Lead Scoring Assignment

Case Study Group

1. Tushar Rajput
2. Tanisha Grover
3. Vansh Tandon

Business Context & Objective

- **What is Lead Scoring?**
  - Prioritizing leads based on their likelihood to convert.
- **Business Objective:**
  - X Education sells online courses to industry professionals.
  - X Education receives many leads, but its conversion rate is low (only 30% conversion).
  - The company wants to identify high-potential leads (Hot Leads) to improve efficiency.
  - By focusing on these leads, the sales team can improve the conversion rate.

Goals of the Case Study

- **Objective:**
  - Identify the most promising leads.
  - Build a logistic regression model to assign a lead score between 0-100.
  - Higher score = Higher likelihood of conversion (Hot Lead), Lower score = Less likelihood (Cold Lead).
  - The CEO aims to achieve a lead conversion rate of 80%.
  - Ensure the model accounts for future constraints like peak time actions, workforce utilization, and post-target strategies.
- **Deployment Goal:**
  - The model should be scalable for future use and adaptable to changing business requirements.

Solution Methodology

- **Data Cleaning & Preparation:**
  - Check and handle duplicate data.
  - Handle missing values (drop or impute as needed).
  - Remove irrelevant columns with excessive missing data.
  - Identify and handle outliers.
- **Exploratory Data Analysis (EDA):**
  - Univariate Analysis – Value counts, distribution of variables.
  - Bivariate Analysis – Correlation coefficients, feature relationships.
- **Feature Engineering:**
  - Dummy variable creation and encoding of categorical data.
  - Feature scaling for numerical variables.

Solution Methodology

- **Model Building & Validation:**
  - Logistic regression for classification.
  - Recursive Feature Elimination (RFE), R-squared, VIF, and p-values for feature selection.
  - Training-Test Split (70%-30%).
  - Model validation using accuracy, precision, recall, F1-score, and AUC-ROC.
- **Making Predictions:**
  - Assigning lead scores based on logistic regression probabilities.
  - Interpreting the results for business decisions.

Data Understanding

- **Dataset Details:**
  - **Total Entries:** 9,240
  - **Total Columns:** 37
  - Included categorical, numerical, and text-based features.
  - Presence of missing values, redundant columns, and uninformative variables.

Data Preprocessing

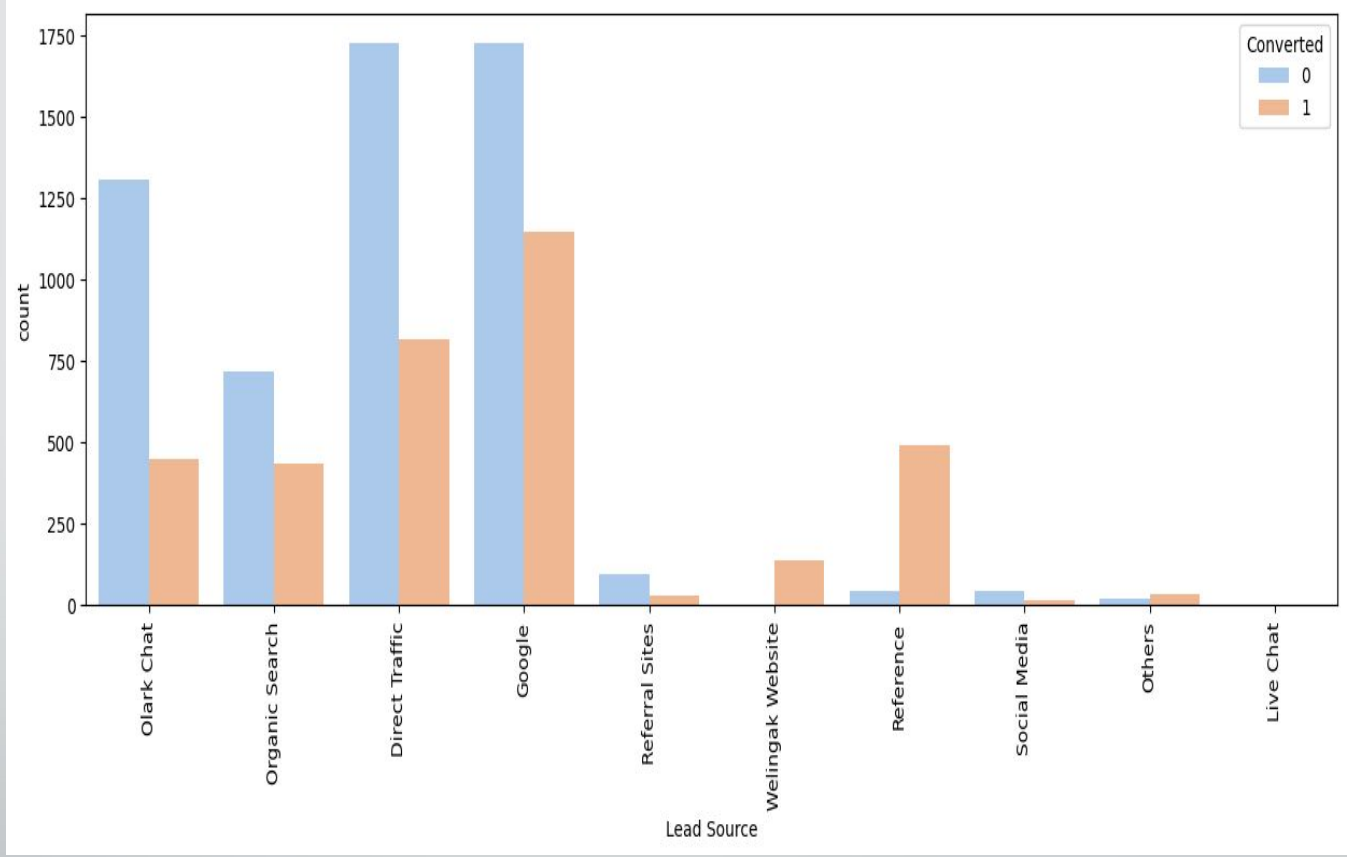
- **Handling Missing Values:**
  a. Columns with excessive missing values (e.g., 'Country', 'Lead Quality' etc) or no variance were dropped..
  b. Remaining missing values were either filled with appropriate statistics or removed.
- **Feature Selection:**
  a. Removed unnecessary features such as 'Prospect ID', 'Lead Number', and other redundant categorical fields.
- **Encoding Categorical Data:**
  a. Converted categorical variables into numerical format where necessary.
- **Feature Scaling:**
  a. Standardization applied for numerical columns.

Processed Data

- **Final Dataset After Cleaning:**
  - **Total Entries:** 520
  - **Total Columns:** 16
  - Key features retained: 'Lead Origin', 'Lead Source', 'Converted', 'TotalVisits', 'Total Time Spent on Website', and others.
  - Dataset is now ready for further analysis and modeling.

