Lead Scoring Case Study Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company requires us to build a model wherein we need to assign a lead score 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.

Goals of the Case Study:

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads.
2. Need to make recommendations based on the model.

Solution Approach:

I. Data Preparation, Cleaning and EDA

- ✓ Step 1: Importing Data
  Reading data from the CSV files into the python notebook.
- ✓ Step 2: Inspecting the Dataframe
  Dimensional, datatype and statistical analysis of dataframe
- ✓ Step 3: Data Cleaning
  Missing Values Analysis and Imputation.
  Handling of 'Select' Values in columns
- ✓ Step 4: Univariate Analysis
  Single variable analysis and removal of columns which are not necessary for the model.
  Outlier Analysis & Removal
- ✓ Step 5: Bivariate Analysis
  Analysis of the variables along with Target Variable
- ✓ Step 6: Multivariate Analysis
  Correlation among the variables.
- ✓ Step 7: Data Preparation
  Creation of Dummy Variables

II. Test-train Split and Scaling

- ✓ Step 1: Test-Train Split
  Splitting in Test & Train data in 70:30 ratio
- ✓ Step 2: Feature Scaling
  Using Fit_transform to scale the Numeric variables in train dataset

III. Model Building

- ✓ Step 1: Feature elimination based on correlations
  Removing highly correlated dummy variables (Absolute correlation > 0.8)

- ✓ Step 2: Feature selection using RFE (Coarse Tuning)
  Considering 15 Variables
- ✓ Step 3: Manual feature elimination (using p-values and VIFs)
  Eliminating variables with VIF > 5 or p-value > 0.05

IV. Model Evaluation

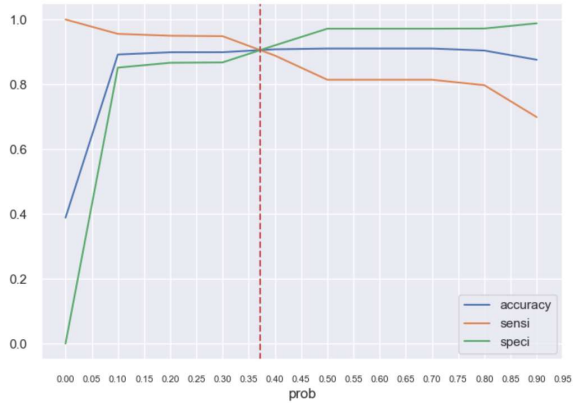
- ✓ Step 1: Accuracy

- ✓ Step 2: Sensitivity and Specificity

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

- ✓ Step 3: Optimal cut-off using ROC curve



- ✓ Step 4: Precision and Recall

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

V. Predictions on the Test Dataset

The predictions are made for test dataset on the basis of Training Model.

VI. Conclusion

✓ Final Observation:

- Train Dataset:

Accuracy	0.91
Sensitivity	~ 0.81
Specificity	0.97
Precision	0.94
Recall	0.81

- Test Dataset:

Accuracy	0.90
Sensitivity	~ 0.95
Specificity	0.87
Precision	0.80
Recall	0.95

✓ Recommendations

X-Education will have to mainly focus important features responsible for good conversion rate which are as follows:

- *Occupation_Working Professional*: The company should focus on working professionals as these leads have higher lead conversion rate.
- *Last Activity_SMS Sent*: Leads whose last activity is sms sent can be potential leads for company.
- *Occupation_Unemployed:* Occupation is 'Unemployed' feature has a good conversion rate. Company can target this feature as in order to get employed people want to upskill themselves.
- *Tags will revert after reading the mail*: This feature also has a good Conversion rate.