Summary of Lead scoring case study

This analysis is done for "X" online Education Company which sells online courses to industry professionals. The provided dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which tells us how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate. Our goal is to identify the ways to increase the conversion rate using logistic regression. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%. We need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Below are the steps performed for this analysis and model building:

1. **Cleaning data:**
   The data was partially clean except for a few null/missing values. The option 'select' had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

2. **EDA:**
   A quick EDA was done to check the summary of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

3. **Dummy Variables:**
   The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler to normalize the data.

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

5. **Model Building:**
   a) Firstly, RFE was done to attain the top 15 relevant variables. (The goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through any specific attribute or callable. Then, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached)

   b) 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). The Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables. It is calculated by taking the the ratio of the variance of all a given model's betas divide by the variane of a single

beta if it were fit alone.

6. **Model Evaluation:**
   A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.
   ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds.
   Accuracy is the proportion of correct predictions over the total number of predictions. Accuracy = (TP + TN) / All Predictions
   Sensitivity (aka Recall) means "out of all actual Positives, how many did we predict as Positive", which can be explained as: Sensitivity (Recall) = TP / (FN + TP)
   Specificity (aka Selectivity or True Negative Rate, TNR) means "out of all actual Negatives, how many did we predict as Negative", and can be written as: Specificity = TN / (TN + FP)
   Precision (aka Positive Predictive Value, PPV) means "out of all predicted Positive cases, how many were actually Positive", or Precision = TP / (TP + FP)

7. **Prediction:**
   Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

8. **Precision – Recall:**
   This method was also used to recheck and a cut off of 0.41 was found with Precision around 73% and recall around 75% on the test data frame

   Conclusion: It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
   a. Google
   b. Direct traffic
   c. Organic search
   d. Welingak website
4. When the last activity was:
   a. SMS
   b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.