Summary

This evaluation is carried out for X Education and to locate approaches to get greater enterprise specialists to join their courses. The primary records supplied gave us a lot of data about how the potential clients go to the site, the time they spend there, how they reached the website online and the conversion rate.

The following are the steps used:

Data cleaning:

We dropped the missing values which having the percentage greater than 50

And we successfully dealt with the missing values having percentage less than our threshold

EDA:

Majority of the customers has not seen the add at X education forum , Newspaper Article , Not searched , Not seen digital advertisement,Through Recommendation, or Newspaper

Customers who wants updates via call,Email are more likely to be a lead

leaders show following behavior :

- They do not want a free copy of mastering the interview
- Medium asymmetrique activity index
- high asymmetrique profile index
- They want better career prospectus
- They might have lead quality and Landing page submission
- They are unemployed and from mumbai and thane and belong to india
- They heard from online search and students from same school
- have asymmetrique profile score ranging from 15 -18,Asymmetrique activity score ranging from 14 - 16
- they recently did the modification , opened their email and sent the SMS and probably did the same as a last activity
- mejority of them are specialized in projects management ,finance management ,human resource management ,marketing management.
- they comes from sectors like banking , investment and insurance and operation management
- mejority lead sources from google ,direct traffic and organic search
- they do revert after reading the email that shows their interest
- they visit the website atleast 50 times they visit the page at-least 10 times and spend at least 500 s on the website

Dummy variables :

Then we created dummy variables and we did one hot encoding

Class-imbalance :

We used SMOTETomek to deal with the class imbalance and we successfully dealt with that.

Train test split and scaling :

- We scaled the model using min max scaling and then performed the split
- For the splitting we split the original data frame into train and validation data frame and after that we even split the tarin data into test data as of 80,20 split

Model building:

- We used RFE for selecting the best possible features and after from that we dropped the features one by one who had the p value greater than 0.05
- We used GLM from statsmodels library for the model building

Model evaluation:

- On test data we got accuracy of 92.52 percentage.
- Further we did model decomposition with the help of PCA and got accuracy of around 97.05 percentage on the test data
- On the validation data we got the accuracy of 89.58 percent and after applying the PCA we got the accuracy of around 95.61 percentage.

Top three positively correlated features are :

I. Total Time Spent on Website

II. Tags_Will revert after reading the email

III. Lead Origin_Lead Add Form

Top three negatively correlated features are :

I. Last Activity_Olark Chat Conversation
II. Last Notable Activity_Modified
III. Tags_Ringing