Lead Case Study - Summary

In this Lead Case study assignment, we had a Problem statement to identify the most Promising leads for X Education that will most likely convert into Paying customers. We have to assign a Lead score to each customer from 0 to 100 that will help the client understand the hot leads to target.

To begin we have imported the 'Leads.csv' data file and check its info and shape of the dataset. The dataset has 9000+ rows and 37 columns.
It also has many null values, so we begin by dropping columns that have high number of missing values ('Lead Quality' and 'Asymmetrique__X'. After checking other values in columns, it is found that data has many rows with 'Select' as values, which is basically a null value as the leads have not chosen any option and thus it is seen as select by default. So, imputing these with np.nan. Then again checking missing values and dropping columns with more than 35% null values and then imputing remaining columns with their modes.

After missing value treatment is done, we do mapping of the entries wit Yes/ No values to 1 and 0 Respectively and then created dummy variables for the remaining categorical variables. Then we found out had created over 90+ columns, then we checked into the data and found that the data is skewed and hence replaced values with lesser occurrences to 'Others' and again created dummy variables. After concatenating these with main dataset we had around 37 columns to work with. Then we dropped Outliers in the Total Visits and Pages Views per visit column.

We divided the dataset into train and test sets keeping 'Converted' as the dependent variable and other columns as independent variables. Then we checked the Correlation matrix to drop variables having higher correlation to avoid multi-collinearity.

Model Building- We decided the top 15 feature using Recursive Feature elimination and then built the model using statsmodel and eliminated variables having higher p values and higher VIF. Then predictions done on train set which gave us probability of conversion. We selected 0.5 as arbitrary above which probabilities are chosen as converted and below which probabilities are not converted. Then we derived a confusion matrix based on actual and predicted conversion values and got the accuracy of the model to be 80% . Sensitivity found out to be 67% which could be improved and Specificity was 88%.

Evaluation metrics – We plotted the ROC curve and Area under curve was 0.87 which means our model is good. Next to find optimal threshold we plotted the Sensitivity and Sensitivity at different thresholds ranging from 0 to 0.9 and found the optimal threshold to be at 0.3. Next, we made predictions at these new thresholds on the train set and then again evaluated the model. The accuracy was slightly decreased to 78% , But Sensitivity was increased to 83% and specificity decreased to 75% which is what we aimed for.

Predictions on Test set – We used this model to make predictions on test set. We got a model with Accuracy 77%, Sensitivity 85% and Specificity 73%. Overall, we have achieved very good performance from the model.

To End we multiplied the Conversion probabilities with 100 to get the Lead scores and created Student column from Lead Number to generate a table with Student and Lead scores for the client.