Lead Scoring Case Study – Summary

Problem Statement

An education company named X Education sells online courses to industry professionals. The current lead conversion rate at X education is around 30%. The company requires a model wherein lead score is 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, with a target lead conversion rate to be around 80%.

Analysis Procedure

- 1. Load and inspect the dataset
- 2. Null values handling
- 3. Analysis of numerical columns
- 4. Analyzing categorical type columns
- 5. Train-test split
- 6. Scaling the features
- 7. Feature selection using RFE
- 8. Model building
- 9. Finding the optimal cut-off point
- 10. Training data predictions on final model
- 11. ROC curve
- 12. Making Predictions on test dataset

Data Preparation

Null Values Handling

The columns with more than 30% of null values were straightaway dropped. Further, the columns with null values less than 30% were analyzed closely. Imputation was done only in the case that it would not affect data distribution in column.

Observations:

- The company is selling courses designed for working professionals, but majority of the targeted audience (leads) is unemployed.
- About 70% of the leads are Indian.
- A quarter of the leads had never visited the company's website.

Numerical Column Analysis

Outliers in the numerical column were removed by trimming the column data to maintain mean +- 3 standard deviation data, which corresponds to 99.7% of data. If business interaction was available a threshold could be agreed with them.

Categorical Column Analysis

Columns were analyzed one by one and below are the observations:

- Majority of leads came in via Google followed by direct traffic and then Olark chat.
- Almost all of the leads preferred not to be called.
- Majority of the leads hadn't seen any ad of the company on any means. This is a key thing that must be highlighted with advertisement team.
- Majority of leads do not want updates on the company's courses including curses on Supply chain and DM.
- Payment of fee via cheque is not a preferred choice of majority of the leads.

Proceeding further, data was split into train-test datasets. Scaled using the MinMaxSclaer. Followed by selection of 25 features using Recursive Feature Elimination for model building.

Model Building

The model was built using statsmodels library. And after selecting the features for first model using RFE manual elimination of redundant features was performed iteratively till the p-value and Variance Inflation Factor of all variables was under 0.05 and 3 respectively. Values were chosen based on ideal industry standard as they were not provided beforehand.

Features were dropped in below order of priority:

1. High p-value & high VIF
2. Low p-value & high VIF
3. High p-value & low VIF

The model was concluded and probability of conversion of each lead was calculated.

Model Evaluation

The final cut-off probability for marking a lead a converted was determined using a plot between the values of accuracy, sensitivity, & specificity at different values of probability. Based on this the cut-off was chosen as 0.35. ROC curve for the model was drawn to visualize the model performance.

Finally, the model was tested using the test dataset wherein the accuracy, sensitivity and specificity were calculated which came out to be 78.86%, 78.67%, and 78.97% respectively.

Hence, the model satisfies the given target ballpark sensitivity of 80%.