Summary

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate. It involves developing a logistic regression model to assign lead scores, facilitating targeted customer outreach and addressing future adaptability needs.

Here are the steps employed in the analysis:

- **Data Cleaning and EDA:**
    - During the cleaning process, the 'Select' values in categorical variables were treated as null. Additionally, null values were replaced with 'not provided' to retain data integrity. Location data was standardized to include categories for 'India,' 'Outside India,' and 'not provided' to streamline analysis. The majority of leads are from India, with Mumbai contributing the highest number among cities.

Once we completed the exploratory data analysis (EDA) and data cleaning phase, we proceeded to the model building stage.

- **Dummy Variable :**
    - Dummy variables were generated to represent categorical data, and subsequently, dummies with 'not provided' entries were eliminated. Additionally, numeric values were scaled using MinMaxScaler.
- **Train and Test Split:**
    - The dataset was split into training and testing sets, with 70% allocated for training and 30% for testing.
- **Model Building:**
    - Initially, Recursive Feature Elimination (RFE) was employed to identify the top 15 relevant variables. Subsequently, the remaining variables were manually removed based on their Variance Inflation Factor (VIF) values and p-values, with variables exhibiting VIF < 5 and p-value < 0.05 being retained.
- **Model Evaluation:**
    - A confusion matrix was generated, followed by the determination of the optimal cutoff value using the ROC curve. This process yielded an accuracy, sensitivity, and specificity of approximately 93% each.

After Model Evaluation we further looked at the data and found the top three variables in model which contribute most towards the probability of a lead getting converted was

- Tags
- Total Time Spent on Website
- Lead Source

"Tags_Closed by Horizzon" has the largest coefficient of approximately 7.33, indicating a significant positive contribution to the probability of conversion.