Summary of Lead Scoring Analysis for X Education

Objective:
The goal was to find ways to attract more industry professionals to join X Education's courses by analyzing potential customers' behavior on their website and identifying key factors influencing conversion.

Steps and Key Findings:

1. Data Cleaning:
  - Addressed null values and replaced 'Select' options with null as they provided no useful information.
  - Null values were changed to ‘not provided’ to retain data, later removed during dummy variable creation.
  - Categorized geographical data into 'India', 'Outside India', and 'not provided' for clarity.

2. Exploratory Data Analysis (EDA):
  - Performed a quick EDA to assess data quality.
  - Found many irrelevant elements in categorical variables, while numeric data had no significant outliers.
  - Removed the Outliers present for some categorical variables namely total visits ,total time spent on website and page view per visit.

3. Dummy Variables:
  - Created dummy variables for categorical data, excluding those labeled 'not provided'.
  - Standardized numeric variables using MinMaxScaler.

4. Train-Test Split:
  - Split data into 70% training and 30% testing sets.

5. Model Building:
  - Used Recursive Feature Elimination (RFE) to select the top 15 relevant variables.
  - Removed variables with Variance Inflation Factor (VIF) > 5 and p-value > 0.05 to avoid multicollinearity and ensure statistical significance.

6. Model Evaluation:
  - Developed a logistic regression model.
  - Evaluated with a confusion matrix and optimized cut-off value using the ROC curve, achieving approximately 80% for accuracy, sensitivity, and specificity.

7. Prediction:
  - Applied the model to the test dataset with an optimal cut-off value of 0.35, maintaining about 80% accuracy, sensitivity, and specificity.

8. Precision-Recall Analysis:
  - Rechecked the model using precision-recall analysis, determining a cut-off of 0.41, resulting in approximately 73% precision and 75% recall.

9. ROC curve:
  - From the ROC curve we got the cut-off point between 0.3 to 0.35 which comes around 0.33 as cutoff probability.

Key Variables Influencing Conversion (in descending order):
1. Total time spent on the website.
2. Total number of visits.
3. Lead source:
  - Google

  - Direct traffic
  - Organic search
  - Welingak website

4. Last activity:
  - SMS
  - Olark chat conversation
5. Lead origin as Lead add format.
6. Current occupation as a working professional.

Recommendations:
- Enhance Data Collection:
  - Mandate selections for critical fields like Customer Occupation and Specialization to avoid 'Select' entries.

- Improve Website Engagement:
  - Encourage more time spent on the platform and increase the number of visits to enhance conversion rates.

- Targeted Marketing:
  - Focus on leads interested in Better Career Prospects, particularly those with specializations in Finance Management, HR, Finance, and Marketing Management, as they show higher conversion probabilities.

- Boost Customer Engagement:
  - Enhance engagement through emails and calls, as interactions with emails and SMS show higher conversion rates.
  - Increase efforts to engage unemployed leads, as they form a significant portion of potential customers.

By leveraging these insights and strategies, X Education can better target potential customers and improve their conversion rates, ultimately increasing their customer base.