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

The following are the steps used:

1. **Cleaning data:**
   The data was partially clean containing no duplicates except for a few null values .As it can be seen, there are select values in many columns. This means that the person did not select any option for the given field. Hence, these were takes as NULL values. Few of the null values were changed to 'Not Sure' so as to not lose much data. Although theywere 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 condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. For 'TotalVisits', the 95% quantile is 10 whereas the maximum value is 251. Hence, we should cap these outliers at 95% value. There are no significant outliers in 'Total Time Spent on Website' . For 'Page Views Per Visit', similar to 'TotalVisits', we should cap outliers at 95% value.
  We don't need to cap at 5% as the minimum value at 5% value are same for all the variables.

3. **Dummy Variables:**
   The dummy variables were created and later on the dummies .For categorical variables with multiple levels, we create dummy features (one-hot encoded).

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

5. **Model Building:**
   Firstly, RFE was done to attain the top 15 relevant variables. 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).

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. Area under curve (auc) is approximately 0.95 which is very close to ideal auc of 1.

7. **Prediction:**
   Prediction was done on the test data frame and with an optimum cut off as 0.2 with accuracy, sensitivity and specificity of 80%.

8. After trying out saveral models, our final model has following characteristics:
   All p-values are very close to zero.VIFs for all features are very low. There is hardly any multicollinearity present. The overall testing accuracy of 90.78% at a probability threshold of

0.05 is also very good. The **optimal threshold** for the model is **0.20** which is calculated based on tradeoff between sensitivity, specificity and accuracy. According to business needs, this threshold can be changed to increase or decrease a specific metric.

9. Twelve features were selected as the most significant in predicting the conversion:
Features having positive impact on conversion probability in decreasing order of impact:
Tags_Lost to EINS|
|Tags_Closed by Horizzon|
|Tags_Will revert after reading the email|
|Tags_Busy|
|Lead Source_Welingak Website|
|Last Notable Activity_SMS Sent|
|Lead Origin_Lead Add Form|