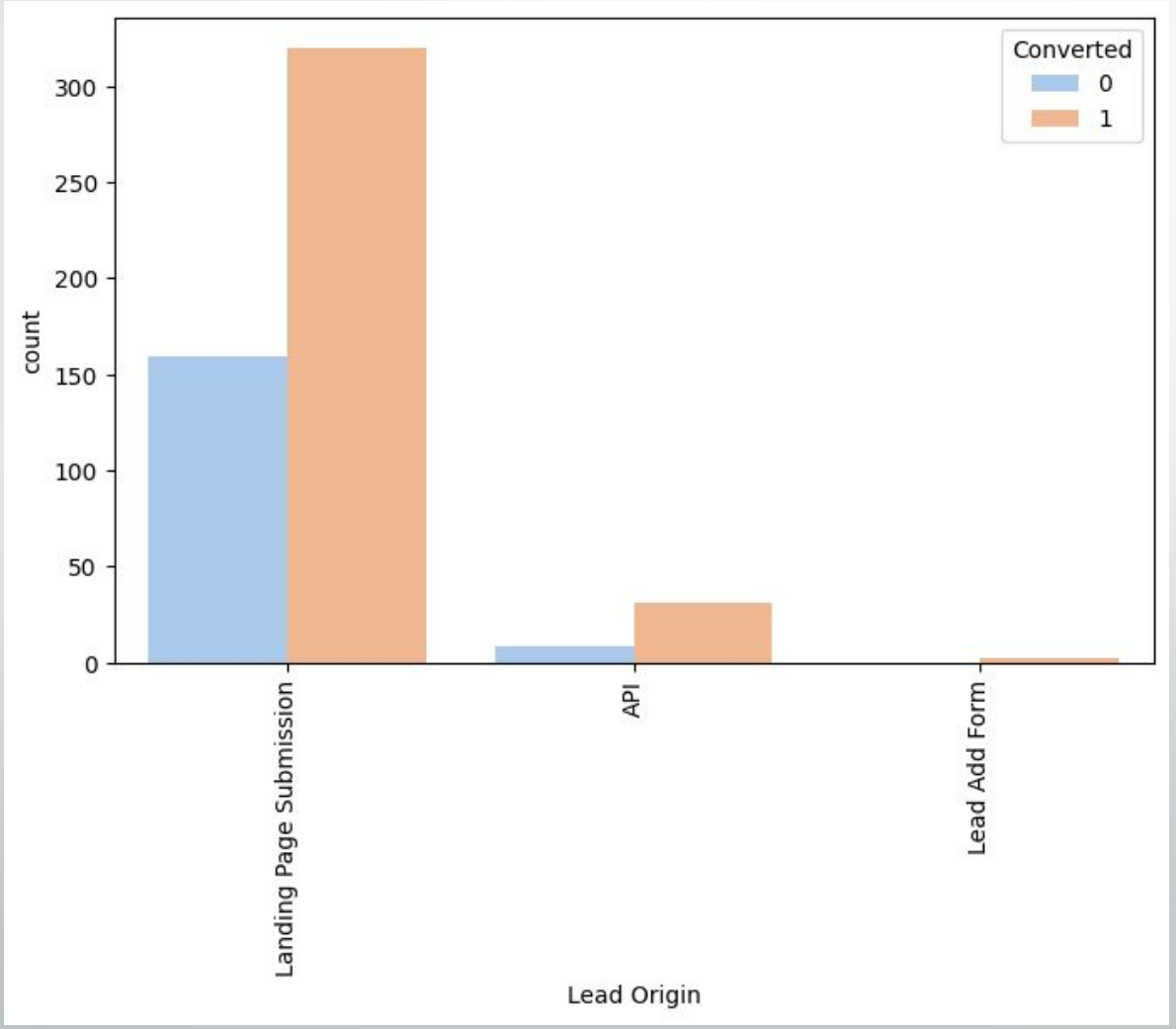
Data Visualization & key Insights

Lead Source



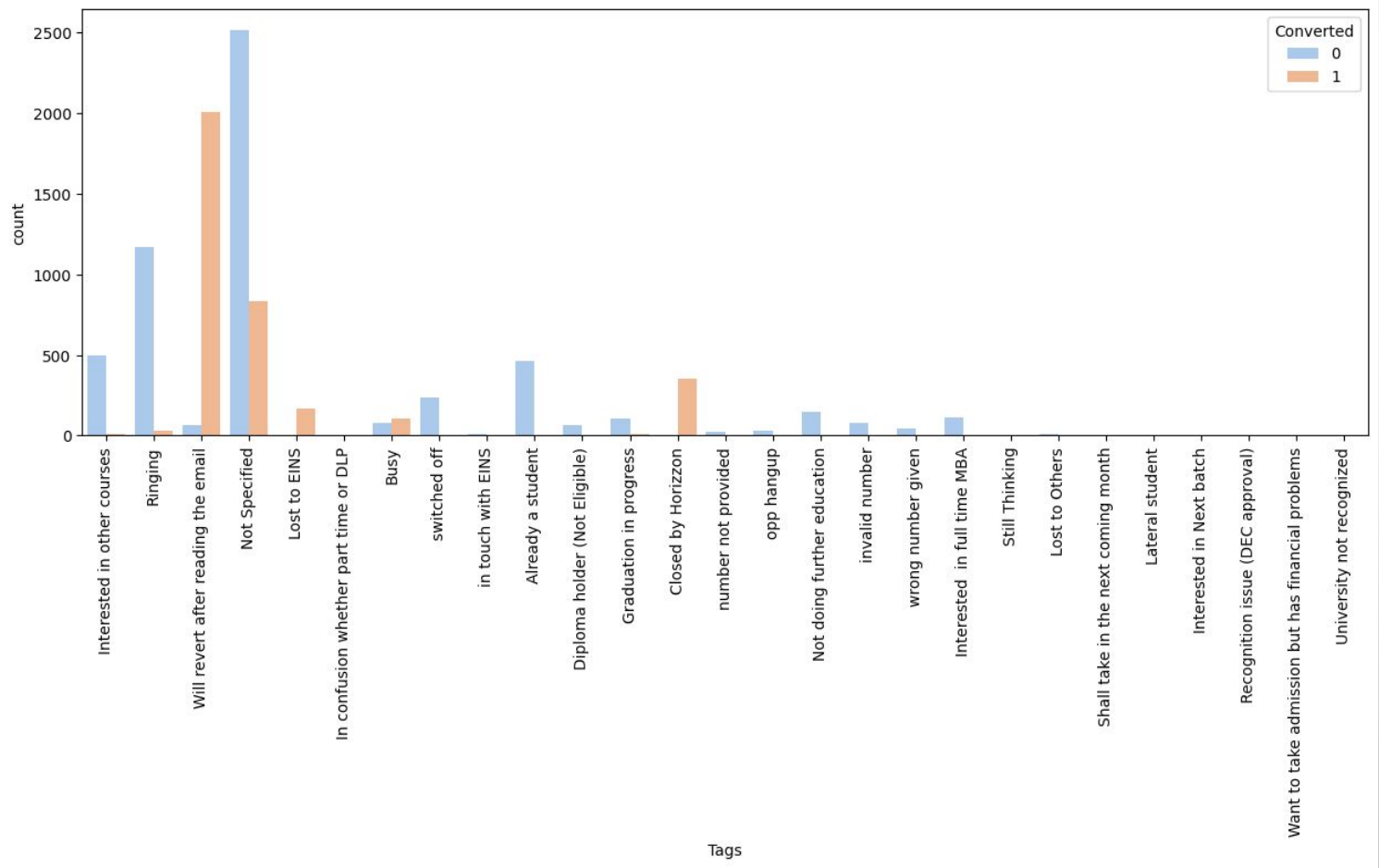
- **Maximum lead generation** comes from **Google and Direct Traffic**.
- **High conversion rates** are observed for **Reference leads** and **Welingak Website leads**.
- To **improve overall conversion**, efforts should focus on **enhancing conversion rates** for **Olark Chat, Organic Search, Direct Traffic, and Google** while also **generating more leads** from **Reference and Welingak Website**.

