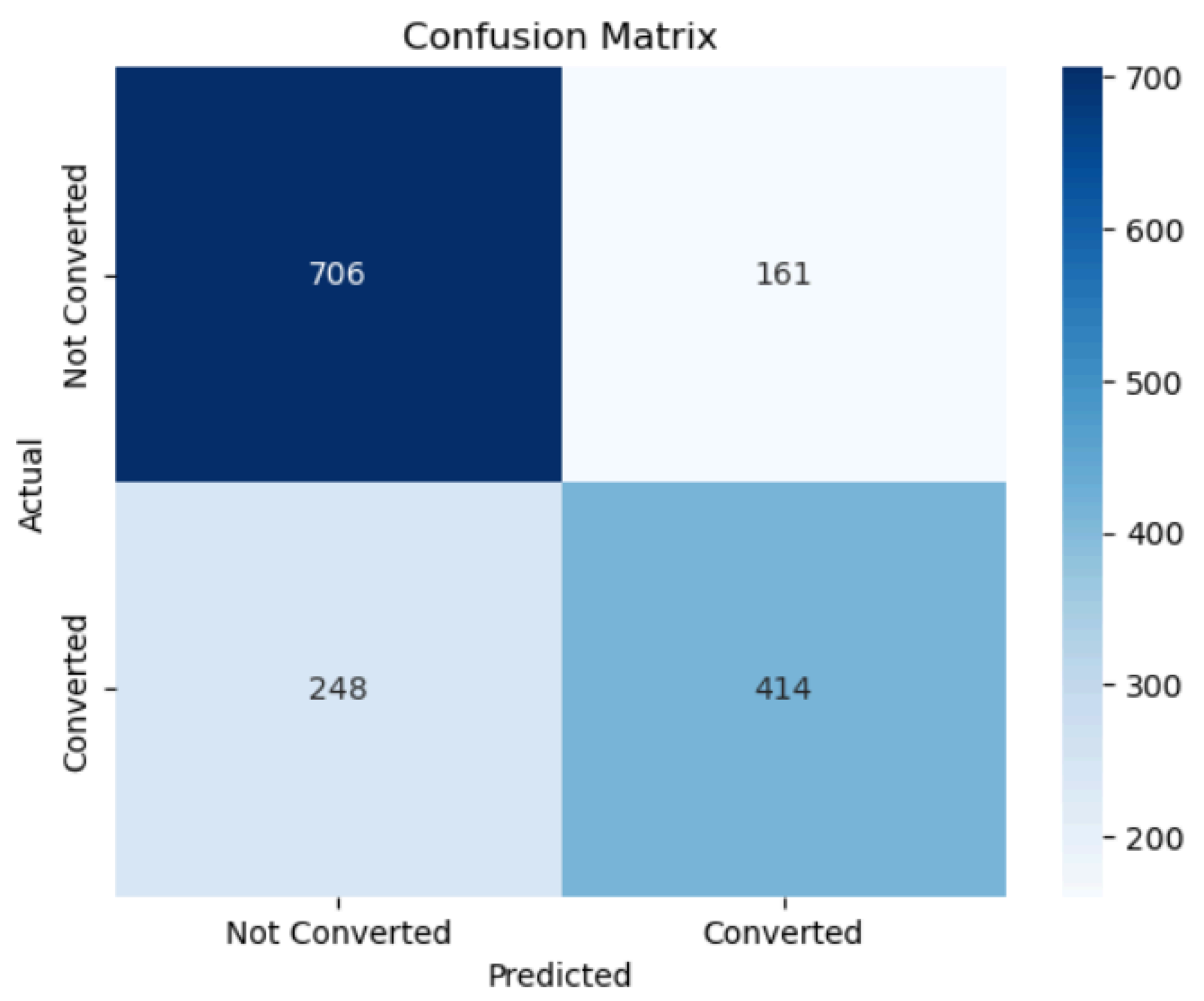
3.Total Visits vs. Total Time Spent:

There seems to be a positive correlation between Total Visits and Total Time Spent on Website. Higher numbers of visits correlate with a longer time spent on the website, particularly among converted leads. This may suggest that engaging with the content leads to a higher likelihood of conversion.

