upGrad

Summary Report

Problem Statement:

An education company, X wanted to build a model where they can assign a lead score to each of the leads so that they can pursue the customers with higher score who have more conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Approach & Outcome:

The lead scoring case study has been handled through building a logistic regression model which models the probability of an event occurring by having log-odds for the lead conversion be linear combination of one or more independent features to meet the constraints as per business requirements.

Reading and understanding the data:

The entire data set was imported to a pandas data frame to read and understand the data completely. The number of rows, columns along with information on data types and number of records with non-null data is reviewed.

Cleaning of data:

Exploratory Data Analysis (EDA) was performed to handle the below critical scenarios,

a) Identify and remove the duplicate rows
b) Whenever the potential customer had not chosen any option for a given attribute then it was considered as a missing value
c) Any columns with more than 40% of the data as missing values then entire column is deleted.
d) The columns with significant level of missing values with less than 40% was handled using imputing technique.
e) Less important columns and columns with skewed data were discarded for analysis

Visualization of the data:

Univariate and bivariate analysis was performed to see the distribution of the data for the critical variables. And Outlier analysis was performed and statistical outliers was removed from the numerical columns.

Data Preparation:

Converted the yes/no to 1/0 binary values. Then conversion of the categorical variables to dummy variables with numerical values as 1/0 was performed.

Split the data to train-test set:

The target and feature variables were separate to individual data frames. And the train-test data was split in the ratio of 70-30.

Scaling:

The numerical columns were scaled for seamless and meaning model building exercise.

Model Building and Metrics Validation:

The RFE technique was used for selection of variables as Identifying correlation between variables is difficult from heatmap. The logistic regression model was built with arbitrary cut-off probability point of 0.5 to find the predicted labels.

The accuracy score was good to begin with but the columns which higher P-values were removed and better models were built iteratively checking for both P-values and VIF score.

For final model built, other metrics sensitivity, specificity, FPR were found to be lower than the target score of 80%.

Plotting ROC curve and finding optimal cut off point:

Final model had area under ROC curve as 0.87 which is good. And optimal cut-off point was 0.35

Model Evaluation:

Sensitivity, Specificity and Accuracy values were above 80% thus model was fit for the requirement

Make Predications on Test data set:

For test data also model performed well, meeting the target Sensitivity, Specificity and Accuracy scores of 80%