Summary

The study is conducted for X Education in an effort to increase the number of customers visiting their site enrolled in their courses. It aims to build an analytical model that can estimate the lead score for the new customer based on learning from past data.

Dataset

For the construction of the analytical model, a dataset of historical leads including around 9000 data points is supplied. In this instance, the target variable is the column 'Converted,' which indicates whether a previous lead was converted or not.

Methodology

1. Load the Dataset

The leads dataset has 9240 Rows and 37 Columns including the target. The target has 2 classes 0 and 1 which represent “not converted” and “converted” respectively. The Plot shows that the dataset is balanced between two classes.

2. Clean the Dataset

The following procedures are performed during the data cleaning process:

- Remove columns with the constant value
- Remove columns with unique values
- Remove columns with more number constant values
- Identify missing values
- Drop columns having a high percentage of missing values
- Impute columns with less percentage of missing values

Once the cleaning is done the dataset contains 6373 rows and 12 columns.

Explore and Analyse the Dataset

The analysis section involves outlier detection, numerical variable analysis, and categorical variable analysis. The results of the outlier analysis showed no presence of considerable outliers. And the analysis result showed all the numeric variables “TotalVisits”, “Total Time Spent on Website”, and “Page Views Per Visit” have a significant relation with the Target variable “Converted”.

Prepare the Dataset

Group Categories have less number of values together under the “other” category. Encode categorical variables using the one-hot encoding technique. At the end of the data preparation stage, there are 6373 rows and 67 columns in the dataset.

Split the data into Train and Test Sets

Split 70% of the data as Trainset and 30% of the dataset as Testset randomly in order to train the model independently on the trainset and test it using the unseen test set.

Standardize the Dataset

First, the training dataset is standardized using MinMaxScalar which squashes the value range between 0-1. Since the one-hot encoded values are already in the standard range, only the numerical variables require standardization. The scalar model which is fit using the training dataset is utilized to transform the test dataset during the prediction.

Building Model

The following procedures are followed during the model building:

- Do automated feature elimination using reduce the number of week predictors in the input dataset.
- Build the initial model using a selected set of features.
- Analyze the statistical significance of the feature and remove the insignificant feature
- Iteratively do the feature elimination manually until there are no more insignificant features to eliminate
- Build the model with the final set of features

The final model contains 11 features.

Model Evaluation

Initially, the model is evaluated with a default probability cutoff of 0.5.

- Accuracy - 79%
- Sensitivity - 74%
- Specificity - 83%

Then optimal probability cutoff is determined by the intersection point of the sensitivity and specificity curve which is 0.42.

Finally, the model is evaluated with an optimal probability cutoff of 0.42.

- Sensitivity - 78%
- Specificity - 79%

Results

Three features contribute the most to the lead conversion probability.

- Total Visits
- Total Time Spent on Website
- Lead Origin Lead Add Form

Two features that negatively influence the likelihood of lead conversion.

- What is your current occupation Student
- What is your current occupation Unemployed