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 customer visit the site, the time they spend there, how they reached the site and the conversion rate.

Below are steps used for analysis:

1. **Problem statement:** A company called X-Education provides professional online education courses and internet marketing through adverts. The company uses a variety of ways to obtain information, and it calls leads who inquire about a particular degree of schooling. Lead conversion is usually 30% of specific education. The company also uses specific criteria to identify Hot Leads. The ratio of leads converted to enrolments is lower. company provided Aim for 80% of the total enrolled.

2. **Goal:** Constructing a logistics regression model to help the company locate leads and reach its goals. backup plan or Alternative approach should be prepared in case the company's needs change in the future and should be adaptable.

3. **Reading and understanding the data**: read the data from the source covert data into a clean format appropriate for analysis by eliminating duplicate data and outlier treatment.

4. **EDA**: EDA was done to check the condition of out data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good.

5. **Data Preparation for training:** splitting data into training and testing set

6. **Model Building:** RFE-based feature selection Utilising logistic regression and calculating the lead scores

7. **Evaluate model:** Ascertain the best model by computing a number of measures, including accuracy, sensitivity, specificity, precision, and recall, then assess the model.

   Breif:
   We finally identified the fifteen most important variables. It was also discovered that the initial VIFs for these variables were good.
   We then constructed the data frame with the converted probability values, starting from the presumption that a probability value greater than 0.5 indicates a positive outcome or a negative outcome.

Based on the aforementioned supposition, we computed the confusion metrics and the model's overall accuracy. To determine the model's dependability, we also computed the sensitivity and specificity mat

The model was further validated when we attempted to draw the ROC curve for the features. The curve showed a respectable area coverage of 88%.

Next, we produced the probability graph for various probability values for the variables "Accuracy," "Sensitivity," and "Specificity." The graphs' intersection was thought to be the ideal probability cutoff point, and it turned out to be 0.35.

We could see from the new number that the model had almost 80% of the values correctly predicted.

The updated accuracy, sensitivity, and specificity scores were also visible to us: 80%, 79%, and 80%, respectively.

Additionally, the lead score was computed, and it was found that the goal lead prediction of 80% was given by the final predicted variables.

The accuracy value was discovered to be 80% sensitivity=79% specificity 81% after we applied the learning to the test model and conversion probability based on the sensitivity and specificity measurements.