# **Problem Statement**

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:



Lead Conversion Process - Demonstrated as a funnel

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc. ) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Summary

## **The steps followed in developing logistic regression model is as follows:-**

1. Importing necessary libraries We have used Numpy, Pandas for dataframe, Matplotlib and Seaborn for visualization and SKLearn for TrainTest Split, Scaling and Model evaluation for building the Logostic Regression.
2. Importing and Describing After importing the dataset, we replace all the ‘Select’ values in the Dataframe with others since they are all NULL values where the user has not selected anything from the Drop-Down menu.We drop the columns with more than 70% Null. After the imputations are complete, we remove those rows which have less than 2-5% null value in their columns. Finally, we have a Dataset with zero null values.
3. Handling anomalies and outliers Columns such as “Total time spent on visit” and “Total Visits” has quite a few outliers,thus we remove all the datapoints which are going beyond upper fense. Also, we drop those columns which are highly imbalanced and does not play as part in deciding the outcome of the converted column.
4. EDA We compare all the columns with respect to ‘Converted’ column to analyse which attributes play a part in deciding the outcome of a Lead Conversion. We decide which column to use in the Model building. Countplot has been used predominantly for this purpose. For most of the columns we have checked the conversion rate on the basis of that column with respect to converted column.
5. Splitting the data into Train and Test Data The final dataset is separated randomly into a 70-30% split dataset for Training and Testing purpose. This is done using SKLearn’s train\_test\_split function.
6. Scaling the data Since the dataset has data of different dimensions, we need to scale the data in order to make it suitable for a regression model. Since there are several dummy variables created from the categorical values, we opt for Normalization or Standard scaling so that the 1 and the 0s remain as it is. For this we use SKLearn’s StandardScaler function. We fit and transform the train dataset only.
7. Building the model Using the logistic regression model.We use vif (variance inflation factor) for checking multicollinearity and those columns which are showing high vif, we drop those columns.
8. Evaluating the model We determine the Confusion Matrix and the parameters like Sensitivity, Specificity, etc. We plot the ROC curve ,the Accuracy, Sensitivity and Specificity plot and got an accuracy of **91**%.
9. Making predictions on the test dataset We scale the test dataset with only transform and then predict the probabilities using the final model. On the test dataset we get an accuracy of 91%.
10. Generating Lead Scores for the sales team for the full dataset We provide that lead score in a range of 1-100 based on the probability determined by the final model.

## **Learnings**

In case of X Education, the sales and marketing team must target leads who have,

* Spent more time on their website
* Visited their website a greater number of times
* Their Last activity was SMS or Email Following these traits would increase the lead conversions to a higher percentage.