

Collection of raw data

The information was extracted with SAS to a CSV. Original information from the company DDBB which was later anonymized. Population and other information related to customers location was taken from INE website.

DataFrame(notice this dataframe uses “;” delimiter).Also used chunksize for breaking of the file into chunks

Originally 400k instances and 51 features.. Filter and delete non residents to avoid noise in the analysis. Furthermore 7 features were dropped as considered not relevant for the objective of this study.

The dataframe used contains 395.937 instances and 44 features. 388.932 customers without personal loan and 7.005 with personal loan as dependent variable.

21 of the 44 features have, at least, one null value and non duplicates were found.

Eventhough the initial information included in the dataset was real, in order to present this work and to complain with all the current regulation, all data has been anonymized.