Summary

X Education sells courses online, the analysis done here is to find ways to get more conversion rate for professionals to join their courses. The data provided gave us information about how the customers visit their site, the time they spend, how they reached the site and their conversion rate.

We started with going through the data given and understanding the requirement that is asked.

Then we started with the python file and followed a stepwise process to complete the analysis.

The Steps followed are as below:

- **Reading and Understanding the Data:** Reading the data in a dataframe, and understanding the shape and size of the dataframe
- **Data Cleaning:** The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to ‘not provided’ so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to ‘India’, ‘Outside India’ and ‘not provided’.
- **EDA:** Performed Univariate Analysis for numerical and categorical variables. No outliers were found and time spent on the website shows a positive impact on lead conversion.
- **Creating Dummy:** The dummy variables were created and later on the dummies with ‘not provided’ elements were removed.
- **Splitting data into train and test set:** The split was done at 70% and 30% for train and test data respectively.
- **Building Model by Logistic Regression:** Firstly, RFE was done to attain the top 20 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 ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.
- **ROC Curve:** From the ROC curve we came to optimal cut off is at 0.35
- **Precision and Recall:** With precision and recall a cut off of 0.4 was found with Precision around 75% and recall around 76% on the test data frame.
- **Prediction on test set:** Using the final model from the model building scaling and prediction was done on test data set
- **Conclusion:** Lead score was assigned to find the leads that should be contacted
- **Recommendations:** From the leads found out following recommendations were given to the company

Recommendations given:

Company should contact the following leads as they are more likely to get conevrted

- Leads coming from the lead sources "Welingak Websites" and "Reference".
- Leads who spent "more time on the websites".
- Leads who are the "working professionals".
- Leads whose last activity was "SMS Sent".
- Leads coming from the lead sources "Olark Chat".

Company should not contact the following leads as they are not likely to get conevrted

- Leads whose last activity was "Olark Chat Conversation".
- Leads who chose the option of "Do not Email" as "yes".