Assignment - Lead Scoring using Logistics Regression

Problem Statement:

An education company X Education sells online courses to industry professionals. It markets itself on various channels and once a prospect fills certain data on its website, its sales agents call these prospects and try to get them enrolled for a suitable course.

This process involves calling all leads while the **lead conversion is very low – about 30%.** Tough the investment is high; the returns are low. X Education is **unable to identify Hot Leads** that can be part of a focussed target group to call instead of making calls to all prospects. This way with lesser investment, X Education can have better results.

To solve this problem, X Education has invited data science experts to analyse their data and score leads. Calls will be made to leads having high score to ensure a higher conversion ratio.

Approach:

As data is derived from an application having drop down selection, some columns have default values. Impute them with null. Then delete all columns that have a high percentage (> 45%) of missing values. For lower, impute mean/median/mode. Delete records that have more than one column having null values if such records are very low (<2%). If a column has high imbalance, then delete it. Finally, data is ready for model building. In categorical columns, low frequency values can be combined into single groups.

Create Dummy variables for categorical columns. Then split the data into training and test sets (70:30). Scale data features that are necessary. Build logistics model. Either manually remove columns or use RFE. In any case, keep checking 'p' values and VIF scores to reduce columns so that model scores are not compromised and features are also not too much. Multiply the percentage assigned to Converted column by 100 to get the lead score. Build ROC curve and decide on a suitable cut off to classify record as converted or not. Then build confusion matrix and use sensitivity as the primary measure to decide on model.

Apply the model on test data. Do Principal component analysis to get important features.

Learning for X Education

Hot Leads are people

- That spend high amount of time on their website
- Whose leads have come from Google or Organic search methods
- That engage in communication exchange like SMS

These are people who are motivated as they are doing some research on their own. For leads less hot, X Education can also focus on non-call based communication strategy such as email and SMS. These leads can be engaged for a longer period of time using such communication channels.

Learning as a developer:

- Most of users leave non mandatory columns blank
- Data can be highly imbalanced
- Model must be built iteratively. In this case sensitivity is most important measure as we are interested in knowing positive cases, in some other specificity or accuracy could be.
- Cut off values must be found using ROC curve as that has mathematical basis. A judgement of 50% does not apply most of the times
- Principal Component Analysis is the right approach to identify the features contributing most to the target variable