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 that would improve their efficiency by focusing their sales team on the most promising prospects.

Step 1: Reading and understanding the data

Read and store the data from the CSV file into a data frame (leads_df) and perform basic analysis.

Step 2: Data preparation and cleaning

- Drop columns with more than 35% missing values.
- Handling few of the columns that have 'select' value in them
- Dropped columns that were heavily skewed.
- Dropped rows that had missing data and since they made up for a very small percentage of the total data (1.48%)

Step 3: Exploratory Data Analysis (EDA)

- Data imbalance checked - only 39.09% leads converted.
- Performed univariate and bivariate analysis for categorical and numerical variables. ‘Lead Origin’, ‘Current occupation’, ‘Lead Source’, etc. provide valuable insight on effect on target variable.
- Time spend on website shows positive impact on lead conversion.
- Further work was done on handling numerical variables and addressing the outliers.

Step 4: Data Preparation

- Converted some binary variables (Yes/No) to 0/1.
- Created dummy variables for categorical variables with multiple levels.

Step 5: Test – Train split

Split the data into train and test data with 70:30 split ratio. We also scaled the data using standardscaler and applied 'fit_transform' to scale the 'train' data set.

Step 6: Building the Logistic Regression Model

We used RFE for feature elimination and proceeded with 15 variables for our model.

We built three versions of the logistic regression model to iteratively improve it by checking the p-value which had to be lesser than 0.05 and the VIF score which had to be lesser than 5. After eliminating certain columns due to high p-values, the above conditions were attained

Step 7: 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 and specificity were balanced, with values of around 78-81%, meaning the model was effective at identifying both converted and non-converted leads.

Step 8: Key Learnings and Insights From the analysis, several key features stood out as strong predictors of conversion:

- **Lead Origin - Lead Add Form**: A strong positive predictor, leads coming from this form were more likely to convert.
- **Lead Source - Olark Chat**: A significant positive effect, indicating leads from this source are promising.
- **Total Time Spent on Website**: Higher engagement on the website strongly correlated with conversion likelihood.
- **Phone Conversations and SMS Sent**: Leads with these activities had a higher chance of conversion.
- **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 9: 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.