Lead Scoring Case Study Summary

In the case study the data of an education company named X Education sells online courses to industry professionals is provided. The study is asked to build a model having an accuracy of 80%. The dataset contains 9241 rows and 37 columns.
We need to build a model which tells about the driving factors contributing to conversion probability and works well on both the test and train data sets.

- **Reading and Understanding Data**.
  Read and analyze the data.
- **Data Cleaning**
  We dropped the variables that had high percentage of NULL values in them. Most of the categorical columns were imputed by mode. Those columns with unique and skewed values were dropped.
- **Data Analysis**
  After cleaning the data we need to analyze each columns (both numeric and categorical variables) and makes useful insights using different visualization tools.
- **Creating Dummy Variables**
  we went on with creating dummy data for the categorical variables.
- **Test Train Split**
  The next step was to divide the data set into test and train sections with a proportion of 70-30% values.
- **Feature Rescaling**
  We used the Min Max Scaling to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.
- **Feature selection using RFE**:
  Using the Recursive Feature Elimination we went ahead and

selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.

- **Plotting the ROC Curve**
  We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 85% which further solidified the of the model.
- **Building the model and evaluation**
  Model is build using the important variables and evaluated based on the pvalues and VIFs
- **Rebuilding the model**
  Rebuild the model again and again until a stable one obtained
- **Predicting the probability**
  Predict the probability of the target variable using the model.
- **Evaluation of metrices**
  Build the confusion matrix and check for the sensitivity, specificity and accuracy for the model
- **Plotting the ROC and find the optimal cut off**
- **Assigning the lead score to each customer**
- **Test the model on test data set and evaluation of the confusion Metrix.**
- **Checking for the coherence in performance of the model in test as well as on train dataset.**

Analysis:

It is found that the variables which contributes to the conversion of the leads are:

- **Total time spent on website.**
- **Lead Origin.**
  - **Lead Add Form**
- **What is your current occupation.**
  - **Working Professional**

The performance of the model in test and train set is given by:

On train dataset:
accuracy: 79.34446505875077
specificity: 80.4847576211894
sensitivity: 77.65612327656123
precision: 77.72325809617271

On test dataset:
accuracy: 78.96825396825396
specificity: 79.90459153249851
sensitivity: 77.53424657534246
precision: 71.58516020236088

From the evaluation it is came to conclude that the model is reasonably good.