Project Report Summary

The Solution for Education X , starts initially with importing data set and start pre-processing of the data. Key steps were to:

1. drop ‘Select’ from the dataset, as it was of no significance
2. Checking for duplicate Prospect ID.

Also, we check the nulls for all the columns. All the features with null values ranging from 27 to 52% are being tackled by either replacing low occurring feature classifiers to a majority occuring feature classifier, Or dropna.

Further we move to Exploratory Data Analysis . After which we dropped features such as 'Lead Number','What matters most to you in choosing a course','Search','Magazine','Newspaper Article','X Education Forums','Newspaper', 'Digital Advertisement','Through Recommendations','Receive More Updates About Our Courses','Update me on Supply Chain Content', 'Get updates on DM Content','I agree to pay the amount through cheque','A free copy of Mastering The Interview','Country'

Further Binary conversion for two variables are done and created dummy variables for categorical variables. After that the duplicate values were dropped. Post this X and Y variables are defined. Data is then split into train and test in 70-30 format.

Feature scaling is performed for continuous variables 'TotalVisits','Total Time Spent on Website','Page Views Per Visit' and check the conversion rate. We also check the correlations amongst the features. We make the logistic regression model using GLM. But using RFE selection we remove the variables and use the 13 variables for the same. Further model is evaluated based on StatsModel. Based on p-value columns 'Tags_invalid number' and 'Tags_wrong number given' are dropped.

Converted probability is calculated after prediction. All probabilities above 0.5 is marked 1 else 0. Post confusion matrix we check variance inflation factor , and none found to be removed so far as VIF reading was below 4. Further saw the metrics beyond accuracy such as sensitivity , specificity, false positive rate, positive predictive value, Negative predictive value.

Drawing ROC Curve between True Positive Rate and False Positive Rate. And find optimal cutoff point. Cut off is found to be 0.2 after seeing the intersection point for accuracy, sensitivity and specificity. Post this we assign the lead score on the training data using converted_probability from 1 to 100. Post that we check the overall accuracy and found positive and negative predictive value to be above 90% which is good.

Precision and Recall are also evaluated after creating confusion matrix which is 93% and 85% respectively. The tradeoff plot between the both are also seen .

After making the prediction on the test set, we checked the overall accuracy which comes out to be 90%. Whereas the 91.9% is the overall accuracy for the training set. So, not much of difference between the both. Further we again calculated the confusion matrix at the end of making prediction on test set, for sensitivity and specificity which is 84% and 94% respectively.