

## Mining in Financial Databases

### Assignment 3

**Submission deadline and oral presentation: April 5-6, 2017**

The objective of this assignment is to design a predictive model that discovers customer churn behavior and reduces churn effects using data mining approach. In financial institutions, the churn is related to the number of customers who decide to stop using a service offered by one bank and to use another bank, usually we because it offers a better service or lower cost. The issue of customer loyalty is very important not only in the banking sector due to the losses resulting from the departure of the client, and large costs and the difficulty of acquiring new customers.

Churn models (also called models of customer loyalty or of retention or of attrition) are used to predict the likelihood that a person will remain our customer. In particular, disloyal customers can be the target of preventive measures (such as offering discounts on subscription). In contrast, for the loyal customers we offer direct purchase of new products (*cross-selling*), or the extension of current cooperation (*up-selling*).

More specifically, the aim of the assignment is to develop a predictive model and decision rules that allow to find these customers for which the probability of churn is the highest. In your report you should also give the interpretation of the model, encode it in the form of Excel functions and to assess the quality of the prediction.

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

There are four datasets:

- 1) *bank-additional-full.csv* with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
- 2) *bank-additional.csv* with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
- 3) *bank-full.csv* with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
- 4) *bank.csv* with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

*The smallest datasets are provided to test more computationally demanding machine learning algorithms.*

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

The dataset is available at UCI Machine Learning Repository;

<http://mlr.cs.umass.edu/ml/datasets/Bank+Marketing>

#### To do:

- 1) Edit the results in Power Point, and submit your compressed presentation as file

**DMF-nazwisko.zip** to: [jerzy.korczak@ue.wroc.pl](mailto:jerzy.korczak@ue.wroc.pl)

The presentation should include the following information and data: experiment description, description of features, data sets used, data mining schema, visual representations of the results, conclusion of the experiments.

- 2) Oral presentation has to be given by selected authors of the *assignment*; *10min per project*.

The assignment can be done by max two students

#### References:

Lecture notes (<http://www.korczak-leliwa.pl>)

Dębski T. *Drzewa klasyfikacyjne w przewidywaniu migracji klientów (churn), systemy IT*, s. 54-57.

Analiza satysfakcji i lojalności klientów, 2003, StatSoft Polska.

Moro S., Cortez P. and Rita P., *A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems*, Elsevier, 62:22-31, June 2014

Phelan S., *Customer Information as a Strategic Asset*, DM Review Online ([www.DMReview.com](http://www.DMReview.com)), 2002