Summary

Sayed Raheel and Debanjan Biswas

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. This 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, their occupation, last activity and many more with conversion rate.
The following approach have been used:

1. **Reading and understanding data**
   Basic data shape and size, its columns and its data types.

2. **Data Cleaning**
   Data had missing values and some columns had more than 50% of their values as missing values so we removed them. Other column have around 30%/40% of missing values so we first categorize the low performing categories of columns and imputed values in missing values proportionally so the data and the future model doesn't get skewed. Example- column 'What is your current occupation' we categorize low performing categories and got three major categories and we imputed values according to that proportions.
   We also checked unique identifiers for repeated values and later dropped them.

3. **EDA**
   Analysis was done to check the data, we did data imbalance check initially and found that lot of columns have major data imbalance, we also found that some columns have only one category so we dropped them.
   We got a lot of insights from data when we performed uni/bi-variate analysis on numerical and categorical variables. It was observed where most of the leads are coming from and where most leads are getting converted.
   We performed outlier treatment and capped the variables which had outliers in them so that model doesn't get skewed in the future. Also looked at correlations

4. **Data Preperation**
   We started preparing the data for model making. We made dummy variables.

   1. Dummy variables
   We made dummy variables of column that had only two categories with 0 and 1 and then we went to get dummy variables for all the categorical columns. Most of them irrelevant columns will be removed during model building

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

   3. Feature scaling
   The scaling was done on numerical columns via StandardScaler.

Summary

Sayed Raheel and Debanjan Biswas

5. **Model building**
   Initially RFE (Recursive Feature Elimination) was done to attain top 15 relevant variables. Later the rest of variables were removed manually depending on the P-value and VIF (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 which came to be around 80% each.

7. **Prediction**
   Prediction was done on the test data frame which was transformed by scaling with an optimum cut off as 0.33, with accuracy, sensitivity and specificity of 83%, 85% an 80%

8. **Precision and Recall**
   With same cutoff of 0.33 Precision- 74% Recall- 86%

Conclusion:

It was found that the variables that mattered the most in the potential buyers are(in descending order):
1.Lead from Origin_Lead Add Form ,Welingak Website and Lead Source_Olark Chat

2. Tags_Will revert after reading the email

3. Last Notable Activity_SMS Sent, of leads who's Activity_Unreachable, Converted to Lead, Olark Chat Conversation and Email Bounced

4. Total Time Spent on Website
5. Their current occupation_Working Professional
6 . Tags_Ringing- Their current status call not picked

Keeping these in mind the X Education can increase their lead conversion if they follow all variables our model gave and contact leads based on this prerequisite displayed in descending order. Also X education can tweak model sensitivity and specificity according to their strategy, manpower, time and budget contact those accordingly via Lead Scoring