Summary of the Case Study

Below 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 their conversion rate.

Steps followed for the solutions:

- **Cleaning data:** The data received was not clean, it had value called Select, which was default for any dropdown field. It had to be replaced with a null value since it did not add any valuable information to data. All column datatype was checked to make sure there was no mismatch between datatype & value available. The columns have more than 70% Null values were dropped from the table. We analysed the data inside the column to suitable select if value imputation needed to be Mean / Median or mode of it. If null percentage for any columns after imputations is more than 2% , we drop them then.

- **EDA:** A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant, hence dropped. The numeric values had few outliers, which were handled using percentile capping at 25% & 75%. Heatmap were plotted to understand which if columns were highly correlated. Barplots helped us to understand how columns values were distributed in graph with Converted value. Countplots helped us to visualize the distribution along with some of the violinplots were used for the outlier treatment.

- **Data Preparation:** The data was prepared to do the further analysis to give us insights.

- **Dummy Variables**: Dummy variables were created for categorical variables & first values are dropped. Then it ‘as added to main data with original categorical variables dropped then.

- **Feature Scaling:** For numerical values, we used StandardScaler standardized the features by subtracting the mean and then scaling to unit variance.

- **Train-Test split:** The split was done at 70% and 30% for train and test data respectively.

- **Model Building:** RFE was done to attain the top 18 relevant variables. Logistic regression model was prepared with available features. Few variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.001 were kept) & model were reiterated till required VIF & p-values were met.

- **Prediction:** Prediction was done on the test data frame and with an optimum cut off as 0.34 with Accuracy, Sensitivity and Specificity of more than 80%.

- **Precision, Recall & F1 Score:** This method was also used to recheck and a cut off of 0.34 was found with Precision around 79%, recall around 66% & F1 Score around 0.74 on the test data frame.

- **Lead Score:** Calculated the lead score using formula Converted_Prob * 100 , As per requirement made, a table was created & sorted descending on Lead Score where Prospect ID, Lead Number & Lead Score were available.