SUMMARY

The analysis is done for X Education to find ways to get more industry professionals to join their online courses. For this analysis we were provided with the potential data of the customers like how many times they visited the site, how they reached to them ,what is their specialization , country, city living and many more. In the data we were given a column which says that people are converted or not.

The following steps are taken in the analysis:

1.Data Cleaning :

The data given has null values ,columns with appropriate data types. Here null values in column are replaced with appropriate values and select option in the column are to be replaced with the values that are appropriate i.e mean. After checking the null values percentage we can drop the column with higher null value percentage (>40%). The remaining null values are replaced with appropriate values and few rows with null values are dropped.

2.Exploratory Data Analysis (EDA) :

A quick EDA was done to check the condition of the data. The data contains more categorical variables and few numeric variables that are has no outliers.

3.Data Preparation :

To build a logistic regression model on this dataset ,first we need to create the dummy variables for categorical variables. Scaling the dataset using standard scalar for numeric variables. Creating the X ,y variables with feature variables, target variable respectively. And then split the dataset to train dataset (70%) and test dataset (30%).

4.Model Building :

Build the model by RFE feature selection. Then variables which are having high p values and high vif values are removed manually. The columns which low p value and low vif value are used to build the model. So ,finally 12 variables are used to build the model.

Confusion matrix, accuracy, sensitivity, specificity, precision and recall score for the model with optimal cut off of 0.34 using ROC curve.

5.Model Evaluation :

The model is test on the test dataset. Predicted the customers who can be targeted to increase the lead conversion rate. Confusion matrix, accuracy, sensitivity, specificity, precision and recall score for the model with optimal cut off of 0.34 using ROC curve.

The variables that are used in model building are:

- Do Not Email
- Total Time Spent on Website
- Lead Origin_Landing Page submission
- Lead Source_Olark chat
- Lead Source_Reference
- Lead Source_Welingak Website
- Lead Activity_Olark Chat Conversion
- Lead Activity_Other Activity
- Lead Activity_SMS Sent
- Specialization_Others
- What is your current occupation_working Professional
- Last Notable Activity_Modified

Observations :

Train data:

Accuracy : 81.1%

Sensitivity : 81.8%

Specificity : 80.5%

Precision : 79.3%

Recall : 81.8%

Test data:

Accuracy : 80.4%

Sensitivity : 80.3%

Specificity : 80.5%

precision : 70.2%

Recall : 80.3%

1.The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.

2.The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.

3.The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.

4.The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.

5.The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.

6.The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.

7.The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.

8.The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.

9.The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.