Lead Scoring Case Study Report

The first step that we took while performing the case study was to inspect the data. Upon inspection we found that the data had multiple null values. The columns that had null value percentage greater than 70% were directly dropped. These columns were ‘How did you hear about X Education' and 'Lead Profile’. After dropping these two columns we noticed that there was still a lot of null values that needed to be handled. We then dropped two columns ‘Prospect ID’ and ’Lead Number’ as they were just to provide unique identification for customer and served no purpose to our analysis. We then noticed there were still as few columns that needed to be imputed as their null value percentage was not high enough for these to be dropped directly from the dataset. We then imputed all these column values with the highest occurring value in each column. We also dropped Magazine’, ‘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' columns as they had only single values and didn’t serve our analysis. We then handled all the variables which had either Yes or No and converted them to numerical data . We also dropped a few values as they weren’t contributing to the variance of the dataset nor were they helping in finding the conversion rate. These columns were Search Newspaper ArticleX , Education Forums, Newspaper, Digital Advertisement Through Recommendation. After dropping all these values we created dummy variables for all the categorical variables and dropped the original columns of each. WE then divided the dataset into test and train datasets respectively. We then ran RFE on all the variables and picked the 15 top ranked variables. We then performed logistic regression on this newly picked columns or predictor variables. We then checked the VIF value for each of the columns as well as the p value. We dropped iteratively columns with the highest p value as VIF remained constant. After dropping the variables iteratively, we finally came with a conversion rate of 88% with an accuracy rate of 88.88%. We then found the lead score of each Customer. Based on lead score we found that a cut off 32 provided with a conversion rate of 88.2%. Thus we created a relatively good model with 11 variables at the end which provided a conversion rate of 88% on the test dataset as well.