4.Page Views Per Visit:

The Page Views Per Visit variable does not seem to have a strong differentiation between converted and non-converted leads. There is a mix of colors (blue and orange) at lower page views, indicating that both converted and non-converted leads had similar page views per visit.

Confusion Matrix Before Optimisation



True Negatives (TN): 706 - The number of instances correctly predicted as not converted (class 0).

False Positives (FP): 161 - The number of instances incorrectly predicted as converted (class 1) when they are actually not converted.

False Negatives (FN): 248 - The number of instances incorrectly predicted as not converted (class 0) when they are actually converted.

True Positives (TP): 414 - The number of instances correctly predicted as converted (class 1).

Model has an accuracy of 73%

Confusion Matrix After Optimisation

- True Negatives (764): 764 leads were correctly classified as not converted.
- False Positives (103): 103 leads were incorrectly classified as converted, but they weren't.
- False Negatives (201): 201 leads were incorrectly classified as not converted, but they actually converted.
- True Positives (461): 461 leads were correctly classified as converted

Model has an accuracy of 80%

Summary

1.Data Exploration and Preprocessing:

- The dataset was analyzed for missing values, and appropriate strategies were employed to handle them. Columns such as 'What matters most to you in choosing a course' and 'How did you hear about X Education' were imputed with their respective modes.
- Categorical variables were converted to dummy variables, and boolean values were transformed to binary (0 and 1) for model training.

2.Feature Selection and Model Training:

- Logistic regression was used to build the lead scoring model, with careful consideration of the features included based on their correlation with lead conversion rates.
- The model was trained and evaluated using cross-validation to ensure its robustness and reliability.

Summary

3.Model Evaluation:

- The model achieved an accuracy of 73% on the test set. After optimizing the hyperparameters, the accuracy improved to 80%, indicating a significant enhancement in predictive performance.
- A confusion matrix and classification report were generated to assess precision, recall, and F1 scores for both classes (converted and not converted).

4.Lead Scoring:

- Based on the predicted probabilities from the logistic regression model, lead scores ranging from 0 to 100 were generated. These scores indicate the likelihood of conversion, with higher scores denoting "hot leads."
- A new column for lead scores was added to the dataset, enabling the sales team to target leads more effectively.

Future Recommendations

1. **Scalability**
   - Adaptation to Larger Datasets: The logistic regression model is scalable and can efficiently handle larger datasets as the company grows. Regular updates to the model with new leads will ensure continued accuracy in predictions.
   - Incorporating Additional Features: The model can be enhanced by integrating new features or variables as they become relevant, allowing it to adapt to changing business environments and customer behaviors.

2. **Dynamic Scoring System**
   - Adjustable Scoring Thresholds: The lead scoring threshold can be adjusted based on the company's evolving business strategy. This allows the sales team to dynamically classify leads as hot or cold, improving the targeting of outreach efforts.
   - Real-Time Scoring Updates: With proper infrastructure, the model can generate real-time lead scores, enabling immediate prioritization of leads based on the most current data.

Future Recommendations

3. Model Retraining and Maintenance

- Periodic Retraining: The model should be retrained periodically (e.g., quarterly or biannually) with new data to reflect the latest trends and maintain accuracy. This is essential to account for changes in lead behavior and market conditions.

- Performance Monitoring: Ongoing monitoring of model performance metrics (e.g., accuracy, precision, recall) will help identify when retraining is necessary. If performance drops, a review of the model and its features will be conducted.

4. User Feedback Integration

- Feedback Mechanism: Establishing a feedback loop with the sales team to gather insights on lead quality can provide valuable information for model adjustments and improvements.

- Incorporating Qualitative Data: Including qualitative feedback from sales interactions can enhance the model by providing additional context to lead characteristics.

THANK YOU