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 except for a few null values and the option 'Select'. We ensured that the columns that does not provide gets removed so that the machine can learn only the useful insights and not unnecessary data.

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. The numeric values seems good and no outliers were found.

3. Dummy Variables: The dummy variables were created. For numeric values we used the MinMaxScaler.

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, the optimum cut off value (using ROC curve) was found via the following metrics i.e. Accuracy, Sensitivity-Specificity View & Precision-Recall View.

7. Prediction: Prediction was done on the test data set with the optimum cut-off values for Sensitivity-Specificity view & Precision-Recall View and below is the comparison:

Sensitivity-Specificity View

Accuracy

- Training Set - 79.08
- Test Set - 78.45

Sensitivity

- Training Set - 79.33
- Test Set - 77.95

Specificity

- Training Set - 78.84
- Test Set - 78.91

Precision-Recall View

Accuracy

- Training Set - 78.95
- Test Set - 78.66

Precision

- Training Set - 78.40
- Test Set - 78.28

Recall

- Training Set - 77.71
- Test Set - 76.75

Conclusion

Based on the model that is created and its performance, it is concluded that the variables that mattered the most in the potential buyers are as below:

- Total number of visits
- Total time spend by the customer on the Website.
- When the lead origin is Lead add format
- When the lead source was:
    - Olark Chat
    - Welingak Website
- When the last activity was:
    - Phone Conversation
- When the current occupation stands out as Student

Keeping all the above points into consideration X Education can convert most of the leads and convince the potential buyers to buy their courses.