Summary of the Lead Scoring Case Study

OBJECTIVE :

- To build a logistic regression model for an education company named "X Education" by assigning a lead score between 0 and 100 for targeting particular leads to be converted or not.
- Adjusting the company's requirement for further changes in the future.

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 markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor.

X Education has appointed you to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

METHODOLOGY:

1. DATA INSPECTION AND CLEANING :

- First of all, the data is loaded along with information, numerical description and shape. Names were changed for some of columns.
- For the data cleaning, missing values were checked and those variables consisting of more than 40% missing values and skewed columns are removed.

2. EDA AND DATA PREPARATION:

- Graph plotting is done with Univariate, Bivariate (taking "Converted" as target variable) and Heatmap for checking the colinearity between the variables.
- Dummy variables were created and the initial variables and repeated cols were removed.
- Data splitting is done into train and test along with data scaling.

3. BUILDING A LOGISTIC REGRESSION MODEL:

- Model building is done along with feature selection (RFE) and VIF check for multicollinearity.

- After the removal of unnecessary variables, further modelling is conducted by taking 15 variables as an output. Irrelevant features may affect the model without getting the actual output in the end.

- Features with high P- Values (0.05+) and VIF (more than 5) were removed eventually by building the model again.
- Lastly, the model with 10 features and low P and VIF values are good to go for further predictions.

4. MATRIX SCORE TEST:

- Further process like Confusion matrix, Model Accuracy, Sensitivity and specificity been done.
- ROC curve shows the trade-off 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.
- As per the graph in python notebook, the ROC curve covered almost 91% of the total area.
- The values of accuracy, sensitivity, and specificity were used for finding out the optimal threshold and 0.3 is the optimum point to take as a cut-off probability.
- In the precision and recall curve, there's seems a jumpy curve for precision. Both curves meet at 0.4. Moreover, the value of precision is 0.750 and recall is 0.843.

5. PREDICTION ON TEST DATA:

- Lastly, the model evaluation of the test data is done by doing the data scaling, making predictions, checking the model accuracy, sensitivity and specificity.

- Following values after model evaluation:

  - True Positive : 884

  -True Negative : 1472

  - False Positive : 195

  -False Negative : 218

  - Model Accuracy(Correctly predicted labels / Total no. of labels): 0.850

  - Sensitivity (TP / TP + FN) : 0.802

  - Specificity (TN / TN + FP) : 0.883

6.CONCLUSION:

- Though the page per visit is high, the customer will be in a dilemma related to the course like comparing this course with others because of which conversion ratio is low.
- As per the business aspect, working professionals have probable chances of getting converted.
- Important features that contributed most towards the probability of a lead getting converted are : Time Spent, Lead Origin_Other, and TotalVisits.
- Features that need to be focused more : Last Activity 0_Other, Page Views Per Visit and Tags_Ringing.