Summary Report

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basis 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.

The following are the steps used:

1. **Cleaning Data:**

The data was partially clean except for a few null values and the option 'Select' had to be replaced with a null value since it did not give us much information. Columns with more than 40% null values were dropped. We have accordingly imputed values for columns having Null Values less than 40%. We have also dropped some columns due to data imbalance.

2. **EDA:**

A quick EDA was done to check the condition of our data. We have performed univariate and bivariate analysis on the continuous and categorical variables. It was concluded that in numerical columns the distribution of data was not normal. Also, we have treated the outliers.

3. **Dummy Variables:**

The dummy variables were created for the categorical columns. For numeric values, we used StandardScaler function.

4. **Train-Test Split:**

The split was done at 70% and 30% for train and test data respectively.

5. **Model Building:**

We used logistic regression model as it was a classification problem. Firstly, RFE was done to attain the most relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF <= 5 and p-value < =0.05 were kept). We created ROC Curve to show the tradeoff between sensitivity and specificity. Then we created a line graph to find the Optimal cutoff probability where we get balanced sensitivity and specificity. We also find the tradeoff between precision and recall. Finally, we made the predictions on test data. We created the confusion matrix for both train and test data also, calculated the matrix like Accuracy, Sensitivity, Specificity, Precision, Recall, F1 Score, etc.

6. **Conclusion:**

- People spending more than average time are promising leads, so targeting and approaching them could be helpful in conversions.
- SMS messages can have a high impact on lead conversions.
- Landing page submissions can help find out more leads.
- Marketing management, human resources management has high conversion rates. People from those specializations could be promising leads.
- References and offers for referring a lead can be a good source for high conversions.
- An alert messages or information has seen to have high lead conversion rate.
- The threshold has been selected from Accuracy, Sensitivity, Specificity measures and precision, recall curves.
- The train and test data has an approx accuracy of 88% which concludes that Logistic Regression model is a good fit.
- The model shows high accuracy which is 88% and a recall rate of 83% with precision around 87%.
- This implies the model is around 83% good in predicting the positive class (Hot Leads).
- Overall, this model proves to be accurate.