Lead Scoring Case Study Summary

1. Problem Statement

X Education is struggling with the conversion of customers though there are some potential leads. Model needs to be created for locating hot leads which can be targeted for conversion. Feedback on what should be focused on for future prospects should be provided for lead conversion

2. Data Handling

Data has been provided with 36 features, 9240 entries.Features about Asymmetrique indexes and lead profile dropped due to null values.Features based on search result: “Search”, “Magazine” etc. have too less variance and so combined under 1 feature Features like Country binned to create 2 major classes due to imbalance. “Prospect Id” and “lead number ” dropped as they are primary keys

3. EDA

Univariate and Bivariate Analysis has been performed to check distribution of features, the correlations among themselves and impact on the target feature. Several imbalances have been discovered, which were handled by Binning and bagging. Outliers also have been removed

4. Data Modeling

Numerical data like Total Visits and Total time spent on website min-max scaled. Categorical features are one hot encoded. Total 50 features after encoding.Reduced to 40 based on low correlation with target. Finally reached a data set of 15 features using Recursive feature extraction. After manual Elimination based on significance of the feature to prediction and VIF final of 9 features reached

5. Predictive Modeling

Logistic Regression Model established
70:30 - train: test data established
Final model input size : 9 features
Established cutoff at 56% probability for prediction

6. Model Evaluation

Train Data

79% accuracy on Train data set
78% precision and 75% recall on training data

Test data

78% accuracy on Train data set
79% precision and 75% recall on training data

7. Recommendations

X education should focus on:

a. Unemployed people
b. People coming from forms
c. Olark Chats, SMS and Emails
d. Maximizing people's visit time on website
e. Maximize total visits for people