Summary Report

This is a logistic linear regression model build to assign a lead score between 0 and 100 to each of the leads for an education company named X Education which sells online courses to industry professionals. This lead score is used by the company to target potential leads and maximize the lead conversion rate.

The analysis comprises of the following steps:

Step 1: Importing and Understanding the dataset

We read the dataset, inspect the number of rows and columns and datatypes of each column. We found out the number of rows to be 9240 and the number of columns to be 37.

Step 2: Data Cleaning

Unwanted columns are dropped, null values are handled along with exploratory data analysis. After removing the unwanted information, we were left with 14 columns.

Step 3 : Data Preparation

Includes encoding the data, train-test split and rescaling the features. Two binary variables were converted to numeric and dummy variables were created for seven categorical variables.

Step 4: Look at the Correlations

Finding and deleting highly corelated variables through heatmap. High correlations were found but were not too high to be dropped.

Step 5 : Feature Selection using RFE

Selecting the top 20 feature variables by using Recursive Feature Elimination method.

Step 6 : Model Building

Manually dropping the variables on the basis of their high p values and high VIF values and building our final model. Two columns were dropped and we got our final model as the variables had good p-values and VIFs.

Step 7 : Making Predictions on Train Set

Prediction values are calculated and the model's potential is tested by calculating accuracy, sensitivity and specificity. Arbitary cut-off was selected as 0.5.

Step 8 : ROC Curve

ROC Curve of the model is plotted to check the performance of the model at different thresholds (true positive rate and false positive rate).

Step 9 : Finding Optimal Cut-off Point

Optimal cut-off point is calculated by balancing sensitivity, specificity and accuracy. Optimal cut-off was found to be 0.36.

Step 10 : Making Predictions on the Test Set

Prediction values are calculated and the model's potential is tested by calculating accuracy, sensitivity and specificity but on the test set.

Results :

Train Set - Accuracy : 80.1 %

Sensitivity : 79.8 %

Specificity : 80.2 %

Test Set - Accuracy : 80.8 %

Sensitivity : 74.8 %

Specificity : 84.7 %

Thus, we achieved the ballpark of the target lead conversion rate to be around 80% given by the CEO of the company.

The feature variables which are important to the company were found out to be as follows:

Lead Origin_Lead Add Form
Last Activity_Other_Activity
Last Notable Activity_Unreachable
Last Activity_SMS Sent
Lead Origin_API
Time Spent
Occupation_Student
const
Occupation_Other
Occupation_Businessman
Lead Source_Google
Specialization_Hospitality Management
Lead Source_Referral Sites
Lead Source_Organic Search
Last Notable Activity_Modified
Email
Lead Source_Direct Traffic
Last Notable Activity_Olark Chat Conversation
Specialization_Other