Report: Lead Scoring Assignment

Submitted by:
Preksha Binakia
Keval Dhodiya
Achuta Mukund Harsha

Summary

Problem Statement:

X Education want to find which leads are most likely to convert into customers. The company want us to build a model which assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and vice-versa.

1. Data Processing:

The data is having (9240 rows, 37 columns)

- The option 'select' is replaced with null value since it is a missing field.
- Few of the null values were changed to top repeated feature while the rest to 'Unknown', so as to not lose much data.
- Random fields like Lead ID, and other fields like Magazine, Cheque are dropped as they were 95+% homogeneous.
- Columns with missing values above 40% are dropped as they alter the results.

2. Visualising the Data and Dealing with Outliers

EDA was done to check our data.

- The outliers values are removed after 99th quantile.
- The continuous values are bucketed.
- The minority features are clubbed with suitable categories.

3. Feature selection for Machine Learning

The dummy variables were created. For numeric values we binned them based on suitable intervals.

- The Heat-map, and correlation matrix are used to drop features which is correlated 80% or more.

4. Splitting the Data into Training and Testing Sets

The split was done at 70% and 30% for train and test data respectively.

5. Building model using statsmodel, for the detailed statistics

Firstly, RFE was done to attain Model 1 with top 20 relevant variables. Then the selected features are brought down to 15 (Model 2), and 10 (Model 3) to check the change in R-Squared value and the change for 20 to 10 variables is 0.02 which is small.

6. Model Evaluation: Predicting the train dataset with model 3

Although the VIF and P-Values of the selected 20 features are in limits. Only 10 features are choose to keep the model simple.

7. Prediction on the test dataset

The model is used on test data.

A confusion matrix was made, ROC curve is drawn and AUC came to be 95%. Key Metrics are as follows:

Train Data Set metrics:	Test Data Set metrics:
Sensitivity: 89.83	Sensitivity: 88.66
Specificity: 85.36	Specificity: 86.63
Precision: 79.11	Precision: 80.99
Accuracy: 87.07	Accuracy: 87.42

From the calculated Converted Probability Lead score is obtained which is between 0 to 100. Following observations were made:

1. Top 3 features for conversion are:
- Tags
- Lead Origin
- Session Time (Total Time Spent on Website)

2. Top 3 dummy variables are:
- Tags: Will revert after reading the email
- Lead Origin: Lead Add Form
- Session Time 1000+ (Total Time Spent on Website)

3. To make the lead conversion more focus on the leads with high Lead Score.

4. To minimize useless phone calls avoid the following categories:
- Landing Page Submission
- Ocupation_Unemployed
- Ocupation_Unknown.