Lead scoring case study – Approach summary

Problem Statement

- X Education gets a lot of leads, its lead conversion rate is very poor. (about 30%)
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- The company requires to build a model wherein a lead score will be assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Step 1 : Importing the dataset

Step 2 : Checking for null values

- Identify all columns that have more than 45% null values
- Drop the columns from the dataset
- Following columns are removed as don't add value to the modeling - 'Prospect ID', 'Lead Number'
- 'Select' is converted to Null for all columns
- *Country and City* columns are removed as they show high class imbalance.
- '*What matters most to you in choosing a* course' is dropped as it has too many null values
- Several columns are dropped as they are skewed or have high null values.

Step 3 : Data Preparation

- Binary categorical variables are converted from Yes/No to 1 and 0
- Created dummy variables for other categorical variables and dropped the parent columns

Step 4 : Test -Train Split

- Data set is split into Train and Test data set in the ration of 70 and 30

Step 5 : Feature Scaling

- Scaled the continuous variables using Minmax Scalar

Step 6 : Checking for correlation

- Checking if there are any highly correlated features that can be dropped from the model

Step 7 : Model Building

- Since this is a binary classification problem, logistic regression is used for modelling.
- RFE used to find the most significant 15 features
- Rebuilt the model based these features
- One variable is dropped as it was found to have high p-value
- All p-values of the variables are in the reasonable range (below 0.05%)
- VIF values are less than 5 for all variables

Step 8 : Model Evaluation

- Initial Confusion matrix calculated for the trained values with 50% as cut off
  - Accuracy : 81%, Sensitivity : 69%
- Area under the ROC Curve : 89%
- ROC shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

Step 9 : Finding Optimal Cutoff

- For various cutoff points accuracy, sensitivity, specificity is plotted
- It was found that 0.38 is the optimum threshold that will be best accuracy and sensitivity.
- Precision and Recall values are also plotted also shows the threshold around 0.38

Step 10 : Making predictions on the Test set

- The test sets are also showing the similar metrics which indicates the model is accurately predicting the outcome
- Finally the lead score (between 1 to 100) is calculated by multiplying the probability by 100
- Recommendations to the management is arrived by analyzing the coefficients of various features in the model

Conclusion : It was found that the variables that mattered the in identifying the potential buyers are :

- Customers who make more visits and spend more time spent on the website have higher chances of conversion
- Customers whose leads came via Olark chat and Welingak website have high conversion probability
- Working Professionals likely to have higher conversion chances.
- People who opted NOT to be emailed have lower chances of conversion.
- Last Activity of SMS or others have higher chances of conversion.