Summary

In this case study we developed a binary logistic regression model for an educational course selling firm called X Education which requires us to find "Hot Leads" out of other leads (industry professionals) who are more likely to purchase a professional course and improve their overall conversion rate i.e. currently at 30% to around 80%. The dataset was provided containing various information regarding the leads like their country, city, current occupation, etc.

1. **Data Preparation (Missing data):**
   The initial data imported had 37 columns within which many had missing values. The value "Select" was present in many categorical columns which was supposed to be treated as null values too. Features with more than 40% missing values were dropped meanwhile other values were imputed using Mean and Mode.

2. **Exploratory Data Analysis**:
   Relations between the binary target variable i.e. 'Converted' and remaining independent variables were visualized in order to understand more about the various crucial predictors and irrelevancy of others. There were more categorical columns compared to numerical ones that had few outliers.

3. **Dummy Variables:**
   Once all the missing values were dealt with and binary "Yes" "No" columns were converted to numerical types the Dummy variables were created with the 'get_dummies' function.

4. **Train-Test Split:**
   The dataset was split 70% for the training set and 30% for the testing dataset following the scaling of the numerical variables.

5. **Model Development:**
   Before beginning to make and train the model we implemented Recursive Feature Elimination (RFE) to retrieve the top 15 features from the preprocessed training dataset. After that through Variance Inflation Factor (VIF) and P-values we eliminated few more features to improve the model's overall performance.

6. **Model Evaluation:**
   After the logistic regression model was done training the confusion matrix was made determining its Sensitivity, Specificity, Recall, Precision along with the ROC curve. The final model accuracy on the training dataset was around 85%.

7. **Prediction:**
   Finally, the trained and evaluated model was used on the testing dataset to make predictions and determine which leads would be converted into permanent customers of the company. An accuracy of 84% was recorded with 67% sensitivity and 96% specificity.

8. Conclusion:

We found out that the most important factors which could tell whether a lead would be converted were as follows: -

a. Tags
b. Last Notable Activity
c. When occupation was a working professional
d. Lead Origin from Lead Add form

Keeping in mind all this X Education can improve their profit margins significantly while not wasting a lot of resources and efforts on unnecessary, redundant practices which would not really cause a lead to convert in the future.