Lead Score Case Study Summary for X Education

Problem Statement

X Education is an online course provider for industry professionals. Company markets its courses on multiple websites and search engines like Google. The aim of company is to increase its lead conversion for potential leads from 30% to 80% .They want to target leads who have higher chances of conversion.

Below are the brief summary steps:

Data Understanding and Cleaning

- Importing the Dataset using Pandas Library.
- Check different parameters associated with the Dataframe like info(), describe(), shape
- Imputing the value 'Select' with 'NaN'
- Dropping columns
    - Having 1 unique value
    - Having more than 40% null values
    - Having very high imbalance in binary values.
    - Columns created by Sales Team.
- Numerical Columns treatment:
    - We were left with 4 numerical columns
    - Capping the outliers in the range of 0.1 to 0.99 percentile
    - Imputing missing values with mean
    - Plotting a heat map to check for correlation and drop highly correlated column
- Categorical Columns treatment:
    - Checking the value counts for each categorical column
    - Categorizing single digit response as 'Unknown' or 'Other'
    - Imputing 'NaN' missing values with mode or categorizing as 'Unknown' or 'Other'
    - Plotting graphs against target variable('Converted') to analyze
    - We are left with 12 categorical columns

Data Preparation

- Dummy variable creation
    - Created dummy variables for Categorical columns and dropped insignificant category column and original columns

Model Building

- Train-Test Splitting
    - Split the dataset into 70:30 ratio
- Feature Scaling
    - Used Min-Max scaler to standardize numerical column values
- Feature Selection using RFE
    - Run RFE with 15 columns
- Model validation using statsmodels.api
    - We have used p-value and VIF to select and drop features

    - After 7th iteration we have got a model with p-values's < 0.05 and VIF's < 5
- Making a Prediction on Training dataset
- Plotted ROC curve to check model performance
- Finding optimal probability cut-off
    - We have found accuracy, sensitivity and specificity with respect to probability from 0 to 0.9
    - Optimal probability cut-off is found as 0.34
    - Different metrics found are:
        - Accuracy: 79.94
        - Sensitivity: 73.51
        - Specificity: 83.90
        - Precision: 73.78
- Making Prediction on Test dataset
    - Different metrics found are:
        - Accuracy: 79.47
        - Sensitivity: 73.05
        - Specificity: 83.66
        - Precision: 74.48
    - We have found our Train and test dataset are approximately matching
- Creating a final equation
    - We have found final equation as:

      Odds(Converted)=(4.52)*Total Time Spent on Website + (3.73)*Lead Origin_Lead Add Form + (2.57)*Occupation_Working Professional + (2.11)*Lead Source_Welingak Website + (1.27)*Matters_Better Career Prospects + (1.03)*Lead Source_Olark Chat + (0.86)*TotalVisits + (0.18)*Lead Origin_API + (-1.3)*Do Not Email + (-3.27)

Assigning Lead Score to Original Dataframe

- Concatenating Train and Test dataset using LeadID
- Joining it with original dataframe using Lead_Score

Learnings:

- Selecting the right variables for creating model which have a significant impact on prediction is important.
- Doing proper missing value treatment and outlier treatment.
- Creating a flexible model which can adjust to the changing needs.