LEAD SCORE CASE STUDY

Assignment Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor.
To make this process more efficient, the company wishes to identify the most potential leads. As the CEO has given a ballpark of the target lead conversion rate to be around 80%.

Solution:

Objective: We have predictor variables, with the help of these we have select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

We have used Logistic Regression model to find out the conversion rate at which students can be converted and join the course.

Below are the 5 steps which we have followed for our analysis.

1.Reading and understand the Data:
Here we have read the data into a df and checked the various aspects of df such as it's shape, dimensions, descriptions, etc.

2.Data Cleaning and Redefining:
In this step we did three main tasks:
A. Remove Redundancy : We have removed some of the columns which were redundant i.e. providing similar information. For example columns "last activity" and "last notable activity".
B. Rename Columns: Then we have removed few column names for better understanding in further analysis.
C. Check missing values: Then we have calculated the percentage of missing values in each columns, we have dropped the columns having more than 30% missing values. Also we have dropped those columns which were having same values i.e. not distinct values. For example " I agree to pay the amount through check".

3.Data Preparation:
This is the most time consuming task of whole case study. Here we performed number of tasks:
A. We mapped yes/no values to 1/0.
B. Then we checked the count of Boolean values in each Boolean columns, if the count of any value say 1 or 0 is far greater than other values then we dropped such columns because such columns contains similar values and won't help further.For example: below columns "Newspaper Article", "Magazine", etc
C. Next step is to create dummy variables for the categorical variables present in the df. We performed outlier treatment also for the columns having many distinct categorical values.
D. Rescaling is the next step which includes the conversion of numerical values with huge difference in df into one standard form. Also we checked the conversion rate which was 38%.
E. Last step of Data preparation was to check the correlation among all the variables present in the df.

4.Applying PCA (Principal Component Analysis):

Next step is to apply PCA on the df to find out the most important components of df. Here we performed below tasks:

A. Splitted the data into train and test set.

B. Calculated the cumulative variance to find the 20 components which are most important.

C. Then we created PCA model with 20 components. Also we checked the correlation matrix by plotting heat map.

D. Finally we applied transformation in test set.

5.Model Building & Evalution:

This is the actual model building part where we created the logisitic regression model and applied to test set. Based upon conversion probability we assigned 1/0 values to each rows. We checked accuracy and redefined the conversion probability. We started with the conversion probability of 80%, and check ed the accuracy . Then we goes further as 75% ,70%. At 70 % conversion probability we got the best accuracy. Finally we have plotted and calculated the ROC to find the overall AUC score which was 88%.