UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Business Statistics

Assignment 5

Customer churn, or attrition, refers to when a customer or subscriber ceases his or her relationship with a company. A customer is said to churn when he or she changes provider. Telephone service companies and internet service providers perform customer churn analysis, as the cost of retaining an existing customer is far less than acquiring a new one.

The dataset churn.xls, taken from the UCI Repository of Machine Learning Databases at the University of California, Irvine, deals with $N = 3333$ customers of of a telecom company and contains information relating to the telephone calls they make.

Here is a short description of the variables:

State	Categorical variable, for the 50 states and the district of Columbia
Area code	Categorical variable
Phone number	Customer identification
Account length	Discrete variable for how long the account has been active
Int'l Plan	International Plan: dummy variable
VMail Plan	Voice Mail Plan: dummy variable
VMail Message	Number of voice mail messages
CustServ Calls	Number of calls to customer service
Day Mins	Total day minutes: number of minutes customer has used the service during the day
Day Calls	Total day calls
Day Charge	Total day charge
Eve Mins	Total evening minutes: minutes customer has used the service during the evening
Eve Calls	Total evening calls
Eve Charge	Total evening charge
Night Mins	Total night minutes: minutes the customer has used the service during the night
Night Calls	Total night calls
Night Charge	Total night charge
Intl Mins	Total international minutes
Intl Calls	Total international calls
Intl Charge	Total international charge
Churn	Target variable

The target variable is `Churn`, with two levels, 1 and 0 (no churn). Your objective is to build an accurate and reliable classification model or method for predicting customer churn. You should compare logistic regression, with variable selection according to forward stepwise, and classification trees, in terms of missclassification error in the test sample and by plotting the ROC.

The dataset can be will be divided into two parts. The first, containing 2,500 observations will be used for training, whereas the second, consisting of 833 records, is used for validation.

The partitioning is done by random sampling. The `seed` for the sample selection (which is the argument of the `set.seed()` function), must be set in the usual way.

Please upload your report at https://www.dropbox.com/request/lpcCMqZGK3gsCRsvT3xv before 20:00 PM on 17/12/2019.