Summary report (assignment solution and learnings)

The basic data provided gave us a lot of information about how the potential customers visit the site, how long they spend there, how they found the site, and the conversion rate. The following steps were used in this analysis, which is being done for X Education and to determine how to get more industry professionals to enroll in their courses:

1. **Data processing:**

The data was not completely clean, some choice based columns had default options selected which were replaced by nan. There were some columns in particular that had huge % of null values but instead of dropping them, we utilised it by making a new category for null values. For the country column, 'na' was imputed as 'unknown' category and other than the top countries, others were labelled as 'other.'

2. **EDA:**

   For univariate analysis, countplot, histograms and box-plot were used. Multivariate analysis was done by means of a heatmap to study correlations among variables.

3. **Dummy Encoding:**

Dummy data-frame was created by dropping the first category and encoding others using the get_dummies function. This df was concatenated with the original df and then certain columns were dropped.

4. **Splitting data into Train-test datasets:**

Sci-kit learn module was used to split train-test data as 70% and 30%. The MinMax scaler was used to scale the numeric values.

5. **Building the Model :**

Logistic regression from Scikit learn was used to build the model. It was fitted on training data and RFE was used.

6. **Feature selection:**

Due to many variables, RFE was used to get top relevant ones after which statsmodels library was used to get summary and p-values along with calculating VIF to manually select features.

7. **Evaluation of model, Optimising cutoff:**

After initial predictions, the metrics module was used to calculate confusion matrix and accuracy on training data and it was found that specificity was around 94% and sensitivity around 90%.

8. **Making predictions :**

An optimal cut off was found using ROC function further improving the accuracy and precision and recall. These improved the metrics further by 2-5 %.

9. **Result (metrics):**

Upon making predictions on the test set, the model's accuracy was close to 93% and precision recall were near about 90%.

Conclusion and learnings:

The key factors influencing potential buyers, ranked from most to least important, are:

1.The total duration spent on the website.

2.The overall number of visits.

3.The lead source, particularly when it comes from:

- Google
- Direct traffic
- Organic search
- Welingak website

4.The last recorded activity, specifically:

- SMS interactions
- Olark chat conversations

5.The lead origin being from a Lead Ad format.

6.The individual's current occupation being a working professional.

Considering these factors, X Education has a strong opportunity to convert nearly all potential buyers into actual customers, significantly boosting their course enrollments.