Summary Report : Lead Scoring Case Study

- Analysis by Ankit Goregaonkar

The problem statement for the lead scoring case study is that we need to help X Education, an education company (which sells online courses to industry professionals) with an efficient way to handle lead conversion. Till now, their lead conversion rate is 30% which is quite inefficient. For this, we made a Binary logistic regression model which would predict if a lead would convert or not basis the features we have in the dataset and see how it performs as compared to the actual conversion values we have. Then, assigned each lead with a score which would tell if a lead is hot or cold.

The case study had been analyzed as per the following steps -

Data Reading – Dataset was read. Some prima facie checks made on the dataframe obtained.

Data Cleaning – This has been the most crucial and time taking step of the whole process because we had to think thrice or even more before dropping any part of data. Initially, We could see that the dataset had value "Select" for all columns and we could infer that this value had been a default for "Null" value for all the columns. While filling the form, if employees didn't select any option, then they land with a "Select" for that column.

Treatment of Null Values –

Columns with more than 50 % null values – We found out that columns "How did you hear about X Education", "Lead Quality", "Lead Profile" had percentages of null values greater than 50%. Since, we do not have any metrics to derive values for these columns; also half of the data has nothing to say, imputing it would just mean exaggeration and having results in a wrong direction.

Columns with null values in a range of 0 – 50% - 'Asymmetrique Activity Index','Asymmetrique Profile Index', 'Asymmetrique Activity Score','Asymmetrique Profile Score' – These columns had 45% of null values and had an index score assigned to each customer based on their activity and their profile by sales team. The metrics followed to assign the scores has not been specified and seems vague. Thus, we have dropped these columns.

City, Specialization and Tags columns were deleted too, citing very few true observations, the inability to derive the null values, business problem in hand or the vague metrics to derive values(for tags).

Skewed columns – Including these in model would mean that it is highly probable that the data points from category with highest count would be picked and hence, imputing some biasedness in the model.

.Rows with more than 5 null values were deleted too.

Outlier treatment for TotalVisits' and 'PageViewsPerVisit' was then done. We decided to keep the outliers to have our model prepared for the outlying values for these columns since they seem to be good outliers.

Data preparation

The dummy variables were derived for categorical columns to ensure we don't define any priority order for the same and to see impact of each category in a column on the target variables. The data was then split into test and train datasets and then test data. **Scaling -** Since the Outliers for the 'TotalVisits' and 'PageViewsPerVisit' were important and we therefore used used MinMaxScaler and not StandardScaler.

Heatmap of the dataframe was analyszed and any highly correlated columns were dropped.

The model was built using RFE approach wherein we started with 15 columns. On the basis of p values and corresponding VIF scores, the columns were removed and model rebuilt again. We finalized the model with 10 columns wherein p values for all the columns were 0.05 and highest VIF being 2.93.

We then defined the optimal cutoff point made by plotting sensitivity, specificity and accuracy for various probabilities and chose it as 0.34. Then, the predictions were made on the test data and confusion metrics derived for the same again.

For the final model, the evaluation metrics came out to be as following –

	Train Data	**Test Data**
Accuracy	80.24%	78.48%
Sensitivity/ Recall	81.29%	79.07%
Specificity	79.58%	78.11%
Precision	71.26%	69.77%

We then evaluated the model basis on F1 score and ROC curve. F1 score being 0.7594 and Area under ROC curve being 0.87 which is a good indicator.