With the data supplied by X Education and the need of the client to identify most convertible candidates as the objective the first task involved understanding of the data and cleaning of the same.

During cleaning a good number of columns with more than 45% of the data missing were dropped as inferring any relationship with the target variable would be impossible. Also the data high degree of class imbalance in categorical variables was dropped as they indicate an impact of one class. The categorical columns containing too few of data points representing a category were clubbed together as others as this would lead to generation of too many dummy variables making the model complex.

After which outlier analysis showed not many data points as outliers and if they did the data points were capped to 99 percentile. Missing categorical data was imputed using mode of the column and continuous variables were imputed using mean of the column.

The data was now split into test and train with 70% as train data. Minmax scaler was used to scale the continuous variables and dummy variables generation was done for categorical variables. The train data underwent fit_transform operation on the continuous variables.

Setting up the model was done using RFE to identify the top features and first model was set up using theses features. The pvalue and VIF was calculated for these variables were calculated and model features with pvalue greater than 0.05 was dropped one at a time starting with the highest and new model was developed. If no feature exhibited higher pvalue than the one then the feature with the highest VIF was dropped. This process was continued till no featured showed a pvalue greater than 0.05 and no feature had VIF score greater than 5.

The final model was taken and cutoff probability of 0.5 showed lower sensitivity. As identifying more true positives became the key sensitivity, specificity and accuracy curve vs probability cutoof curve showed optimized model at 0.34.

The final model with optimized cutoff probability was developed and ran against test data which showed near same values of evaluation metrics as the train data thus showing that the model was generalizing as against memorizing the data points.