Summary – Lead Scoring Case Study

(Submitted by Abhinav Phirani and Abhishek Sharma – EPGPDS - C53)

Problem Statement

X Education - an online edtech company selling courses is currently running at ~30% lead conversion rate. It currently uses several websites and google for bringing people to websites and then these leads are converted through watching videos, filling form, calls, SMS etc. Company also uses referrals. Company wants to identify more leads that have higher conversion rate, calling them 'Hot Leads'.

Goal

To develop a logistic regression model to provide lead score to each lead. Targeting conversion rate to be higher than 80%. In addition, there are 4 subjective questions that need to be answered separately.

Solution:

8 stage process followed to define a robust logistic regression model:

Stage 1: Importing libraries and read data

Import all necessary libraries and read the raw data from **'Leads.csv'** file.

Stage 2: Data Cleansing

Clean the data, starting with removal of unique values 'Prospect ID' and 'Lead Number'. Drop the columns with > 30% null values. Remove the imbalanced variables. Remove the rows with Outlier values. Fix null values and case sensitivity issues. Bivariate analysis using correlation matrix to identify dependency on variables. Eliminate variables with high correlation.

Stage 3: Data Encoding

Create Dummy variables for categorical variable. Create 0,1 coding for binary variables.

Stage 4: Train-test split

Divide total data between 70:30 ratio, 70% being training data. Move converted variable being as Y and all other variables as x's.

Stage 5: Feature Scaling

Convert Numerical variables to standard scaler.

Stage 6: Bivariate Analysis

Again, run the correlation matrix with all the variables. Drop the variable with >80% correlation.

Stage 7: Model Building

Create logistic regression model using sklearn. Use Recursive feature elimination to limit number of variables to 15. Review the P-values and remove the variables with p-value >0.05 iteratively. Validate VIF for balance variables.

Stage 8: Model Validation

Calculate accuracy, sensitivity and specificity of the model. Plot ROC curve and validate the area under the curve. Area under ROC curve comes out to be 0.86. Calculate the Optimal cut-off points, by plotting accuracy, specificity and sensitivity for various probability cut-offs. Optimal cut-off point is near 0.3.

Stage 9: Predicting outcome on test data

Start with scaling test data using standard scaler and select the required variables. Predict the values of y. Compare the converted ratio at optimal cut-off probability of 0.3 between train data and test data. Both are ~80% hence the test data is a good fit on the model. Also the hot leads are meeting the CEO criteria of ~80%.

Stage 10: Validating the performance parameters for test data

Calculate Accuracy, sensitivity and specificity for test data which are similar to train data.

There for the final model has a conversion rate of 83.8% on test data which meets the CEO expectations. Prospects with Lead score of >30 can be considered as hot lead for building the outreach strategy.