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. **EDA:**
    - Dropping ‘lead number’ and ‘prospect ID’ because they have unique values.
    - Quick check was done on % of null value and we dropped columns with more than 45% missing values.
    - We also saw that the rows with the null value would cost us a lot of data and they were important columns. So, we replaced the NaN values with ‘Not Specified’.
    - Since India was the most common occurrence among the non-missing values, we imputed all not provided values with India.
    - Then we saw the Number of Values for India were quite high (nearly 97% of the Data), so country column was dropped.
    - We also worked on numerical variable, outliers and dummy variables.
    - Handle the outliers for using outlier technique such as removing top & bottom 1% of the Column Outlier values.

2. **Train-Test split & Scaling :**
    - The split was done at 80% and 20% for train and test data respectively.
    - We will do min-max scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

3. **Model Building**
    - RFE was used for feature selection.
    - Then RFE was done to attain the top 15 relevant variables.
    - A confusion matrix was created, and overall accuracy was checked which came out to be 91.68%.

4. **Model Evaluation**

Confusion Matrix-
True positive= confusion[1,1]=582
True Negative= confusion[0,0]=1060
False positive= confusion[0,1]=61
False Negative= confusion[1,0]=88
582+1061=1643 cases, the predicted values matched with the actual values.
61+88=149 cases, the predicted values did not matched with the actual values.
582+1060+61+88=1791 total population

- **Sensitivity – Specificity**

  If we go with Sensitivity- Specificity Evaluation. We will get :

  Accuracy 91.68%
  Sensitivity 94.55%
  Specificity 86.86%

- **Precision – Recall:**

  If we go with Precision – Recall Evaluation

  Accuracy 91.68%
  Precision 91%
  Recall 91%

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.

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.