Project 1

Data preprocessing.

`Deadline`: Monday September 6 at 11 am.

1 Assignment: BaseFuga Preprocessing

Consider the problem faced by a financial institution that has high voluntary churn rates. This institution is not sure what is the characteristic profile of churners or what are the reasons why they leave the company.

Within this context, the task is to preprocessing the financial institution database to see the behavior of the data and then be able to make predictive models with them (using the same Jupyter Notebook, you must go to File and make a copy of this file before you can modify it).

1.1 Variables

A complete description of the variables collected for each customer is presented below:

1. ID: Customer ID.
2. Genre: Customer genre.
3. Age: Age in years.
4. NIV_Educ: Educational level.
5. E_Civil: Marital status.
6. City: Office City.
7. D_Marzo: March Debt.
8. D_Abril: April Debt.
9. D_Mayo: May Debt.
10. D_Junio: Debt of June.
11. D_Julio: Debt of July.
12. D_Agosto: Debt of August.
13. D_Septiembre: September Debt.
14. M_Moroso: Months in arrears.
15. Amount: Pre-approved Amount.
16. Insurance: Whether the customer has insurance or not.
17. Churn: target variable. Past

2 Submission

You will have to make a report explaining the previous tasks. Please deliver your code in .ipynb and a pdf with your report.

2.1 Code Guideline 70%

You have to:

1. Perform a descriptive analysis of the data. Support your analysis using graphs. Carry out a descriptive analysis of the data. At this point you must calculate the main statistical metrics for a better description of the data, for example, descriptive statistics, frequency tables, etc. Support your analysis using charts. (15%)
2. Replace/correct missing values. Indicate what problems you detect in the database and take measures to correct them. You must indicate the procedure you followed to resolve them clearly and precisely.(10%)
3. Group inconsistent/under-represented categories. For example widow/widower can be merged with another category, or make a dummy for Santiago versus the rest of the regions. (15%)
4. Data transformation. Transform the available variables in order to find new variables, generated from the original ones, that can have a good predictive performance based on your knowledge of the business (at least 5 transformations). For example, adding up all the debts of the months or use logarithm in the debt variables. You can create other variables as bonuses. (10%)
5. Selection and ranking of attributes to choose relevant variables. Analyze your variables. Do they discriminate between churners and non-churners? Study them at the three levels seen in class (uselessness, redundancy, and irrelevancy). Create a relevance ranking of the variables. Draw conclusions based on this analysis.(15%)
6. Normalize the dataset.(5%)

2.2 Report Guideline 30%

The report developed must contain at least the following points:

1. Cover, name of the members, index. (2%)
2. Introduction: A general description of the problem to be solved, a brief description of the report and a clear visualization of objectives, both general and specific, must be given.(5%)
3. Statistical analysis of the data. (8%)
4. Description of the preprocessing methodology. (8%)
5. Selection of attributes and explanations based on the findings. (7%)

The report should not have more than 7 pages (Without considering the cover, members and index), some works will be selected for the English revision.