Lead Origin



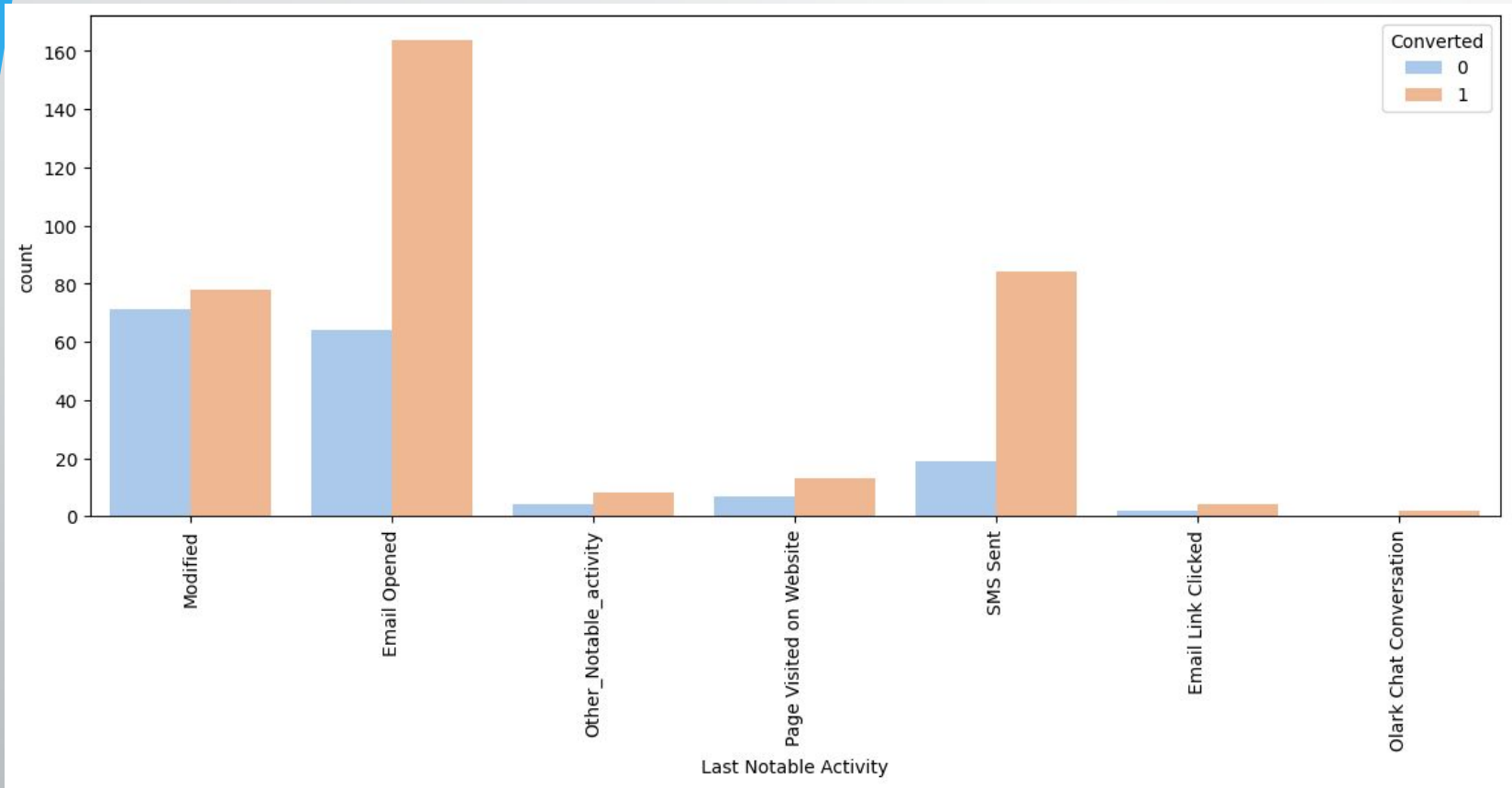
- **API and Landing Page Submission** bring the highest number of leads and conversions.
- **Lead Add Form** has a **very high conversion rate** but generates fewer leads.
- **Lead Import and Quick Add Form** receive very few leads.
- To improve the **overall lead conversion rate**, efforts should focus on improving the conversion of **API and Landing Page Submission** while generating more leads from **Lead Add Form**.

Tags



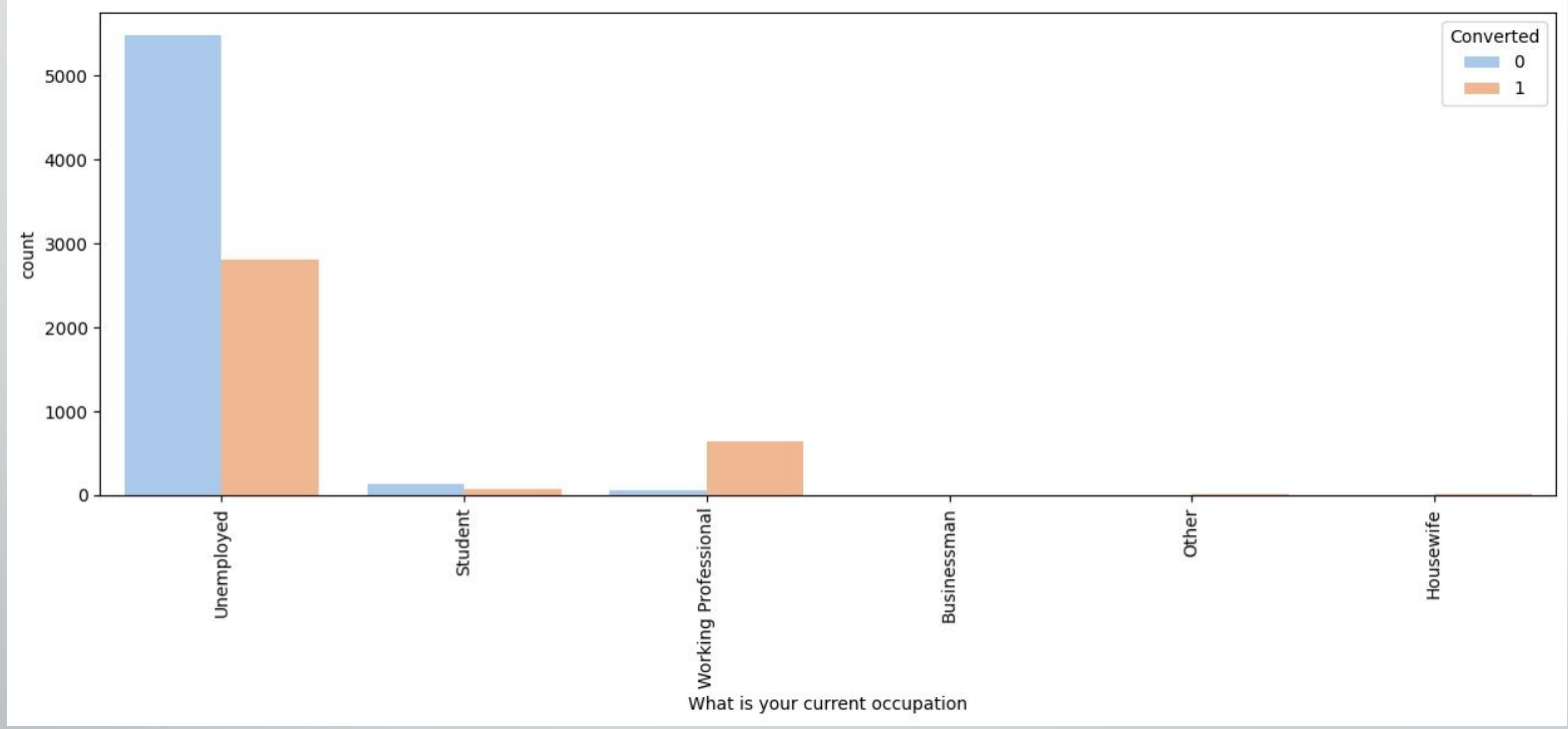
- The majority of leads fall under the **"Not Specified"** and **"Will revert after reading the email"** categories.
- **Conversion is relatively higher** for leads tagged as **"Closed by Horizon"**, indicating strong intent.
- To improve conversions, focus on **engaging and following up** with leads in the **"Will revert after reading the email"** category.

