Summary:

This analysis is done for X Education which sells online courses to industry professionals. The main objective of this logistic regression model building is 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' had to be replaced with a null value since it did not give us much information. And variables having missing values more than 45% were dropped. Few of categorical variables had vivid categories of data we have combined the less % of categories into 'Others' so as to not lose much data. Although they were later removed while making dummies.

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 less outliers were found.

3. Dummy Variables: The dummy variables were created and later on the dummies with 'others' elements were removed. For numeric values we used the Standard Scaler.

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 23 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 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 and with an optimum cut off as 0.34 with accuracy, sensitivity and specificity of around 80%.

8. Precision – Recall: This method was also used to recheck and a cut off of 0.34 was found with Precision around 80% and recall around 69% on the train data frame.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

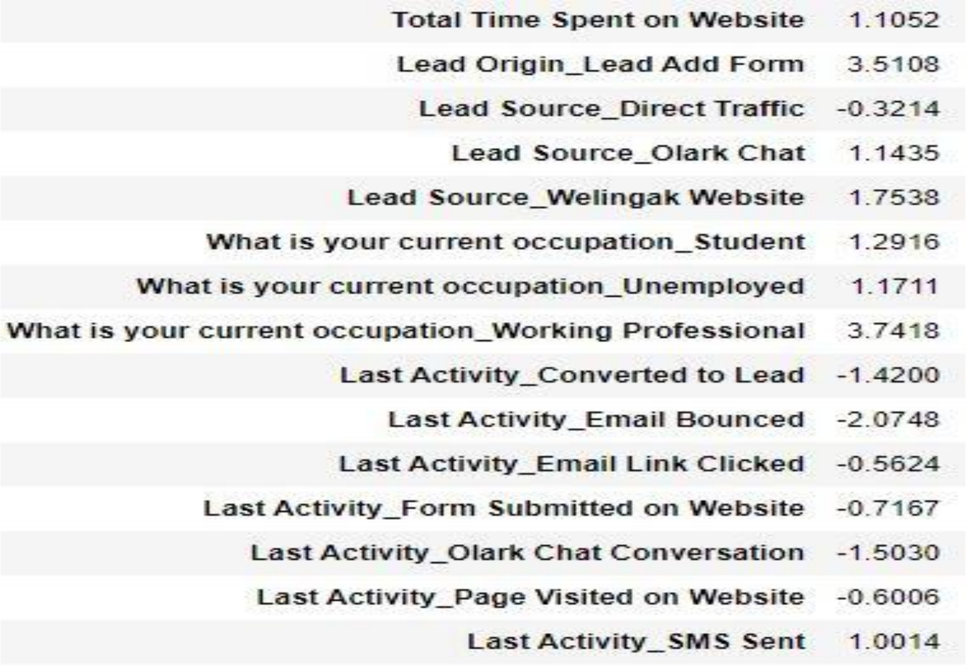
Through our model we got the top variables which needs more focus,

1. Occupation – Working Professionals
2. Where the origin of the Leads is from Lead Add form
3. Lead Source- Wellingak Website
4. Students
5. Unemployed
6. Olark Chat as Lead Source
7. Time spent by the Lead on website
8. When the last activity was: a. SMS b. Email_bounced

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

- The final model on the train dataset is used to make predictions for the test dataset
- The train data set was scaled using the scaler.transform function that was used to scale the train dataset.
- The Predicted probabilities were added to the leads in the test dataframe.
- Using the probability threshold value of 0.34, the leads from the test dataset were predicted if they will convert or not.
- Got the accuracy of 80% and Sensitivity of 80% on the test data.
- The train and test dataset is concatenated to get the entire list of leads available.
- The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.

- Higher the lead score, higher is the probability of a lead getting converted and vice versa,
- Since, we had used 0.34 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of 34 or above will have a value of ‘1’ in the Final predicted column.
- 15 features have been used by our model to successfully predict if a lead will get converted or not.
- The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.
- Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.
- Similarly, features with high negative beta values contribute the least.



Total Time Spent on Website	1.1052
Lead Origin_Lead Add Form	3.5108
Lead Source_Direct Traffic	-0.3214
Lead Source_Olark Chat	1.1435
Lead Source_Welingak Website	1.7538
What is your current occupation_Student	1.2916
What is your current occupation_Unemployed	1.1711
What is your current occupation_Working Professional	3.7418
Last Activity_Converted to Lead	-1.4200
Last Activity_Email Bounced	-2.0748
Last Activity_Email Link Clicked	-0.5624
Last Activity_Form Submitted on Website	-0.7167
Last Activity_Olark Chat Conversation	-1.5030
Last Activity_Page Visited on Website	-0.6006
Last Activity_SMS Sent	1.0014