PROBLEM STATEMENT

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%.

The data set has been uploaded in the python notebook which has 9240 rows and 37 columns. There are few columns which have null values greater than 70% which are dropped completely from the dataset. The columns which have null values around 30% are imputed as unknown column and the columns which have null values less than 10% are replaced with mean/median/mode. After dealing with null values, the columns having constant value throughout the dataset are dropped as they do not contribute in model building. Once the data is cleaned, the dataset is verified for any outliers and the outliers are dealt. Thus, the final dataset has 9029 rows and 18 columns. Once it is done, the Boolean features are dealt, where they are converted into binary features. The categorical features are converted into dummy variables and dropped one of the dummy variables as it does not affect the model analysis and main categorical features are dropped as they are required anymore. After all these, the dataset has 9029 rows and 73 columns. Since, data cleaning, data analysis and data preparation for modelling is done, the data is split into train-test datasets. After which the numerical features in the train dataset are scaled using fit_transform. Before starting the modelling, the percentage of converted customers was found to be 38.5%. Initially the model is built using all the variables and found to have 6258 residual, p-values 0 for all the features. However, it has few features with negative coefficients. This could be overfitting model. In order to find the accurate model with significant p-values and VIF, we use RFE to chose the variables and build the model. Once the RFE chosen the variables, almost 7 models are built by removing the features having large p-values, 7^{th} model is finalised as it has significant p-values. The VIF values of the final model are calculated and found to be significant. The metrics like sensitivity, specificity, false positive rate, Positive predictive value, negative predictive value, precision and recall are checked and are found to be significant. Thus, the model finalised is found with optimal cut off of 0.37 with ROC curve. The same metrics are performed on test dataset and both train and test datasets show good performance with 80% accuracy as per the problem statement.