**Summary Report**

In this assignment, we aimed to build a predictive model to improve the lead conversion rate for X Education, which sells online courses to industry professionals. The company faced challenges with low lead conversion rates (around 30%) and sought to identify high-potential leads (hot leads) that would improve their efficiency by focusing their sales team on the most promising prospects.  
  
**1. Exploratory Data Analysis (EDA)**

* A quick check for missing values was conducted, and columns with more than 35% missing values were dropped.
* Since dropping rows with null values would result in significant data loss in important columns, we replaced the NaN values with 'not provided.'
* As "India" was the most common value in the dataset, we imputed all 'not provided' values in the relevant column with "India."
* Upon observing that nearly 70% of the data had "India" as the value, the column was dropped due to redundancy.
* Further work was done on handling numerical variables, addressing outliers, and creating dummy variables.

**Step 2: Understanding the Problem** The goal was to assign a lead score based on historical data so that leads with a higher score would have a greater likelihood of conversion. The dataset provided around 9,000 leads with various attributes such as Lead Source, Lead Origin, Last Activity, and Total Time Spent on the Website. Our target variable was the binary variable "Converted," indicating whether a lead converted into a paying customer.

**Step 3: Data Preprocessing** The dataset contained multiple categorical variables, some of which had a "Select" category that acted as a null value. We began by cleaning the data, including handling null values, removing irrelevant columns, and converting categorical variables into dummy variables for use in a logistic regression model. We also dropped variables that were either not significant (p > 0.05) or had high multicollinearity.

**Step 4: Building the Logistic Regression Model** We built three versions of the logistic regression model to iteratively improve it:

* **Model 1**: The initial model included all the relevant variables after cleaning and encoding. It gave us a good starting point but included insignificant predictors, such as What is your current occupation\_Housewife, which was later dropped.
* **Model 2**: After removing What is your current occupation\_Housewife due to its high p-value, the model’s performance slightly improved. However, further refinement was needed, particularly with variables like Lead Origin\_Lead Add Form, which had an overly strong effect but was statistically less significant.
* **Model 3 (Final Model)**: After removing Lead Origin\_Lead Add Form, the final model was built. This model gave us more interpretable results, with a pseudo R-squared value of 0.4052, indicating that it explained about 40% of the variance in lead conversion.

**Step 5: Model Evaluation** The final model was evaluated using metrics like accuracy, precision, recall, and F1-score on both the training and test datasets. The model achieved an accuracy of 80.94% on the training set and 79.91% on the test set, indicating it generalized well. Sensitivity (recall) and specificity were balanced, with values of around 78-81%, meaning the model was effective at identifying both converted and non-converted leads.

**Step 6: Key Learnings and Insights** From the analysis, several key features stood out as strong predictors of conversion:

1. **Lead Origin - Lead Add Form**: A strong positive predictor, leads coming from this form were more likely to convert.
2. **Lead Source - Olark Chat**: A significant positive effect, indicating leads from this source are promising.
3. **Total Time Spent on Website**: Higher engagement on the website strongly correlated with conversion likelihood.
4. **Phone Conversations and SMS Sent**: Leads with these activities had a higher chance of conversion.
5. **Occupation - Working Professional**: Working professionals were more likely to convert.

Negative predictors, such as Do Not Email and leads from Landing Page Submissions, were less likely to convert, suggesting areas to focus on or reconsider.

**Step 7: Recommendations** To improve lead conversion, X Education should focus on driving leads from sources like Lead Add Form and Welingak Website, target working professionals, and prioritize SMS communication and phone conversations. Additionally, improving website engagement will further increase the conversion rate.

This assignment helped us gain insights into the importance of data preprocessing, variable selection, and understanding the impact of key features on conversion, along with interpreting logistic regression models to drive actionable business decisions.