LEAD SCORING ASSIGNMENT

Assignment Summary

An education company named X Education sells online courses to industry professionals. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. Thus, the business objective is to promising leads, i.e. the leads that are most likely to convert into paying customers. It clearly reflects the case of supervised machine learning. It is a classification problem Therefore, in this analysis, we have used Logistic regression to identify the promising leads. Before building a model, data preparation processes were applied to make the data interpretable so that it can provide significant business value by improving Decision Making Process. The data preparation processes include following steps :

1. There are some columns in dataset which have a lvalue called 'Select'. This could be qowing to non mandatory fields with drop downs options for the customer to choose from. Amongst the dropdown values, the default option is probably 'Select' and since these aren't mandatory fields, many customer might have chosen to leave it as the default value 'Select'
2. Assigning a Unique Category to NULL/SELECT values
3. Deleting the columns having more than 455 missing values
4. Imputing the null values with the respective measure such as Median and Mode.For example, the columns 'TotalVisits' and 'Page Views Per Visit' are continuous variables with outliers. Hence the null values for these columns were imputed with the column median values.
5. Caping the high range outliers to keep intact the information about the prospective leads who are visiting the page again and again

Thereafter dummy variables were created and the original dataset was split into train and test dataset Scaling was performed on the numeric variables to normalize the data. After using the RFE, we were able to find the best performing subset of features. Then , we manually eliminated some variables based on p value and VIF such as Specialization_Human Resource Management, City_New_Cities and left with 15 best performing variables. Finally, we run the logistic regression with GLM to find the conversion probabilities. We have look at the various measures such as accuracy, sensitivity, precision and specificity to check the reliability of our model. On the actual Train model, we got the model with high accuracy (=80%) and high specificity (=82%), but medium sensitivity (=77%). Similarly, on the Test model, we again have an accuracy of 80%. and sensitivity of the test model is 77% . Since the company want the high conversion probability, the 77% of sensitivity ensures that almost all leads who are likely to Convert are identified correctly.

Thus, based on the coefficient values the final model, the following are the top three variables that contribute most towards the probability of a lead getting converted:

a) Working professional (from what is the current occupation)

b) Other_References (from Lead Source)

c) Others Activity -SMS sent (from the Lead Notable Activity)

Finally we have also calculated the lead score by multiplying the conversion probability with 100. Higher the lead score, higher is the probability of a lead getting converted and vice versa,