Last Notable Activity



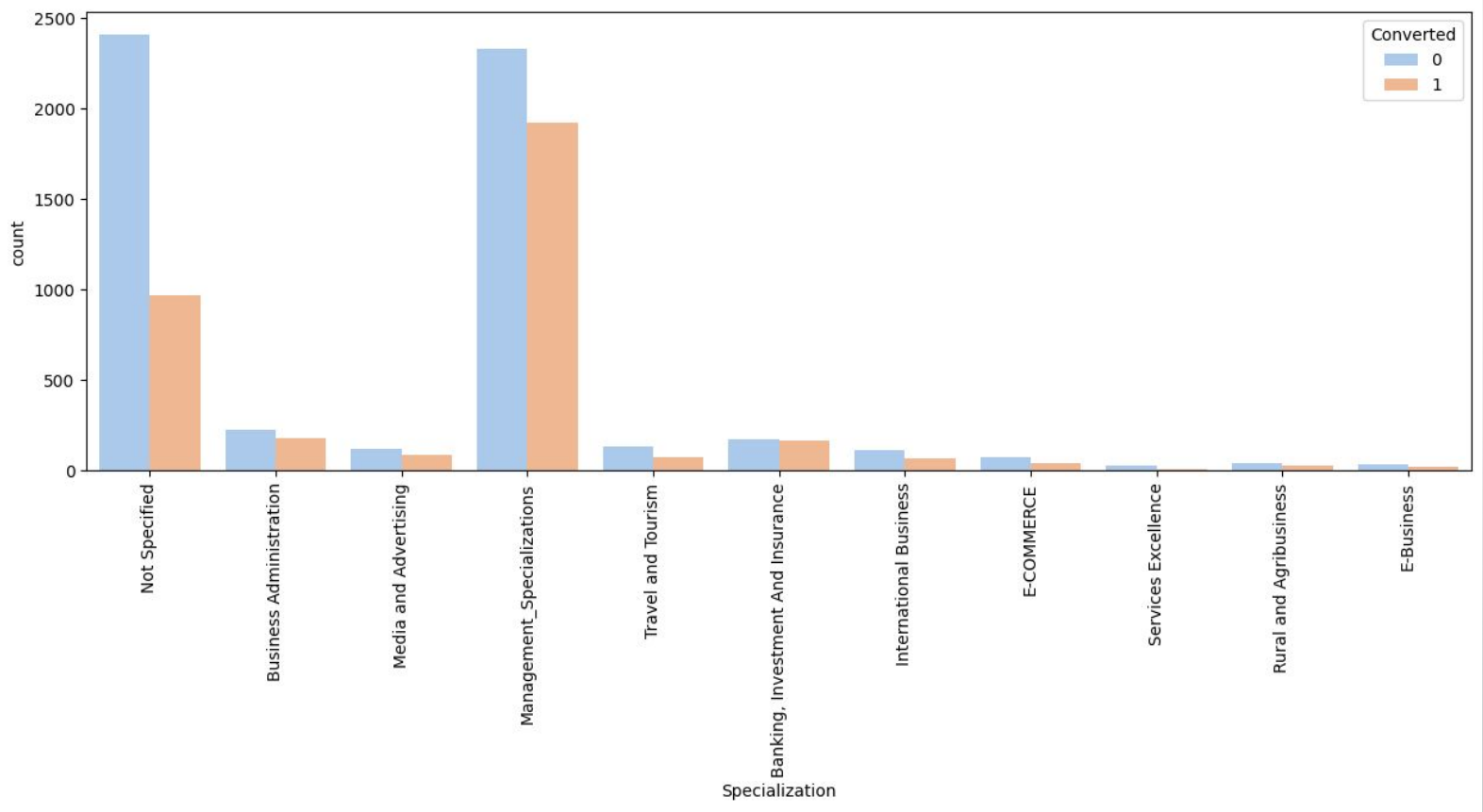
- **Leads who open emails** have a high probability of conversion.
- **SMS communication** also increases the likelihood of conversion.
- **Website visits and chat conversations** show minimal impact on conversion.

What is your current occupation



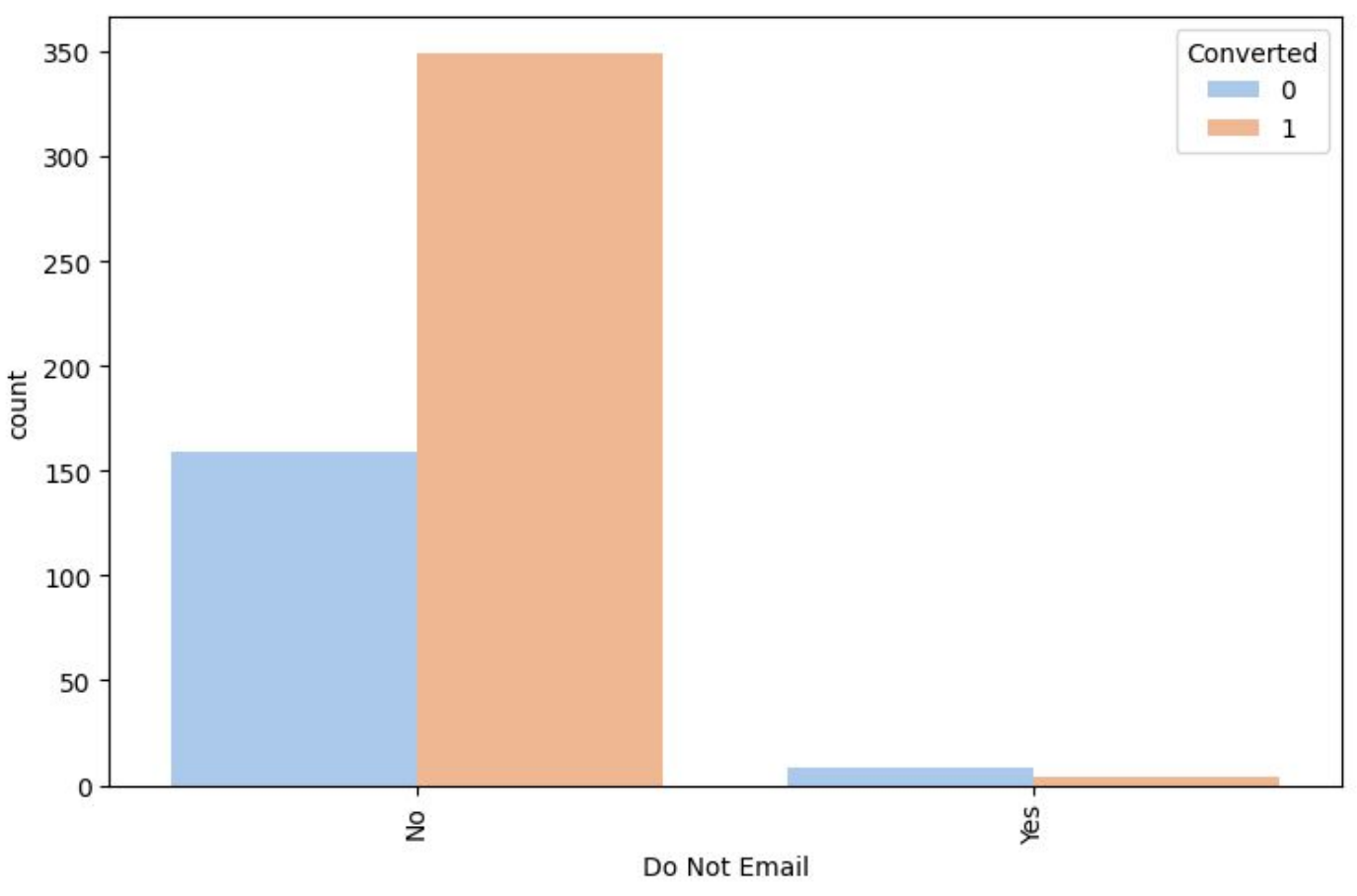
- **Unemployed individuals** show the highest interest in joining the course, with a significant number converting.
- **Working professionals** also contribute to conversions, but in smaller numbers.
- **Students and other categories** have minimal impact on lead conversion.

