Leads Scoring Case Study

A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

Answer:

Below are the steps how we have proceeded with our assignments:

1. **Reading the data:**
    a. Importing all the necessary libraries.
    b. Reading the csv file and inspecting the dataset.
    c. Checking the shape, info, null value, summary of statistical columns etc.

2. **Data Cleaning:**
    a. First step is to check the missing values and duplicate values present in the dataset.
    b. After this we found that some columns are having label as 'Select' which means the customer has chosen not to answer this question. The ideal value to replace this label would be null value as the customer has not opted any option. Hence, we changed those labels from 'Select' to null values.
    c. Removed columns having more than 45% null values
    d. For remaining missing values, we have imputed values with maximum number of occurrences for a column.
    e. We found for one column is having two identical label names in different format (capital letter and small letter). We fixed this issue by changes the labels names into one format.
    f. After treating the missing values next step is outlier treatment and visualizing the variables after the treatment.
    g. Analyzing and visualizing some categorical columns to get few more insights and patterns from the dataset for model building.
    h. Performing the **EDA** with univariate and bivariate analysis of numeric and categorical variables.

3. **Data Preparation:**
    a. Changed the variables from yes or no category to 0 and 1 with the help of binary map function
    b. Created the dummy variables for categorical column.
    c. Removed all the redundant and repeated columns.

4. **Test and Train Split:**
    a. Importing the sklearn libraries
    b. Split the dataset into train and test dataset and scaled the dataset.

c. After this, we plot a heatmap to check the correlations among the variables.

d. Found some correlations and they were dropped

5. Model Building:

a. Running the Training set for model building

b. We created our model with rfe count 18 and 15 and compared the model evaluation score like AUC and choose our final model with rfe 15 variables as has more stability and accuracy than the other.

c. For our final model we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity.

d. We found one convergent points and we chose that point for cutoff and predicted our final outcomes.

e. We checked the precision and recall with accuracy, sensitivity and specificity for our final model and the tradeoffs**.**

f. Prediction made now in test set and predicted value was recoded.

g. We did model evaluation on the test set like checking the accuracy, recall/sensitivity to find how the model is

h. We found the score of accuracy and sensitivity from our final test model is in acceptable range.

i. We have given lead score to the test dataset for indication that high lead score are hot leads and low lead score are not hot leads.

6. Conclusion:

We have successfully built a Logistic Regression Model with below Evaluation scores:

Train Set:

- sensitivity	84%
- specificity	78%
- Accuracy	80%

Test Set:

- sensitivity	83%
- specificity	78%
- Accuracy	80%

1) In Order to increase the lead conversion rate ,the sales team can follow up with customers who were reached out by phone call as the last activity.

2) Customers who belong to the working Professional as their current occupation

3) Concentrate on customers for whom the lead source was Welingak Website

4) Customers for whom the lead source was reference

5) If Customers opt for 'Do not email' as Yes , then their conversion rate is low , as they do not want to be reached out via email, so the possibility of conversion is low

In combination with the lead score and above mentioned columns , the conversation rate for X education company can be increased.

Problems encountered:

This dataset had multiple categories in various columns. Also the missing values in each of the columns were high , so we could not simply impute them with mode. We created various categories to tackle this. Also few columns for example Specialization had 19 categories , if we would have not combined them , we would have 18 dummy variables to represent those and likewise we would have crossed 100 variables in total before we began with the RFE process.

Also the dataset had some highly skewed column which had to be taken care of. As it would add bias to the model and would give wrong results.

Key Learning:

1) How to merge categories to reduce dummy variables
2) Cleaning data which is highly skewed.
3) While imputing missing values , have a business understanding and not just impute mode values as it will not always be correct and can skew the data , for example in column ‘Country’ , imputing missing values with value ‘India’ skewed the data.
4) Selecting columns which are relevant to business is extremely important.
5) Build a scalable model that can be tweaked to handle business requirements.