Université de St-Etienne.

Data Mining for Big Data: Project

Assignment: Due date Monday January 22nd 2018

All material is available on claroline connect
Project groups of at most 4 students

1 Project Objectives

The objective of this project is to achieve a large scientific analysis of a dataset by means of both data mining, machine learning and data analysis techniques (that you have learned so far in any course!).

You must take this project according to the following context: Suppose you work for an Information Technology company specialized in data science. Your boss is in touch with the Groupama insurance company and he asks you to study a dataset of this institution. The final goal is to study the problem of this company and answer their questions on the data (see section 4.).

You must present your study in a report which is to be delivered to the Groupama company. The quality of the presentation, of the writing and of your scientific conclusions are very important. The best reports will be delivered to the Groupama company.

You are free to define the structure of your report, but you can base your work on the following guidelines:

- Provide a cover page with title, name of the contributors, logos (e.g., from the university); do not forget to provide a table of contents.
- Provide an introduction that will present your work and the structure of the report.
- Provide a section describing the dataset.
- Provide a section describing your objectives and the different studies made.
- Provide a section for each study made: you must clearly state the objective of the section, present clearly the experimental setup (preparation of the data, algorithm(s) used, any relevant information), gives the results in a neat way (table of results, plots, curves, graphs, . . . , do not forget to comment them) and give your conclusions for this section.
- Provide a final section summarizing what you have done, developing your conclusions with personal remarks and/or suggestions about the problems/questions asked by the company.

To help you, you can find on claroline some reports made by students of previous years.

You can use anything you think relevant for doing this analysis. You can use data mining, some data analysis approaches and possibly machine learning methods (depending on your background).

You are free to use any existing software, implementation, library, platforms... free of use, but you have to mention all the tools used in the report.

You must of course take the time to read the documents related to the dataset and take the time to understand what is in this dataset. Do not hesitate to be creative, propose ideas, ...

2 Context of the Study

The Groupama company (an insurance company) wants to make a study about the satisfaction of its customers and the effect of its actions on the satisfaction of the customers. The main objective is to find how to improve the customer satisfaction and the income of the company.

There are two kind of customers: person and others (other companies, associations etc.).

Each customer may have several insurance contracts with the Groupama company. The company store several information about the customers (age, category, category of home area,...). It also stores the customers requests and complaints. In reaction to these requests and complaints, the company can take actions or give the customer an advantage (price reduction for instance).

The company also regularly do satisfaction survey on its customers. The results of these surveys are also provided.

3 Data

The data consist of several tables. Each table is stored in a CSV file (Comma-Separated Values, actually the field separator is a semicolon ";") using UTF-8 encoding of characters.

The data consist in two types of data:

- a relational database concerning the customers who sent a complaint or a request to Groupama. This database is composed of six tables describing the customers, their requests and complaints, the actions done by Groupama to answer these requests and complaints;
- and a set of 16 tables with the results of customer satisfaction surveys.

In the database, a customer is identified by the field ID-GRC. In the satisfaction surveys, he is identified by the field IDENTIFIANT. Consequently, when a customer send a request and answers to a satisfaction survey, the value for the field ID-GRC and Identifiant corresponds. However, all the customers who send a request or a complaint, and consequently who appear in the database, do not necessarily answer to a survey and, reciprocally, a customer can reply to a survey without sending a request or a complaint.

The data is described in the files "description_of_data.pdf" and "PPT_Fouille de données.pptx" available on claroline connect.

4 Questions

The Groupama company has several questions:

1. Which customers are satisfied or not?
   - Are there some areas where customers are more/less satisfied? The area is defined by the "typologie" field.
2. Understanding why the customer is not satisfied:
   - Analysis of customer complaints by customer category, profile and area. Are there categories, profiles, areas or combinations of these attributes for which customers are less satisfied.
   - Analysis of customer comments in complaints. E.g., what words / sequences of words in comments explain the best the customer satisfaction / dissatisfaction. Find correlations between customer evaluations and their comments. The aim is to explain why the customer are not satisfied with the analysis of their comments and make the link between the reason of their dissatisfaction and their characteristics.
   - Propositions to improve customers satisfaction: Using your analysis of previous questions, point out the most important reasons why customers are dissatisfied.
3. Measure the effect of commercial relationship between the company and customers on the satisfaction. A customer may be associated to several requests, complaints and actions. The objective is to study what is the effect of these actions on the customer satisfaction and the income of the company from the customer.
   - What is the temporal evolution of the evaluation of the company in the different satisfaction studies? You can take into account the different characteristics of customers.
   - Measure the correlation between the evolution of the income of the company from a customer and its global satisfaction.
   - For the customers having a complaint, study the effect of customer advantages on the satisfaction and the evolution of income from this customer.

5 Reproducibility of Your Results

Your report must be a scientific report. This means that every result presented in your report (in the text or in a graphic, table, etc.) can be replicated by anyone having the datasets. You must therefore describe what you have done with sufficient details such that anyone reading your documents (for instance, your professors) can do the same and get the same results as you. You must, for instance, include the code you have written, explain which software you used (with the version number), give all the parameters that you used each time you did an experiment and so on.

This information can be put in the report itself or in an additional document. You can include a log file of every command you launch as we did in the practical sessions.

Do not wait until the end to write this document! Do it each time you do an experiment, otherwise you will forget some details or parameters.

6 Assignment

Your work (zip file) must be uploaded on claroline connect for **Monday January the 22nd, 23:00** in the Ressources section of the workspace "M2 CS Big Data". Your work must contains:

- the report;
- anything (code, documentation) necessary for the reproducibility of your results.

The work can be done by groups up to 4 students.

7 Evaluation

Your work will be evaluated based on:

- the report;
- the number of questions that you have studied and the quality and depth of your studies;
- the answers to the questions and the provided evidence to support your claims;
- the reproducibility of your results.

8 Plagiarism

Plagiarism is considered a fraud and will lead to the exclusion from the master. We use an automatic tool to detect it. Plagiarism is the fact to *"copy texts or take ideas from someone else's work and use them as if they were one's own"*[1]. It means that everything in your report, document, code, ... that is not your own work must be properly cited: for instance, use a different typeface and/or indentation and give the precise reference of the source (web site, book, paper, ...).

[1] definition from Cambridge Dictionary, `http://dictionary.cambridge.org`