Specialization



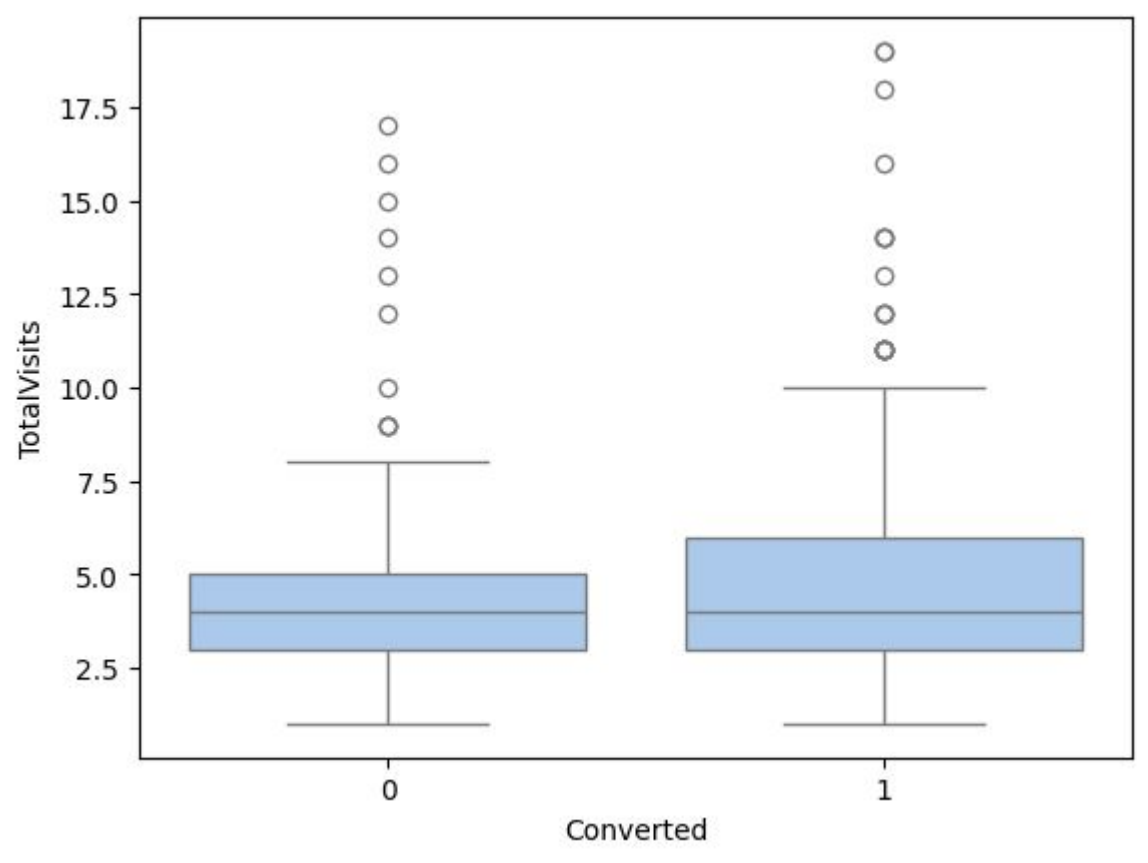
- **Management Specializations** have the highest number of leads and conversions.
- **Not Specified** category also has a significant number of leads but a lower conversion rate.
- Other specializations contribute minimally to overall conversions.

Do Not Email



- Leads who **do not opt out of emails** have a significantly higher conversion rate.
- Leads marked as **Do Not Email** have very few conversions, suggesting emails are an important communication channel for lead conversion.

TotalVisits

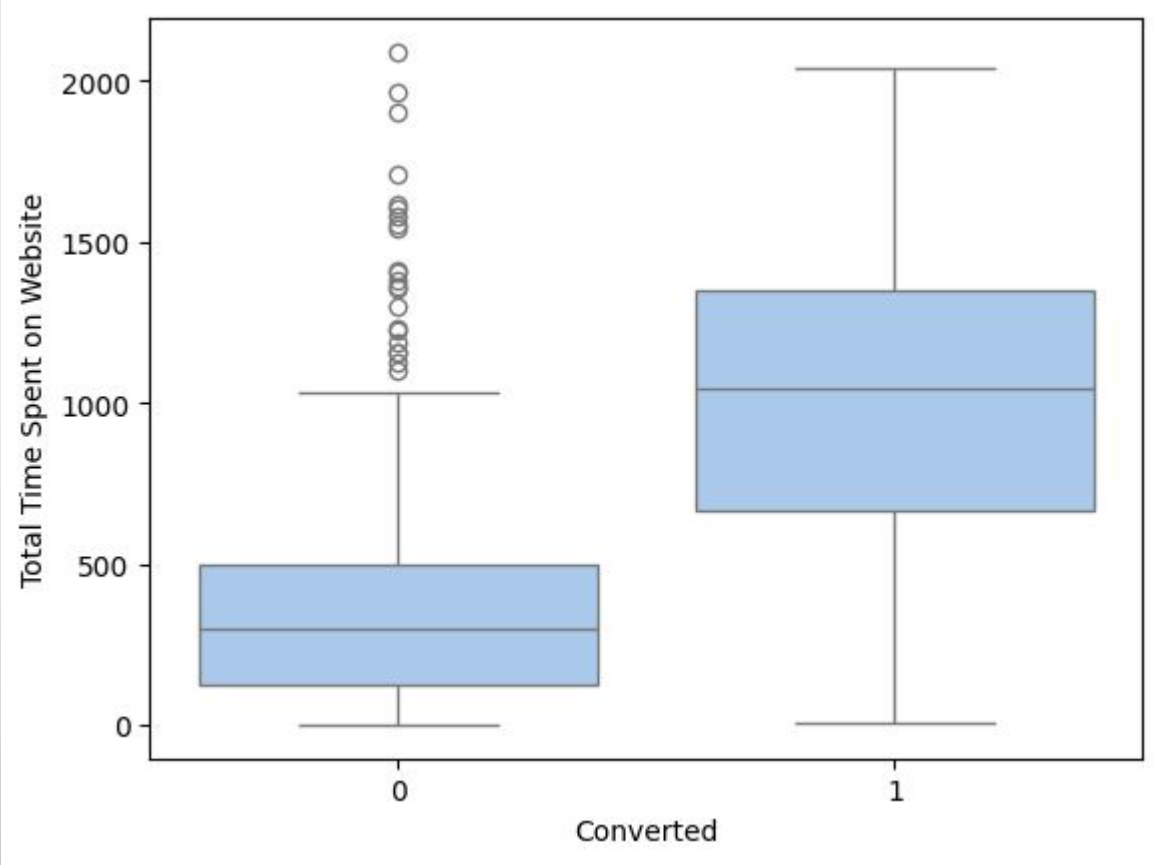


The box plot compares the distribution of "Total Visits" for converted and non-converted leads. Both groups have a similar median, and the range of values, including outliers, overlaps significantly.

