This directory contains the churn dataset used in the book Data Science for Business, by Provost and Fawcett. It is important to note that it is a realistic---BUT NOT REAL---dataset. For this reason it may be used solely for teaching and education. It may NOT be used for research purposes or as the basis for published results.

Questions about this dataset should be directed to Foster Provost <fprovost@gmail.com>.

-- Foster Provost and Tom Fawcett, March 2014.

DATA DICTIONARY

The data file here, churn.arff, is in Attribute-Relation File Format (ARFF). The ARFF file format is described precisely at: http://www.cs.waikato.ac.nz/ml/weka/arff.html . The format is basically a Comma-Separated Value (CSV) file preceeded by a concise description of the variables.

The dataset consists of 20,000 examples (lines, rows) over 12 variables (fields, columns). The dataset constitutes a two-class supervised learning problem. The class variable, LEAVE, is the last variable on each line, and its legal values are LEAVE and STAY. The header of churn.arff describes the legal values of each variable. Informally, here are their meanings:

COLLEGE : Is the customer college educated?
INCOME : Annual income
OVERAGE : Average overcharges per month
LEFTOVER : Average % leftover minutes per month
HOUSE : Value of dwelling (from census tract)
HANDSET_PRICE : Cost of phone
OVER_15MINS_CALLS_PER_MONTH : Average number of long (>15 mins) calls per month
AVERAGE_CALL_DURATION : Average call duration
REPORTED_SATISFACTION : Reported level of satisfaction
REPORTED_USAGE_LEVEL : Self-reported usage level
CONSIDERING_CHANGE_OF_PLAN : Was customer considering changing his/her plan?
RESULT : (Class variable) whether customer left or stayed