SUMMARY REPORT

X Education is an education company which sells online courses to industry professionals. When the people search for the courses through various forms on their website and provide certain details like mail address, cell numbers are classified to be a lead. We have been provided with a leads dataset from the past which has around 9000 data points to analyse the data and help them select the most promising leads (Hot leads) with higher chances of converting. The main aim of the analysis is to increase the lead conversion rate from the current 30% to around 80%.

First we imported the dataset and inspected it for the number of records and the attributes present in it .The data set has around 9000 records of past leads and consisted various attributes such as Lead Source, Specialization, Total time spent on the website, Total visits, City, Countries etc. We then proceeded with the cleaning up of the dataset to make it prepared for performing the analysis to get the desired result.

We checked for the null values percentage for the attribute columns and dropped the columns with very high percentage of null values as they will be of no help for our analysis. The attributes with moderate null value percentages were handled by imputing the null values with appropriate values such as mode, median etc. by using various approaches. For the categorical columns we find the mode for the column and impute the missing values with mode and for numerical variables we checked for the presence of outliers and for the columns with outliers, null values were imputed with median. By performing exploratory data analysis we visualised the data with respect to our target variable “converted”. Many columns in the dataset were highly skewed, so we dropped them as it might negatively affect our analysis. We also dropped the attributes which we felt not so relevant for our goal. The columns which were added by the sales team were also dropped in the analysis.

We performed standard scaling for the numerical variables and created dummies for all the categorical attributes before proceeding with the model building. For

the model building we split our datasets into two parts with a 70:30 ratio. The dataset with 70% of the records is used for training our ML model and the remaining 30% data is used as our test model.

The logistic regression model was built and run using the train dataset. We used RFE automation technique for feature selection to get the top 15 variables of our model and later assessed the model with stats Models. We ran the model using the 15 attributes filtered from the rfe technique and looked for the correlation, P-values and VIFs. Based on their Correlation coefficients, p-values and VIFs we filtered out the attributes until we reached 10 variables. Then by plotting ROC curve and the specificity, sensitivity and accuracy curve we found out the optimal cut-off point for the converted column and assigned lead score for every lead in order to find out the most promising leads. We then ran the model with the test dataset and the sensitivity (the accuracy of identifying the hot lead correctly) i.e. lead conversion rate turned out to be pretty good and very similar in both train and test datasets. So this helps the organisation to achieve a better lead conversion and results in the growth of the business.