Inference:

1. The median for converted and non-converted leads is close, indicating no major difference in visit behavior.
2. No clear conclusion can be drawn about the impact of "Total Visits" on conversion.

Total Time Spent on Website

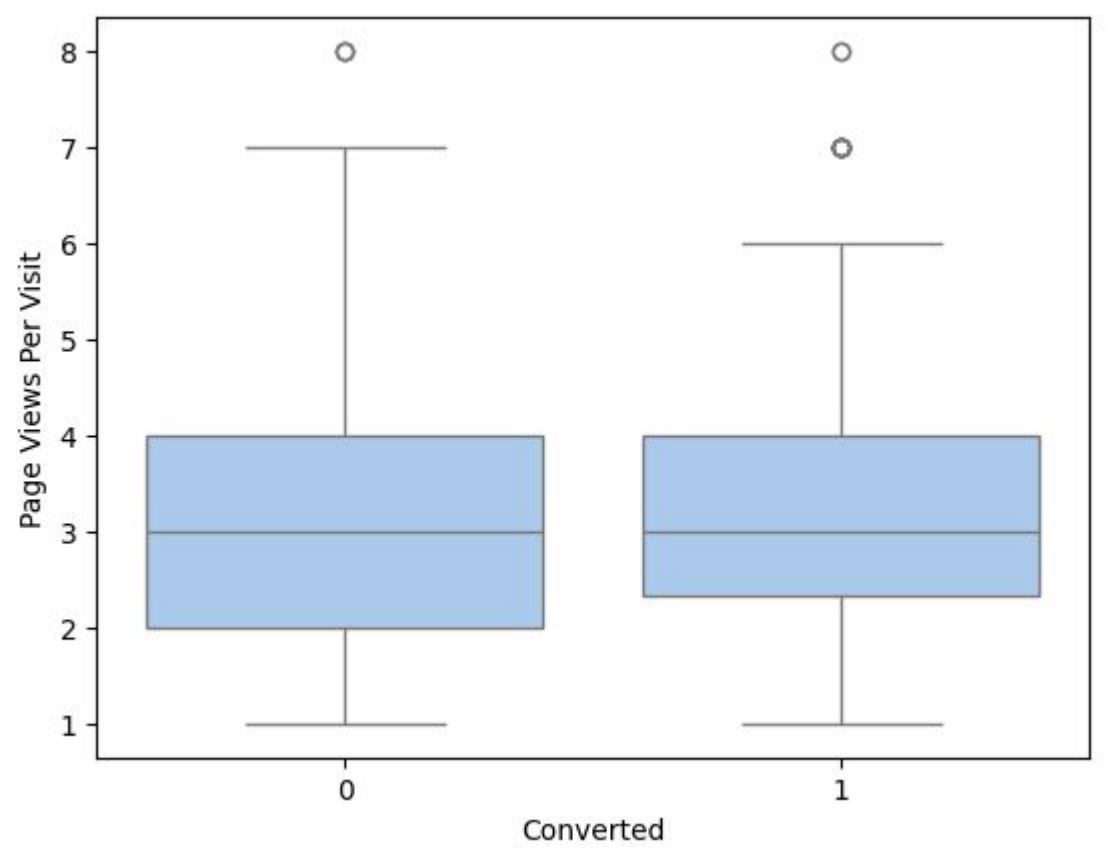


The box plot shows that converted leads tend to spend significantly more time on the website compared to non-converted leads. The median time spent is much higher for converted leads.

Inference:

1. Leads spending more time on the website are more likely to convert.
2. Enhancing website engagement could help increase conversion rates.

Page Views Per Visit

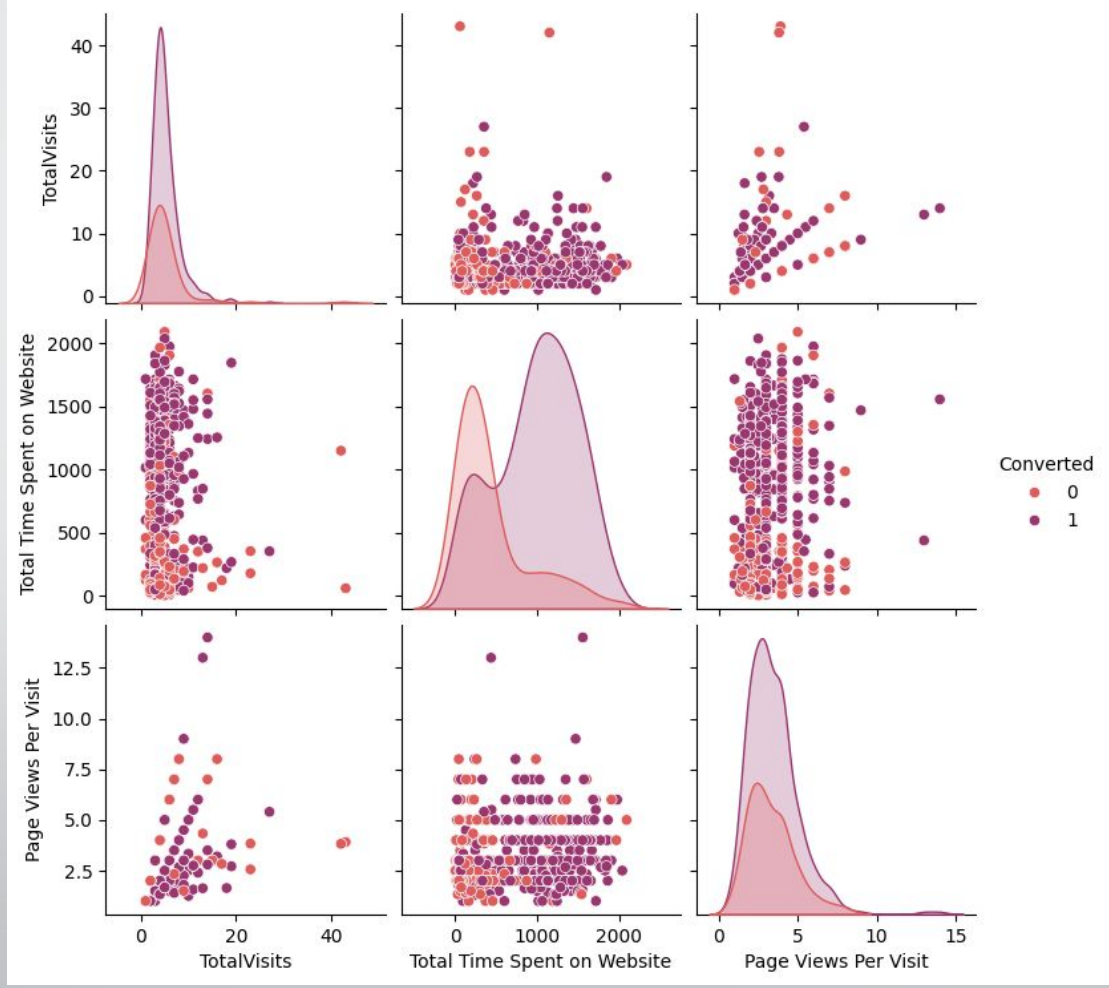


The box plot shows that the distribution of Page Views Per Visit is similar for both converted and unconverted leads. The median values and interquartile ranges are almost identical.

