Summary

This analysis focuses on X Education's goal of increasing enrollments from industry professionals into their courses. The analysis utilized various data points on user behavior, site engagement, and conversion rates to develop a predictive model. Below is a detailed overview of the steps taken:

1. **Data Cleaning:**
   - **Initial State:** The dataset was mostly clean but had a few null values. Some categorical variables, such as options that provided little useful information, were replaced with 'null' or 'not provided' to preserve data integrity.
   - **Categorical Refinements:** For geographic data, categories were simplified to 'India', 'Outside India', and 'Not Provided' to improve clarity and usability.
2. **Exploratory Data Analysis (EDA):**
   - **Overview:** A preliminary EDA revealed that some categorical variables contained irrelevant elements, while numeric values were generally in good shape with no outliers detected.
3. **Dummy Variables:**
   - **Creation and Refinement:** Dummy variables were created from categorical data. Categories labeled as 'Not Provided' were removed. Numeric values were scaled using MinMaxScaler to standardize the range.
4. **Train-Test Split:**
   - **Data Partitioning:** The dataset was split into training (70%) and testing (30%) sets to evaluate model performance.
5. **Model Building:**
   - **Feature Selection:** Recursive Feature Elimination (RFE) was employed to identify the top 15 relevant variables. Additional variables were removed based on Variance Inflation Factor (VIF) and p-values, retaining those with VIF < 5 and p-value < 0.05.
6. **Model Evaluation:**
   - **Confusion Matrix and Metrics:** A confusion matrix was used to assess model performance. The optimum cutoff value, determined via the ROC curve, achieved accuracy, sensitivity, and specificity of approximately 80%.
7. **Prediction:**
   - **Performance on Test Data:** Predictions were made on the test set using an optimal cutoff value of 0.35, resulting in accuracy, sensitivity, and specificity of around 80%.
8. **Precision-Recall Analysis:**
   - **Reevaluation:** Precision-Recall analysis suggested a cutoff value of 0.41, with Precision at about 73% and Recall at about 75% for the test data.

Key Variables Influencing Conversion (In Descending Order):

1. **Total Time Spent on the Website:** Higher engagement correlates with a higher likelihood of conversion.
2. **Total Number of Visits:** Frequent visits increase the probability of conversion.
3. **Lead Source:**
   - Google
   - Direct Traffic
   - Organic Search
   - Welingak Website
4. **Last Activity Type:**
   - SMS
   - Olark Chat Conversation
5. **Lead Origin:** Specifically 'Lead add format'.
6. **Current Occupation:** Being a working professional significantly impacts the likelihood of conversion.

By focusing on these key factors, X Education can strategically enhance their approach to engage potential leads more effectively and increase course enrollments.