# Report

## PROBLEM STATEMENT

The business problem that we were given to work on was of an X education company that required us to build a logistic regression model for the company. This education company 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. When these people fill up a form providing their email address or phone number, they are classified to be a “lead”.

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%.  As it is clear that the conversion rate is very poor and so the company wants us to solve this problem for them by building a logistic model where we would 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 an estimate of the target lead conversion rate to be around 80%.

So, we were required to build a logistic regression model using the given dataset and make predictions for the company to help them further increase the conversion rate and bring it to about 80%.

## APPROACH

The approach that we used was to go through the usual steps that are necessary for building any logistic regression model.

Foremost, we did the data cleaning by getting all the missing values & duplicate values and then dropping the columns that were not necessary for the model building followed by checking outliers.

In order to understand our dataset better we did the data analysis where we made important inferences from the plots that we derived for different variables which were lead origin, lead source, email, call, last activity, current occupation, search, newspaper article, forums, digital ad, recommendations, mastering interview and last notable activity.

We then did data preparation where we created dummies for the columns and removed unnecessary ones.

For model building we then had to split our dataset into test and train set. Then we rescaled the features with minmax scaling so that no variable is dominated by the other.

We achieved the logistic regression model that we were looking for.

We further rebuild the model and then we checked the confusion metrics and accuracy, VIFs and metrics as sensitivity, specificity, false true rate, positive prediction value & negative prediction value.

Further we plot the ROC curve to determine the best cut off value. Then, we found the optimal cut off point and finally made the predictions on the test set.

## CONCLUSION

We concluded that we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.

Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%

Hence overall this model seems to be beneficial for the company.