Inference:

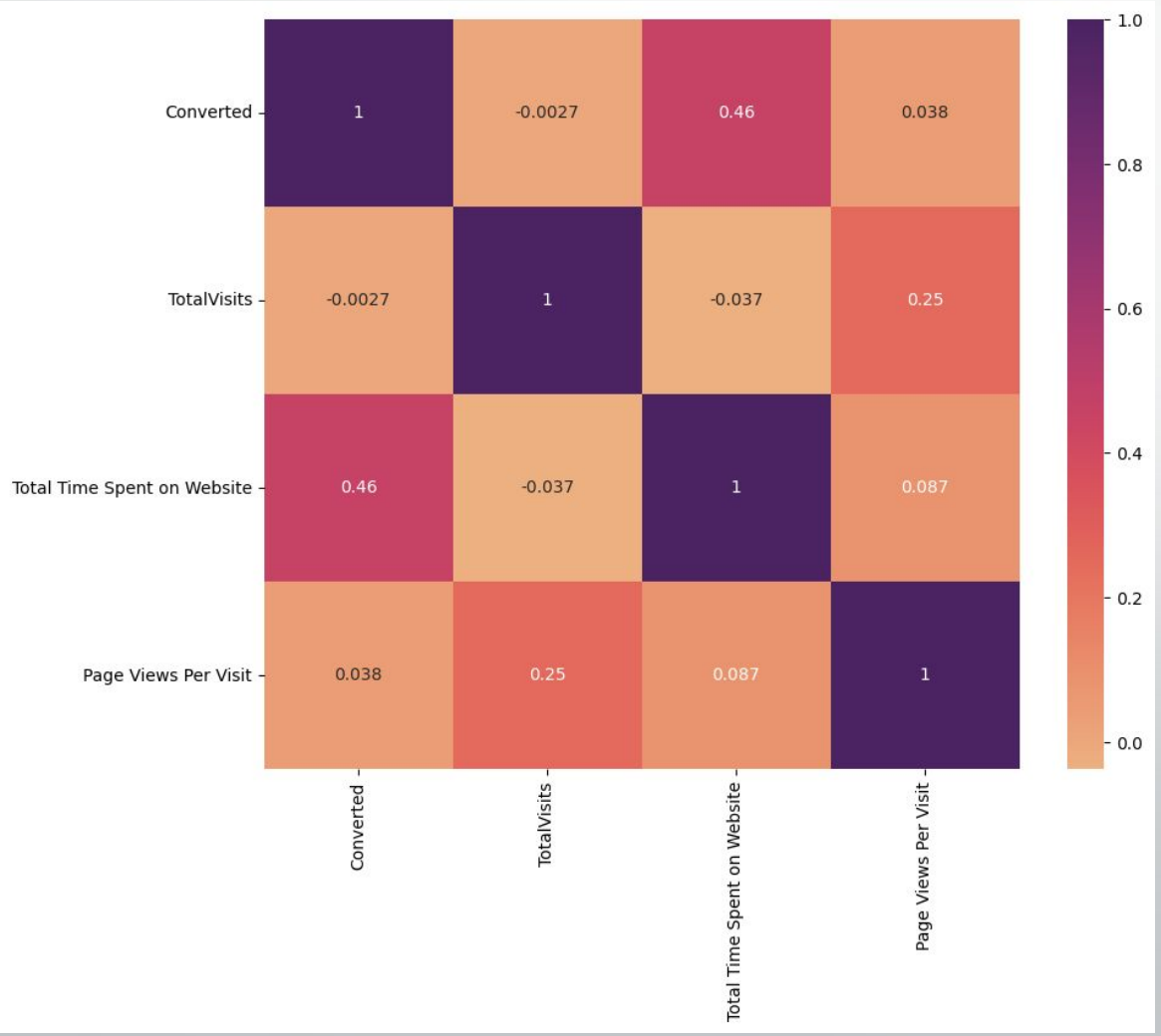
1. The median for converted and unconverted leads is the same.
2. Page Views Per Visit does not appear to be a strong indicator of lead conversion.

TotalVisits, Total Time Spent on Website, Page Views Per Visit



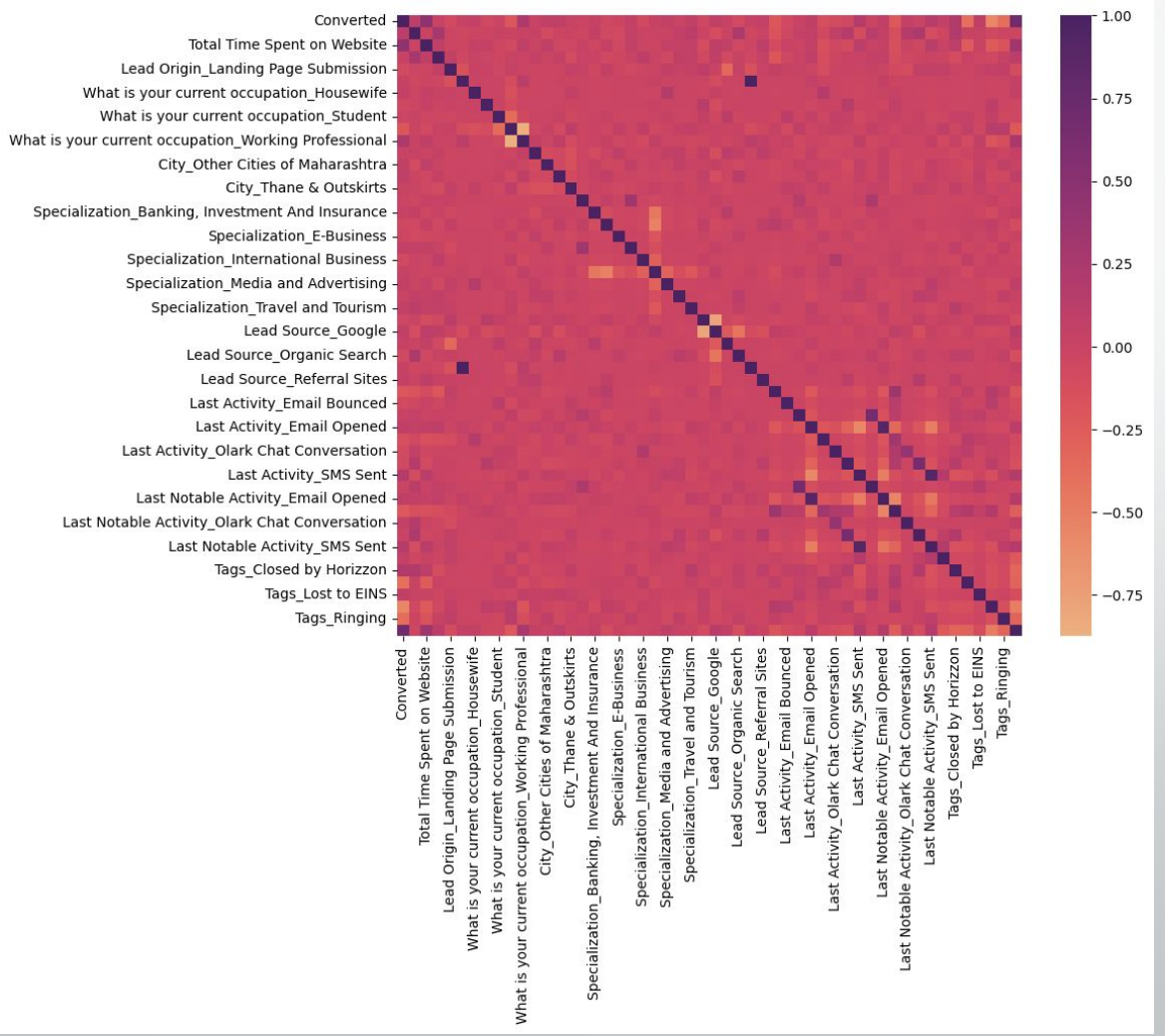
- Leads who **spend more time on the website** tend to have a higher conversion rate.
- **Total visits** do not show a strong correlation with conversion, as both converted and non-converted leads have similar visit patterns.
- **Page views per visit** have a slight influence, but the trend is not very distinct.

