Summary Report

The problem was to create a multi-feature logistic regression model to classify whether certain leads will get converted or not, i.e. will they buy a course at X Education. After loading and reading the Leads.csv dataset, we conducted the data preprocessing steps: Removing columns with over 3000 null values; removing columns with a singular value (*'Do Not Call', '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', 'What matters most to you in choosing a course', 'How did you hear about X Education', 'Lead Profile', 'Country', and 'City'*); and accounting for the 'Select' values, which can be regarded as null values. This was followed by data visualization and exploratory data analysis to get an idea of the features relevant to lead conversion. Finally, the index columns, i.e. 'Prospect ID' and 'Lead Number' were removed as these do not contribute to the analysis. All the remaining categorical variables were split into dummy variables and the original columns were removed from the dataframe. The entire dataframe was split into 70% train and 30% test sets and the train set was fitted and transformed using the MinMaxScaler. Using recursive feature selection, we obtained the 15 most important features. Based on the chosen features and the statsmodel library, we trained the logistic regression model and analyzed the summary statistics. Further, we removed the features with low significance ($p > 0.05$ and VIF > 5), and were left with the 10 most important features which increased the probability of lead conversion. For this model, after setting an initial conversion threshold (0.5), we checked the area under the ROC curve, which came out out be 0.86 (a value closer to 1 indicates a good model). Then we checked the conversion rate for a range of conversion probabilities (0.0-0.9). Finally, based on the optimal coinciding values for sensitivity-specificity (and precision-recall), we fixed the optimal conversion probability to 0.44. Based on the resultant confusion matrix, we obtained the following metric values: accuracy = 0.79, sensitivity/recall = 0.78, specificity = 0.80, and precision = 0.78. Upon testing the constructed model for the test data set, we obtained the following metric values: accuracy = 0.78, sensitivity/recall = 0.78, specificity = 0.79, and precision = 0.77.