Université de St-Etienne.

Data Mining for Big Data: Project

Description of Data

These data describe the interactions between the Groupama company and its customers.

1 Relational Database

The schema of the database is in the file "PPT_Fouille de données.pptx". Only the main fields are described below. You will find information on the other fields in the "PPT_Fouille de données.pptx" file on claroline. These are the most important fields to use in your studies. In the document file "PPT_Fouille de données.pptx", the important fields are tagged by a yellow star.

All tables (except "Structure_Commerciale" are linked by the customer ID (field "ID_GRC"). This Id uniquely identifies each customer in the database. Customer data has been anonymised (no name, address, etc.).

Each customer may have several requests, complaints, actions and advantages.

1.1 BASE_donnees_clients

This is the customer table. It contains one row for each customer. the main fields are:

- ID_GRC: this is the unique identifier of the customer. It is also found in other tables.
- TRANCHE_AGE: the ages of customers are grouped in intervals. For instance "45 - 64 ANS" means age between 45 and 64.
- NATURE_PERSONNE: category of customer: "PP" means person, "PM" means other (e.g., company, association, etc.)
- MARCHE_PSO: category of customer by activity: e.g., "Particulier" and "Retraité" are people, "collectivité" is local communities, "Agricole" is related to agricultural activities, "Entreprise" is a company and "ACPS" is for small companies with generally just one person (like artisan, storekeeper,...).
- TYPOLOGIE: category of living area (there are 6 categories like rural area, city center, ...)
- CD_COMMERCIAL_CHARGE: ID of the Groupama employee in charge of this customer. This links to the field "CD_POSTE" in table "structure_commerciale".
- MT_IARD_2015: income from this customer for 2015: how much this customer paid for its insurance contracts in 2015.
- MT_IARD_2016: same for year 2016
- MT_IARD_2017: same for year 2017

1.2 BASE_Structure_Commerciale

This table give information on the employees in charge of customers. Each employee belongs to an agency which belongs to an area which itself belongs to a region.

The main fields in this table are:

- CD_POSTE: the Id of the employee
- REGION_COMMERCIALE: region of the employee
- SECTEUR_COMMERCIAL: area of the employee
- AGENCE_COMMERCIALE: agency of the employee

This table is useful to study the satisfaction of customers in different regions or areas in addition to the TYPOLOGIE field which specifies their living area.

1.3 BASE_Demandes_clients_hors_reclamations

This table contains request from customers (complaints are in another table).

The main fields are:

- ID_GRC: Id of the customer (refers to the same field in the Base_donnees_clients)
- NUM_DEM: Id of the request. Linked to the same field in Base_actions_rattachees_demandes and in the satisfaction surveys.
- DATE_CREATION and DATE_CLOTURE: date of the request and date when the request is closed (i.e., an answer has been given to the request).
- delai_TT and statut_delai_TT: number of days between request and closure.

Each request is linked to an action in the following table.

1.4 BASE_actions_rattachees_demandes

- ID_GRC: id of customer
- NUM_DEM: id of request. Linked with same field in previous table.
- TYPE and SOUS_TYPE: category and sub-category of the action.
- DATE_CREATION: date of action.
- COMMENTAIRES: text comments on action.

1.5 BASE_Reclamations_clients

This table contains the complaints of customers. Once a complaint has been answered it is closed. It can be re-open if the customer is not satisfied with the answer. Then, it is closed a second time after the second answer has been given.

- ID_GRC: id of customer
- NUM_DEM: id of the complaint (linked with the satisfaction survey)
- DATE_CREATION: date of complaint
- DELAI_TT: total number of days between complaint and final closure of complaint (1st closure or 2nd closure if the complaint was re-opened).
- TYPE and SOUS_TYPE: complaint are grouped into types and sub-types.
- RECLAM_QUAL_CLOT: if the answer is favorable to the customer or not.
- REPONSE_RECLAMATION and REPONSE_RECLAMATION_N2: transcript of the answer given to the customer (and 2nd answer if the complaint was re-opened
- COMMENTAIRE_DEMANDE: transcript of the customer complaint
- MOTIF and SOUS_MOTIF: words in the transcript of the complaint that characterize the complaint.

1.6 BASE_Avantages_clients

A customer can be given several advantages (after a request or a complaint).

- ID_GRC: id of customer.
- CODE_AVG: code of the advantage given to customer.
- STATUT: status of the advantage (used or not by customer).
- CLIENT_FIDELE: fidelity of customer: 1: true; 0: False

It is important to point out that the table BASE_Reclamations_clients concerns customers who are not satisfied since they sent a complaint whereas the table BASE_Demandes_clients_hors_reclamations concerns in majority customers who can be satisfied and, eventually not satisfied if they sent a complaint in addition to the request described in this last table.

2 Satisfaction Surveys

The satisfaction surveys results are in 16 files (”SATISFACTION_XXXX”).

Each time a customer send a request or a complaint, it is asked, a few days later, to fill a survey to check if he is satisfied with the way it was handled (there is however a limit on how frequently a particular customer is surveyed). Moreover, the satisfaction survey may also concern customers who do not sent a request or a complaint. Each of these files contains the answer of the customers to these surveys (only if the customer filled the survey). Each file corresponds to a category of insurance contract or to a kind of request.

Each of these file contains:

- Date of survey
- Answers to questions (depends on the file). The first question is always the global evaluation of the customer on the Groupama company. It is an integer between 1 and 10 (higher is better). The next answer is a text where the customer can explain is global evaluation. The choices for other questions are generally in the set ”very satisfied”, ”satisfied”, ”not satisfied”.
- After the questions, there are a number of attributes about the customer and the request / complaint that triggered the survey. The most important one is:
- The Id of the customer (field ”IDENTIFIANT”). It corresponds to ID_GRC in the database.
- For the files ”RECLAMATION” (complaints) and ”DEMANDES” (requests), there is also the request / complaint Id (”numero_reclamation” or ”numero_demande”) which correspond to the same field in the tables ”BASE_Demandes_clients_hors_reclamations” and ”BASE_Reclamations_clients”.