# SUMMARY

**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%. 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%.

## AIM

The aim of our analysis to find ways to get more industry professionals to join their courses. To build a predictive logisitc model to identify the potential leads that can be converted. The 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.

## EXECUTION

**Cleaning data**: The data provided seems to be in good shape, except for few null values. The columns with null values more than 30 percent are dropped. Various methods were adopted to treat null values carefully like deleting the rows where null values are less in number. Replacing the null values with mean or mode. With all these methods we could establish 69 % of data.

**Exploratory Data Analysis** A quick EDA was performed to check the condition of our data. It was found that there are few columns with unique values. These columns with unique value count < 2 are dropped as they may not be essential in our prediction model. A lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

1. **Dummy Variables**: The dummy variables were created and merged with main data. For numeric values we used the Min Max Scaler.

2. **Train-Test split**: The split was done at 70% and 30% for train and test data respectively.

3. **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).

4. **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.

5. **Prediction**: Prediction was done on the test data frame and with an optimum cut off as 0.42 with accuracy, sensitivity and specificity of appx 80%

6. **Precision – Recall**: This method was also used to recheck and a cut off of 0.45 was found with Precision around 80 and recall around 75% on the test data frame

**Other Major observations:**It was found that the variables that mattered the most in the potential buyers are:

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was: a. Google b. Direct traffic c. Organic search d. Welingak website
4. When the last activity was: a. SMS b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional. 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.

## Conclusion

In order to get a higher lead conversion, first, sort out the best prospects from the leads you have generated. 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted. Then, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies. Monitor each lead carefully so that you can tailor the information you send to them. Carefully provide job offerings, information or courses that suits best according to the interest of the leads.

A proper plan to projecting the needs of each lead will go a long way to capture the leads as prospects. Focus on converted leads. Hold question-answer sessions with leads to extract the right information you need about them. Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses.