Lead Scoring Case Study

Summary Report

This report is intended to mention the process we followed and our learnings on Logistic Regression Model building on Lead Scoring Case Study dataset provided.

- ✓ Read the dataset provided.
- ✓ Inspect the dataset in different dimensions such as knowing its shape, describe, info.
- ✓ Replaced the 'Select' values in dataset with nulls where user did not provide any value while filling the application. So, it is equivalent to null value.
- ✓ Calculated missing value percentage and removed couple of columns where their missing percentage was more than 70%.
- ✓ Imputed the missing values in each column. Columns of data type object are imputed with the most frequent value in column. And columns of other types are imputed with mean value of column.
- ✓ Dropped the columns with same values where assuming such columns do not add any value for model.
- ✓ Converted Yes/No column values to 1/0 respectively.
- ✓ Created dummy variables for all categorical columns with multiple values using one-hot encoded technique. Then dropped the actual categorial columns.
- ✓ Checked for outliers in dataset, identified outliers in couple of columns and removed them.
- ✓ Scaled the numerical column data (Feature scaling).
- ✓ Checked correlation among variables and dropped highly correlated variables by considering correlation value 0.6 as cut-off.
- ✓ Implemented Train-Test split on dataset where considered 70% for training and 30% for testing the model.

- ✓ Logistic Regression Model built using both automatic and manual processes i.e., Recursive Feature Elimination (RFE) and Statistics Model. Built multiple models repeatedly until getting model with appropriate p-values (considered as every feature should have less than 0.05) and VIFs (considered as every feature should have less than 5).
- ✓ Performed model evaluation metrics such as Accuracy, Simplicity and Specificity to assess the model. Plotted ROC curve, Precision-Recall Tradeoff and identified optimal cutoff point for evaluations.

Output and Answers

- ✓ Created a “Lead Score” column and assigned a score from 0-100 for every lead based on the conversion probability calculated.
- ✓ Identified top 3 variables contributed towards lead conversion probability.
- ✓ Identified top 3 variables where business should concentrate for more lead conversions.