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:

- **Import Libraries and Data :**

  Imported the required libraries and Dataset to work on.

- **Understanding the Data:**

  We referred to the Data Dictionary to know the details of the variables. The functions like info,Describe, and shape helped us to understand the no of rows n columns and percentiles and mean, data type, etc details.

- **Cleaning data:**

  1. We converted the data type of variables where needed.
  2. The data had a few columns with more than 45% null value, we checked if the columns are significant in terms of data analysis. We didn't find the same hence we removed it.
  3. We checked if there are any duplicate entries. There were not otherwise we would have removed it.
  4. Converted the categorical variable values to binary.
  5. Checked if there are any rows where the Target variable is missing.
  6. Checked if there are any rows with more than 50% missing values.
  7. Few variables had 'Select' mentioned instead of any value. This may be because the user forgot to select a specific value. It's as good as missing data so converted Select to NaN.
  8. Dropped the columns with unique values as they won't have any effect on analysis.
  9. Then again checked the missing value percentage. Dropped the columns with more than 70% missing data after EDA. We selected all categorical values and found the value counts. We imputed Categorical variables' missing values with Mode.
  10. Checked if data is imbalanced which was not. So proceeded with the analysis.
  11. We then checked the relation between the Categorical variable and the Converted Variable which is our Target Variable.
  12. We checked if there are any incorrect labels for categorical variables and corrected the labels.

- **Exploratory Data Analysis:**

  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 had outliers. Outliers were treated. After EDA the Variables we had found earlier having more than70% missing data were removed. We also removed the variables which were not suitable for further analysis.

- **Retained Data and its Analysis**

96.23%of the leads provided by the company have been used for analysis. We analyzed the variables in Re tained data & tried to find the conversion rate. We found that the Lead Add Form from Lead Origin (93%) Reference From Lead Source (92%), Welingak Website From Lead Source (98%), and Working Professional from current Occupation (91%) have a high impact on conversion ratio.

Variables 'Last Notable Activity' & 'Last Activity' seem to have similar descriptions. From the data, Last Not able Activity seems like a column derived by the sales team using Last Activity. Since this insight might not be available for a new lead, we looked at the possibility of dropping Last Notable Activity. However, when we tried dropping the Last Notable Activity along with its dummies, the model was overfitting. Hence we d idn't remove it. Though they seem to be similar there are few categories in each variable that are different from categories in another variable. e.g. Modified, Form Submitted on Website, View in browser link Click ed. So we let the two variables be as is.

- **Dummy Variable creation and Feature Scaling.**

The dummy variables were created. For numeric values, we used the MinMaxScaler.

- **Train-Test split:**

The split was done at 70% and 30% for train and test data respectively

- **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).

- **Model Evaluation:**

A confusion matrix was made. Later on, the optimum cut off value (using the ROC curve) was used to find the accuracy, sensitivity, and specificity which came to be around % each

- **Prediction:**

Prediction was done on the test data frame and with an optimum cut of 0.36 with accuracy, sensitivity, and specificity of %.

- **Precision-Recall:**

This method was also used to recheck and a cut off of 0.41 was found with Precision around % and recall around % on the test data frame

- **It was found that the variables that mattered the most to the potential buyers are:-**

1. Lead Source_Welingak Website
2. Last Activity_Email Bounced
3. Last Activity_Had a Phone Conversation
4. Last Activity_SMS Sent
5. Lead Origin_Landing Page Submission
6. Last Notable Activity_Modified
7. Last Notable Activity_Olark Chat Conversation
8. Last Notable Activity_Unreachable

9. What is your current occupation Working Professional?
10. Specialization_Other_Specialization
11. Total Time Spent on Website
12. Page Views Per Visit

Conclusion:*

A logistic regression model is created using lead features. To arrive at the list of features that significantly affect conversion probability, a mixed feature elimination approach is followed. Most important features are obtained through Recursive Feature Elimination and then reduced to 15 via the p-value / VIF approach. The dataset is randomly divided into train and test set. (70 - 30 split).

Then a cutoff of the probability is used to obtain the predicted value of the target variable.Here, the logistic regression model is used to predict the probability of conversion of a customer.

The optimum cut off is chosen to be 0.34 i.e. any lead with greater than 0.34 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.34 or less probability of converting is predicted as Cold Lead (customer will not convert)

For final Logistic Regression Model, we selected below 15 features.

1. Lead Source_Olark Chat
2. Lead Source_Welingak Website
3. Last Activity_Email Bounced
4. Last Activity_Had a Phone Conversation
5. Last Activity_Olark Chat Conversation
6. Last Activity_SMS Sent
7. Lead Origin_Landing Page Submission
8. Lead Origin_Lead Add Form
9. Last Notable Activity_Other Activity
10. What is your current occupation_Student
11. What is your current occupation_Unemployed
12. What is your current occupation_Working Professional
13. Specialization_Other_Specialization
14. TotalVisits
15. Total Time Spent on Website

The top three categorical/dummy variables in the final model concerning the absolute value of their coefficient factors.

1. Welingak Website From Lead Source (98% conversion rate)
2. Lead Add Form from Lead Origin (93% conversion rate)
3. From Lead Source (92% conversion rate)

Our final Logistic Regression Model is built with 12 features.

Features used in the final model are

1. Lead Source_Welingak Website

2. Last Activity Email Bounced
3. Last Activity Had a Phone Conversation
4. Last Activity SMS Sent
5. Lead Origin Landing Page Submission
6. Last Notable Activity Modified
7. Last Notable Activity Olark Chat Conversation
8. Last Notable Activity Unreachable
9. What is your current occupation Working Professional
10. Specialization_Other_Specialization
11. Total Time Spent on Website
12. Page Views Per Visit

The model seems to be performing well. The ROC curve has a value of 0.87, which is good.

We have the following values for

Train Data:

- Accuracy : 79.86 %
- Sensitivity : 65.46%
- Specificity : 88.75%

Test Data:

- Accuracy : 81.44%
- Sensitivity : 81.55%
- Specificity : 81.38%

The model seems to be performing well. Can be recommend this model in making good calls based on this model.