TotalVisits, Total Time Spent on Website, Page Views Per Visit



- **Total Time Spent on Website** has the highest positive correlation (0.46) with conversion, indicating that users spending more time are more likely to convert.
- **Total Visits** and **Page Views Per Visit** show very weak correlations with conversion, suggesting they are not strong predictors.
- There is a moderate correlation (0.25) between **Total Visits** and **Page Views Per Visit**, meaning users who visit more tend to explore more pages.

Overall Correlation



- **The heatmap shows correlations between various features, with darker shades indicating stronger relationships.**
- **"Total Time Spent on Website" has a notable positive correlation with "Converted," suggesting higher engagement increases conversion chances. Other categorical variables like "Lead Source" and "Last Activity" exhibit weaker correlations.**

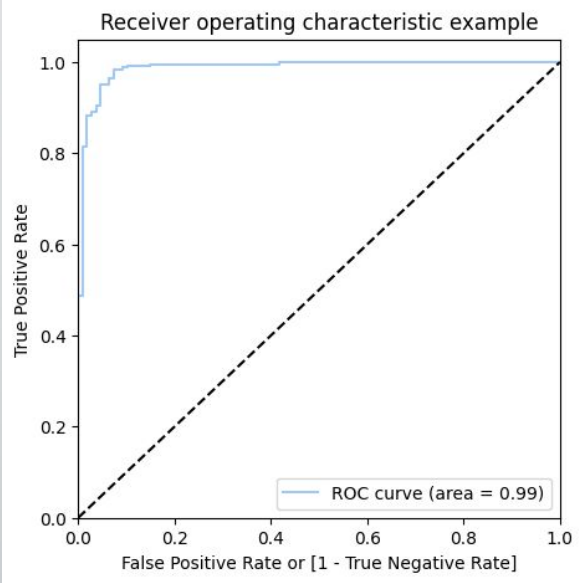
Model Building

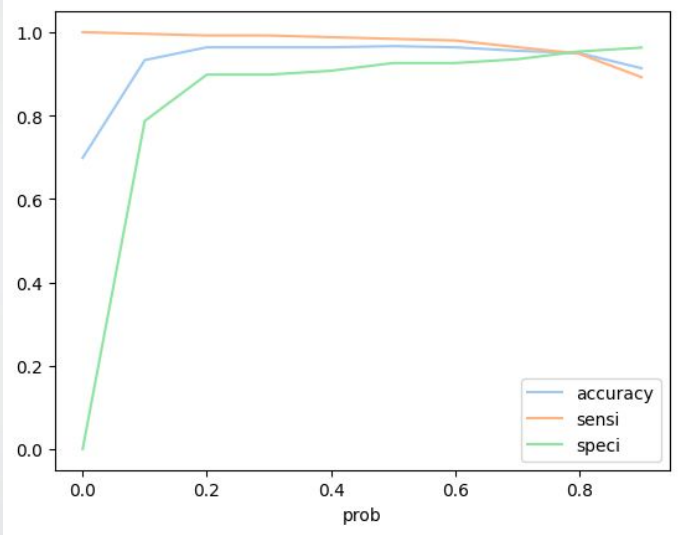
- **Model Building & Validation:**
  - Logistic regression for classification.
  - Recursive Feature Elimination (RFE), R-squared, VIF, and p-values for feature selection.
  - Training-Test Split (70%-30%).
  - Model validation using accuracy, precision, recall, F1-score, and AUC-ROC.
- **Making Predictions:**
  - Assigning lead scores based on logistic regression probabilities.
  - Interpreting the results for business decisions.

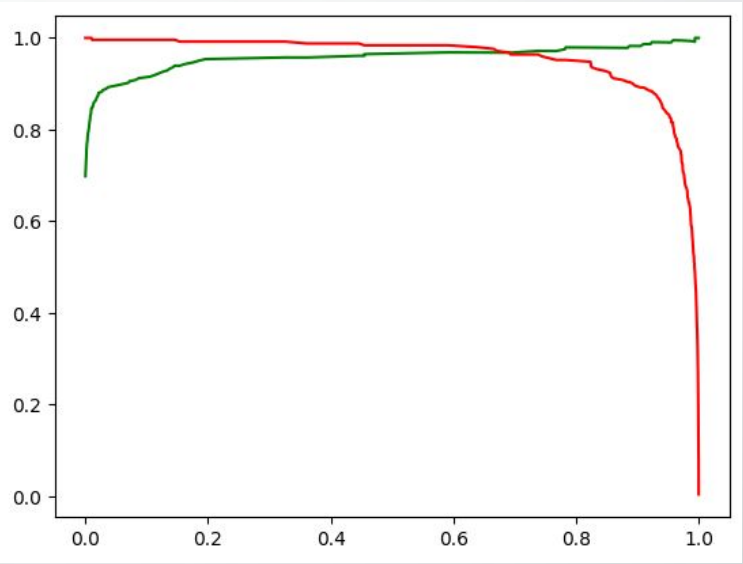
Model Evaluation

- **Model Building & Validation:**
  - Logistic regression for classification.
  - Recursive Feature Elimination (RFE), R-squared, VIF, and p-values for feature selection.
  - Training-Test Split (70%-30%).
  - Model validation using accuracy, precision, recall, F1-score, and AUC-ROC.
- **Making Predictions:**
  - Assigning lead scores based on logistic regression probabilities.
  - Interpreting the results for business decisions.

ROC CURVE







- The ROC curve shows an **AUC of 0.99**, confirming the model's strong ability to differentiate between converted and non-converted leads.
- The **optimal tradeoff between Precision and Recall is at 0.8**, making it a suitable threshold for classification.
- Thus, any **Prospect Lead with a Conversion Probability higher than 80% can be considered a hot Lead**, ensuring a focused approach to lead prioritization.

Observations

So as we can see above the model seems to be performing well.
The ROC curve has a value of 0.99

Train Data

Accuracy : **96.6%**

Sensitivity : **98.2%**

Specificity : **92.6%**

Test Data

Accuracy : **94.9%**

Sensitivity : **94.8%**

Specificity : **95.4%**

Final Feature List

- Total Visits
- Total Time Spent on Website
- Page Views Per Visit
- What is your current occupation_Working Professional
- City_Other Cities of Maharashtra
- Specialization_Management_Specializations
- Last Activity_Email Opened
- Last Notable Activity_Modified
- Tags_Interested in other courses
- Tags_Other_Tags
- Tags_Ringing
- Tags_Will revert after reading the email