Università di Pisa	A.A. 2012-2013
Data Mining II	

Project assignments / Part 2

Market transactions

General information

Objective of this project is to perform some data analysis steps that (in part) make use of the preprocessing operations done in the part 1 of the project. The same rules adopted for part 1 hold here, in particular:

1. the project can be performed by single students or groups up to 3 persons each;
2. each group should perform the analyses indicated in the text, trying to answer to each request. Any spontaneous addition to that is welcome yet optional, and cannot replace the original TODO list;
3. each group should summarize the work done in a short report (indicatively 5-15 pages), loosely following the guidelines of the CRISP model;
4. each group is totally free to choose the tools and software it prefers;
5. any question, suggestion or request related to the project can be addressed to Mirco Nanni (mirco.nanni@isti.cnr.it).

The dataset

Use the same dataset adopted for part 1 of the project.

Objectives

Given the specific category "X" of products assigned to the group, perform the following steps:

1. Customer segmentation: use and (if needed) integrate the set of aggregates computed for each customer to extract homogeneous clusters of customers. Try to characterize each cluster by analyzing the distributions of the attributes of its

customers.

2. Event characterization: choose one of the two events detected in the preprocessing phase (churn and focusing). Then, build a classifier that is able to predict whether a customer is affected by that event or not, based on the aggregates computed in the previous task (further integrations with additional aggregates and variables are welcome). Describe and comment the most interesting relations that emerge between the event and the aggregates.
3. Innovators: using the time series extracted in the first part of the project, find the innovators (if they exist) for the selected product. It is allowed (and welcome) to perform the analysis on more than one product.
4. [OPTIONAL] Frequent patterns: select the 3 top clusters found in task 1, and search frequent patterns (itemsets, rules or sequences, at your choice) that characterize each cluster against the others.
5. Evaluate the privacy issues that might arise in this application, and discuss possible counter-measures to adopt in order to remove or limit them.