Lead Scoring Case Study

Problem Statement

The company aims to attain these goals:

- To select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- A ballpark of the target lead conversion rate to be around 80% has been given

Solution Process

Data Cleaning-Handling Missing Values and Outliers

- Dropped variables with 40 percent or greater missing values.
- Changed Select to Null Values which indicates missing values
- Dropped variables with a greater than 40 percent select and missing values
- Checked for outliers and found outliers in two numerical variables
- Imputed Missing values in certain cases such as with median in the case of numerical variables due to presence of outliers and mode for the categorical variables
- Dropped variables with values belonging to mostly or only one category.

Exploratory Data Analysis

- Created Count plots for relationship between independent variable and target variable

Dummy Variable Creation

- Created Dummy Variables for the categorical variables, used drop first=True command to prevent dummy variable trap.

Divide Data into Train and Test Set

- The dataset is divided into train and test set with a proportion of 70%-30% of the values.

Feature Scaling

- The Min Max Scaling was used to scale the numerical variables.

Model Building

- Recursive Feature Elimination (RFE) was used to select the relevant independent variables. Stats model package was used to generate the statistics of the logistic regression model. The variables were dropped first according to high p values and low VIF and then according to high VIF and low p value and the variables were dropped in four steps and the fourth model was the one with significant variables and VIF below 5. 20 statistically significant variables were identified. The ROC Curve was plotted the AUC of the ROC was found to be 0.97(an area coverage of 97%) and it was away from the 45-degree line. The model was evaluated on the basis of measures such as Accuracy, Sensitivity and Specificity and also Precision, Recall. The optimal probability threshold was calculated on the basis of the Sensitivity-Specificity Trade and the Precision Recall Trade Off. The optimal probability threshold calculated was 0.3 and 0.38 respectively. We decided to stick with the cutoff from Sensitivity and Specificity Trade Off hence we calculated the lead score which indicated the target variable conversion rate for this optimal probability cutoff.

Implementation on Test Set

- The learnings from the train data model of the optimal probability cutoff based on sensitivity specificity trade off were applied to test set and the lead score which indicated the target variable conversion rate was also calculated for the test set

Results and Conclusions

- The goal of a lead conversion rate has been met with a lead conversion rate of 91.28% for the train set and 92.79% for the test set.

Metrics	Train test	Test Set
AUC of ROC	0.97	0.93
Accuracy	92.36%	92.53%
Sensitivity	91.28%	92.79%
Specificity	93.03%	92.37%
Precision	88.97%	88.81%
Recall	91.28%	92.79%
Lead Conversion Rate	91.28%	92.79%