The problem statement is X Education team wants to identify the most potential leads which are most likely to convert student into the full-time student.

The solution that we followed is: We have prepared our data and Exploratory Data Analysis tasks which include - data cleaning, univariate analysis, and bivariate analysis for a better understanding of data.

In the data cleaning part, some interesting things we get to see that many of the categorical variables have a level called 'Select' which needs to be handled because in certain categorical variables by default 'Select' is saved as an option to choose or write the answer. So, we convert to Nan than we will again look for whole Nan values. After this, we drop columns with a very high percentage of missing value and for less percentage of missing value, we use some imputation techniques such as mode, median, and mean. Then we identify all those categorical columns that are very highly skewed and drop because we want our model to learn some interesting facts. To handle so many categories with a very little percentage of rows, we combine into a new category name 'Other'. Finally, we check the percentage of rows retained in the data cleaning process which came to be 5374 out of 9240 means we still have around 58% of the rows.

We create dummies for the remaining categorical columns and perform a train-test split by dropping our target variable (Converted).

After scaling, we use the recursive feature elimination technique to choose 25 variables, and using the manual approach we drop some variables based on p-value and VIF, and the final model we have 12 variables. We found the optimal cutoff point is 0.4. Using this threshold, we perform metrics calculation from the Confusion Matrix. And also perform tradeoff where we get to see high recall and high precision which means the model is a very good means we don't have unwanted variables (noise) in our model. And also 95% under the curve had a chance that the model can `distinguish` between `positive and negative class`. At last, our model is able to determine the probability of students getting converted into full-time students.

The Major Problem we faced that in many categorical variables the same label/row name (like Other, SMS sent, Olark Chat Conversation) is used when we create a dummy for that variable we failed to identify which label belongs to which variable and also for that correlation it is very high (VIF came to be infinitive). To handle that we used prefix to assign a similar label name with a respected variable name.

Major Learning is it is really necessary to understand the data thoroughly. So, that little mistake cannot abide you to deviate from building a better model

As per business requirement, the potential leads are those on which at a certain threshold the recall/sensitivity more. And in the future also the company can handle the model smoothly by changing the threshold according to their requirements.